Technische Universität München
TUM School of Engineering and Design

A data-analytical approach to reduce temporal complexity of energy system models

Inga Maria Fladerer

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung einer

Doktorin der Ingenieurwissenschaften (Dr.-Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr. rer. pol. Christoph Goebel

Prüfende der Dissertation:

1. Prof. Dr. rer. nat. Thomas Hamacher
2. Prof. Dr.-Ing. Marco Pruckner

Die Dissertation wurde am 30.01.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 12.11.2024 angenommen.

Acknowledgements

An dieser Stelle möchte ich allen Menschen danken, die mich während der Doktorarbeit begleitet und unterstützt haben.

Besonders danken möchte ich meinem Betreuer Professor Thomas Hamacher für die Möglichkeit, mein eigenes Forschungsthema entdecken und diesem nachgehen zu können. Ich danke meinem Mentor Philipp Kuhn für die vielen Diskussionen und die daraus resultierenden Impulse. Mein Dank geht auch an meinen ehemaligen Kollegen Konrad für die gemeinsame Grundsteinlegung dieser Arbeit und an meine ehemalige Kollegin Magdalena für den methodischen Austausch in der Anfangszeit dieser Arbeit.

Danken möchte ich auch meiner Familie, die mich auf meinem Lebensweg unterstützt und immer für mich da ist. Ein großer Dank geht an Katharina und Quirin, die mich durch alle Höhen und Tiefen dieser Arbeit begleitet haben.

In Liebe danke ich meinem Mann Martin, der mir auf seine wundervolle Art Stärke verleiht. Mit ihm an meiner Seite erscheint nichts mehr unmöglich zu sein. Und ich danke meinem Sohn Jasper, der mich einfach nur durch sein Dasein erdet und meinen Fokus auf das Wesentliche lenkt.

Abstract

The aggregation of time series is a common approach to reduce the temporal complexity of energy system models. However, the modeling results of aggregated time series are not always robust and vary depending on the defined energy system, the aggregation method and the data used. So far, we have only limited knowledge about the interaction of (different) time series and energy system models, which would allow to understand the deviations of the modeling results or to adapt the aggregation methods. In this thesis, a data-analytical approach for aggregating time series is presented, which can be divided into two methods. The first method is based on clustering and nested regression (CNR) and analyzes the interaction between time series and modeling results. For this purpose, the information from time series is converted into various time series parameters and relevant parameters are identified. These are transferred to the second method, the profiling. Already aggregated time series are iteratively adjusted to the relevant time series parameters of the original time series. The results show that profiling leads to a better representation of the original time series and that systematic deviations in the modeling results are significantly reduced. This approach therefore complements existing aggregation methods. In addition to the general further development of the CNR and profiling methods, directions of future research are discussed (e.g., the transfer of CNR and profiling to more complex energy system models).

Zusammenfassung

Die Aggregation von Zeitreihen ist ein verbreiteter Ansatz, um die zeitliche Komplexität von Energiesystemmodellen zu reduzieren. Die Modellierungsergebnisse von aggregierten Zeitreihen sind allerdings nicht immer robust und variieren in Abhängigkeit des definierten Energiesystems, der Aggregationsmethode und der verwendeten Daten. Bisher haben wir nur ein begrenztes Wissen über die Wechselwirkung von (unterschiedlichen) Zeitreihen und Energiesystemmodellen, welches uns ermöglichen würde, die Abweichungen der Modellierungsergebnisse zu verstehen bzw. die Aggregationsmethoden anzupassen. Im Rahmen dieser Doktorarbeit wird ein datenanalytischer Ansatz zur Aggregation von Zeitreihen vorgestellt, der sich in zwei Methoden unterteilen lässt. Die erste entwickelte Methode CNR (engl: clustering and nested based regression) analysiert die Wechselwirkung zwischen Zeitreihen und Modellierungsergebnissen. Dazu werden die Informationen von Zeitreihen in diverse Zeitreihenparameter überführt und relevante Parameter identifiziert. Diese werden in der zweiten Methode, dem Profiling, aufgegriffen. Bereits aggregierte Zeitreihen werden iterativ an die relevanten Zeitreihenparameter der ursprünglichen Zeitreihe angeglichen. Die Ergebnisse zeigen, dass Profiling zu einer deutlich besseren Repräsentation der ursprünglichen Zeitreihe führt und systematische Abweichung in den Modellierungsergebnissen signifikant reduziert werden. Somit stellt dieser Ansatz eine Ergänzung zu bestehenden Aggregationsmethoden dar. Neben der allgemeinen Weiterentwicklung der CNR und Profiling Methode, wird weiterer Forschungsbedarf diskutiert (z.B. der Transfer von CNR und Profiling auf komplexere Energiesystemmodelle).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, a data-analytical approach is proposed to reduce the temporal complexity of energy system models and therefore the complexity of models in general. The approach can be divided into two complementary methods - CNR and profiling - that are described and evaluated separately in two peer-reviewed publications: The *identification paper* and the *profiling paper*. Within this thesis, these publications are merged into a coherent monograph supplemented by theoretical foundations, in-depth analyses, and extended discussions. In addition, the data basis is unified, and analyses are renewed in order to achieve improved transferability and comparability. As major parts of the thesis are based on these publications, explicit reference is made to the respective paper in the beginning of chapters and sections concerned.

1.1 Background[1]

Energy system models based on optimization are a common tool for analyzing climate political strategies and targets as well as deriving recommendations for action (e.g., system configurations [74] and flexibility requirements [12]). To provide reliable results and allow comprehensive analyses, models must represent the technical details of the energy system but at the same time avoid high computational effort or infeasible problems. With the transformation of the power system from conventional to green-house-gas emission free technologies the share of intermittent renewable energies (iRES), such as photovoltaic (PV) and wind power, has increased [55] and requires specific modeling in both temporal and spatial detail [42]. Time-dependent systematic and stochastic patterns of iRES can be represented by time series that, firstly, have a high time resolution to include hourly and daily fluctuations and, secondly, cover multiple years to consider intra-annual fluctuations [47]. To take geographical dependencies into account, for example, intra-country differences in the distribution and correlation of iRES [22] as well as their annual generation potential [60], modeling on a regional or sub-country level is needed. However, detailed energy system models with a high temporal and spatial resolution are not always feasible due to limited computation resources [21] and require simplifications without compromising the quality of the model and its results by losing relevant information.

[1] This section is based on the *profiling paper* - Chapter 1 [40].

In the literature, the complexity and related simplification methods are split into three categories: (1) technical dimension, (2) spatial dimension and (3) temporal dimension (see, for example, [36, 42]). Technical relationships are already simplified in many existing models, for example, to avoid non-linear constraints or binary variables (e.g., *GENeSYS-MOD*, [11], default version of *calliope* [48], *urbs* [34]. An analysis of model capabilities indicates that among other system components, the distribution grid, the demand and the technical flexibility are insufficiently represented [54]. A comparison of spatial and temporal aggregation [21] shows that for energy systems with a high share of iRES a reduction in temporal resolution is preferable to a reduction in spatial resolution. The potential of temporal aggregation becomes also evident in the high number of publications focusing on this topic (e.g.,[6, 36, 42, 47, 51]). In addition, sector coupling, that is, the integration of the heating and mobility sectors (e.g., [5, 41]) becomes increasingly relevant in the modeling of future energy systems. Their demand patterns can also be described by time series and suggest a comprehensive investigation of time series and their aggregation possibilities in general.

1.2 State of research[2]

In the literature, four main approaches of time series aggregation (TSA) are discussed, see, for example, [42, 47, 51]: downsampling (i.e., decreasing of temporal resolution, e.g., from one to two hour time steps), heuristic (i.e., selecting (contiguous) days based on defined criteria, e.g., maximum demand or wind power), clustering (i.e., selecting representative (contiguous) days including a weighting factor), and optimization (i.e., selecting (contiguous) days by minimizing an error indicator). Further, combinations of these four approaches (e.g., heuristic and clustering [47]) as well as more complex techniques are proposed such as directly including TSA in energy system models by decomposition [6]. However, a comprehensive analysis that compares up to 30+ TSA approaches, 25 years of time series input data, and three model scenarios, shows that there is not a one-fits-all TSA approach [47]. The performance of TSA depends on model scenarios (e.g., storage availability and system design, see also [36]) and the selected year of the time series. A closer look into clustering approaches indicates that these are not always robust and that their performance does not monotonically improve with the number of clusters but can have alternating patterns (e.g., [36, 42]).

Consequently, researchers and practitioners of complex energy system models face the challenge of identifying TSA methods suitable for their models and data. A detailed comparison of the TSA approaches comprising multiple years and scenarios is costly or not possible depending on the model complexity. Hence, a pure time series based evaluation of TSA approaches is favored. For example, Poncelet et al. [51] defined four indicators (e.g., normalized root mean squared error) to measure the degree of representation of aggregated time series. However, when comparing time series indicators and modeling results there is no strict correlative relationship between them: A good representation of defined time series indicators can result in poor model results and vice versa (see, e.g., [30, 36])[3].

[2] This section is based on the *identification paper* - Chapter 1 [39] and the *profiling paper* - Chapter 1 [40].

[3] For a more in-depth description of the state of research in relation to TSA for energy system modeling, see the *profiling paper* [40] - Chapter 2.

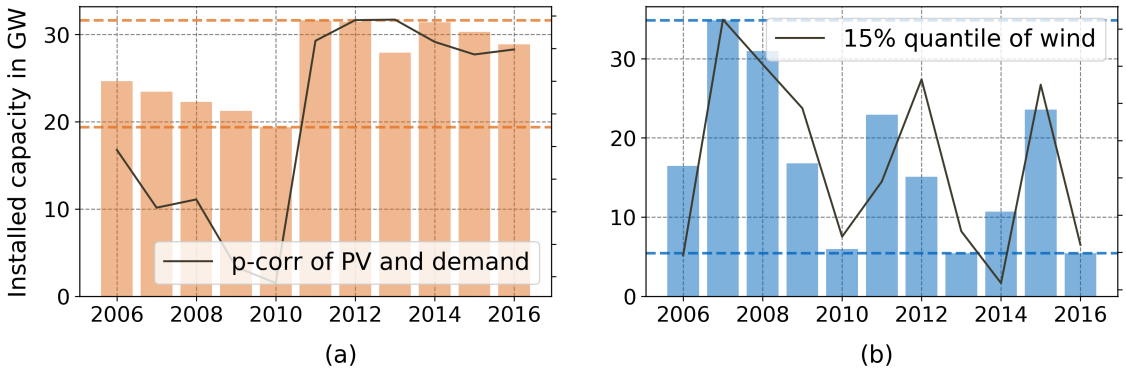

Figure 1.1: Resulting installed capacity of PV (scenario PV) and wind power (scenario WIND) with the Pearson's correlation between PV and the electricity demand and 15 % quantile of wind as the respective selected time series parameter.

1.3 Motivational example

The following example visualizes the complex interaction of time series and models. The data and model used are described in detail in Section 3.1. For now, we assume a one-node energy system with an electricity demand and four power generation technologies: PV and wind power as iRES as well as one flexible and one inflexible power plant. The energy system is described by a linear optimization model including expansion and dispatch planning. The demand as well as iRES are represented by annual time series (collectively referred to as time series bundle) with an hourly time resolution for in total eleven years. We model three different scenarios (i.e., PV, WIND, PV+WIND) to explore the impact of PV and wind power separately without pre-installed power generation capacities.

For each scenario and each year we get the installed capacities for the four power generation technologies as respective modeling result. Exemplary, the installed capacity of PV and wind power derived from the PV and WIND scenario is shown in Figure 1.1. In parallel, an exploratory data analysis discovers two parameters. First, the correlation between PV generation potential and electricity demand and second, the 15 % quantile of the wind time series. When comparing these annual values (shown as line in Figure 1.1) to the installed capacities of PV and wind power we find matching shapes, respectively. In other words, there is a correlation between the detected time series parameters and modeling results.

In a next step, we transfer this knowledge to the selection of a representative year in a very simplified way. Note, that this is equivalent to an aggregation of time series from eleven years into a one-year time series. We assume that the two parameters are relevant to project all capacities (including those of the flexible and inflexible power plant). For the PV and WIND scenario the respective parameters are taken to measure the representativeness of the annual time series bundle. For the PV+WIND scenario the parameters are summarized to one parameter per year. Across all years the average is calculated and the normalized absolute deviation of the annual parameter from the average is derived. The same procedure is done for the resulting installed capacities. Thus, the closer the summarized time series parameter or installed capacity is to zero, the more representative is the annual time series bundle or modeling result. However, visualizing these values for each year in Figure 1.2 shows that there

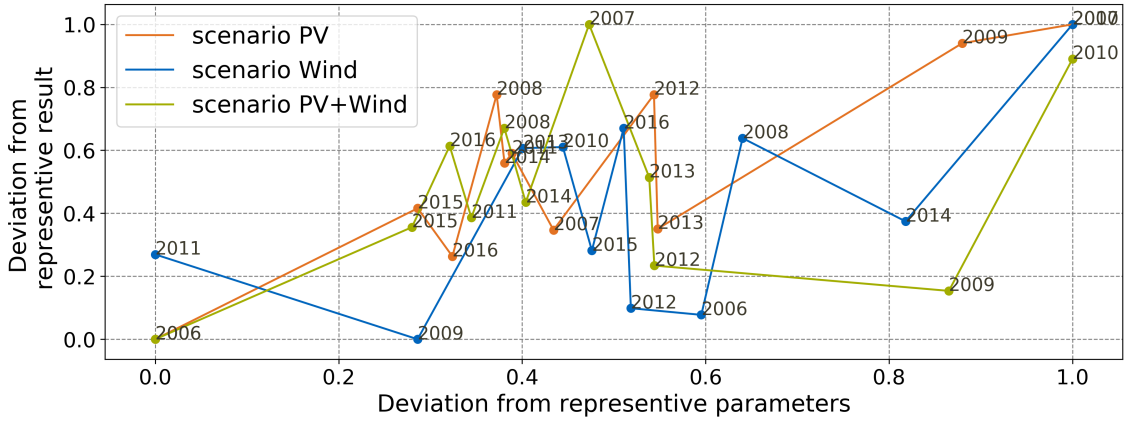

Figure 1.2: Comparing time series parameters and modeling results of three scenarios with regard to their representativeness. The representativeness is defined as deviation from the related mean.

is no clear link between parameters and model results. For example, in the PV and PV+WIND scenario, the most representative time series (2006, x = 0) also have the most representative result (y = 0). However, the second-best representative time series (2015, x = 0.29) leads to a significant deviation (y = 0.4). Similar or less deviations of modeling results can be achieved with lower representativeness of the time series parameters. For example, in the PV+WIND scenario, the year 2009 is the second-worst representative year regarding the parameters (x = 0.87), however, it results in the second-best modeling results (y = 0.15). In addition, there is no overall representative time series bundle. Rather, it depends on the scenario.

Combining the observations from the literature and the motivating example, two causes may explain the observed missing link between parameters and model results: First, the applied parameters do not represent (all) relevant time series characteristics, and second, the applied parameters are of different relevance.

1.4 Contribution

So far, we do not know which information of a time series is relevant in the context of energy system modeling that can also be used to measure the representativeness of time series. However, the relevant information of time series must be understood before developing and applying aggregation methods. A more systematic approach is required to better understand the interaction between time series and energy system models to identify relevant time series information that has to be included in aggregation and selection algorithms.

The following two research questions are derived in the thesis to close this gap:

RQ1: What are relevant time series information in the context of energy system modeling?

The identification of relevant information is based on a comprehensive analysis of time series. Various time series parameters are calculated to transfer the hidden information of a time series

into several one-dimensional values. By deriving a large variety of time series (obtained, for example, through aggregation and manipulation) a large data set of parameters and modeling results and allows to make statements about the relevance of individual parameters. Therefore, a mathematical model is developed based on clustering and nested regression (CNR). The model and its evaluation are presented in detail in the *identification paper: Feature Selection for Energy System Modeling: Identification of Relevant Time Series Information* [39].

RQ2: How can existing aggregation methods be extended to better represent relevant time series information?

Based on the identified parameters aggregated time series can be adjusted to better reflect the original information. Therefore, a profiling algorithm is developed that iteratively aligns the information in terms of time series parameters of the aggregated time series with that of the original time series. The performance of the profiling algorithm is independent of the aggregated time series and can therefore be considered a reasonable extension for aggregation approaches. The algorithm and its evaluation are presented in detail in the *profiling paper: Energy System Modeling with Aggregated Time Series: A Profiling Approach* [40].

Answering these questions also contributes to a comprehensive understanding of time series and their interaction with energy system models. Researchers and practitioners of energy system models thus gain further evidence to interpret results from a data perspective. To this end, the final discussion translates the findings on time series into a modeling recommendation that includes sensitivity analyses based on time series as well as a statistical evaluation of modeling results. Further possible applications, for example, the selection of representative time series from a growing amount of data, and current limitations of the proposed identification and profiling methods are discussed.

1.5 Outline

A brief overview of proposed method and resulting structure is given by Figure 1.3. The thesis can be divide into three parts:

Part I: Introduction and theory

- Chapter 1 provides an introduction to the complexity reduction of energy system models.
- Chapter 2 introduces the theoretical fundamentals that form the basis for CNR and profiling.

Part II: Method development

- Section 3.1 provides an overview of the underlying model and applied data.
- Section 3.2 describes the method of identifying relevant time series information (RQ1).
- Section 3.2 describes the profiling method based on identified time series information (RQ2).

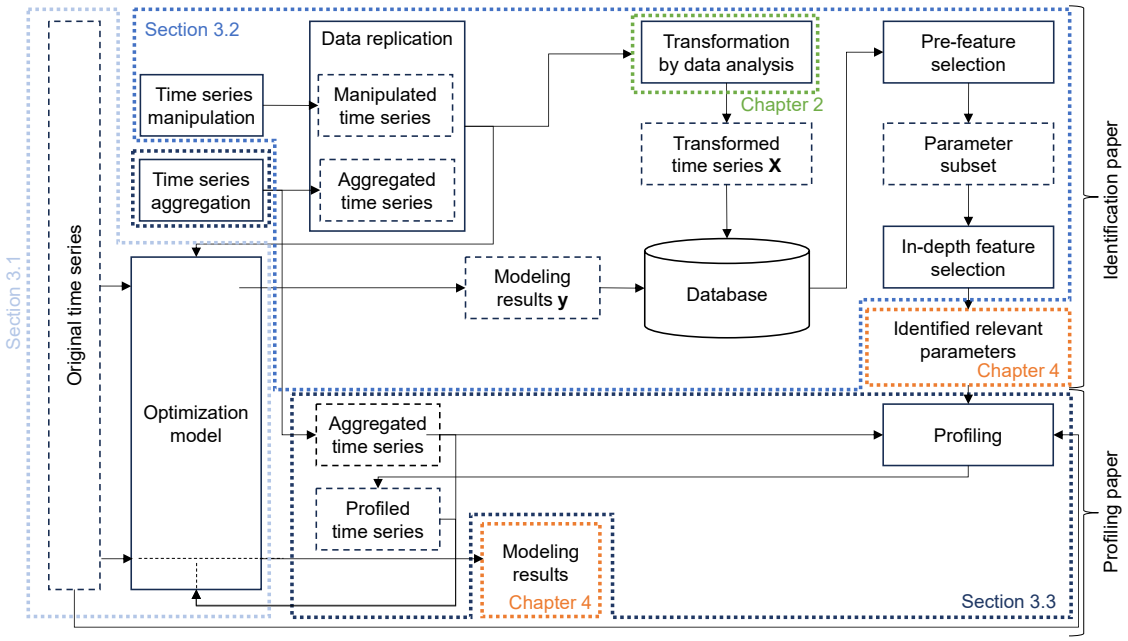

Figure 1.3: Brief overview of the proposed method that can be divided into the theoretical part (green), the method (blue) including the identification and profiling approach, and the evaluation (orange).

Part III: Evaluation and transfer

- Chapter 4 provides the results of identified time series parameters as well as modeling results of aggregated and profiled time series.
- Chapter 5 discusses, categorizes and transfers the results to further energy system modeling challenges.
- Chapter 6 concludes the discussed findings.

Chapter 2

Theoretical background

This chapter presents the theoretical basics of data analysis (methods) needed for the holistic understanding of both the identification of relevant time series parameters and the profiling of time series. Therefore, a nomenclature for the methods valid within this framework is introduced. Different terms and concepts of data analysis used in the literature are summarized in Section 2.1. Time series parameters and analyses that are relevant in this work are described in Section 2.2. Approaches beyond this are only roughly outlined or reference is made to further literature. In Section 2.3 and 2.4 the related fields of regression and feature selection are described in more detail. Finally, Section 2.5 provides a brief overview of clustering analysis.

2.1 Terms and concepts of data analysis

Data analysis is a broad field that has developed over many hundreds of years ("[...] since human life began" [73, p.4]). In its simplest form data analysis is based on collecting, documenting and aggregating information to use it inter alia for decision-making processes (e.g., records and surveys as population census [3], fiscal and military matters [43]). In recent years, the volume and complexity of data available has exponentially increased as well as the technologies to store and process the data (e.g., [18, 53]). Some even say that "[...] we are actually living in the data age" [26, p.1]. When applying classical statistical methods to the now amounted data, "[...] analysts could bring computers to their “knees” [...]" [43, p.30]. Thus, new approaches to data analysis have been developed leading to new methods (e.g., clustering, classification) and terms (e.g., data mining, knowledge discovery) that are briefly described in the following sections. A detailed description of terms and methods can be found in [26, 43, 73].
The field of data analysis has rapidly developed in recent years. According to [43] five phases of data analysis can be identified which are (1) the Classical Bayesian Statistics, (2) the Classical Parametric Statistics, (3) Machine Learning (4) Statistical Learning and (5) Distributed Analytical Computing. As a result, the analysis of data has turned into an interdisciplinary field dominated by statistic, machine learning (as a subset of artificial intelligence) and computer science [4, 26, 53]. Moreover, depending on the discipline, new terms have been introduced including data mining, knowledge extraction data analytics and knowledge discovery in databases (KDD) or knowledge discovery from data. The definitions of these terms are similar but have a

different focus derived from the perspective of the disciplines. However, the distinction between the terms becomes increasingly blurred.

Data analysis, data mining and knowledge discovery

In 1996, Fayyad et al. [19] introduced definitions for both, data mining and KDD aiming, for a unifying framework. KDD, as term often used by the machine learning community is defined as "[...] the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [19, p.83]. The process is further described as interactive and iterative procedure, which includes selection, preprocessing and transformation of data, data mining as well as interpretation and evaluation of the mined results. Thus, data mining, as term often used by statisticians and data analysts, "[...] is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data [...]" [19, p.83]. However, these definitions of data mining and KDD are not established but are continuously modified and extended. For example, data mining can be specified as an (semi-)automated process [43] or KDD can include additional steps like data cleaning and data integration [26], data exploration [43] or knowledge presentation [26]. Further data mining definitions clarify the applied methods and the intention behind as "[t]he use of machine learning algorithms to find faint patterns [...] in [...] data sets, which can lead to actions to increase benefit [...]" [43, p.17]. Moreover, the term data mining has become a synonym for KDD [26]. Thus, data mining is described as a process of discovering patterns that results in meaningful information [26, 73]. As a result, the short term KDD is also translated with knowledge discovery and data mining [43]. Further terms like data analytics are similar to the extended form of data mining. In [53] the definition of data analytics emphasizes "[...] the application of computer systems to the analysis of large data sets for the support of decisions" [53, p.2]. The process covers the preparation, pre-processing, analysis, and post-processing of data.

In summary, different definitions can be found in the literature, that are rarely discussed in the context of the general term data analysis. Within this thesis, the focus is on methods and less on the process. Thus, data mining methods are considered as a part of data analysis. They are particularly suitable for large amounts of data where patterns are to be discovered without a specific target or hypothesis (see [4]). Thus, data mining methods have an exploratory focus and are assigned here to the exploratory data analysis. In contrast, data analysis includes further parts such as descriptive and confirmatory analysis (e.g., [2, 4, 67]). Tuckey [67] also describes that for scientific and technical questions, the exploratory analysis forms a basis for the confirmatory analysis by, among other things, deriving the questions and the design. The data analysis approaches should therefore be combined rather than used separately. This can be transferred to the descriptive analysis, which can be a supplement by creating a basic understanding of the data.

Statistics and machine learning

Data analysis as an interdisciplinary field not only contains vague terms, the methods also overlap. The disciplines involved in data analysis are defined differently in the literature and include, for example, system theory [53], operation research [43] or data base systems [26].

However, statistic and machine learning are named in almost all literature but the question arises where statistics ends and machine learning begins.
Historically, statistics and machine learning have different origins. In simple terms, the focus of statistics lied in hypothesis testing, whereas machine learning - having the perspective from computer science or artificial intelligence in particular - concentrated on "[...] formulating the process of generalization as a search through possible hypotheses" [73, p.28]. However, different methods such as the classification and regression trees or the nearest-neighbor methods for classification have been independently developed in both disciplines so that statistical and machine learning methods have merged [73]. Methods such as regression and classification include aspects of both and cannot be assigned unambiguously to just one discipline. In other words "[...] between machine learning and statistics [...] there is a continuum — and a multidimensional [line] — of data analysis [methods]" [73, p.28]. Considering the continuum of statistics and machine learning the analysis methods presented in the following sub-sections are divided by functionality and not by disciplines. To supplement this, a general overview of statistics and machine learning is provided.
Statistics can be defined as science which applies and designs methods and models to collect, prepare and analyze data [3]. The statistical methods can be divided into three parts:

1. The descriptive statistics includes the representation of data in tabular and graphical form as well as the characterization of data as parameters, for example, mean and median [3, 18].
2. The exploratory statistics supplements the descriptive statistics [3]. The purpose is to discover new patterns or to generate new hypotheses. It is often applied when the suitable statistical model is unknown [18].
3. The inferential or inductive statistics includes probability theory. By applying stochastic models results derived from a data set are transferred to the population [18] under consideration of uncertainty [3].

In [26] machine learning is described as discipline that "[...] investigates how computers can learn (or improve their performance) based on data" [26, p.24]. Machine learning methods can be divided into four parts:

1. The supervised learning that deals with labeled data [26]. The outcome for a training data set is provided [73].
2. The unsupervised learning that only includes unlabeled data [26]. Thus, the outcome is not included in the data set.
3. The semi-supervised learning that is a mix of both supervised and unsupervised learning [26].
4. The active learning that involves the users [26].

Elements of data sets

So far, the focus has been on the analytical or methodological part of the term data analysis. In the following part, a shared understanding of data or data sets and their terminology is elaborated. "[A data set] is represented as a matrix of instances versus attributes [...]" [73, p.42].

$$\text{instances} \quad \underbrace{\begin{Bmatrix} \begin{bmatrix} x_{1,1} & \dots & x_{1,m} \\ \dots & \ddots & \dots \\ x_{n,1} & \dots & x_{n,m} \end{bmatrix} \end{Bmatrix}}_{\text{attributes}}$$

In general, an instance is individual and independent from other instances in the data set [73]. Other terms to describe the data points are, for example, observation, object, sample or example [26, 53]. Instances are described by attributes [73] which can be nominal (also named categorical, e.g., green, yellow), ordinal (e.g., small, medium, large) or numeric (also named metric) [3, 26]. A data set including one attribute is named univariate. A bivariate or multivariate data set includes two and more attributes, respectively [3, 26]. A univariate data set is equivalent to a data vector and is labeled in this thesis with a small, bold letter, for example, $\boldsymbol{x}$. Correspondingly, a multivariate data set can be understood as a matrix and is labeled with a large bold letter, for example, $\boldsymbol{X}$. An observation i of an attribute j or other scalar values are indicated with a small letter, for example, $x_{i,j}$. Further, data sets can be labeled or unlabeled. Unlabeled data sets only consist of attributes, whereas labeled data sets include attributes and the related outcome or class [26, 73]. Depending on the scope of application, synonyms for attributes are features [26, 73], (known, independent, exogenous) variables [3, 26] or predictors [27], whereas the outcome is also named response [27], (unknown, dependent, endogenous) variable [3]. A spacial format of data sets are time series. The instances are depended as they are recorded periodically or even equidistantly [62]. Within this thesis, only time series data sets or data sets derived from time series are applied. The data is discussed in more detail in the respective chapters.

2.2 Descriptive analysis

Various location and dispersion parameters can be calculated for univariate data, whereas for a bivariate data set correlation parameters can be determined [14]. Selected parameters relevant to this work are summarized from [3] and [14]. For a detailed description of descriptive analysis [3] and [14] are recommended. Note: A time series analysis is not considered in this thesis. Since a time series analysis seems obvious when using time series data, the relevance of time series analysis for this thesis is classified in Appendix A and an overview with reference to further literature is provided.

Location parameters

Location parameters represent the position of the entire data set or its distribution in compressed form [3]. In the following, we assume a one-dimensional data set as vector $\mathbf{x} = [x_1, x_2, \dots x_n]^T$ with n observations of one attribute. The *arithmetic mean (mean)* value is defined as:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.1}$$

The arithmetic mean is sensitive to statistical outliers. By calculating the *trimmed mean (tmean)*, in which α of the smallest and largest values are excluded, a robust mean value is obtained:

$$\overline{x}_\alpha = \frac{1}{n-2g} \sum_{i=1+g}^{n-g} x_i \tag{2.2}$$

with $g = int(\alpha n)$. In addition, the *geometric mean (gmean)*

$$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{2.3}$$

or *harmonic mean (hmean)* can be determined

$$\overline{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{2.4}$$

Furthermore, quantiles such as the median, can be calculated. The general expression for an α *% quantile* is:

$$x_\alpha = (1-f)x_q + f x_{q+1} \tag{2.5}$$

with $q = int(\alpha(n+1))$ and $f = frac(\alpha(n+1))$. The *median* (Q50) for an even or uneven n results in:

$$x_{50} = \tilde{x} = \frac{1}{2}(x_{\frac{n}{2}} - x_{\frac{n}{2}+1}) \;\; or \;\; \tilde{x} = x_{\frac{n}{2}+1} \tag{2.6}$$

Dispersion parameters

A visual overview of the scattering of parameters can be obtained using a histogram or boxplot. The latter includes the visualization of the *interquartile range (IQR)* defined as difference between the 25 % and 75 % quantile:

$$IQR = x_{75} - x_{25} \tag{2.7}$$

The empirical or theoretical *variance (var)* and *standard deviation (STD)* can be included in the histogram and are defined as:

$$var_{emp} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \;\; or \;\; var_{theor} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.8}$$

$$s_{emp} = \sqrt{var_{emp}} \;\; or \;\; s_{theor} = \sqrt{var_{theor}} \tag{2.9}$$

Note, that the empirical variance as mean of squared distances is equivalent to the 2. moment. The *k-th moment* is generally defined as:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k \tag{2.10}$$

The *skewness (skew)* (normalized 3. moment) and *kurtosis (kurt)* (normalized 4. moment) describe the shape of the distribution:

$$skew = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{S^3} \tag{2.11}$$

$$kurt = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{S^4} \tag{2.12}$$

If $skew > 0$, the distribution is right-skewed. With $kurt - 3$ the kurtosis is normalized according to the normal distribution (also called Fisher's kurtosis or excess). If $kurt > 3$ ($excess > 0$) the distribution is leptokurtic (higher peak). In addition to the mean of squared distances, the *mean absolute deviation (MAD)* can be determined in relation to the median (or mean):

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}| \tag{2.13}$$

Relation parameters

The description of a relation between two attributes can be divided into the type and strength of the relation. However, linear measures are commonly used and, if necessary, non-linear attributes are linearized by a transformation. Similar to the variance (Equation 2.8), the *covariance (cov)* can be determined for bivariate data set $\mathbf{X} = [\mathbf{x}, \mathbf{z}]$ with $\mathbf{x} = [x_1, x_2, ...x_n]^T$ and $\mathbf{z} = [z_1, z_2, ...z_n]^T$:

$$s_{\mathbf{xz}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z}) \tag{2.14}$$

By normalizing with the STD s_x and s_z we get the *Pearson's correlation (pcorr)*:

$$r_{\mathbf{xz}} = \frac{s_{xz}}{s_x s_z}, \quad -1 \leq r_{\mathbf{xz}} \leq 1 \tag{2.15}$$

In addition, the *Spearman's correlation (scorr)* and *Kendall's correlation (kcorr)*, which apply to ordinal data and can therefore also be used for metric data, can be calculated. As values in the data set can occur multiple times, ties need to be included. The scorr is defined as:

$$r_s = \frac{\sum_{i=1}^{n}[R(x_i) - \overline{R}(x)][R(z_i) - \overline{R}(z)]}{\sqrt{\sum_{i=1}^{n}[R(x_i) - \overline{R}(x)]^2 \sum_{i=1}^{n}[R(z_i) - \overline{R}(z)]^2}}, \quad -1 \leq r_s \leq 1 \tag{2.16}$$

with $R(\cdot)$ as list rank of sorted values and $\overline{R}(\cdot)$ as arithmetic mean of rank numbers. The kcorr is defined as:

$$r_k = \frac{P - I}{\sqrt{(\frac{n(n-1)}{2} - T)(\frac{n(n-1)}{2} - U)}}, \quad -1 \leq r_k \leq 1 \tag{2.17}$$

with P as sum of proversion pairs (concordant pairs, $R(z_i) < R(z_j), i < j$), I as sum of inversion pairs (discordant, $R(z_i) > R(z_j), i < j$) and T and U as length of ties of $\mathbf{x}$ and $\mathbf{z}$.

2.3 Regression analysis[1]

Regression analysis can be considered as an extension of correlation analysis (see Equation 2.17), in which not only a correlation but also a causal direction is determined. When multiple independent variables are used to describe a depended variable the regression is

[1] This section is partly based on the *identification paper* - Section 3.2 [39].

specified as multivariate regression [3, 14]. In the following, we assume a matrix of independent variables (also called feature matrix):

$$\mathbf{X} = \begin{bmatrix} x_{1,1} \cdots x_{1,m} \\ \vdots \ \ddots \ \vdots \\ x_{n,1} \cdots x_{n,m} \end{bmatrix} \tag{2.18}$$

and a vector as dependent variable

$$\mathbf{y} = (y_1, ..., y_n)^T \tag{2.19}$$

that consist of $i = 1, ..., n$ observations and $j = 1, ..., m$ attributes, that are referred to as features in the following. The associated linear regression model is:

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i \tag{2.20}$$

where β_0 is the offset coefficient, vector $\boldsymbol{\beta} = (\beta_1, ...\beta_m)^T$ represents the coefficients and vector $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_m)^T$ is the error. Equation 2.20 can be transformed to the Ordinary Least Square estimator (OLS estimator) to obtain the regression coefficients $\hat{\beta}_{OLS}$ by minimizing the residual sum of squares:

$$\hat{\beta}_{OLS} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 \right\} \tag{2.21}$$

An extended linear regression model is ridge regression as a shrinkage method introduced by [28]. The shrinkage of coefficients is achieved by linking the coefficients with a penalty term of L_2 norm $\|\boldsymbol{\beta}\|_2 = \sum_j (\beta_j)^2 \leq t$. The resulted ridge estimate $\hat{\beta}_R$ is:

$$\begin{aligned} \hat{\beta}_R &= \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 \right\} s.t \sum_j (\beta_j)^2 \leq t \\ &= \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_j (\beta_j)^2 \right\} \end{aligned} \tag{2.22}$$

where $t \geq 0$ and $\lambda \geq 0$ are interrelated tuning parameters controlling the shrinkage.

Different criteria can be used to evaluate the regression model determined (see, e.g., [59]) and are usually included in the output of implemented statistical packages such python statsmodels package [57]. The coefficient of determination (R^2) indicates the proportion of explained variance):

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{2.23}$$

where $\hat{y}$ is the estimate of the regression model and $\overline{y}$ the average of y. The mean absolute error (MAE) ca be defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{2.24}$$

Further, the goodness of fit can be set in relation to the number of features m represented by the Akaike's Information Criterion (AIC):

$$AIC = \log\left(\frac{\sum_{i=1}^{n} \hat{y}_i - \overline{y}_i}{n}\right) + \frac{n+2m}{n} \tag{2.25}$$

or the Bayesian Information Criterion (BIC):

$$BIC = \log\left(\frac{\sum_{i=1}^{n} \hat{y}_i - \overline{y}_i}{n}\right) + \frac{m \log n}{n} \tag{2.26}$$

Besides model related criteria, the significance of an features is represented by the p-value (p) using a two-tailed t-test.

2.4 Feature selection[2]

The OLS estimate $\hat{\beta}_{OLS}$ (Equation 2.21) returns coefficients for all features, whereas the ridge estimate $\hat{\beta}_R$ (Equation 2.22) shrinks some coefficients but does not set them to zero. For a large number of features whose relevance is not clear, the use of these estimates is not useful. Instead, feature selection can be applied. "Feature selection can be defined as the process of detecting the relevant features and discarding the irrelevant and redundant ones with the goal of obtaining a subset of features that describes properly the given problem with a minimum degradation of performance" [8, p.14]. Feature selection methods can be categorized by two criteria (see [77]). Firstly, the methods are divided into three categories according to their identification algorithm: filter models (e.g., Chi-Squared, Correlation-Based Feature Selection [8]), wrapper models (e.g., Wrapper Subset Eval, see [8] and [25]) and embedded models (e.g., Recursive Feature Elimination for Support Vector Machines, see [8] and [24]). Methods involving a filter model focus on characteristics of features usually assuming that they are independent and the evaluation of features does not include any learning algorithm (classifier). As consequence, these methods are often fast, simple and easy to understand. Both, wrappers and embedded methods, include a learning algorithm in the selection process which is implemented separately or integrated. Unlike filters, wrapper and embedded methods consider dependencies between features [8, 77]. Secondly, feature selection methods are categorized by their outcome which is either an ordered ranking of all features (named feature weighting algorithms [77], for example, correlation coefficient, multivariate methods, [43]) or a subset of relevant features (named subset selection algorithm, for example, almost all wrappers [77]). In contrast to feature weighting algorithms, subset selection approaches consider the interaction between features and evaluate those in context of each other [23]. In the following, three methods from the literature and applied within this thesis that are describes in more detail: LASSO [65], LASSOLARS and ElasticNet [80]. These methods are derived from linear regression and have been widely applied in the energy research field achieving improved (prediction) results (e.g., [33, 35, 37, 79, 78])[3].

[2] This section is based on the *identification paper* - Chapter 2 and Section 3.2 [39].

[3] For a more in-depth description of the state of research in relation to feature selection in the energy research field, see the *identification paper* [39] - Chapter 2.

LASSO and LASSOLARS

LASSO (*least absolute shrinkage and selection operator*) is an embedded feature selection method introduced by [65]. LASSO is based on linear regression and OLS estimate and combines the advantage of general feature selection models (interpretability) and ridge regression (stability). By replacing the L_2 penalty term of Equation 2.22 by a L_1 penalty term $\|\beta\|_1 = \sum_j |\beta_j| \leq t$ the coefficients of features with low benefit on RSS are set to zero. In other words, features with non-zero coefficients present the selected feature subset [65]. The resulted LASSO estimator $\hat{\beta}_L$ represents a quadratic optimization problem with linear inequality constraints and can be expressed as follows:

$$\begin{aligned}\hat{\beta}_L &= \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 \right\} s.t \sum_j |\beta_j| \leq t \\ &= \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j| \right\}\end{aligned} \tag{2.27}$$

With $\lambda = 0$ ($t \to \inf$), Equation 2.27 represents the standard OLS estimator providing coefficients for all features, whereas with $\lambda \to \inf$ ($t = 0$) all coefficients are set to zero [65]. Although, LASSO is built on linear regression using the OLS estimator, the method is more general and can be applied in other statistical models such as generalized regression models or tree-based models as well [65]. Moreover, regarding to [68] the LASSO regression itself can be considered as a generalization of a linear regression model. In practice, different optimization algorithms are used to solve the LASSO estimator. In this thesis the python scikit-klearn package [46] is used which applies – inter alia – the coordinate decent algorithm and the Least Angle Regression (LARS) [15] (see also [27]). Both algorithms can involve cross validation (CV) for selecting the best model by setting λ. Note: The LASSO estimate in the [46] is defined as:

$$\hat{\beta}_{L^*} = \arg\min \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j| \right\} \tag{2.28}$$

with $\bar{y} = \beta_0 = 0$. Corresponding to [27] there is no difference between using the factors $\frac{1}{2n}$, $\frac{1}{2}$ or 1, but it leads to different λ and influences the comparability of different data set sizes.

Elastic Net

Empirical studies have shown that LASSO does not provide stable results when features are highly correlated [27]. This behavior is improved by adding a L_2 norm as second penalty term on the regression coefficients of LASSO leading to the ElasticNet approach developed by [80]. As a result, ElasticNet combines both, LASSO and ridge regression and thus theirs strengths: Removing features with low relevance (LASSO) and being more robust to correlated features (ridge regression) [35, 80]. By adding the L_2 norm to Equation 2.27, we get the ElasticNet

estimate $\hat{\beta}_E$:

$$\begin{aligned}\hat{\beta}_E &= \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right\} \\ &\qquad s.t.\ \alpha\sum_j|\beta_j| + (1-\alpha)\sum_j(\beta_j)^2 \leq t \\ &= \arg\min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right. \\ &\qquad \left.+\lambda\left(\alpha\sum_j|\beta_j| + (1-\alpha)\sum_j(\beta_j)^2\right)\right\}\end{aligned} \tag{2.29}$$

where $\alpha \in [0,1]$ wights the the L_1 and L_2 penalty terms. With $\alpha = 0$ the ElasticNet estimate becomes a LASSO estimate and $\alpha = 1$ leads to ridge regression. In the python scikit-klearn package [46] the coordinate decent algorithm is used to solve the ElasticNet estimator. Like LASSO, the algorithm can involve cross validation (CV) for selecting the best model by setting λ. Note: The ElasticNet estimate in the [46] is defined as:

$$\begin{aligned}\hat{\beta}_{E*} = \arg\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m}\beta_j x_{ij}\right)^2\right. \\ \left.+\lambda\left(\alpha\sum_j|\beta_j| + \frac{1-\alpha}{2}\sum_j(\beta_j)^2\right)\right\}\end{aligned} \tag{2.30}$$

Besides the factor $\frac{1}{2n}$ an additional term $\frac{1}{2}$ is added for the L_2 penalty. Corresponding to [69] this improves the efficiency and intuitiveness of the soft-thresholding operator in the optimization.

2.5 Clustering analysis[4]

"Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets" [26, p.444]. The observations are grouped by maximizing the similarity within a cluster and minimizing the similarity between clusters [26]. The distance function $dist(x_i, x_j)$ is usually used to quantify similarity between to objects $\mathbf{x}_i$ and $\mathbf{x}_j$ [18]. Clustering analysis is applied on unlabeled data sets, thus, the intention is to discover classes and find labels [26, 53]. Common clustering algorithms are, for example, partitioning methods (e.g., k-means approach) and hierarchical methods (e.g., single-linkage approach) [26]. For the following description of clustering, we can assume a $(n\ x\ m)$ matrix:

$$\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_n] = \begin{bmatrix} x_{1,1}\cdots x_{1,n} \\ \vdots\ \ddots\ \vdots \\ x_{m,1}\cdots x_{m,n}\end{bmatrix} \tag{2.31}$$

[4]This section is partly based on the *profiling paper* - Section 3.2 [40] and is supplemented by [26] if not specifically stated.

k-means clustering

The k-means clustering approach uses the centroid of clusters to quantify the similarity within a cluster and assigns observations $\mathbf{x}_i$ to a cluster. The centroid $\overline{\mathbf{x}}_j$ of a cluster j can be defined by the mean of the observations:

$$\overline{\mathbf{x}}_j = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i \tag{2.32}$$

where C_j is the set of observations of cluster j, $|C_j|$ the number of observations and i an observation in cluster $|C_j|$. The clusters are determined by minimizing the Euclidean distances (or L_2 norm, $||\cdot||_2$ simplified to $||\cdot||$) between the observations $\mathbf{x}_n$ and the centroid $\overline{x}$ across all clusters C, thus, similarity within on cluster is maximized:

$$\min \sum_{j=1}^{c} \sum_{i \in C_j} ||\mathbf{x}_i - \overline{\mathbf{x}}_j||^2 \tag{2.33}$$

Equation 2.33 also represents the *sum of squared error (SSE)*, which is used as an evaluation criterion, for example, to identify an appropriate number of clusters.
The k-mean algorithm mainly consists of two steps: The (re)assignment of the observations to a cluster and the (re)calculation of the centroids. In a first step c observations are randomly selected. They each represent a cluster and serve as initial centroid. The remaining observations are assigned to the most similar cluster according to Equation 2.33. When all observations are assigned, the centroids are recalculated and the observations are reassigned. The iterative process of reassignment and centroid recalculation continuous until the assignment is stable. Note, in [18] an alternative algorithm is described that recalculates the centroid whenever an object changes its cluster affiliation. Finally, the representative for each cluster can be derived as centroid $\mathbf{x}_j^*$ or closest to the centroid $\mathbf{x}_j^{**}$:

$$\mathbf{x}_j^* = \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i \quad \vee \quad \mathbf{x}_j^{**} = \mathbf{x}_{i^*} \;\; , i^* = \underset{i \in C_j}{\arg\min} \left|\left| \mathbf{x}_i - \mathbf{x}_j^* \right|\right| \tag{2.34}$$

In addition to k-means clustering, the k-medoid clustering should also be briefly mentioned. In this method, the medoid $\mathbf{x}_j^m$, i.e. an existing central object of a cluster, is used to calculate the distance instead of the centroid. The absolute error criterion is used as the distance measure $dist(\mathbf{x}_i, \mathbf{x}_j^m) = \left|\left| \mathbf{x}_i - \mathbf{x}_j^m \right|\right|$. This avoids the disadvantage of the k-means method of being sensitive to outliers.

Hierarchical clustering

The hierarchical clustering can be either following the agglomerative (bottom-up) or divisive (top-down) approach. With the agglomerative approach, each observation initially represents a cluster. Step-by-step, most similar clusters are merged until all observations are in one cluster or a termination criteria is fulfilled, for example, a pre-defined SSE. The divisive approach is reversed, thus, in the beginning all observations are in one cluster and in each step a cluster is split into sub-clusters. The process ends when all observations are in separate clusters or, as before, a termination criterion is met.
To evaluate similarity different measures based on distance (linkage measures) or density and

continuity can be applied. Common measures to define the distance between two clusters C_k and C_l are the minimum distance (single linkage):

$$dist_{min}(C_k, C_l) = \min_{i \in C_k, j \in C_l} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{2.35}$$

the average distance (average linkage):

$$dist_{min}(C_k, C_l) = \frac{1}{|C_k||C_l|} \sum_{i \in C_k, j \in C_l} ||\mathbf{x}_i - \mathbf{x}_j|| \tag{2.36}$$

or the Ward's method [72] and described in [42, 53] that measures the increase of SSE (see Equation 2.33) when two clusters are merged:

$$dist_{min}(C_k, C_l) = \frac{|C_k| \cdot |C_l|}{|C_k| + |C_l|} \left|\left| \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{x}_i - \frac{1}{|C_l|} \sum_{j \in C_l} \mathbf{x}_j \right|\right|^2 \tag{2.37}$$

Besides the partitioning and hierarchical clustering, further approaches such as density-based methods (for each observation or data point a minimum number of other data points must exist within a defined radius) or grid-based method (as an extension of the other methods) exists.

Chapter 3

Methodology

This chapter describes the methods developed and applied in this thesis. As CNR and profiling are using the same energy system model and data basis, the model and data are presented in the following section before the individual methods are described in more detail.

3.1 Underlying model and data[1]

Five requirements are derived from the literature findings and the motivating example described in Section 1.2 and Section 1.3 for the underlying energy system model and data basis of this thesis:

1. *High modeling flexibility*: Changes in time series characteristics affect modeling results, which can be mitigate by restrictive constraints, for example, a defined minimum share of iRES [47, 51] as well as a minimum capacity [51], maximum capacity [42] or maximum energy generation [36] of selected power generation technologies. In the underlying model, constraints that severely limit the solution space are omitted, allowing for extreme model results (e.g., 0 % iRES). In addition, the utilized greenfield approach enables the system to be reassembled by eliminating pre-installed technologies (e.g., [42]).
2. *No energy storages*: Energy storages are not considered in this model. Results in [47] indicate a mitigating effect of storages on the scatter of modeling results. For example, the installed capacity of wind offshore varies across 25 years between approx. 51 and 80 GW. By adding energy storages, the scatter is reduced to approx. 32 and 49 GW.
3. *Unbundling of iRES*: The focuses is on wind power and PV, as these technologies are significantly expanded and account for a large share of iRES in the energy system [44]. iRES are modeled separately to specifically understand the interaction between individual time series and the energy system model. The findings from simplified models are applied to a complex model to demonstrate their transferability. Thus, three energy system scenarios are derived to model PV and wind power separately (scenario PV and WIND) and in interaction (scenario PV+WIND). Consistent with (1) and (2), a single-node model is optimized to exclude countervailing effects between regions with the respective iRES.

[1] This section is based on the *profiling paper* - Section 3.1 [40].

4. *Modeling multiple years*: Time series of eleven years (2006-2016) are used. Results in [47] show high variability of optimized costs (LCOE) depending on the selected TSA approach and year. For example, the costs deviations for k-means (5 days, closest) varies between approx. –60 % (2011) and +40 % (2010).
5. *Installed capacity as modeling criteria*: Aggregated time series are compared using installed capacities instead of costs. Firstly, costs can be recalculated by splitting the optimization problem into expansion and dispatch planning (e.g., [6, 51, 63]). Secondly, aggregated time series resulting in low cost deviations can have high differences in system configuration. For example, in [47] the optimized costs (LCOE) of a TSA approach combining heuristic and k-means (ten days, closest) deviates by -0.2 %, whereas the installed capacity of wind power deviates by -10.6 %. Thus, similar system configurations lead to similar costs, whereas similar costs can be derived by different system configurations.

For modeling the energy system according to the described requirements, the linear optimization model *urbs* [34] is used. The open-source model combines unit commitment and expansion planning for multi-commodity energy systems and is written in python using pyomo as optimization modeling language. The optimization is composed of an objective function minimizing the total system costs and linear constraints reflecting technical, economic, and political interrelationships and restrictions. The standard form of a linear programming problems is:

$$\begin{aligned} &min\ \mathbf{c}^T\mathbf{x} \\ &subject\ to: \\ &\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0 \end{aligned} \tag{3.1}$$

where $\mathbf{c}^T\mathbf{x}$ describes the objective function and $\mathbf{Ax} = \mathbf{b}$ represents the constraints of the model. The vector $\mathbf{x}$ is the variable of the model and includes the installed capacities of the four power generation technologies as the central result for this thesis. A schematic overview of the considered power system is shown in Figure 3.1. Besides PV and wind power as iRES, one peak-load power plant (flexible, FPP) and one base-load power plant (inert, IPP) are modeled.

The time series for the electricity demand [16] and power generation potential from PV and wind power [49] are selected for Germany comprising the years 2006 to 2016. The electricity demand is described by absolute values (unit MW), whereas the generation potential of iRES is represented by normalized values (unit MW / MW_{inst}). To build the energy system model further input parameters, for example investment costs and efficiencies are needed. The values of these parameters, shown in Table 3.1, are derived from [52] aiming for a significant solution space. This is characterized by an expansion of all technologies when using the original time series as model input. Thus, changing characteristics of the aggregated time series can result in extreme system configurations of both directions (e.g., one technology is not considered or double represented) and thus, become analyzable.

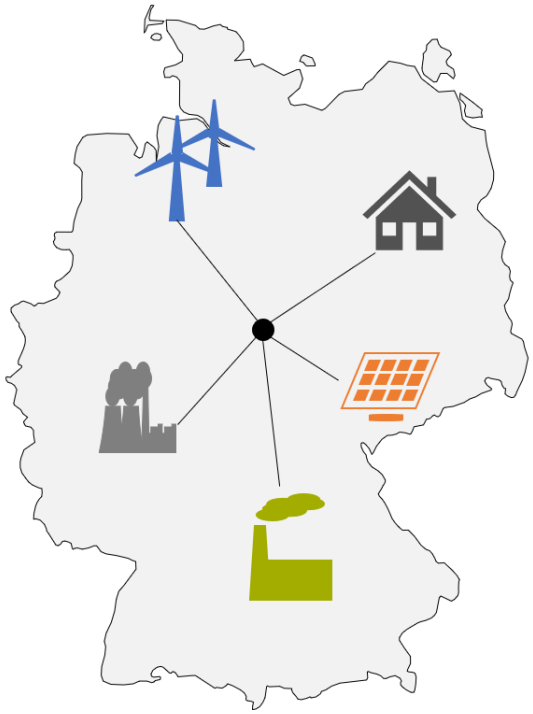

Figure 3.1: The energy system model including PV (orange), wind power (blue), a (flexible) peak-load power plant (FPP, green) and an inert base-load power plant (FPP, gray) to meet the electricity demand (dark gray).

Table 3.1: Input parameter for the optimization model *urbs*, which are derived from [52] for PV, wind power, FPP, and IPP.

		PV	Wind power	FPP	IPP
Costs					
Investment	EUR/MW	1,000,000	1,300,000	700,000	4,500,000
Fixed	EUR/MW/a	10,000	26,000	21,000	270,000
Variable	EUR/MWh	-	-	4	5
Operation	EUR/MWh	-	-	60	5
technical					
Efficiency	%	-	-	60	37
Must run	%	-	-	25	80

3.2 Identification of relevant time series parameters[2]

The presented method for identifying relevant time series parameters is based on the hypothesis that, in accordance with information theory [58], time series can be interpreted as information carriers with low information density. The information is implicit and needs to be converted into time series parameters. It is also assumed that these parameters explain the result of the energy system model. Thus, feature selection can be applied to identify relevant parameters and eliminate irrelevant ones.

In the energy field, feature selection methods are usually applied to improve the performance of prediction models (see also [71]), for example, for energy demand (e.g., [13, 17, 61, 76]), energy prices (e.g., [1, 37, 68]) or energy generation – in particular from wind sources (e.g., [20, 45, 75]) and solar sources (e.g. [33, 50]). These applications have in common that feature selection and the final (prediction) model base on the same data and that selected features serve as direct input for the final model[3]. However, this does not apply to the application in this thesis so that other requirements are placed on feature selection.

Note: In the following time series parameters are referred to as features and the modeling results as response.

3.2.1 Requirements

Five challenges for the feature selection can be derived from the application of time series aggregation or, more precisely, the profiling of aggregated time series:

1. The feature selection and the final profiling are only indirectly linked by the energy system model and selected (or identified) features.
2. The feature selection and the profiling do not build on the same database.
3. The number of selected features must be limited as the time series aggregation and profiling is more complex.
4. The complexity of the profiling method depends not only on the number of features but also on the feature itself.
5. Due to the complexity in (3) and (4) the "best" selected features can be infeasible and rejected by the profiling method and another feature subset or a reduced subset needs to be applied.

Thus, not only identifying relevant features is essential but also a deeper understanding of the effect of removing single features from a feature subset afterwards. So far, existing feature selection methods do not cover all requirements. For example, filter methods provide a ranking of features but exclude feature dependencies. Wrappers and embedded methods usually provide a subset of features. However, the number of selected features can only be set indirectly and therefore iteratively by model settings (e.g., LASSO, where the λ parameter controls the penalty strengths and thus the shrinkage of the features [78]). Moreover, the evaluation of a selected feature subset including subsequent elimination of features or alternative feature sets is expensive for the same reason.

[2] This section is based on the *identification paper* - Chapter 2, Section 3.1 and Section 3.3 [39].

[3] For a more in-depth description of the state of research in relation to feature selection in the energy field, see the *identification paper* [39] - Chapter 2.

3.2.2 Database

To apply feature selection, the required extensive database is derived from the eleven original time series bundles (i.e., electricity demand, PV and wind power generation potential) and an energy system model, both described in Section 3.1. The database includes two parts, which are time series parameters summarized in a feature matrix **X** and the modeling results represented by the response **y**. The relation between features **X** and model responses **y** is given as $\mathbf{y} = f(\mathbf{X})$, whereas each time series bundle represent one observation. With n = 11 observations the database is not meaningful. Thus, the time series bundle is transformed by manipulating the time series bundle (e.g., shifting time series against each other) and by aggregating time series by methods such us clustering and downsampling. This results in different time series characteristics leading to a different modeling result. Overall 9000+ time series bundles are received.

In addition to the time series of the bundle, further time series can be constructed in the form of the residual load. This allows to analyze the interdependence of time series not only by parameters such as correlation but also by calculating location and dispersion parameters of the residual load. Due to an unknown energy system in terms of installed capacity, we can only approximate the residual load by transforming the electricity demand into normalized values using the overall maximum.

Thus, two pre-processing steps are performed before calculating various time series parameters: First, the time series as vector $\mathbf{z}_i$ are normalized by their inter-maximum value of the original time series. As the generation potential of iRES is already normalized the following equation only applies for the electricity demand (dem):

$$\bar{\mathbf{z}}_i = \frac{\mathbf{z}_i}{\max_{t\in T, y\in Y}(z_{i,y,t})} \quad with \;\; i \in [dem], \;\; Y = \{2006, ..., 2016\}, \;\; T = \{1, ..., 8760\} \tag{3.2}$$

Second, the residual load (res) for the PV and Wind scenario s (see Section 3.1) is calculated by subtracting the respective generation potential of iRES from the normalized electricity demand:

$$\bar{\mathbf{z}}_{res} = \bar{\mathbf{z}}_{dem} - \sum_i \bar{\mathbf{z}}_i \;\; with \;\; i \in iRES(s), \;\; s \in \{PV, Wind\} \tag{3.3}$$

For each extended time series bundle 170+ features are calculated using the descriptive parameters described in Section 2.2 and parameters re-calculated with each other (e.g., IQR relative to the mean). Furthermore, additional time series parameters are derived, for example, when calculating extreme values of a time series the related value of other time series are considered:

$$t^* = \arg\max_{t\in T}(\bar{z}_{i,t}) \quad \vee \quad t^* = \arg\min_{t\in T}(\bar{z}_{i,t}) \tag{3.4}$$

$$x_i^* = \bar{z}_{i,t^*} \quad \forall \quad i \in \{dem, iRES, res\} \tag{3.5}$$

The overall feature matrix **X** has a dimension of $n > 9000$ and $m > 170$. Finally, features are normalized by applying the min-max-scaler:

$$\mathbf{x}_j = \frac{\mathbf{x}_j - min(\mathbf{x}_j)}{max(\mathbf{x}_j) - min(\mathbf{x}_j)} \tag{3.6}$$

where $\mathbf{x}_j = (x_{1,j}, ..., x_{n,j})^T$ is the vector of feature j with n observations. A visual overview of the all features is provided in Appendix B

The response **y** is calculated by the energy system model *urbs* using the 9000+ time series bundles as input data. For the PV and Wind scenario we receive the installed capacity for the FPP, the IPP and the included iRES. More precisely, the received response is a matrix. Thus, feature selection is performed three times for each scenario with the modeling result of individual installed capacity as response.

3.2.3 CNR algorithm

The proposed method CNR is based on linear regression and involves clustering to handle correlated data and nested modeling to link sub-models. These method components form the name of CNR as *clustering and nested based regression*. The algorithm of CNR includes two sections: (1) pre-feature selection to screen the features and speed up the algorithm, and (2) in-depth feature selection with a detailed procedure to receive the final results. Both parts involve clustering and nested modeling. In the following, the clustering and nested modeling approach is described first. Afterwards, the algorithm of both, the pre-feature and in-depth feature selection, and the respective evaluation are described.

Clustering

Starting point of the clustering are the initial features F or a subset of selected features $F^* \subseteq F$. The feature matrix **X** is filtered by the selected features F^*. The remaining features $\mathbf{X}^*$ are clustered into c groups by applying the k-means algorithm (TimeSeriesKMeans [64]) using the Euclidean distance to measure the similarity (or disparity) of the features (see Section 2.5). Figure 3.2 shows an excerpt from the clustered features with the observations on the x-axis and the time series parameter value on the y-axis.

Nested regression

The nested regression approach allows the evaluation of features and feature subsets based on the performance of their regression models. The starting point is a parent regression model that contains a defined maximum number of features. The regression is conducted and a performance index *PI* of the model is calculated (see Equation 3.7). By excluding iteratively one feature from the model, feature subsets as children are generated and the regression as well as the performance calculation are repeated.

Example 4.1
Figure 3.3 provides a simple example of three features $F = \{a, b, c\}$ that form the parent model. The resulting performance of the model is assumed to be $PI_{(a,b,c)} = 5$. In the first iteration, three sub-model combinations are derived, that are (b, c), (a, c) and (a, b). The resulted performances are assumed to be $PI_{(b,c)} = 3$, $PI_{(a,c)} = 4.5$ and $PI_{(a,b)} = 4$, respectively. In the second and last iteration, we receive three sub-models including only a single feature a, b and c. The performance of the models is assumed to be $PI_c = 2$, $PI_b = 2.5$ and $PI_a = 3$.

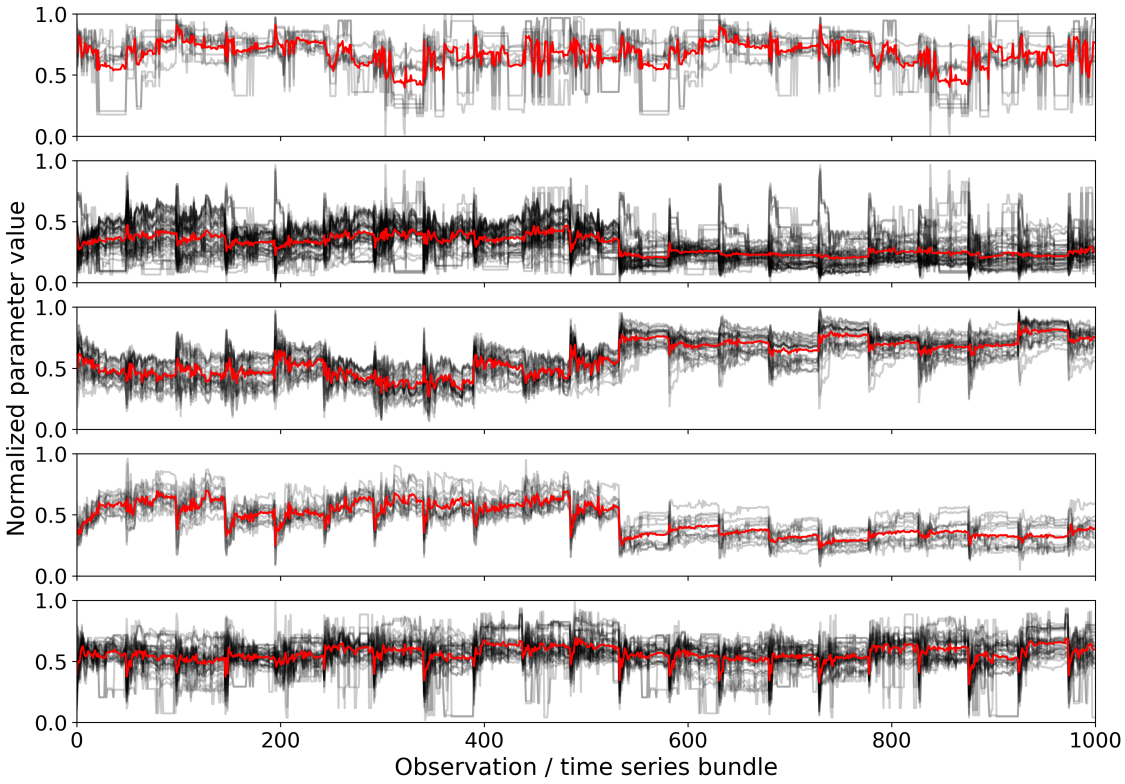

Figure 3.2: Excerpt from clustered features including the cluster centroid (red) with the observations on the x-axis and the time series parameter value on the y-axis.

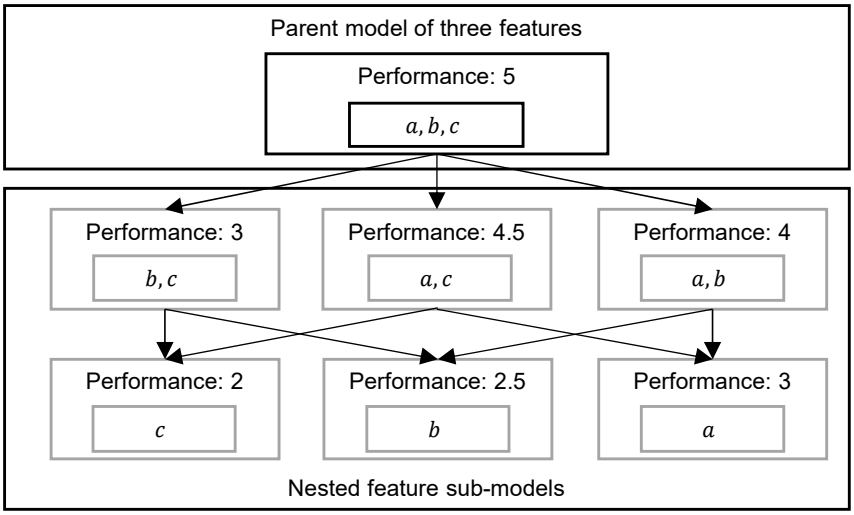

Figure 3.3: Example of the nested modeling approach for a starting model of three features and three derived sub-model combinations. By comparing the performances the significance of excluded features is determined.

In a next step, the excluded features are evaluated based on the performance change ΔPI, respectively. The performance change can be interpreted as significance of the feature described by a significance index (*SI*). As the previous example shows, individual parameters are excluded several times so that an average value is ultimately determined (see Equation 3.8).

Example 4.2
By excluding feature a from the parent model, the model performance decreased by $\Delta PI_{(a,b,c)\setminus a} = \Delta PI_{(a,b,c)} - \Delta PI_{(b,c)} = 2$ indicating that feature a is significant. However, removing feature b the model performance decreases by $\Delta PI_{(a,b,c)\setminus b} = 0.5$ indicating a low significance. The average significance of a results in $SI_a = \Delta\overline{PI}_a = \frac{1}{3}((5-3)+(4.5-2)+(4-2.5)) = 2$.

Besides single features, feature subsets are evaluated in a similar way (see Equation 3.9). Thereby, the performance of a model is compared with the performance of its sub-models that together form a coherent nested model combination.

Example 4.3
Overall, six coherent nested model combination can be derived in reverse order and evaluated by an average performance:

1. $(c) \rightarrow (b,c) \rightarrow (a,b,c)$: $\frac{2+3+5}{3} = \frac{10}{3}$
2. $(b) \rightarrow (b,c) \rightarrow (a,b,c)$: $\frac{2.5+3+5}{3} = \frac{10.5}{3}$
3. $(c) \rightarrow (a,c) \rightarrow (a,b,c)$: $\frac{2+4.5+5}{3} = \frac{11.5}{3}$
4. $(a) \rightarrow (a,c) \rightarrow (a,b,c)$: $\frac{3+4.5+5}{3} = \frac{12.5}{3}$
5. $(b) \rightarrow (a,b) \rightarrow (a,b,c)$: $\frac{2.5+4+5}{3} = \frac{11.5}{3}$
6. $(a) \rightarrow (a,b) \rightarrow (a,b,c)$: $\frac{3+4+5}{3} = \frac{12}{3}$

Thus, the 4. coherent model combination is the best result with the features and feature subsets ordered by the relevance.

The implemented interaction of clustering and nested modeling is shown in Figure 3.4. The considered features F^* are clustered into c groups. Thus, the number of clusters defines the size of the parent model that is $s = c$. All feature combinations cc are calculated by selecting one feature from each cluster. For each parent model, the regression is performed and the performance *PI* is calculated. The regression model applied is a multivariate, linear regression model using Generalized Least Squares (GLS) to estimate the model coefficients. Therefore, the GLS algorithm of statsmodels [57] is implemented. For T top models fulfilling a defined performance threshold the sub-models are derived. As before, the regression is performed for each (sub-feature) combination cc. Based on the (sub-model) performance *PI*, the best T models are selected, and their sub-models are derived. This process continues until the last regression of models with size $s = 1$ (one feature). In the end, the nested modeling results are summarized and evaluated.

The evaluation is derived from the law of parsimony, also known as Ockham's razor (see [31]). According to [7, 38, 56] the evaluation of model fit, especially when comparing nested models, should be based on several performance indices simultaneously. Thus, the performance index *PI* consists of five criteria of model fit representing significance, accuracy

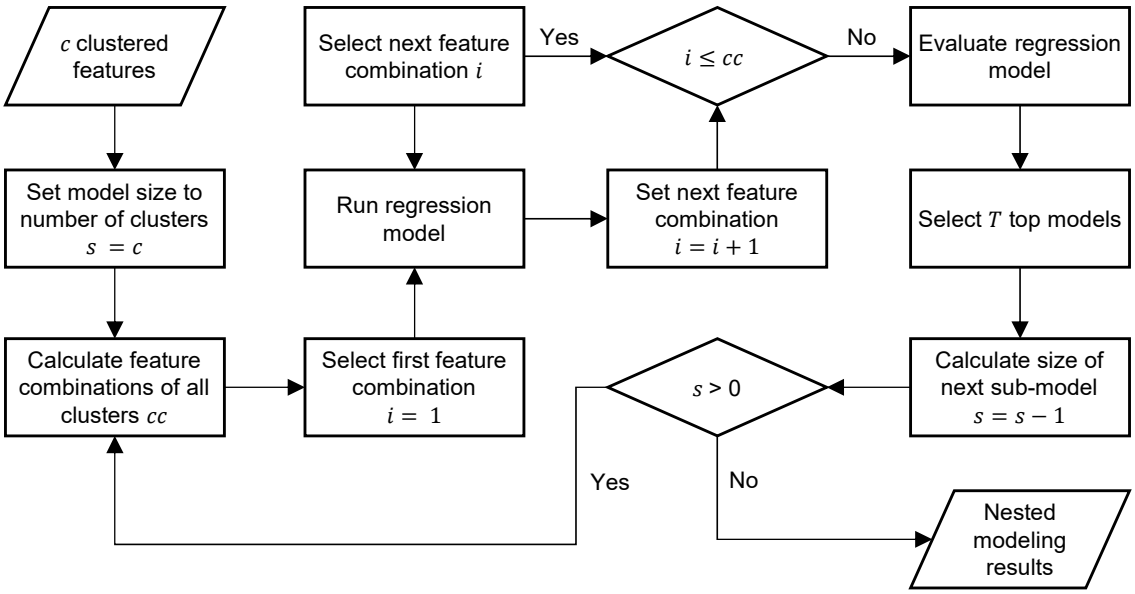

Figure 3.4: Schematic overview of nested modeling. After cc combinations of c clustered features are calculated, suitable models are selected and their subset combinations derived. The procedure is repeated until the combinations contain only one feature.

and information content: The coefficient of determination R^2, the feature significance (p), the *AIC* and *BIC* as well as the *MAE* based on the test and train data set (see Section 2.3). The *MAE* is of particular importance, as this value allows a comparison with modeling results in a later analysis of Section 4.3. The performance index PI_m for a model with a specific feature subset is a standardized value which is defined by the following equation:

$$PI_m = \frac{R_m^2}{\max R^2} + \frac{p_m}{\max p} + \frac{AIC_m}{\max AIC} + \frac{BIC_m}{\max BIC} + 2\,\frac{MAE_m}{\max MAE} \tag{3.7}$$

where the model m is defined by its size s and feature subset combination cc. The significance SI_j of feature j calculated as follows:

$$SI_j = \frac{1}{e}\sum_{i=1}^{e} OPR\left(PI_{m(i)} - PI_{m\tilde{(i)}}\right) \tag{3.8}$$

with

$$\tilde{m} \subset m$$
$$j \in m, j \notin \tilde{m}$$

where $\tilde{m}$ is a nested model of m and e describes how often the feature j is eliminated from a model. Four different operators OPR are implemented which are the max operator (select highest significance), the median operator (select median significance), the mean

operator (calculate the mean significance), and the quantile operator (select defined quantile significance).

The best coherent nested models are determined from size $s = 2$ up to $s = c$. Therefore, for all (sub-) models the best coherent model combinations are selected that fulfills the following criteria:

$$\max \frac{1}{s^* - 2} \sum_{i=2}^{s^*} PI_{m(s,cc(i))} \tag{3.9}$$

with

$$\forall i \in [2, s^* - 1] : \; \tilde{m}(i, cc(i)) \subset m(i+1, cc(i+1))$$
$$PI_m \geq PI_{\tilde{m}}$$
$$AIC_m < AIC_{\tilde{m}}$$
$$BIC_m < BIC_{\tilde{m}}$$
$$MAE_m \leq MAE_{\tilde{m}}$$
$$rank_m \leq i \cdot RL$$

where $\tilde{m}$ is a nested model of m defined by model size i and associated feature combination $cc(i)$ and $s^* \leq c$ the model size fulfilling the criteria. *PI*, *AIC*, *BIC*, *MAE* are derived model criteria, $rank$ is the position of a model compared to models of same size and RL the rank limit. [56] provides a table of recommendations for selected model evaluation criteria which are applied for *AIC* and transferred to *BIC*: As soon as the value does not improve when the model is extended by a feature, the chain of models is no longer continued. The *PI* is an aggregated criterion, thus, the evaluation is derived from *AIC* but relaxed ($\leq$). Additionally, the *MAE* is included as this criterion is most relevant for the subsequent profiling application. Like *PI*, the evaluation is relaxed. The $rank$ limitation RL is introduced to improve the speed of the algorithm following the assumption that the worse the *PI* of a model is compared to others, the less likely it is to be part of an appropriate nested model chain.

Pre-feature selection

The pre-feature selection aims to reduce the number of features with reasonable effort but without affecting the significance of the nested modeling. Thus, a heuristic framework is built around the nested modeling and a feature assessment is added. Figure 3.5 gives an overview of the pre-feature selection algorithm. Starting point is a list of all features F. The features are clustered into c groups. One feature out of each cluster is selected randomly and passed to the nested modeling. The random selection of features and the nested modeling are repeated R times. Afterwards, the nested model results are evaluated, and the significance indices *SI* of the features is derived (see Equation 3.8). One or multiple features having the lowest significance indices are removed from the feature list. The process is repeated until the number of selected features f^* undercuts the abort criterion L.

In-depth feature selection

The in-depth feature selection aims to identify relevant features, feature subsets as well as nested feature subsets used for time series profiling applications. Figure 3.6 gives an overview

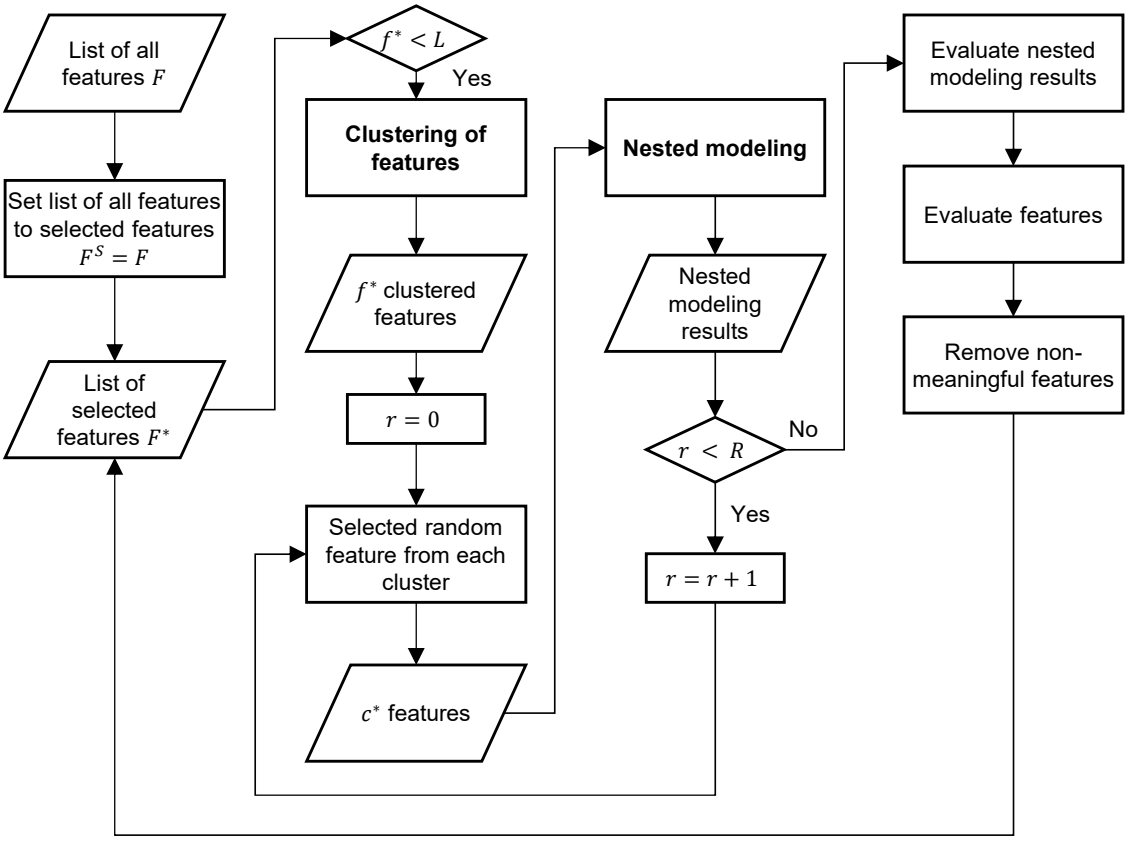

Figure 3.5: Schematic overview of the pre-feature selection process including clustering and nested modeling. Randomly, features from different clusters are combined and evaluated based on their nested modeling results.

of the feature selection algorithm. Starting point is a list with all features F. Based on the pre-feature selection non-meaningful features are removed and the remaining features $f*$ are clustered and passed to the nested modeling. Similar to the pre-feature selection, the nested modeling results are evaluated, and the significance SI of the features is derived. Optionally, the best coherent nested models are determined (see Equation 3.9).

3.3 Profiling of aggregated time series[4]

Within the profiling, findings from the CNR are translated into an iterative algorithm that is applied to already aggregated time series aiming for a better representation of the relevant information of original time series. Therefore, the profiling algorithm supplements existing aggregation methods. In this section, the selected and applied aggregation methods are described first before the developed profiling algorithm is presented.

[4]This section is based on the *profiling paper* - Section 3.2 and 3.3 [39].

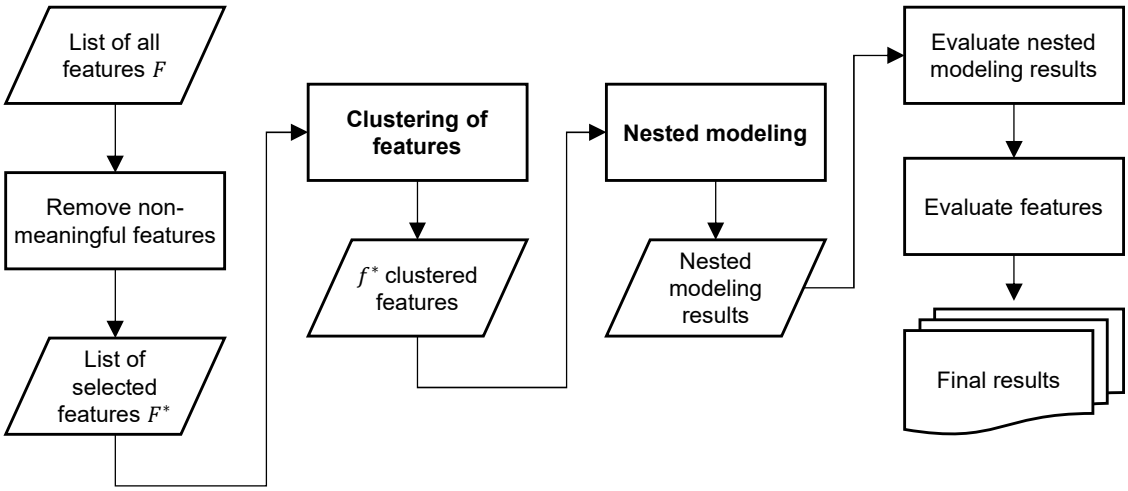

Figure 3.6: Schematic overview of the in-depth feature selection process including clustering and nested modeling. All feature combinations are modeled, and based on the results, features as well as the feature subsets are evaluated.

3.3.1 Time series aggregation

Two simple to implement and commonly applied TSA approaches (similar to [47]) are selected from literature: Clustering as well as a combined method of heuristic and clustering. In addition to profiling algorithm, aggregated time series are also used as data input for the CNR to identify relevant time series parameter (see Section 3.2.2). Contrary to the publication [40], a third TSA method, downsampling, is not applied in this thesis due to limited comparability in terms of aggregation potential: To achieve a reduction as through clustering, the temporal resolution of the original time series would have to be reduced by a factor of 18 to 60. This corresponds to an averaging of three-quarters of a day to 2.5 days. Intraday characteristics would be lost, which would lead to poor modeling results.

Clustering

Two clustering algorithms – k-means and hierarchical clustering using Ward's method – are selected with either the centroid or the closest as daily representative (see Section 2.5). Additional, representatives are weighted individually or uniformly resulting in eight clustering variants, whereby the unweighted time series is only used for the identification of relevant time series parameters. For this purpose, aggregated time series are determined based on one to 50 clusters. For the actual profiling application and evaluation the number of clusters is six to 20. A lower number of clusters are excluded from the analysis as they lead to high outliers, which would not allow a conclusive evaluation.

In a first step, the annual electricity demand time series $\mathbf{z}_{dem}$ are normalized by their maximum value across all considered years y and time steps t to allow valid clustering results (see Equation 3.2). In a second step, the normalized time series z_i[5] of one year are transformed

[5]To improve comprehensibility, the normalized time series is not specifically labeled below.

to one (n x m) matrix $\boldsymbol{X}$, with $n = d$ as the number of days (365) and m as the total number of time steps (24 hours for each time series i), here exemplary with $i \in \{dem, PV\}$:

$$\boldsymbol{z_i} \rightarrow \begin{bmatrix} z_{i,1} \cdots z_{i,(d-1)\cdot 24+1} \\ \vdots \quad \ddots \quad \vdots \\ z_{i,24} \cdots \quad z_{i,d\cdot 24} \end{bmatrix} = \begin{bmatrix} x_{i,1,1} \cdots x_{i,1,d} \\ \vdots \quad \ddots \quad \vdots \\ x_{i,1,24} \cdots x_{i,24,d} \end{bmatrix} = \boldsymbol{X_i} \tag{3.10}$$

$$\mathbf{X}^T = [\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_d] = \begin{bmatrix} x_{1,1} \cdots x_{1,d} \\ \vdots \quad \ddots \quad \vdots \\ x_{m,1} \cdots x_{m,d} \end{bmatrix} = \begin{bmatrix} z_{dem,1} \quad \cdots \quad z_{dem,(d-1)\cdot 24+1} \\ \vdots \quad \vdots \quad \vdots \\ z_{dem,24} \quad \cdots \quad z_{1,d\cdot 24} \\ z_{PV,1} \quad \cdots \quad z_{PV,(d-1)\cdot 24+1} \\ \vdots \quad \vdots \quad \vdots \\ z_{PV,24} \quad \cdots \quad z_{PV,d\cdot 24} \end{bmatrix} \tag{3.11}$$

In the third step, the clustering algorithm is applied on the normalized and transformed time series $\mathbf{X}$ by minimizing the distances across all clusters c (see Equation 2.33 and 2.37) using KMeans from scipy [70] and clusterhierarchy from scikit-learn [46]. In the forth step, the representative for each cluster j is derived as centroid $\mathbf{x}_j^*$ or closest to the centroid $\mathbf{x}_j^{**}$ (see Equation 2.34). Afterwards, the weighting factor of a cluster is calculated based on the number of days in the cluster $|C_j|$ and the total number of days d. Alternatively, clusters are equally weighted:

$$w_c = \frac{d}{|C_j|} \quad \vee \quad w_c = \frac{d}{c} \tag{3.12}$$

In the last step, the representatives $\mathbf{x}_j^*$ (or $\mathbf{x}_j^{**}$) are split into single time series $\mathbf{z}_{i,j}^*$ with $i \in \{dem, iRES\}$ to be re-scaled by their average values derived from the original time series $\mathbf{z}_i$:

$$\widetilde{\mathbf{z}}_{i,j} = \mathbf{z}_{i,j}^* \cdot \sigma_i \tag{3.13}$$

with:

$$\sigma_i = \frac{\sum_{t \in T} z_{i,t}}{d} \frac{c}{\sum_{j=1}^{c} \sum_{t \in T_{C_j}} w_j z_{i,j,t}} \tag{3.14}$$

The rescaled values $\widetilde{\mathbf{z}}_{i,j}$ are re-transformed into single time series $\widetilde{\mathbf{z}}_i$ with $c \times 24$h time steps each:

$$\widetilde{\mathbf{z}}_i = \begin{pmatrix} \widetilde{\mathbf{z}}_{i,1} \\ \vdots \\ \widetilde{\mathbf{z}}_{i,c} \end{pmatrix} \tag{3.15}$$

Combined heuristic and clustering

The general procedure of the combined method is similar to the clustering approach. Between step (1) and step (2) the selection of extreme days from D is added, which are the days with the daily minimum and maximum (average) electricity load and iRES, respectively:

$$\mathbf{z}_{i,j}^{min} = \underset{j \in D}{\arg\min}(\mathbf{X}_i) \tag{3.16}$$

$$\mathbf{z}_{i,j}^{max} = \underset{j \in D}{\arg\max}(\mathbf{X}_i) \tag{3.17}$$

with:

$$\mathbf{X}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,d}] \ \forall \ i \in \{dem, iRES\} \tag{3.18}$$

The selected days represent single clusters with $w_c = 1$ and are excluded from the subsequent clustering.

3.3.2 Profiling algorithm

Profiling selectively adjusts aggregated time series to improve the representation of relevant time series characteristics. The basis for its development are the identified relevant time series parameters that are integrated by deriving specific adjustment algorithms. Single or multiple values of the aggregated time series are modified to achieve a better representation of the relevant characteristics of original time series. Thereby, an improvement of one time series parameter cannot be achieved without effecting other parameters (see, e.g., [51]). For this reason, profiling is performed iteratively and ends (at the latest) when the improvement of all considered time series parameters converges or their deviations are below predefined thresholds.

As shown in Figure 3.7, the profiling approach can be split into three consecutive steps, which are the adjustment of correlation, single values and average values. In the following, the calculations in each step are described – and when useful – exemplary for two normalized original time series $\mathbf{z}_i$ and two normalized profiled time series $\hat{\mathbf{z}}_i$ with $i \in \{dem, PV\}$.

Correlations

The correlation between two time series $\mathbf{z}_{dem}$ and $\mathbf{z}_{PV}$ is adjusted by shifting one time series. As shown in Figure 3.7a, single data points are not affected but the relations to each other (for example, residual load). The shifted time series constellation with the smallest correlation deviation is selected. The correlation error CE between two normalized original time series $\mathbf{z}_{dem}$ and $\mathbf{z}_{PV}$ with $t \in T$ time steps and two normalized profiled time series $\widehat{\mathbf{z}}_{dem}$ and $\widehat{\mathbf{z}}_{PV}$ with $t \in \hat{T}$ time steps is used to derive the required shifting factor s:

$$s_{PV}^{*} = \underset{s \in \{1,...,|\hat{T}-1\}}{\arg\min} |CE_{dem,PV,s}| = |r_{dem,PV} - \widehat{r}_{dem,PV,s}| \tag{3.19}$$

with:

$$\widehat{r}_{dem,PV,s} = \frac{\sum_{t \in \widehat{T}} (\widehat{z}_{dem,t} - \overline{\widehat{z}}_{dem})(\widehat{z}_{PV,t+s} - \overline{\widehat{z}}_{PV})}{\sqrt{\sum_{t \in \widehat{T}} (\widehat{z}_{dem,t} - \overline{\widehat{z}}_{dem}) \sum_{t \in \widehat{T}_s} (\widehat{z}_{PV,t+s} - \overline{\widehat{z}}_{PV})}}$$

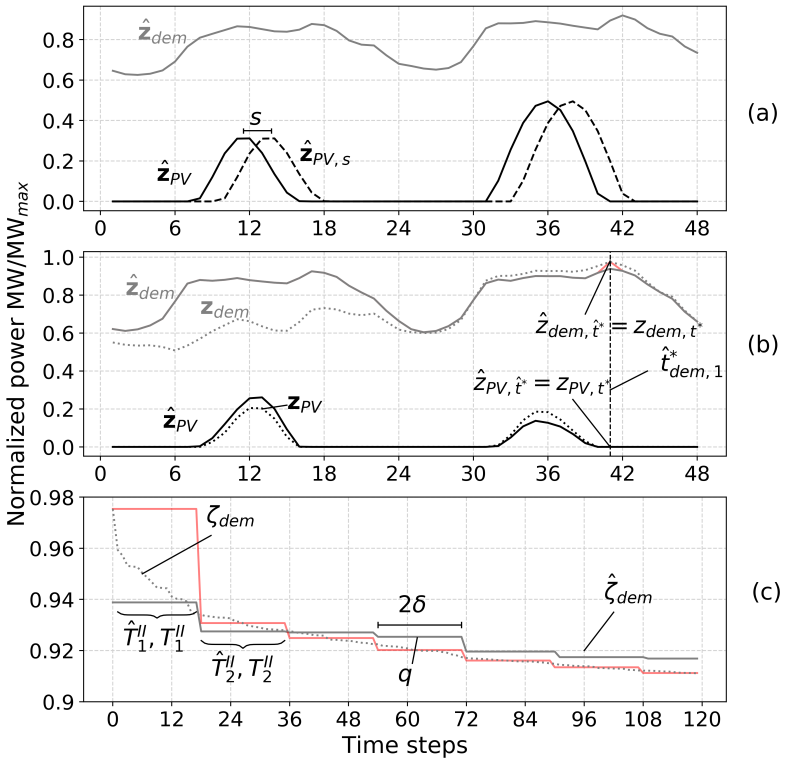

Figure 3.7: Graphical representation of profiling including the fitting of (a) correlation, (b) single values, and (c) average values using the normalized duration curve.

Single values

To reset single values of a time series, the time step t of an extreme value j is determined for both, the profiled time series $\widehat{\mathbf{z}}_i$ and original time series $\mathbf{z}_i$.

$$\widehat{t}^*_{i,j} = \arg\min_{t \in \widehat{T}}(\widehat{z}_{i,t}) \quad \vee \quad \widehat{t}^*_{i,j} = \arg\max_{t \in \widehat{T}}(\widehat{z}_{i,t}) \tag{3.20}$$

$$t^*_{i,j} = \arg\min_{t \in T}(z_{i,t}) \quad \vee \quad t^*_{i,j} = \arg\max_{t \in T}(z_{i,t}) \tag{3.21}$$

As shown in Figure 3.7b, the corresponding values for all aggregated time series are replaced by those of the original time series:

$$\widehat{z}_{i,\widehat{t}^*_{i,j}} = z_{i,t^*_{i,j}} \;\; \forall \;\; i \in \{dem, PV\} \tag{3.22}$$

The time steps $\widehat{t}^*_{i,j}$ are added to a set of blocked time steps $\widehat{B}$ and cannot be changed within the recalculation of average values. At the end of this profiling step, $\widehat{B}$ includes all (unique) time steps of extreme values e:

$$\widehat{B} = \{\widehat{t}^*_{dem,1}, ..., \widehat{t}^*_{dem,e}, \widehat{t}^*_{PV,1}, ..., \widehat{t}^*_{PV,e}\} \tag{3.23}$$

Multiple values

The remaining time steps of the time series $\widehat{\mathbf{z}}_i$ are adjusted to correspond to parameters of multiple values, for example, the variance (*var*). Further, for quantiles, it is not sufficient to adjust only one time series value. Rather, the values around the quantile must be taken into account. Therefore, the duration curve is roughly adjusted before relevant quantiles are modified in more detail. Therefore, the time series are reordered and the resulting duration curves ζ_i and $\widehat{\zeta}_i$ (see Figure 3.7c) are re-indexed with $t' \in T'$ and $\widehat{t}' \in \widehat{T}'$ as well as the blocked time steps $\widehat{B}'$. As the adjustment of duration curves and average values can be derived from that of the quantiles, the generalized formulas are presented in the following. The set of all considered quantiles is defined as:

$$q \in Q = \left\{ \frac{\tau}{|\widehat{T}'|}, \frac{2\tau}{|\widehat{T}'|}, ... \right\} \tag{3.24}$$

For the adjustment of duration curves, the granularity factor is set to $\tau = 1$, thus, each time step represents a single quantile. For specific quantiles, τ can be increased, for example, the 5 % quantiles is obtained by $\tau = 0.05|\widehat{T}'|$ or the set can be determined explicitly, for example, $Q = \{0.35\}$. For each quantile, time sets can be derived for the aggregated and original duration curve:

$$\widehat{T}''_q = \left\{ (q-\delta)|\widehat{T}'| < \widehat{t}' \leq (q+\delta)|\widehat{T}'| \right\} \subset \widehat{T}' \tag{3.25}$$

$$T''_q = \left\{ 1 + \frac{|T'|}{|\widehat{T}'|} \left(\min\left(\widehat{T}''_q\right) - 1 \right) \leq t' \leq \frac{|T'|}{|\widehat{T}'|} \left(\max\left(\widehat{T}''_q\right) - 1 \right) \right\} \subset T' \tag{3.26}$$

The step size δ results from the quantiles (or the granularity factor τ):

$$\delta = \frac{q_i - q_{i-1}}{2} = \frac{\tau}{|\widehat{T}'|} \tag{3.27}$$

The values of the duration curve $\widehat{\zeta}_{i,t}$ are determined similar to the recalculation of the average value in Equation 3.13 and 3.14. The time steps of the profiled time series are equally weighting with $\widehat{w}_q = \frac{|T''_q|}{|\widehat{T}''_q|}$:

$$\widehat{\zeta}_{i,t} = \widehat{\zeta}_{i,t}\sigma_{i,q} \quad \forall \quad t \in \widehat{T}''_q \backslash \widehat{B}' \tag{3.28}$$

with:

$$\sigma_{i,q} = \frac{\sum_{t \in T''_q} \zeta_{i,t} - \widehat{w}_q \sum_{t \in \widehat{B}' \cap \widehat{T}''_q} \widehat{\zeta}_{i,t}}{\widehat{w}_q \sum_{t \in \widehat{T}''_q \backslash \widehat{B}'} \widehat{\zeta}_{i,t}} \tag{3.29}$$

The equations to recalculate the duration curve and the total average can be derived and simplified from Equation 3.24 - 3.29. The derivation is described in Appendix C. The adjustment of the time series values according to the duration curve is simplified to:

$$\widehat{\zeta}_{i,\widehat{t}'} = \frac{|\widehat{T}'|}{|T'|} \sum_{t=1+\frac{|T'|}{|\widehat{T}'|}(t-1)}^{\frac{|T'|}{|\widehat{T}'|}t} \zeta_{i,t} \quad \forall \quad \widehat{t}' \in \widehat{T}''_q \backslash \widehat{B}' \tag{3.30}$$

The alignment of the total average values results with $\widehat{w} = \frac{|T'|}{\widehat{T}'}$ to:

$$\widehat{\zeta}_{i,\widehat{t}'} = \frac{\sum_{t \in T'} \zeta_{i,t} - \widehat{w} \sum_{t \in \widehat{B}'} \widehat{\zeta}_{i,t}}{\widehat{w} \sum_{t \in \widehat{T}' \setminus \widehat{B}'} \widehat{\zeta}_{i,t}} \quad \widehat{\zeta}_{i,\widehat{t}'} \quad \forall \quad \widehat{t}' \in \widehat{T}' \setminus \widehat{B}' \tag{3.31}$$

Finally, the duration curves ζ_i and $\widehat{\zeta}_i$ are transferred back to the time series $\mathbf{z}_i$ and $\widehat{\mathbf{z}}_i$ using the original time index $t \in T$ and $\widehat{t} \in \widehat{T}$. The adjustments – correlation, single values, and average values – are repeated until a termination criterion is met. Individual sections of the profiling are skipped for one or more time series, if the respective deviation is below a defined threshold.

Chapter 4

Results

The results section can be divided into four parts: In the first section, an exploratory data analysis provides insights into the original time series used. The second section presents the results of CNR and the identified parameters. The third section presents the modeling results of all time series before finally comparing the aggregated and profiled time series in the fourth section.

4.1 Exploratory data analysis

As described in Section 3.2.2, the bundle of three normalized time series (electricity demand and generation potential of PV and wind power) from eleven years (2006-2016) is extended by three additional time series that describe the residual load (see Equation 3.3). Thus, a total of 66 time series are explored below using descriptive parameters to provide a basic understanding of the data and the parameter used in CNR.

4.1.1 Time series parameters

Figure 4.1 provides a first overview of the time series including the main location and dispersion parameters. The values of the original time series are shown in the first row of sub-figures and the values of the residual loads in relation to both iRES (Residual iRES) and separately for each (Residual PV and Residual Wind) are shown in the second row. For each year, the single values as normalized power are visualized in the form of scatters. Note, that each sub-figure has individual y-axis values, that might distort the comparison of different time series. However, as the time series have different patterns (e.g., the demand has a minimum relative power of around 0.4 MW/MW$_{max}$, PV and wind power have zero values, and the residual negative values) different y-axes enable improved visual analysis. This applies for all figures within this section. The box plot of each scatter plot shows the IQR (25 % and 75 % quantiles), the mean and the median as well as the whiskers with a maximum length of 1.5xIQR and outliers. Together, these values give a first indication of the distribution. For example, the PV time series show a significant difference between the median and mean (0 < 0.125), which indicates a right-skewed distribution. This conclusion is supported by the asymmetrical length of the whiskers (lower < upper) and the outliers in positive direction. Despite similar characteristics, small difference can be observed for each time series across the years. For

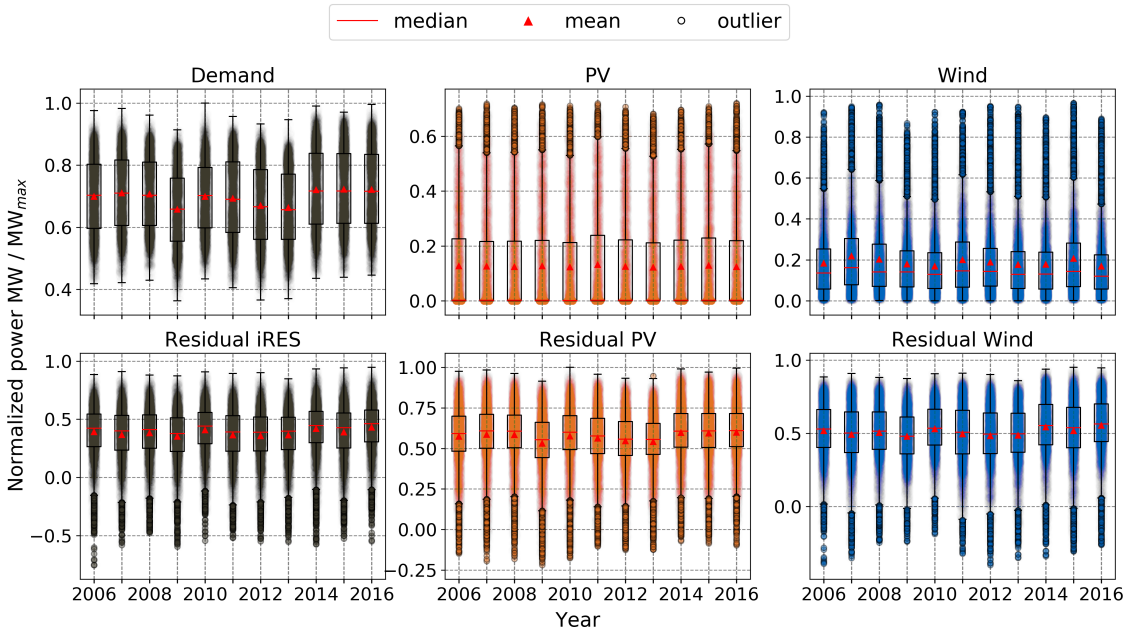

Figure 4.1: Box plot including mean and median values for the original time series as well as recalculated residual loads.

example, the demand time series have similar patterns in terms of overlapping mean and median and missing outliers. However, the exact mean or median values differ as well as the IQR and whiskers. In the following analysis, the individual parameters of location, dispersion and correlation are discussed in more detail.

Location parameters

Figure 4.2 shows the mean, the trimmed mean (tmean) with 5 % cut off of both tails of the distribution and the geometric mean (gmean[1]) as three different mean measures. In addition, the median is shown to classify the mean values. The values are plotted for each time series as line graph over the years to identify differences between the years.
For each time series, the four measures have a similar pattern over the years: The demand time series has overlapping values of the mean, the tmean and median. The values for PV and wind appear to maintain a constant difference, whereas the distance between the values of the residual loads vary (e.g., residual PV 2010 and 2014, residual wind 2009 and 2012). The similar patterns of the measure indicate a correlation or in other words, they represent similar information of the time series. An illustrative example is the mean $\overline{z}$ and sum $\sum_{i=1}^{n} \mathbf{z}_i$. They have a linear relationship $\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$ and the values differ only by the factor n. Thus, these measures have a correlation of 1. The average correlation of all measures ranges between 0.940 (wind) and 0.995 (Residual iRES) with the exception of PV where the median differs from the mean and tmean (0.30 and 0.35) resulting in a lower overall correlation of 0.55.
Besides correlating measures, a second observation can be made regarding the order of

[1]The gmean is only calculated for time series of positive reals with at least one number greater than 0.

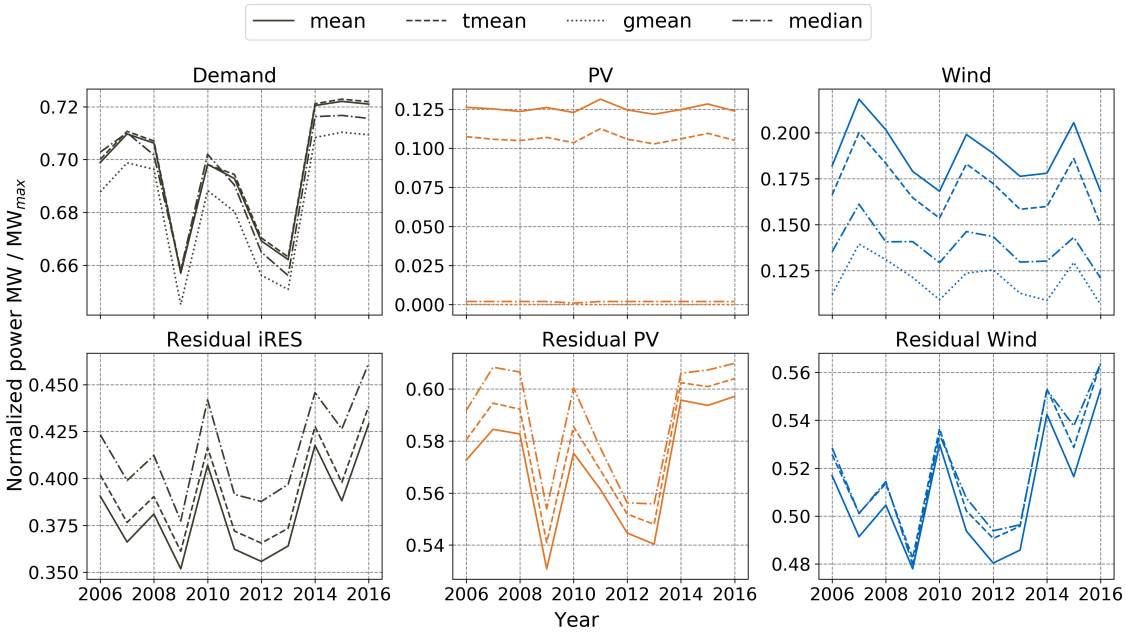

Figure 4.2: Location parameter: Parameters describing the mean of the original time series as well as recalculated residual loads.

measures. The iRES time series have maximum values for the mean, followed by the tmean and median. This order is reverse for the residual load and meets the expectations based on Equation 3.3 in which the iRES are subtracted from the demand. As already described, the order of the measures indicate the distribution that can also be derided from Figure 4.3 showing exemplary quantiles between the minimum (0 %) and maximum (100 %). The quantiles of the demand time series are symmetrical around the median (50 %). The quantiles of PV and wind are asymmetric and show wide distances between the quantiles in direction of the maximum. In direction of the minimum there are no quantiles below the median (PV) or smaller distance between the quantiles (wind). As before, the pattern is reversed for the residual load. In addition, it is less asymmetrical compared to the iRES. Overall, with the exception of demand (0.92), the correlation between quantiles is lower at an average of 0.38 (PV) to 0.76 (residual PV), as the correlation decreases with increasing distance between the quantiles.

Dispersion parameters

The discussed location parameters already give a first estimate of the time series distribution, which is shown in Figure 4.4 as histogram and empiric density function. The location parameter of the demand indicate a normal distribution. However, the distribution has a dip in the middle leading to a negative kurtosis of -1.11. A skewness of 0.05 is consistent with the symmetric quantiles observed. For PV and wind a right-skewed distribution can be confirmed with a skewness of 1.32 and 1.53, respectively. The distribution of residual loads is left-skewed and thus inverse to the distribution of PV and wind. With an average skewness between -0.62 (residual wind) and -0.71 (residual PV), the asymmetry is lower, which quantitatively confirms

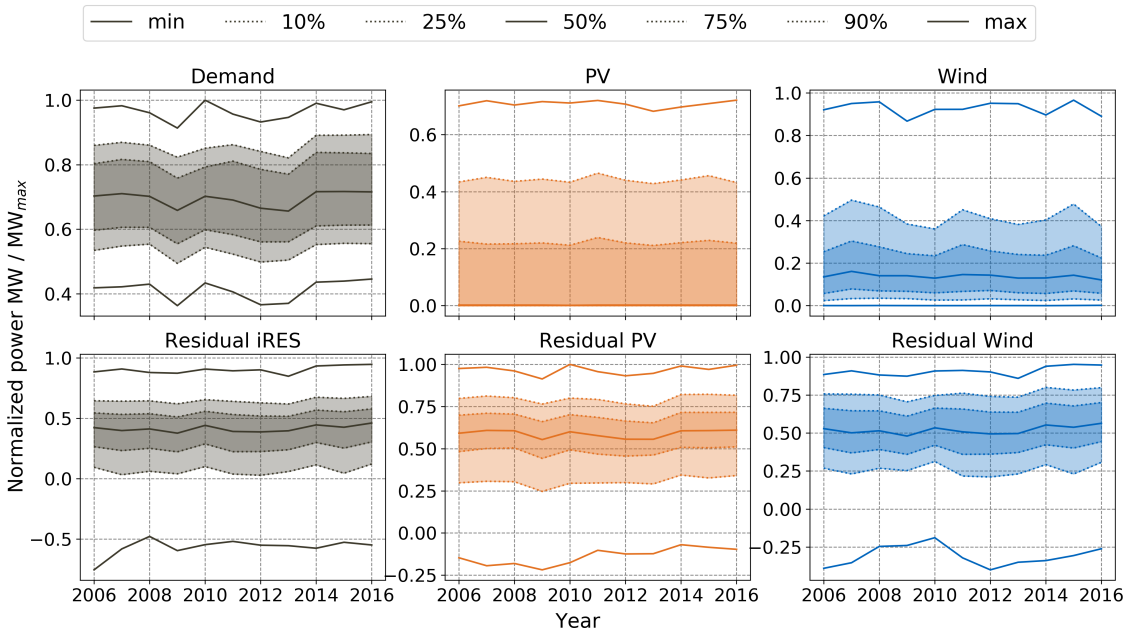

Figure 4.3: Location parameter: Parameters describing quantiles between the minimum and maximum of the original time series as well as recalculated residual loads.

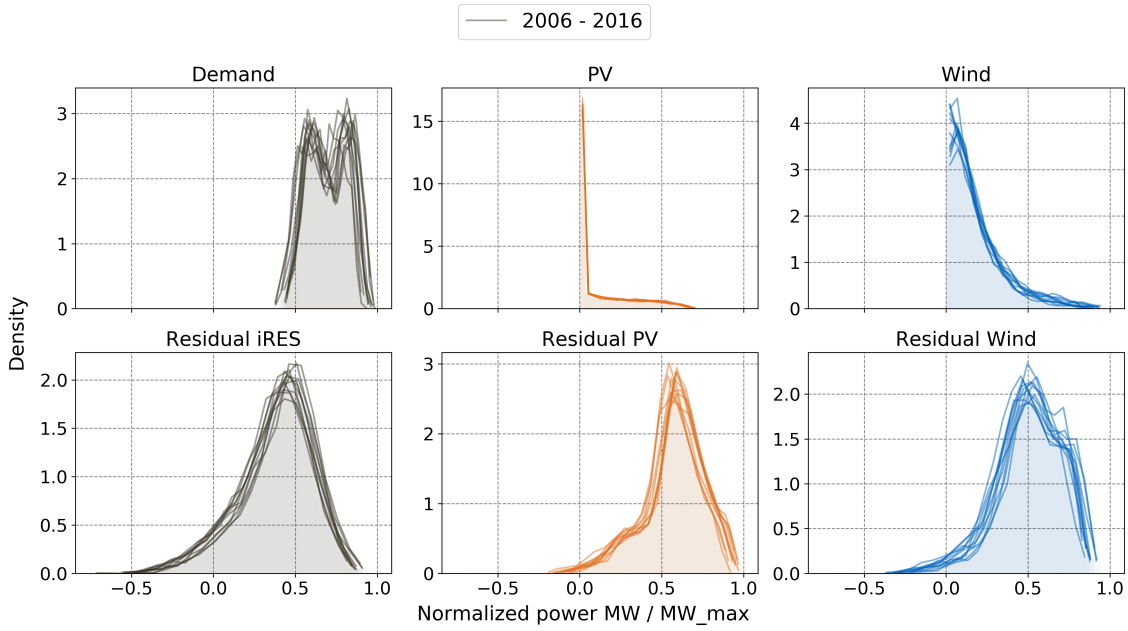

Figure 4.4: Distribution parameter: Histogram and empiric density function of the original time series as well as recalculated residual loads.

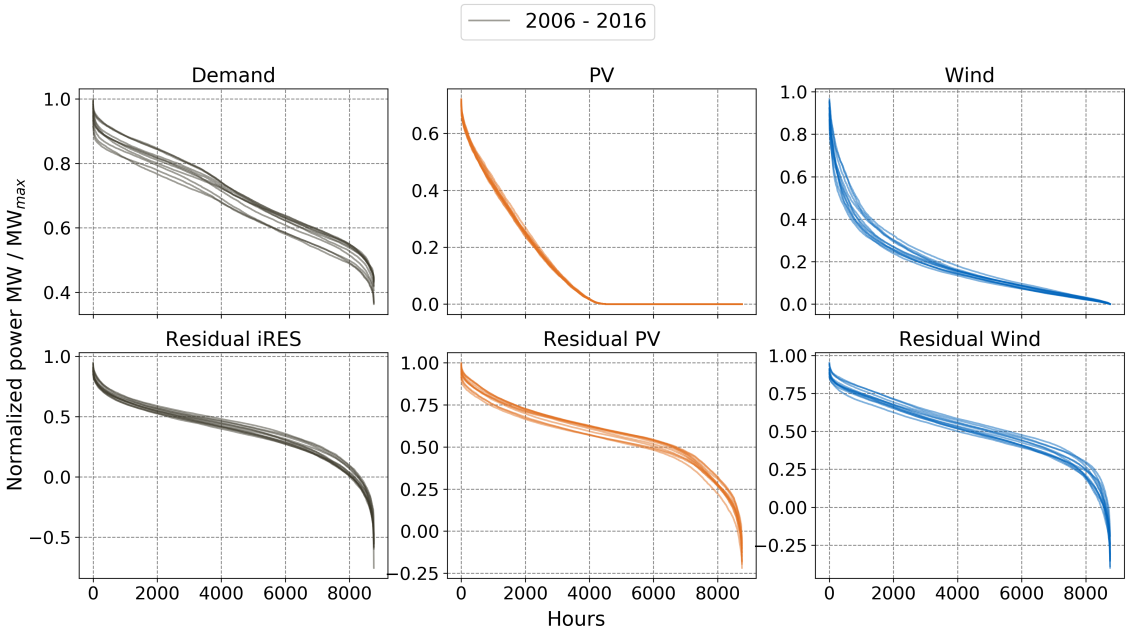

Figure 4.5: Distribution parameters: Duration curve of the original time series as well as recalculated residual loads..

the previous results.
A closer look at the time series of different years shows time series specific deviations. For example, the PV time series overlap and have a slight deviation, whereas the other time series show larger differences between the years. This becomes more apparent by changing the visualization from the empirical density function to the duration curve, that is the sorting of the time series values in descending order, shown in Figure 4.5. The demand as well as the residual load time series including PV or wind show a span over the hole duration curve. The average span is 0.028 for demand and 0.028 and 0.035 for the residual loads, respectively. The wind time series has a spread in the first half of the curve (0.034 vs. 0.008) and the residual load of both iRES has a slightly higher spread in the second half (0.037 vs. 0.027).
Further parameters for describing the distribution are the STD andMAD in relation to the mean or median, which are shown in Figure 4.6. As with the location parameters, the parameters correlate within a time series. With the exception of PV and residual PV, the average correlation is between 0.84 (wind) and 0.97 (demand). The MAD in relation to the median correlates less with the other parameters and averages 0.27 and 0.66 for the PV and residual PV time series, respectively. One cause for this is the deviation between median and mean value, especially for PV time series.

Correlation parameters

In addition to the description of internal time series characteristics, interactions between time series can be quantified using different measures. Figure 4.7 shows the pcorr, the scorr and the kcorr for selected pairs of time series. The first row of sub-figures presents the three

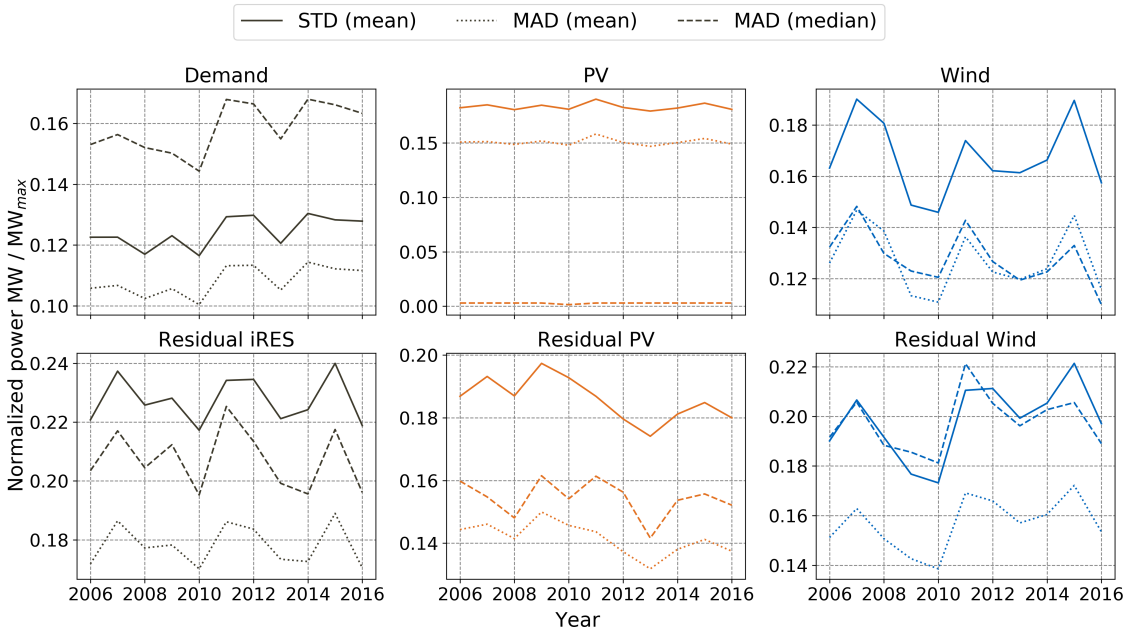

Figure 4.6: Distribution parameter: standard deviation (STD) and mean absolute deviation (MAD) in relation to mean or median.

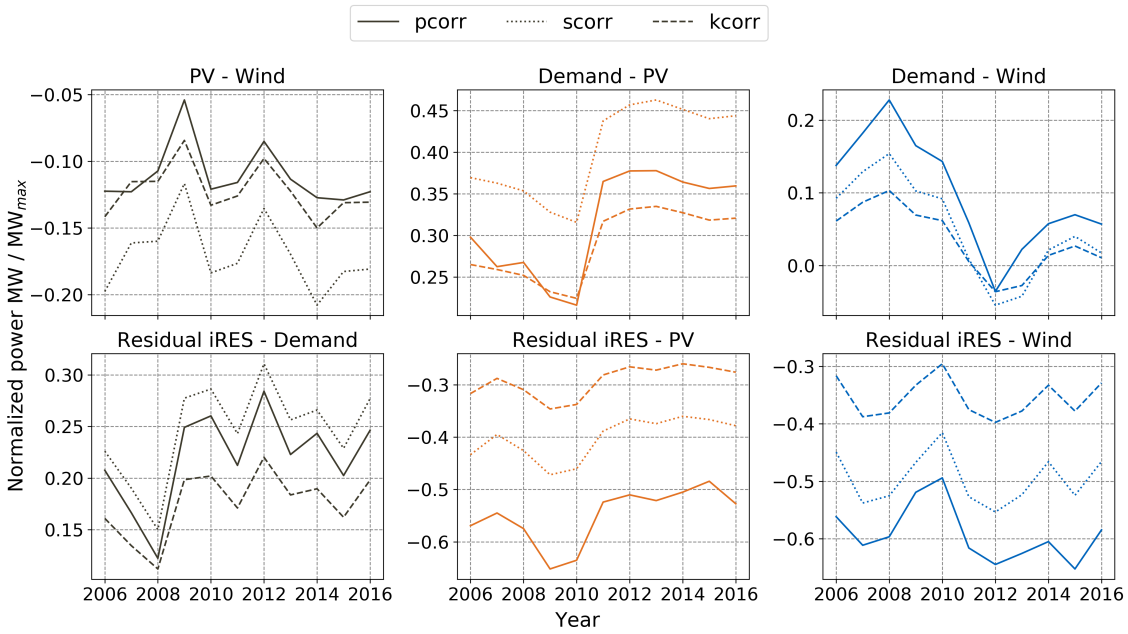

Figure 4.7: Correlation parameter: Pearson's correlation (pcorr), Spearman's correlation (scorr) and Kendall's correlation (kcorr) to describe interaction between selected time series.

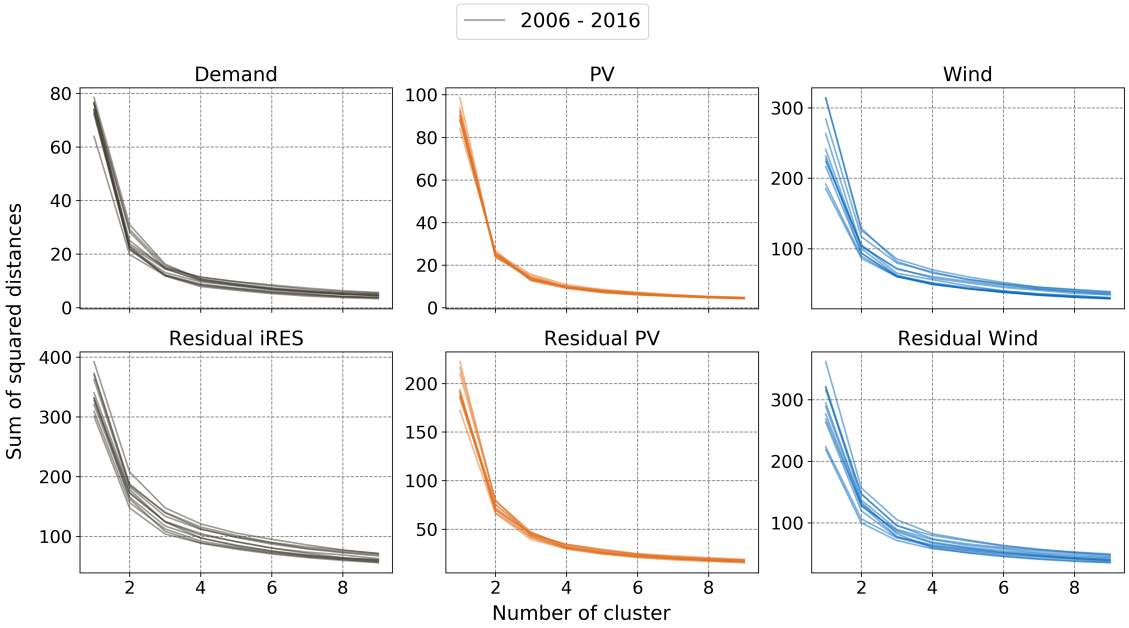

Figure 4.8: Clustering analysis: Development of the sum of squared distances (SSD) for one to nine clusters when the original time series are clustered separately.

combination of the original time series, whereas the second row presents the correlation of the original time series with the residual load of both iRES. Two main findings appear again: First, the correlation of parameters that on average varies between 0.73 (residual iRES - wind) and 0.99 (demand - PV). Second, the reverse patterns that are visible for the time series pairs demand - PV, residual iRES - PV, demand - wind, and residual iRES - wind.

4.1.2 Clustering analysis

In addition to the analysis with classic statistical parameters, time series can be explored using further methods. In the following, the time series are described by applying k-mean clustering. On the one hand, clusters or the number of clusters can be interpreted as a measure of information content, which is of central importance in this thesis. On the other hand, clustering is used in CNR and the time series aggregation, thus, general insights support a later evaluation of the results.
In the following cluster analysis, the focus is on daily patterns and the identification of similar days. Therefore, time series of each year are converted into a $(n \times m)$ matrix with $n = d$ days and $m = 24$ hours (see Section 3.3.1). In a first step, k-mean clustering is applied for different numbers of cluster to derive the relevant number of clusters using the sum of squared distances (SSD) between the days and their cluster centroid. The results are shown in Figure 4.8. Two observations can be made: First, the SSD converges from three or four clusters, and second, the SSD values for time series with wind are higher then those of demand and PV only. For example, with four clusters the average SSD of demand and PV is 10.68 and 6.69, respectively, whereas the average SSD of wind is 48.86. However, it should be noted that around half of the

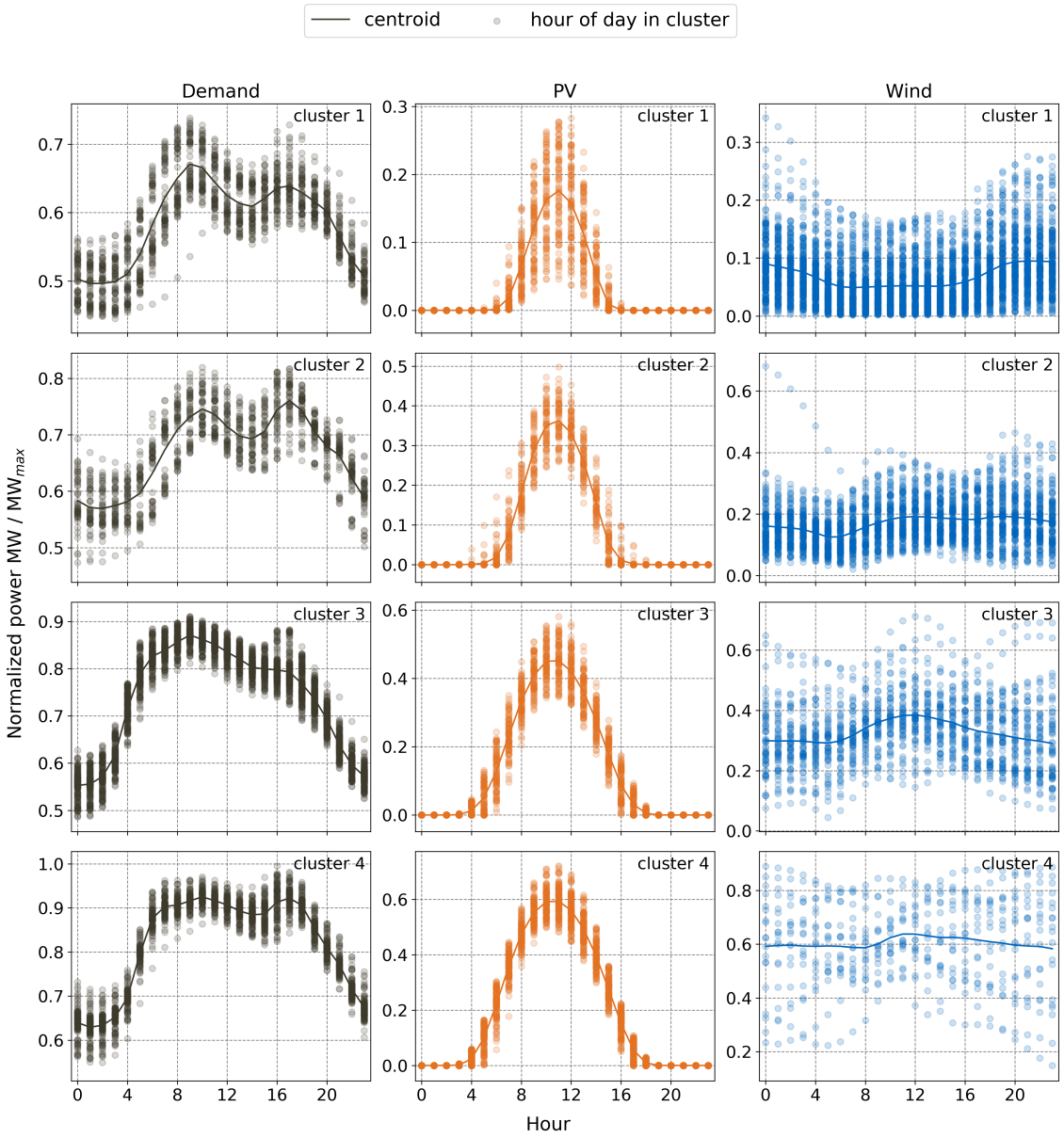

Figure 4.9: Clustering analysis: days clustered into four groups represented by centroids.

PV values are zero and therefore the SSD value is underestimated (approx. by the factor of two).

Figure 4.9 shows the results when days of a single year are clustered separately into four groups. For PV and demand, there are clear clusters corresponding to the SSD ordered by the average value of demand or PV. The daily profiles of PV follow the typical bell-shaped curve with different heights and widths representing the different seasons of the year. The daily demand profiles of the four clusters vary not only regarding the height, but also regarding the shape. For example, the first two clusters have two peaks at around hour 9 and hour 17,

whereas the second peak in clusters 3 and 4 is absent or only slightly present. A closer look into the data indicate that weekend days are more likely to be included in the first two clusters and week days are part of the last two clusters. However, for wind it can be seen that each cluster has a high dispersion of the individual days around the centroid. This complicates the interpretation of the clusters and suggests that daily or seasonal patterns are less pronounced.

All together, the exploratory data analysis illustrates the challenges of selecting representatives or aggregating time series. Although correlating parameters can be identified and summarized, several parameters are necessary to evaluate the selection or aggregation of time series. In particular, wind time series with stochastic rather than systematic patterns make it difficult to aggregate time series by clustering. High SSD values are expected for aggregated time series resulting in an uncertainty regarding the effects on the modeling results.

4.2 Identified time series parameters[2]

A comprehensive data set of 9000+ observations is used to identify relevant time series parameters (see Section 3.1). The data set covers a broad range of parameters (e.g., PV mean: 0.07-0.16, correlation between demand and wind: -0.49-0.67) and installed capacities (e.g., PV: 0-60 GW, wind power: 0-326 GW). The analysis is performed for the PV and WIND scenarios using the *MAE*, *AIC* and R^2 as evaluation criteria. Since the focus of this thesis is on aggregation and profiling of time series to reduce the temporal complexity of energy system models, the following results are reduced accordingly in comparison to the *identification paper* in Chapter 4.

4.2.1 Comparison of feature selection methods

The four methods - CNR, LASSO, LASSOLARS and ElasticNet - are compared qualitatively with regard to the selected features and quantitatively by using selectivity and the three evaluation criteria.

Selectivity

The selectivity describes how many parameters are identified as relevant by the respective feature selection methods. A method has a high selectivity if only a few parameters are selected. In the PV scenario, CNR has the highest selectivity with an average of 38 out of 179 features (79 %). LASSO (57 %) and LASSOLARS (59 %) show rates of a similar amount, whereas ElasticNet achieves the lowest selectivity with 1 %. In the WIND scenario the selectivity of ElasticNet is also 1 %. Smaller selectivity values are observed for the other methods compared to the PV scenario: CNR has the highest selectivity of 75 %, followed by LASSO (31 %) and LASSOLARS (19 %).

[2] This section is based on the *identification paper* - Chapter 4 [39] and the *profiling paper* - Section 4.1 [40].

Identified parameters

In order to enable a meaningful comparison of the feature selection methods with regard to the identified parameters, only the first 40 parameters (corresponds to the selectivity of CNR) of the methods from the literature are used for the comparison. In the PV scenario, 13 to 16 parameters are identified as relevant by at least three feature selection methods. To describe the installed capacity of PV, three parameters are relevant according to all methods, that are the relative variance (to the mean) and the fourth moment of the residual load as well as the (Spearman) correlation between PV and electricity demand. For FPP, four parameters are relevant to all methods that are the maximum, the 60 % quantile and third moment of the residual load as well as the relative range (median) of the electricity demand. For IPP, only the 35 % quantile of the electricity demand and the 30 % quantile of the residual load are identified by all feature selection methods. Similar values can be determined for WIND scenario, in which ten to 18 parameters of at least three feature selection methods are identified. To describe the installed capacity of wind power, mainly two parameters are relevant to all methods: the (arithmetic, trimmed, geometric) mean and the median of wind power. For FPP, four parameters are relevant to all methods that are the MAD, the gmean and the median of wind power as well as the relative range of the residual load (mean). For IPP, the mean, median and relative range (mean) of wind power are identified by all feature selection methods. However, the majority of parameters are only identified by one (38-63 parameters) or two methods (22-42 parameters).

Model performance

The performance of the methods indicates how well selected features explain the model responses. To derive performance indicators for the methods from the literature, a GLS regression is applied using two to ten of the most relevant features. Regardless of the scenario and the technology, the performance of CNR is significantly better on average. For example, in the PV (WIND) scenario the average *MAE* across all technologies and parameter models is 1.0 GW (5.2 GW), whereas the average *MAE* of LASSO, LASSOLARS and Elastic Net is 3.0 GW, 3.1 GW and 3.8 GW (12.1 GW, 12.4 GW and 13.6 GW). The same applies for the *AIC* that is on average 18 (40) in the CNR method and around 113-116 (129-131) in the respective feature selection methods. The average values of R^2 are similar in the PV scenario (0.98-1.00), whereas the differences between CNR and methods from the literature are higher in the WIND scenario (0.97 vs. 0.82-0.84).

4.2.2 Evaluation of relevant time series parameters

In the following, the identified parameters based on CNR for the PV and WIND scenarios are presented in more detail. In the scenario PV, nine time series parameters stand out describing the installed capacity of PV, FPP, and IPP. From a PV time series perspective, relevant characteristics are the mean, the variance and the skewness. Relevant characteristics of the electricity demand are the quantiles around 35 %, the relative MAD (to mean) and the relative range (to median). For the residual load, time series parameters such as the mean, the variance and the maximum values are the most essential. Figure 4.10 shows two selected parameters (black and colored line) describing the installed capacity for each

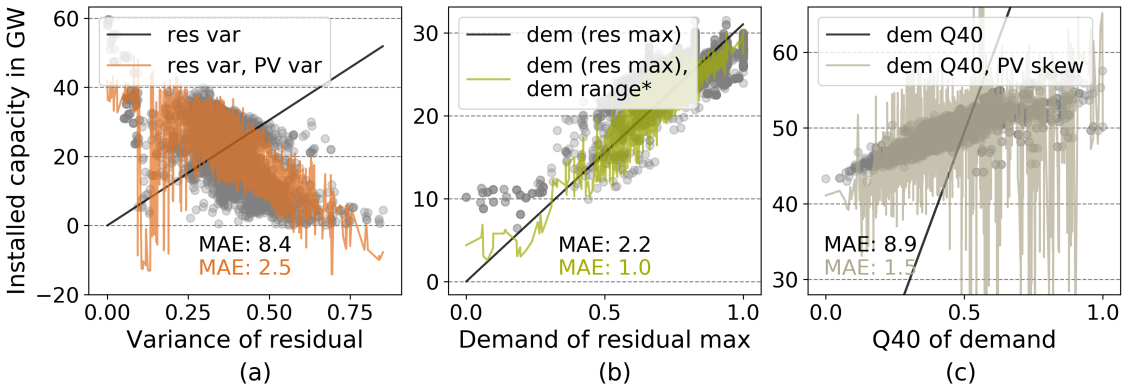

Figure 4.10: Two identified time series parameters of the PV scenario for the installed capacity of (a) PV, (b) FPP, and (c) IPP, respectively.

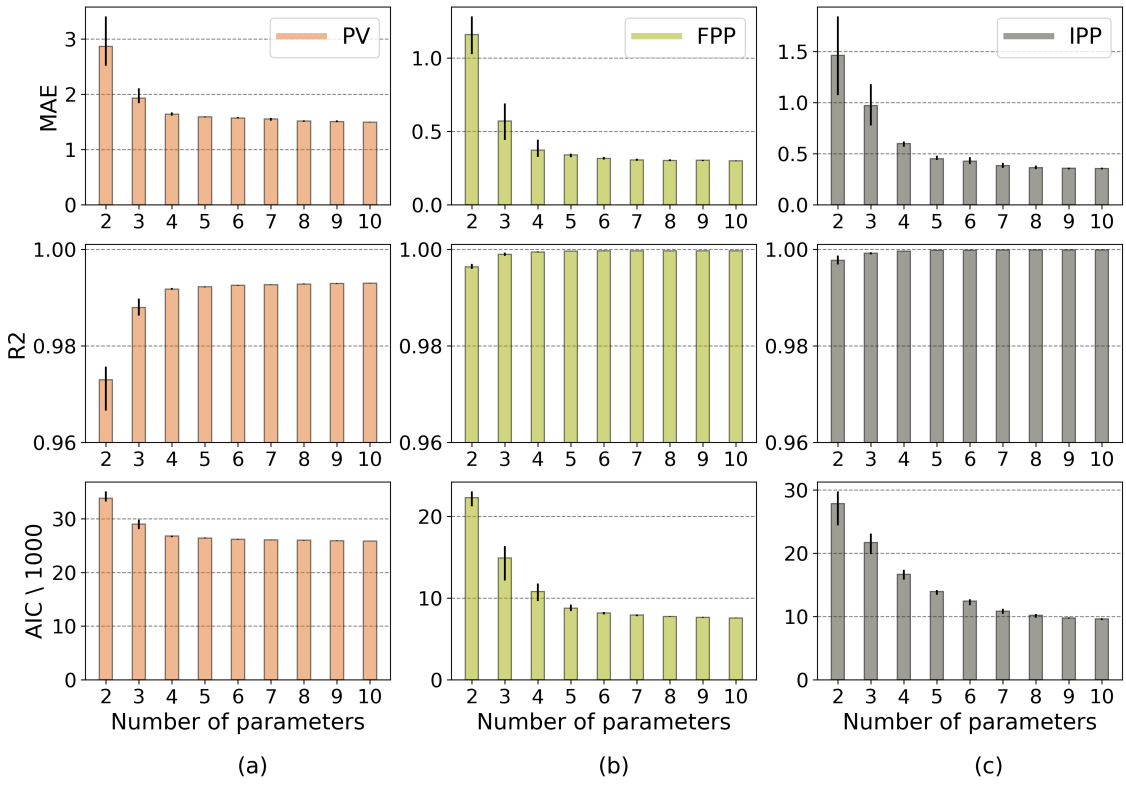

Figure 4.11: Evaluation of top 50 independent models that include two to ten parameters for each technology (a) PV, (b) FPP, and (c) IPP in the PV scenario.

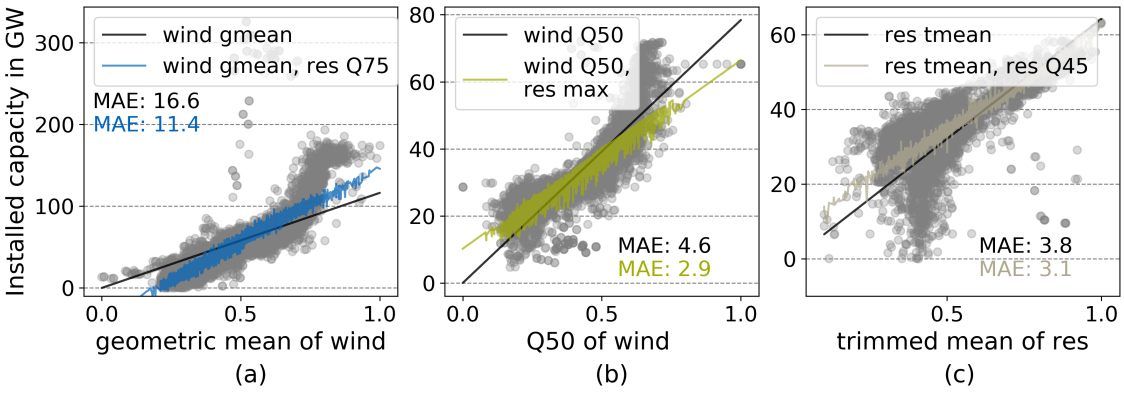

Figure 4.12: Two identified time series parameters of the WIND scenario for the installed capacity of (a) wind power, (b) FPP, and (c) IPP, respectively.

technology. The installed capacity of PV depends on the variance of PV and the residual load. The *MAE* is 8.4 GW when only considering the variance of the residual load and is decreased by 5.9 GW when the variance of PV is added. However, the visualization in Figure 4.10a shows a moderate fit of some modeling results as these are not covered by the recalculated installed capacities (orange line) of the regression model using variances of PV and residual load as model parameters. By adding up additional parameters such as the PV (trimmed) mean or the relative MAD (to mean) and 60 % quantile of the electricity the *MAE* converges to 1.50 $\pm$ 0.01 GW (see Figure 4.11a). The installed capacity of FPP can be represented by the maximum of the residual load and the relative range (to median) of the electricity demand resulting in a *MAE* of 1.0 GW. However, Figure 4.10b shows outliers of low capacity which require further parameters such as the (trimmed) mean and the 60 % quantile of the residual load or the relative IQR of PV. With ten parameters, the *MAE* converges to 0.30 $\pm$ 0.003 GW (see Figure 4.11b). The installed capacity of IPP can be explained by the 40 % quantile of the electricity demand and skewness of PV resulting in similar performance results as for IPP (*MAE*: 8.9 and 1.5 GW). By adding further parameters like (trimmed) mean and variance of PV or the relative MAD of the residual load the *MAE* is reduced to 0.36 $\pm$ 0.01 GW (see Figure 4.11c). Figure 4.11 shows the CNR results of the top 50 independent models. One model consists of an unique parameter subset of two to ten parameters. For all technologies and evaluation criteria we observe converting patterns. The model improvement in terms of *MAE* and *AIC* is on average less then 10 % when a fifth or sixth parameter is included. With regard to the variance explained (R^2), the improvement is already below 5 % on average with the third parameter.

In the scenario WIND, seven time series parameters are particularly relevant for describing the installed capacity of wind power, FPP, and IPP: The means of wind and residual load, the maximum of the residual load and quantiles in the middle and upper area (around 40 % and 75 %) of wind and residual load. As shown in Figures 4.12 and 4.13 the *MAE* across all technologies is higher than in the PV scenario. The installed capacity of wind power can be explained by the gmean of wind and the 75 % quantile of the residual load which together achieve a *MAE* of 11.4 GW. By adding further parameters, such as the 40 % quantile and relative MAD (mean) of wind or the mean and the 30 % quantile of the residual load the

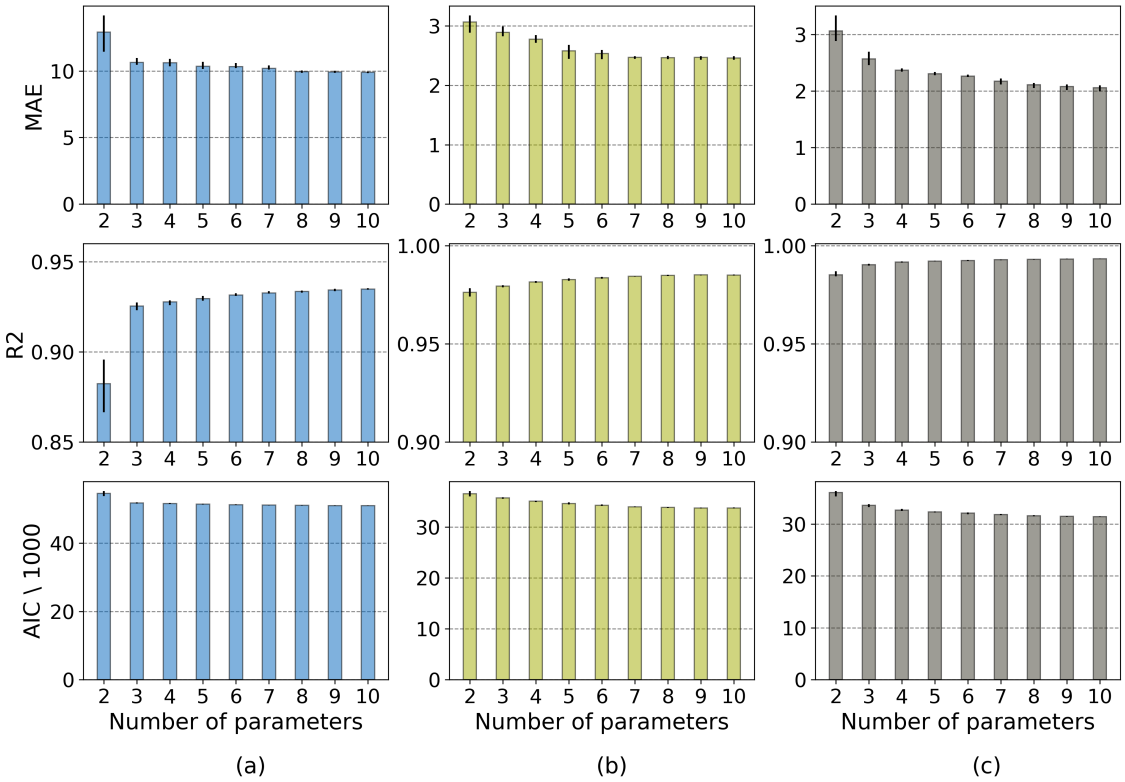

Figure 4.13: Evaluation of top 50 independent models that include two to ten parameters for each technology (a) wind power, (b) FPP, and (c) IPP in the WIND scenario.

MAE converges to 9.92 GW $\pm$ 0.04 GW. The installed capacity of FPP depends on the 50 % quantile of wind and the maximum of the residual load which together reach a *MAE* of 2.9 GW. Additional parameters such as (arithmetic, trimmed) mean of wind or the relative range (median) and the 50-60 % quantile of the residual load can result in a slightly smaller *MAE* of 2.46 GW $\pm$ 0.02 GW. The installed capacity of IPP can be approximated by the trimmed mean and the 45 % quantile of the residual load (*MAE*: 3.1 GW). Including further parameters such as the 50 % quantile and the (trimmed) mean of wind or the (trimmed) mean of the electricity demand can improve the *MAE* up to 2.07 GW $\pm$ 0.03 GW. Figure 4.13 shows the CNR results of the top 50 independent models. As already described for the *MAE*, worse performance values can also be observed for *AIC* and R^2 for the WIND scenario. The model improvement in terms of *MAE* and *AIC* is on average less then 5 % when a third or fourth parameter is included. With regard to the variance explained (R^2), the improvement is already below 5 % on average with the third parameter.

4.2.3 Extended time series complexity

Figure 4.14 highlights another dimension of complexity that is hidden in the time series - in particular the interdependence of time series. Figure 4.14a shows a perfect fit for the duration curve of the aggregated demand time series compared to the original time series. When splitting the duration curve by the median of PV (approx. zero) the curves diverge. For PV

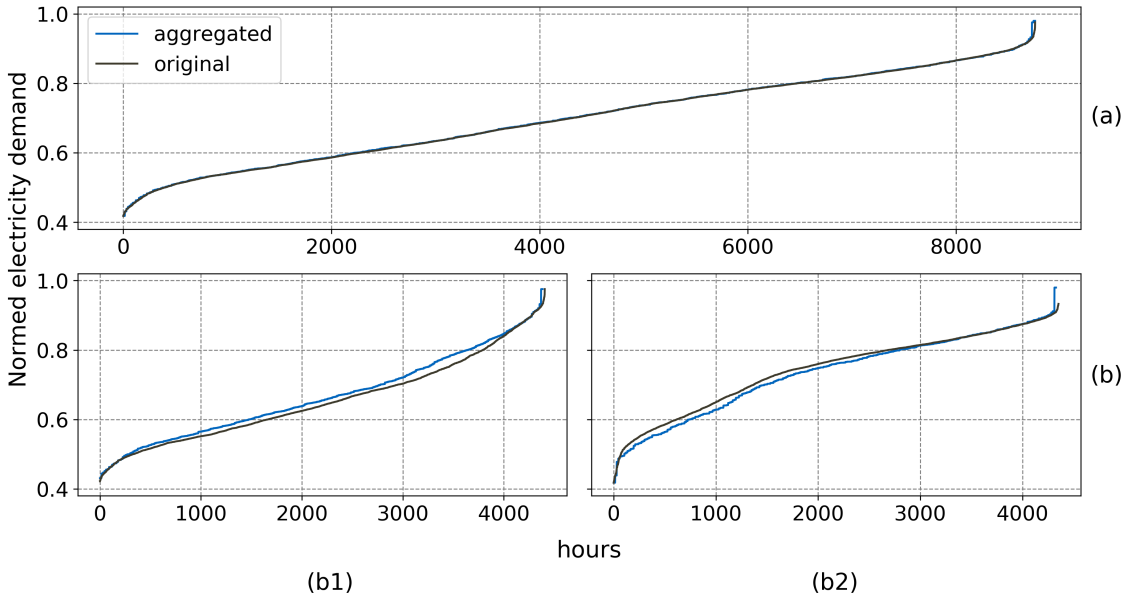

Figure 4.14: Aggregated and original time series of the electricity demand as full duration curve (a) and split duration curve (b) depending on the PV median (below median (b1), above median (b2)).

values of zero (Figure 4.14b1), the aggregated time series overestimates the demand, whereas it underestimates the demand for medium to high PV values (Figure 4.14b2). This deviation contributes to the fact that in the optimization of the energy system the effectiveness and thus the expansion of PV is overestimated by 3.7 GW (15 %).

4.3 Modeling results[3]

The modeling results of the all time series - original, aggregated and profiled - are presented below. The modeling results represent the installed capacities of the generation technologies. For the original time series, the results are given in absolute values (GW). The results of the aggregated and profiled time series are shown as a deviation from the original time series. The deviation can be averaged over several years or time series as mean deviation (or error *ME*), standard deviation (*STD*) or mean absolute deviation (or error *MAE*). The latter allows a comparison with the results from CNR in Section 4.2.2.

4.3.1 Original time series

The eleven original time series (2006-2016) and three defined scenarios (PV, WIND, PV+Wind) result in a total of 33 modeling runs (see Section 3.1). The modeling results as installed capacity of the respective generation technologies (PV, wind, FPP, IPP) are shown in Figure 4.15. Table 4.1 supplements the visual representation of the results with a quantitative evaluation by calculating the mean, median, minimum and maximum values of the technologies. It can

[3]This section is based on the *profiling paper* - Section 4.2 and 4.3 [40].

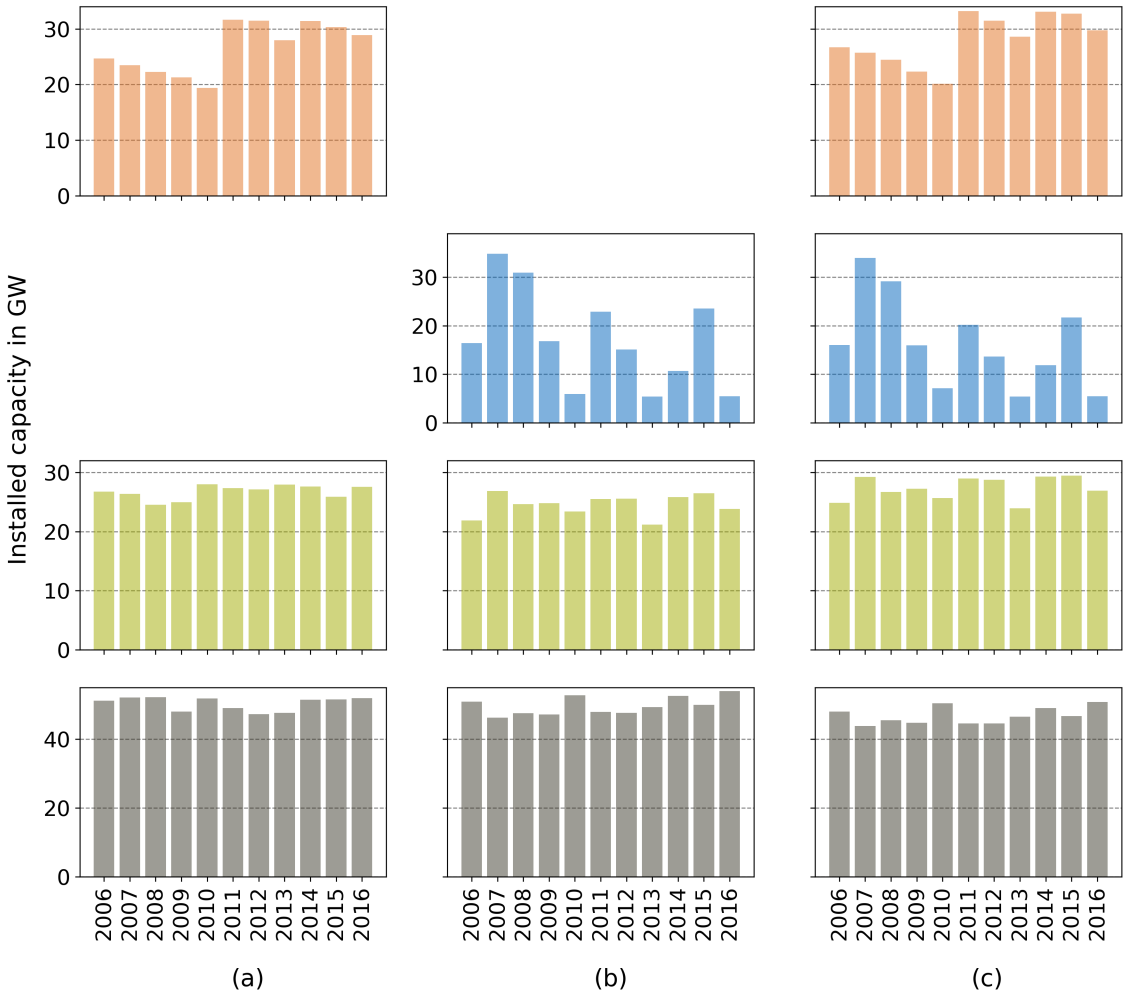

Figure 4.15: Resulting installed capacity of original time series for (a) the PV scenario, (b) the WIND scenario, and (c) the PV +WIND scenario.

be noted that already time series from eleven different years lead to varied results in all three scenarios. For example, the ranges of PV and wind power are 12-13 GW (46-47 % relative to mean) and 28-30 GW (176-181 %). Although the installed capacities of the iRES vary greatly, this only has a minor impact on the installed capacities of conventional technologies. The ranges of FPP and IPP are 3-6 GW and 5-8 GW, respectively. In relation to the mean value, their respective relative range with 13-23 % and 10-15 % is significantly smaller compared to the iRES.

4.3.2 Aggregated time series

The original time series are aggregated by two TSA approaches: clustering and clustering combined with heuristic. Clustering can be divided into k-mean and hierarchical with centroid or closest as representative (see Section 3.3.1). In the following figures, the aggregation variants are labeled as follows: If included, the heuristic selection (heur), the clustering method (kmean,

Table 4.1: Installed capacity resulting from modeling with original time series from eleven years (2006-2016) for the scenarios PV, WIND, and PV+WIND and the related power generation technologies.

Scenario	Technology		Mean	Median	Min	Max	Range
PV	PV	GW	26.6	27.9	19.4	31.6	12.2
	FPP	GW	26.7	27.2	24.5	28.0	3.5
	IPP	GW	50.4	51.5	47.3	52.2	4.9
WIND	WIND	GW	17.1	16.5	5.4	34.9	29.5
	FPP	GW	24.5	24.8	21.2	26.8	5.6
	IPP	GW	49.7	49.3	46.3	54.0	7.7
PV+WIND	PV	GW	28.0	28.6	20.1	33.2	13.1
	WIND	GW	16.4	16.0	5.4	34.0	28.6
	FPP	GW	27.4	27.2	23.9	29.5	5.6
	IPP	GW	46.8	46.6	43.9	50.9	7.0

hier) and the representative (cen, clo). The resulting aggregated time series involve six (seven in the PV+WIND scenario) to twenty days, that corresponds to 144 to 480 data points.

Time series parameters

Figure 4.16 shows selected time series parameters for each scenario and TSA approach as deviation from the original time series. The parameters are obtained from Section 4.2.2. The deviations are derived from normalized time series (MW / MW_{max} x 1000) and are given below without a unit. In the PV scenario, the average deviations of the variance of PV and the 40 % quantile of the electricity demand are in the same range for all aggregation methods (*MAE* $\pm$ *STD*: 0.9 $\pm$ 0.2 and 4.5 $\pm$ 0.7, respectively). However, outliers can be observed for TSA methods including heuristic. For the maximum of the residual load, the average deviation is low for TSA methods including heuristic (12.1) and high for TSA methods based on clustering only (45.5). Similar observations can be made in the WIND scenario. The average deviations of the gmean of wind power and the 40 % quantile of the electricity demand are in the same range for all aggregation method (15.7 $\pm$ 3.1 and 6.3 $\pm$ 0.6, respectively). However, when choosing the centroid as cluster representative, the deviations of the gmean are slightly higher compared to the closest as representative (21.7 vs. 13.1). As in the PV scenario, the average deviation of the maximum residual load is lower for TSA methods including heuristic (35.5) and higher for TSA methods based on clustering only (72.4). The findings of the single scenarios can be transferred to the combined PV+WIND scenario. The variance of PV, the gmean of wind power and the 40 % quantile of the electricity demand is on average 1.9 $\pm$ 0.3, 17.7 $\pm$ 4.3 and 6.6 $\pm$ 0.3, respectively. The average deviation of the maximum residual load is 11.8 for TSA methods including heuristic and 50.3 for TSA methods based on clustering only.

Figure 4.17 compares the duration curve of the original time series and the aggregated time series. The year 2006 is shown as an example and aggregated time series with ten days

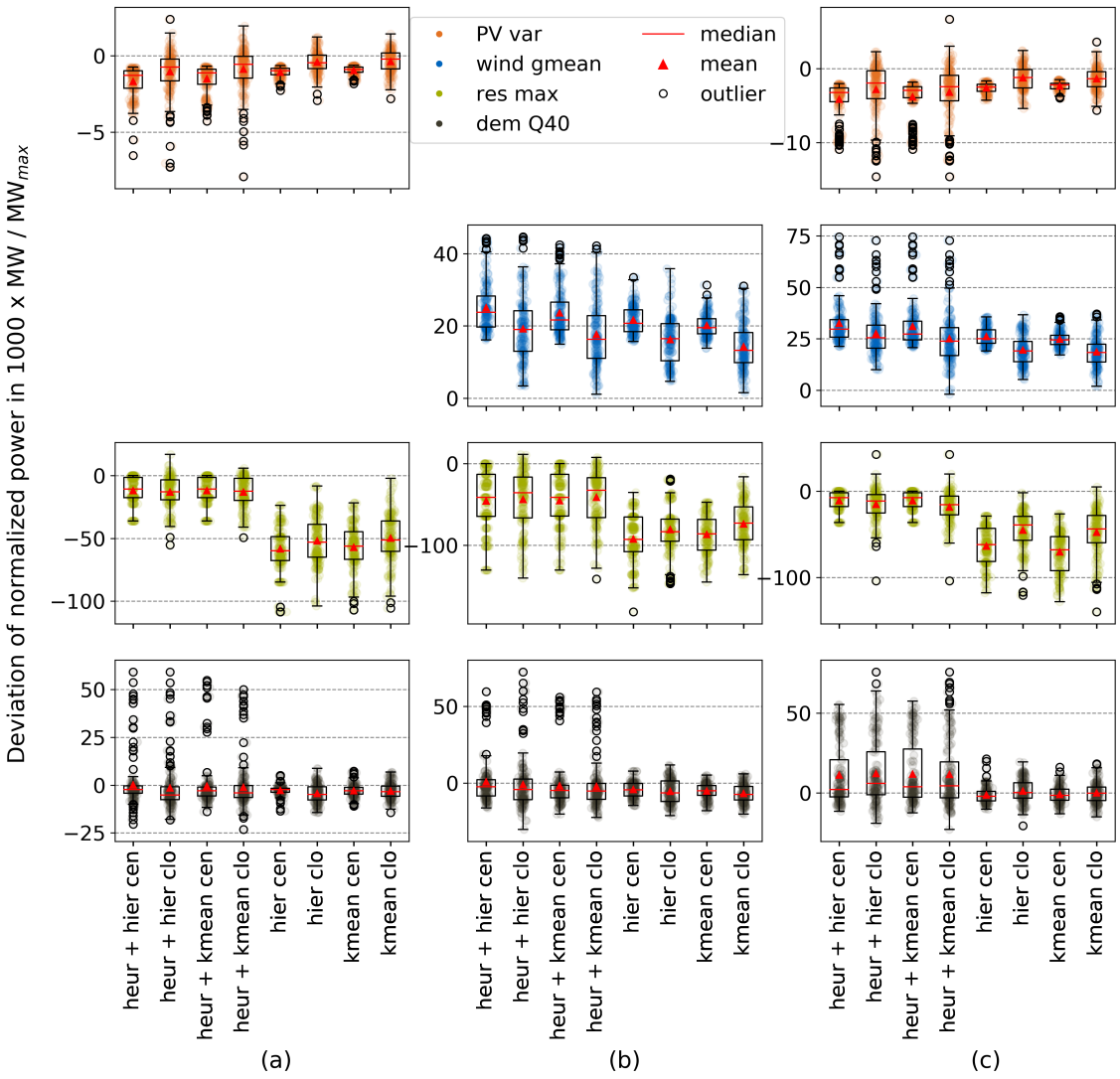

Figure 4.16: Selected time series parameters of aggregated time series as deviation from the original time series shown for each aggregation method and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

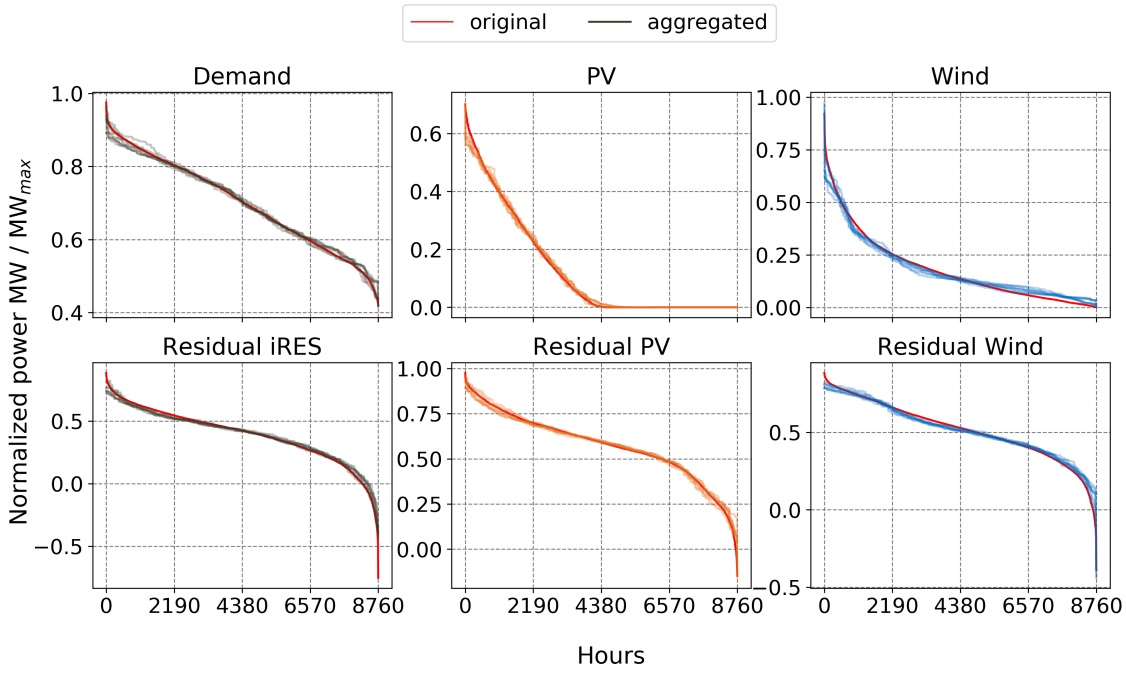

Figure 4.17: Duration curve of aggregated time series compared to the original time series shown for the year 2006, scenario PV+WIND and aggregated time series of ten clusters (days).

are selected for all TSA methods. Overall, we can observe smaller average deviations for the electricity demand (*MAE*: 7.1), the PV generation potential (6.4) and residual load when only PV is included (12.2). For the remaining time series, the deviations are higher and range between 18.0 (residual load including wind power) and 18.6 (wind power).

Modeling results and transferability from single to combined scenario

Figure 4.18 shows the modeling results for each scenario, technology, and TSA approach as deviation from the modeling results of the original time series. The results show that the selected TSA variants either systematically overestimate (PV and wind power) or underestimate (FPP and IPP) the installed capacities. In addition, there are high scattering and outliers (see Section 4.4).

In the PV scenario, the installed capacity of PV is overestimated on average (*ME*: 1.2 GW, *MAE*: 1.8 GW) by all TSA approaches, whereas the installed capacities of FPP and IPP are underestimated by -2.4 GW and -0.3 GW (*MAE*: 2.5 GW and 0.4 GW), respectively. Two patterns are observed for clustering approaches including heuristic: First, smaller mean deviations for FPP, and second, high outliers for PV and IPP (see also the difference between mean and median). A closer look at the results shows a relation between PV and IPP outliers (96 %) which occur predominantly in time series with six days. This suggests that selecting mainly extreme days of individual time series indirectly leads to an overestimation of the residual load and thus higher and lower capacities for IPP and PV, respectively. The decreased mean deviation for FPP (-0.7 GW vs. -4.1 GW) matches the findings in Section 4.2.2 as (daily) extreme values of the PV generation potential and electricity demand are taken into account.

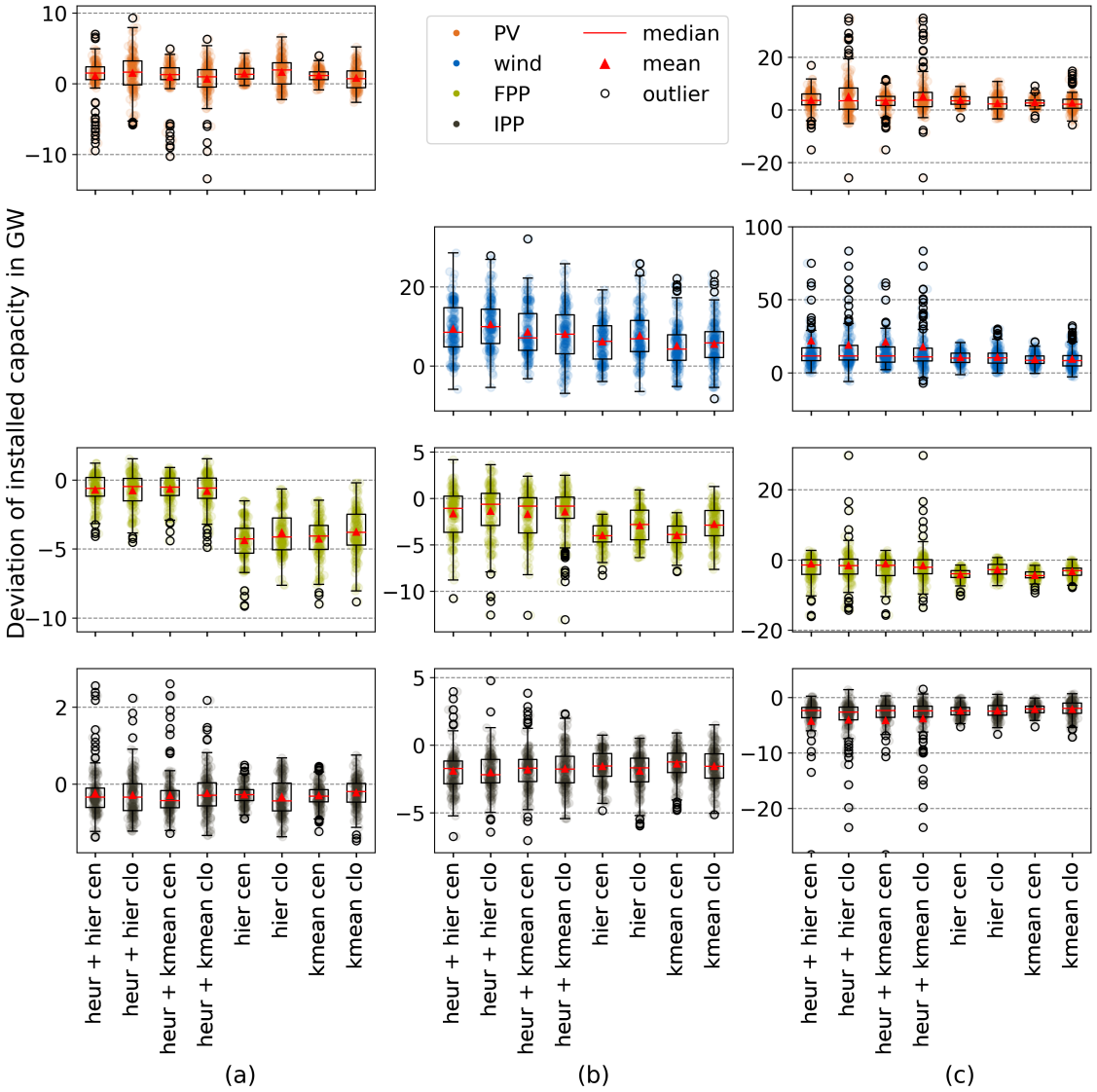

Figure 4.18: Resulting installed capacities of aggregated time series as deviation from results of original time series shown for each aggregation method and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

A comparison of the related *MAE* also underscores this assumption. For PV and IPP, the average *MAE* of all TSA approaches is in a similar range (Δ *MAE* $\leq$ 0.3 GW), whereas the *MAE* for FPP is 2.2 GW higher (with heuristic: 0.7 GW, without heuristic: 3.8 GW).

As already discovered in Section 4.2, deviations of the installed capacities are higher in the scenario WIND. The TSA approaches overestimate on average the installed capacity of wind power by 7.6 GW (*MAE*: 8.0 GW). Thereby, combined clustering and heuristic overestimates more (*ME*: 9.1 GW) than clustering alone (*ME*: 6.1 GW). FPP and IPP are underestimated by -2.5 GW (*MAE*: 2.8 GW) and -1.7 GW (*MAE*: 1.9 GW), respectively. Similar to the scenario PV, the installed capacity of FPP is more likely to be underestimated by TSA based on clustering only (*ME*: -3.4 GW vs. -1.5 GW) as extreme values of time series are not sufficiently represented. However, including heuristic can lead to outliers in negative direction (up to *ME* of 13.1 GW, see also deviation between mean and median). Comparing the *MAE* to the ones derived in Section 4.2, we observe improved values for wind power and IPP (up to 1.9 GW) but worse values for FPP (0.3 GW on average, 1.0 GW for clustering only).

In the combined scenario PV+WIND, the interaction between the iRES time series has an additional impact on modeling results. Three findings from the individual scenarios can be transferred although deviations are higher across all technologies: First, high interrelated outliers (79-83 %, PV scenario), second, smaller mean deviations for FPP (PV and WIND scenario) for clustering approaches involving heuristic, and third, systematic under- or overestimation. As in individual scenarios, TSA approaches overestimate the installed capacity of PV (*ME*: 3.6 GW, *MAE*: 4.3 GW) and wind power (*ME*: 14.8 GW, *MAE*: 14.9 GW), whereas the capacities of FPP (*ME*: -2.5 GW, *MAE*: 3.8 GW) and IPP (*ME*: -3.2 GW, *MAE*: 3.2 GW) are underestimated.

Dependence on year and number of clusters

The previous results of TSA approaches illustrate the complexity of time series and their impact on modeling results. A closer look into single TSA approaches of selected years indicates three additional characteristics, which are shown in Figure 4.19. First, the performance of TSA approaches depends on the year (see D.1 in the Appendix). For example, the k-mean clustering combined with heuristics achieves for FPP in scenario PV (WIND) small deviations of 0.5 GW (-1.6 GW) in 2007 (2009), whereas the deviation in 2006 (2010) is -3.0 GW (-6.0 GW). Second, the courses are alternating with partly large amplitudes. For example, the installed capacity of PV (wind power) ranges from six days upwards between -1.1 and 5.2 GW (-4.9 and 23.1 GW) for the k-mean closest approach in 2007 (2009) making it difficult to choose the right number of clusters (days). Third, the deviations do not always convert towards 0 GW within 20 days. For example, the deviation of FPP converts to -3 GW and -4 GW (heuristic and k-mean closest, 2006 and 2007), whereas the deviation of wind power converts to approx. 10 GW (heuristic and k-mean closest, 2009 and 2010).

4.3.3 Profiled time series

Profiling is applied to aggregated time series as described in Section 3.3.1 and analyzed in Section 4.3.2. For the PV and WIND scenarios, the aggregated time series bundle and the

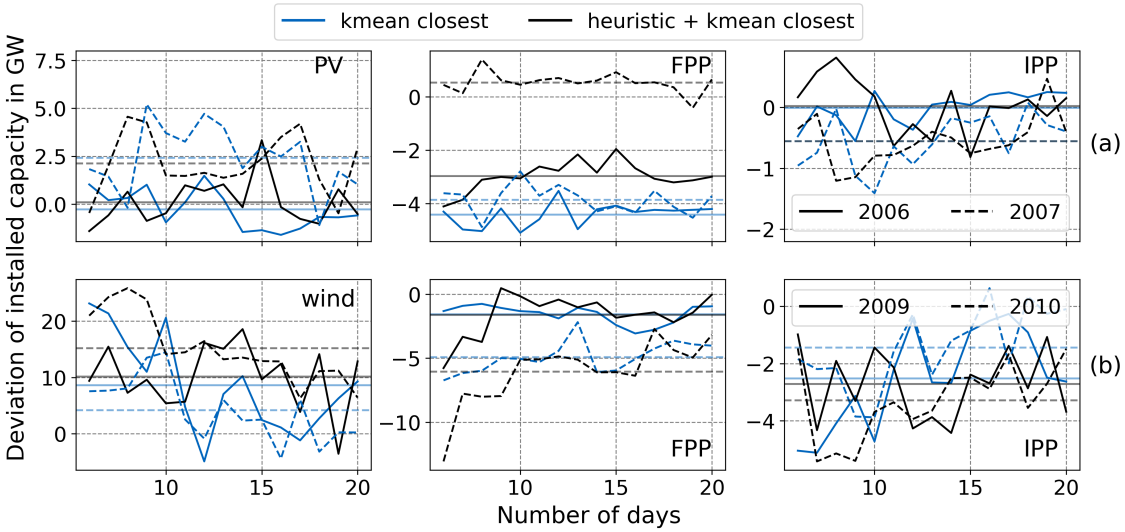

Figure 4.19: Resulting installed capacity as deviation from original results of k-mean closest and combined heuristic with k-mean closest for selected years (solid or dashed line). The results are shown depending on the number of clusters (days) for (a) the PV scenario and (b) the WIND scenario.

respective residual load are iteratively fitted to the information of the original time series. For the PV+WIND scenario, a total of seven time series are considered. The two additional time series represent the residual load of demand and both iRES and the ratio between wind and PV. To adjust the duration curve of the aggregated time series to that of the original time series (profiling step 3), time series are split into two sub-time series (see Section 4.2.3). For all scenarios involving PV, they are divided according to the median of PV (approx. 0), otherwise the median of demand is used as splitting criterion. This is not applied to the two additional time series in the PV+Wind scenario since it is assumed that the splitting of the other time series has already sufficiently covered the median-specific relations of the time series.

Time series parameters

Figure 4.20 shows selected time series parameters for each scenario and TSA approach as deviation from the original time series. As before, the deviation is derived from normalized time series (MW / MW_{max} x 1000) and is given below without a unit. Independent from the time series parameter and scenario, we can observe significantly lower deviations, which are independent from the original TSA method. The average deviation (*MAE*) is 0.4 for the variance of PV, 0.1 for the maximum of the residual load and 1.8 for the 40 % quantile for the electricity demand. The average deviation of the gmean of wind power is low in the wind scenario (1.4) and relatively high in the PV+WIND scenario (12.0). In addition, we can observe outliers for the gmean in the PV+WIND scenario and maximum of the residual load in the WIND scenario visible by the deviation of the median and mean.

Figure 4.21 compares the duration curve of the original time series and the profiled time series. As before, the year 2006 is shown as an example and aggregated time series with ten days are selected for the underlying TSA methods. Overall, we can observe a smaller

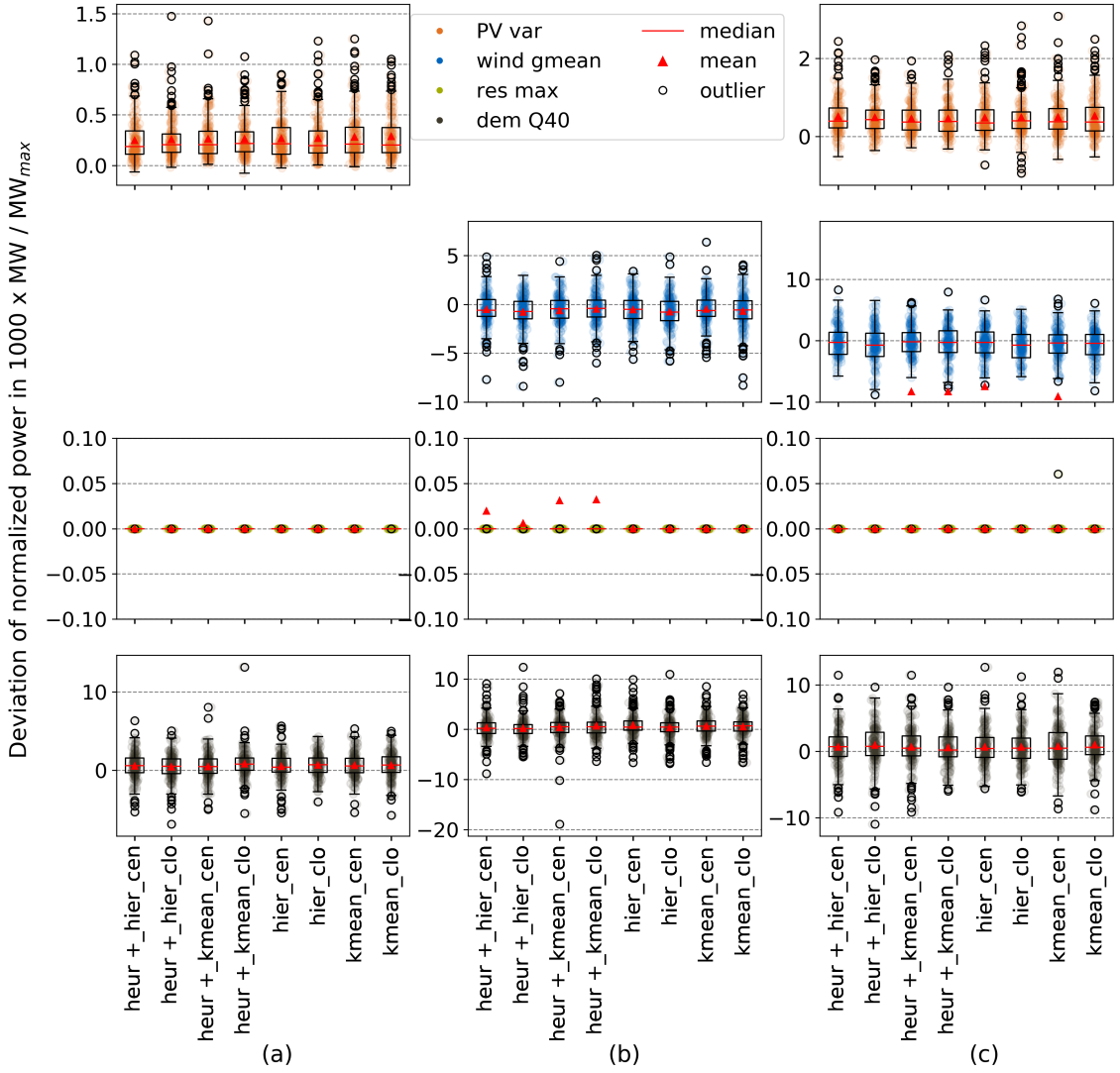

Figure 4.20: Selected time series parameters of profiled time series as deviation from the original time series shown for each aggregation method and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

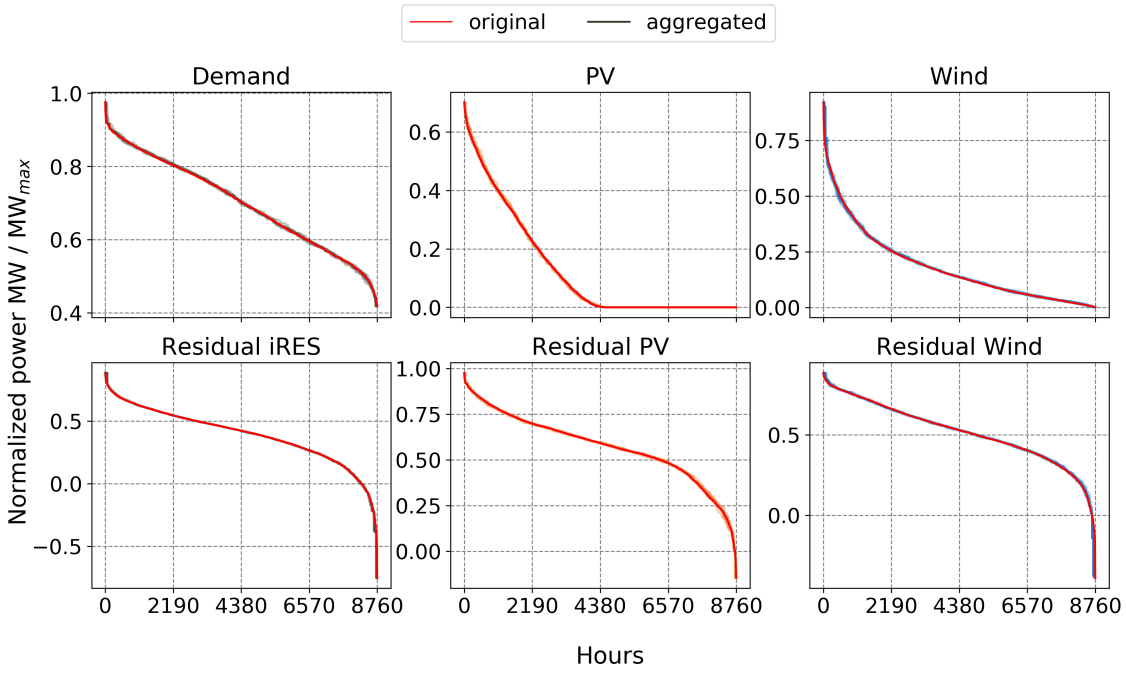

Figure 4.21: Duration curve of profiled time series compared to the original time series shown for the year 2006, scenario PV+WIND and aggregated time series of ten clusters (days).

deviations for all time series that on average range from 2.3 (PV) to 5.9 (residual load including wind power).

Modeling results and transferability from single to combined scenario

Figure 4.22 shows the modeling results for each scenario, technology and profiled time series based on different TSA approaches as deviation from the modeling results of the original time series. Independent of the underlying aggregated time series, the installed capacities of PV are on average slightly underestimated (*ME*: -0.3 GW, *MAE*: 1.3 GW) and the capacities of wind power are slightly overestimated (*ME*: 0.2 GW, *MAE*: 1.9 GW). The dispersion of these mean deviations is low across the TSA approaches with a maximum range of 0.4 GW and 0.6 GW, respectively. The deviations between the individual scenarios PV and WIND and the combined scenarios PV+WIND are also of a similar scale (PV: Δ*ME* 0.03 GW, Δ *MAE*: 0.3 GW and wind power: Δ*ME* 0.2 GW, Δ *MAE*: 0.2 GW). For the installed capacity of FPP and IPP, different directions of deviation are obtained depending on the scenario, which are independent of the profiled time series. FPP is overestimated in the PV scenario (*ME*: 0.2 GW, *MAE*: 0.2 GW) and underestimated in the WIND scenario (*ME*: -0.4 GW, *MAE*: 0.6 GW). This is reversed for the resulting capacity of IPP, which is underestimated in the PV scenario (*ME*: -0.2 GW, *MAE*: 0.2 GW) and overestimated in the WIND scenario (*ME*: 0.1 GW, *MAE*: 0.4 GW). As for PV and wind power, the dispersion of these mean deviations is low across the TSA approaches with a maximum range of 0.3 GW and 1.0 GW, respectively. The comparison of the single and combined scenarios shows that FPP and IPP are underestimated in similar scale (*ME*: -0.3 GW, *MAE*: 0.6 GW and *ME*: -0.10 GW, *MAE*: 0.5 GW, respectively).

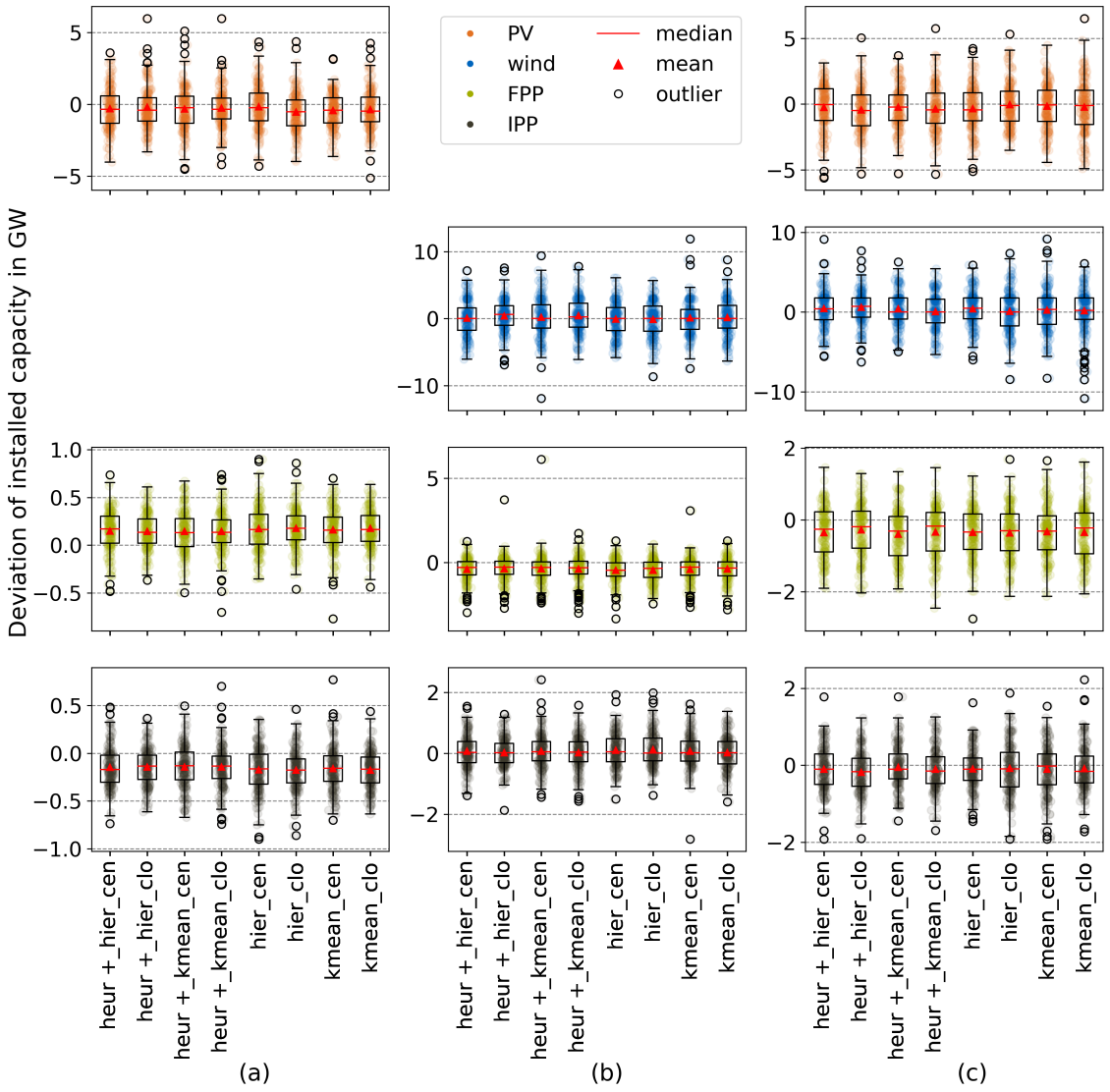

Figure 4.22: Resulting installed capacities of profiled time series as deviation from results of original time series shown for each aggregation method and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

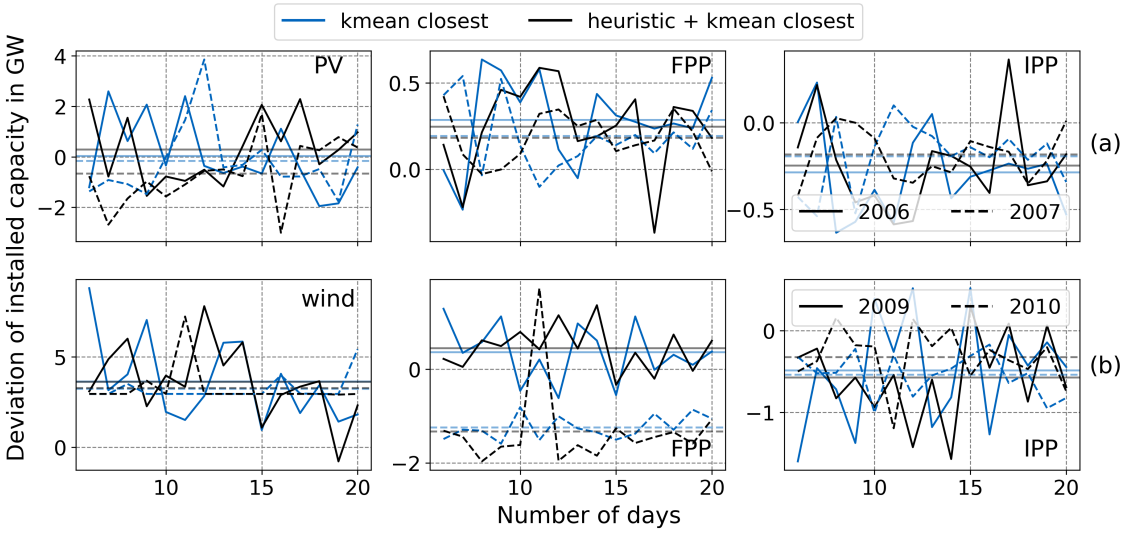

Figure 4.23: Resulting installed capacity as deviation from original results of k-mean closest and combined heuristic with k-mean closest for selected years (solid or dashed line). The results are shown depending on the number of clusters (days) for (a) the PV scenario and (b) the WIND scenario.

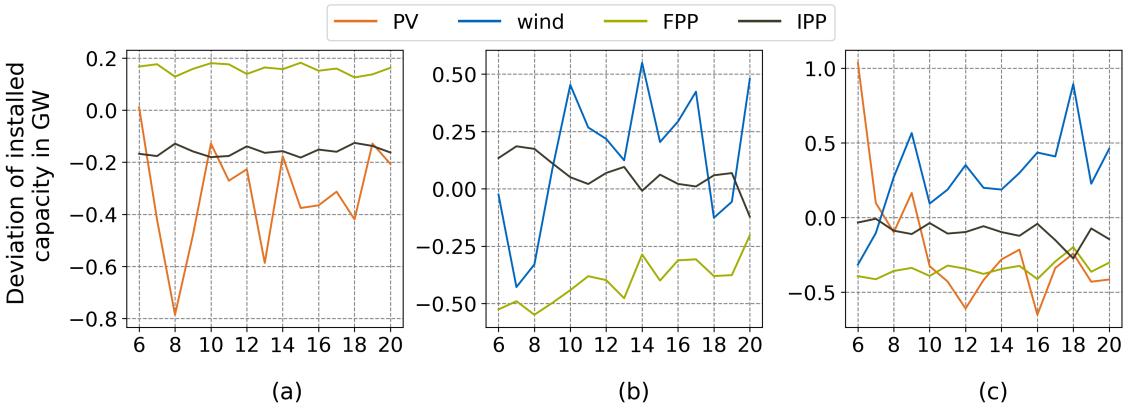

Figure 4.24: Resulting installed capacity as deviation from original results averaged across all TSA and years. The results are shown depending on the number of clusters (days) for (a) the PV scenario, (b) the WIND scenario, and (c) the PV+WIND scenario.

Dependence on year and number of clusters

A closer look into specific profiled time series of selected years indicates similar patterns to the aggregated time series. The performance of profiling can depend on the year (see also D.2 in the Appendix). For example, the profiled time series derived from k-mean clustering with and without heuristics achieves for FPP in scenario WIND smaller average deviations of 0.36 GW and 0.45 GW in 2009, whereas the average deviation in 2010 is -1.24 GW and -1.33 GW, respectively. The deviations are alternating with partly large amplitudes and do not always convert to zero. This becomes also visible in Figure 4.24 showing the average deviation

Table 4.2: Mean deviation *MAE* of the modeling results of aggregated and profiled time series from the original results compered to CNR results of a ten parameter model.

Scenario	Technology		CNR	Aggregated	Profiled
PV	PV	GW	1.5	1.8	1.3
	FPP	GW	1.0	2.5	0.2
	IPP	GW	0.3	0.4	0.2
WIND	WIND	GW	9.9	8.0	1.9
	FPP	GW	2.5	2.8	0.6
	IPP	GW	2.1	1.9	0.4

across all years and TSA approaches per cluster number. The deviations of iRES in particular show larger ranges in all scenarios (0.8-1.7 GW and 1.0-1.2 GW, respectively), whereas the deviations of conventional technologies remain at a constant level (0.1-0.3 GW for both).

Comparison to CNR

Comparing the *MAE* of modeling results obtained from profiled time series to the *MAE* achieved in CNR we can observe a reduction of 2.1 GW on average. In Table 4.2 the averaged *MAE* for the profiled (and aggregated) time series is compared to the average *MAE* of ten parameter models derived from CNR (see 4.2). In the PV scenario, the reduction of the error is between 0.2 GW (PV) and 0.8 GW (FPP). In the WIND scenario, the error reduction is higher and ranges from 1.7 GW (IPP) to 8.0 GW (wind power).

4.4 Comparison of modeling results[4]

Applying profiling to the aggregated time series shows a significant improvement for all technologies and scenarios for both, mean deviation *ME* (39 %-98 %) and standard deviation *STD* (34 %-91 %). As a result, the *ME* is at or below 0.4 GW and the *STD* does not exceed 2.6 GW.

Figures 4.25-4.27 show the distribution of the results as a box plot and histogram to display mean values, medians, and IQR as well as standard deviations. The first diagrams (a) present the variance of original results compared to mean across all considered years. The second row of diagrams (b) shows the results of aggregated time series, whereas the third row (c) presents the results of the profiled time series. Table 4.3 quantitatively supplements the visualization with the summarized *ME* and *STD*.

In the PV scenario, the *ME* and *STD* of the installed capacity of PV is reduced by 72 % and 34 %, respectively. The *ME* for FPP and IPP are also significantly improved by 93 % (*STD*: 90 %) and 39 % (60 %), respectively. The resulting *MAE* for PV is 1.1 GW and is, thus, below the errors of CNR (-0.7 GW) and aggregated time series (-0.4 GW). The *MAE* for FPP and IPP amounts to 0.2 GW and is also below the values previously achieved (see also Table 4.2). The reduction of deviations is accompanied by an improved representation of relevant time

[4]This section is based on the *profiling paper* - Section 4.3 [40].

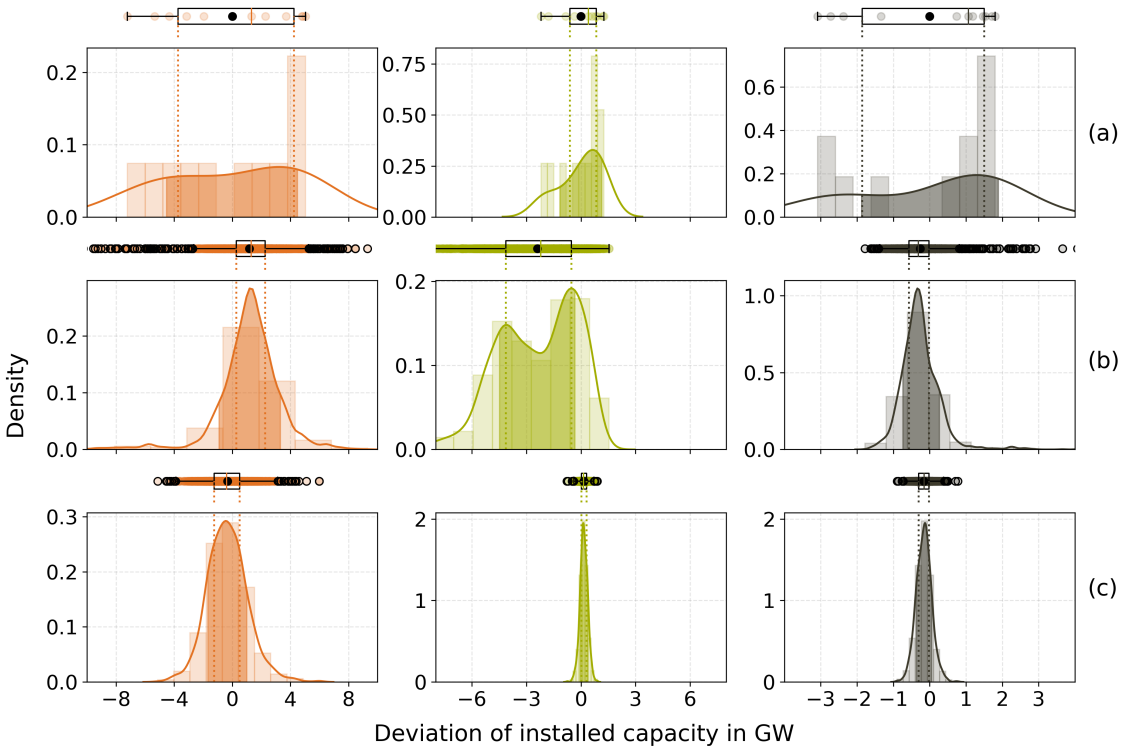

Figure 4.25: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The modeling results of (a) original time series, (b) aggregated time series, and (c) profiled time series are shown for the technologies PV, FPP, and IPP of the PV scenario.

Table 4.3: Mean deviation *ME* of the modeling results of aggregated and profiled time series from the original results and the related standard deviation *STD*. The original results are set in relation to their mean value.

Scenario	Technology		Original	Aggregated	Profiled
PV	PV	GW	± 4.5	1.2 ± 2.1	-0.3 ± 1.4
	FPP	GW	± 1.2	-2.4 ± 2.1	0.2 ± 0.2
	IPP	GW	± 2.0	0.2 ± 0.5	-0.2 ± 0.2
WIND	WIND	GW	± 10.1	7.1 ± 6.6	0.1 ± 2.6
	FPP	GW	± 1.8	-2.6 ± 2.4	-0.4 ± 0.7
	IPP	GW	± 2.6	-1.6 ± 1.5	0.1 ± 0.5
PV+WIND	PV	GW	± 4.5	3.4 ± 4.8	-0.3 ± 1.7
	WIND	GW	± 9.3	14.1 ± 25.8	0.3 ± 2.4
	FPP	GW	± 2.0	- 2.5 ± 5.2	-0.3 ± 0.7
	IPP	GW	± 2.5	- 2.9 ± 5.1	-0.1 ± 0.6

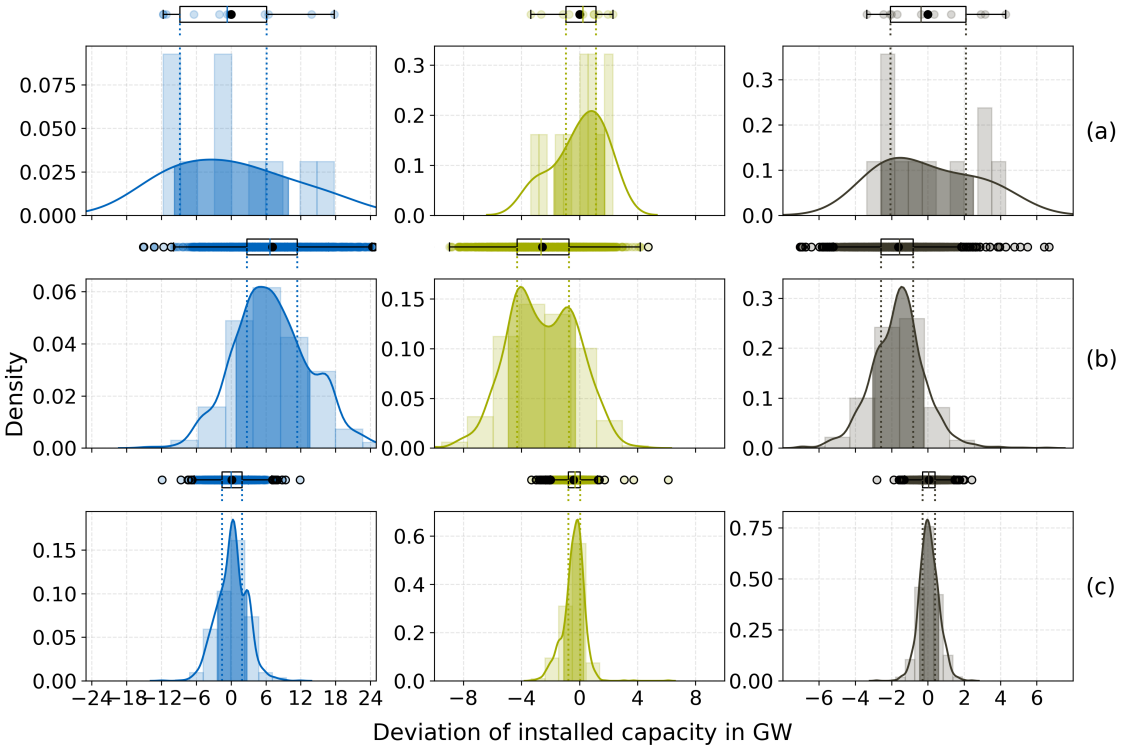

Figure 4.26: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The modeling results of (a) original time series, (b) aggregated time series, and (c) profiled time series are shown for the technologies wind power, FPP, and IPP of the WIND scenario.

series parameters (see, e.g., Figure 4.16 and 4.20). On average, the deviation of these is 18 % for aggregated time series and 2 % for profiled time series. In the WIND scenario the improvements are even higher as the *ME* (*STD*) have been on average reduced by 98 % (61 %) for wind power, 84 % (69 %) for FPP and 96 % (63 %) for IPP. The related *MAE* amount to 2.0 GW for wind power, 0.6 GW for FPP and 0.4 GW for IPP and are, therefore, at least 1.5 GW lower than previous values. As before, the small deviations are accompanied by a better representation of relevant parameters which error decreases from 7 % to 1 %. Also for the more complex PV+WIND scenario, significant *ME* (*STD*) improvements of 93 % (65 %) for PV, 98 % (91 %) for wind power, 87 % (86 %) for FPP and 97 % (89 %) for IPP can be stated. The *MAE* is in the same range ($\pm$ 0.3 GW) as in the individual scenarios and, thus, validates the assumption that findings from simplified energy systems can be transferred to complex ones. Despite stronger interaction between time series or parameters in the profiling algorithm, a significant improvement of the relevant time series from 20 % to 2 % is achieved.

Besides *ME*, *STD*, and *MAE*, a significant reduction of extreme values of 61 % on average can be observed. For example, negative outliers of PV have been reduced from -15.6 GW to -5.1 GW (PV scenario) and -25.7 GW to -5.6 GW (PV+WIND), whereas positive outliers of wind power have been reduced from 32.1 GW to 11.9 GW (WIND) and 242.1 GW to 9.1 GW (PV+WIND).

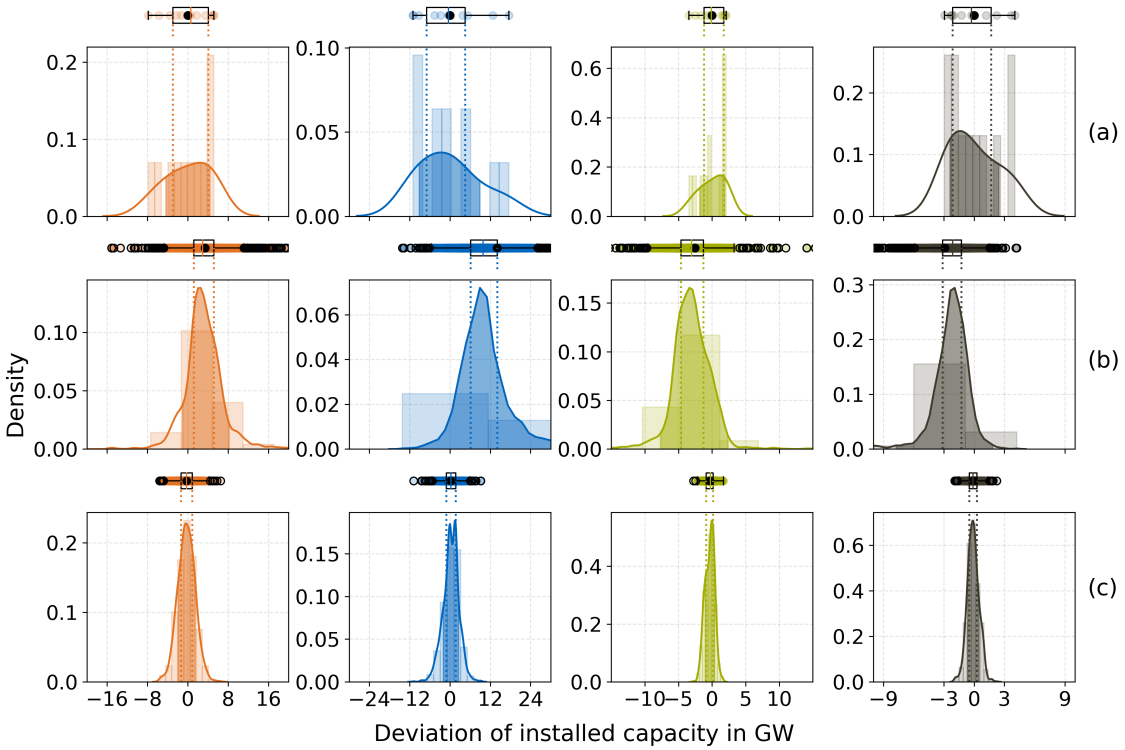

Figure 4.27: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The modeling results of (a) original time series, (b) aggregated time series, and (c) profiled time series are shown for the technologies PV, wind power, FPP, and IPP of the PV+WIND scenario.

Chapter 5

Discussion[1]

5.1 Summary

This thesis addresses a central challenge in energy system modeling, which is the limited reliability of time series aggregation methods that are applied to compensate the increasing complexity of energy system models. To reduce the temporal model complexity annual time series representing the electricity demand and generation potential of iRES are aggregated to time series of several days. However, the results are not always robust but depend on power system configurations and years. So far, we have little knowledge about why aggregation of time series performs better or worse. In this thesis a data-analytical approach is developed that, firstly, improves the understanding of the interactions between time series and modeling results and secondly, transfers the knowledge gained in the form of relevant time series parameters to aggregated time series. The basis for this approach is a highly simplified energy system model that consists of two conventional generation technologies (peak load power plant (flexible, FPP), base load power plant (inert, IPP)) and one or two iRES (PV, wind power) resulting in three scenarios (PV, WIND, PV+WIND). A bundle of three time series (PV, wind power and electricity demand) for a total of eleven years form the data basis.

To investigate the interaction between time series and model results and to identify relevant time series parameters the CNR method is proposed that includes clustering and nested regression. An extensive database of 170+ features for 9000+ time series bundles is derived from the original time series bundle. These are linked to the results of the energy system model from PV and WIND scenario that are the installed capacity of the three power generation technologies. CNR is compared to the feature selection methods LASSO, LASSOLARS and ElasticNet by evaluating selected parameters and their performance using *AIC*, *MAE* and R^2 as goodness-of-fit criteria. The results show a high selectivity for CNR, medium selectivity for LASSO and LASSOLARS, whereas ElasticNet has a low selection rate. Only a few parameters, such as relative variance and the maximum of the residual load or the mean and median of wind power, are identified by all feature selection methods. However, the majority of parameters are only identified as relevant by one or two methods. The transfer of the most important parameters into a regression model enables the quantitative evaluation of the feature selection

[1] This chapter is based on the *identification paper* - Section 4 and 5 [39] and the *profiling paper* - Section 4 and 5 [40].

methods. CNR shows significantly better values for *MAE* and *AIC* than the methods from the literature. A better performance with regard to these criteria is to be expected, as they are implemented in the CNR algorithm that is therefore optimized accordingly. Further, the combined evaluation of different parameters and parameter subsets of CNR leads to a more optimal feature selection at different subset sizes, especially with only a small selection of two to ten parameters. However, the computing time of CNR is significantly longer (one iteration takes about 40 minutes). The combination of CNR with another feature selection method as pre-feature selection can reduce the overall time required. Therefore, CNR complements existing feature selection methods and is suitable for applications where selectivity and the consideration of different parameter sets as well as an in-depth understanding of data are relevant.

Overall, CNR is applied three times to two different data sets (reduced data set without residual load in the *identification paper* [39], extended data set with residual load in this thesis and *profiling paper* [40]). The results can be reproduced in terms of content, meaning that the algorithm is robust despite the random selection of parameters in the pre-selection. Comparing the results to previous findings derived from the data set without residual load, two observations stand out. Firstly, the relevance of the correlation between PV and demand is replaced by parameters of the residual load. Secondly, parameters describing the electricity demand (e.g., range and 20 % or 40 % quantile) are partly replaced by parameters of the residual load (e.g., maximum and 30 % or 45 % quantile). These results indicate that the interactions between time series are highly relevant and cannot be described by correlations alone. Rather, time series parameters describing the residual load are of central importance. A comparison of the results from the PV and WIND scenarios shows that modeling results of the PV scenario can be better explained by a limited number of time series parameters as those of the WIND scenario. This is consistent with the visualizations of the clustering of time series from the exploratory data analysis, in which no clear patterns are apparent in the wind power time series. Without clear patterns, it is therefore only possible to derive time series parameters to explain the modeling results to a limited extent. The results of the parameter analysis illustrate the complex relationship between time series and modeling results. For each technology, the model performance criteria, for example, *MAE* converges and ranges from 0.3 GW (FPP) to 9.9 GW (wind power). This suggests that other characteristics of time series are missing that cannot be described by simple statistical parameters or that a linear approach to describe the relationship is limited. An in-depth analysis emphasizes the complexity of time series and their interaction with energy system models. For example, the total aggregated time series may show a good representation of average values or duration curves, whereas a split time series (according to the median of PV) may show significant differences leading to modeling deviations. Transferred to the aggregation and profiling of time series, it can be assumed that a (mean) residual error remains. Therefore, time series aggregation methods and profiling should focus on minimizing systematic deviations to achieve robust modeling results.

The identified relevant parameters and further findings are transferred to aggregated time series by the proposed profiling method. Thereby, aggregated time series are iteratively adjusted in three steps to align correlations, single and average time series parameters with those of the original time series. The development and evaluation of the method is based on an extensive analysis framework covering three energy system scenarios (with three or four

power generation technologies) and two aggregation approaches with in total eight aggregation variations. Overall, 3700+ time series bundles and modeling results for aggregated and profiled time series are analyzed that allows to make statistically significant and generalized statements.

The modeling results of the original time series show large differences between the years, especially for the installed capacity of PV and wind power. This confirms the required high sensitivity of the defined energy system (see Section 3.1).

The results of aggregated time series validate already known findings from the literature, such as the performance of aggregation methods depends on the underlying year and on time series aggregation method is superior to another. In addition, the deviation of modeling results have alternating courses along the increasing number of cluster (days) without converging to zero deviation. This reinforces the impression that current approaches are not able to represent (all) relevant characteristics of time series. Rather, systematic deviations are found, indicating that some relevant time series parameters are not controlled or covered by the aggregation methods. For example, the installed capacity of PV and wind power are overestimated on average, whereas the capacity of conventional power plants are underestimated on average. On the one hand, aggregation approaches including heuristic tend to have more outliers. On the other hand, these aggregation methods perform better for peak load power plants as extreme values of demand (and residual load), which are identified as relevant parameter, are better represented. In contrast, a simple explanation for systematic deviations cannot be derived from the results. For example, the underrepresented variance of the PV generation potential and overrepresented geometric mean of the wind power generation potential contribute to an overestimation of the installed power of PV and wind power. Both incorrect representations can be attributed to the smoothing effect of clustering, in which the variety - especially for the wind time series - can only be represented inadequately. However, it is also decisive how the deviations are related or in which context they occur. For example, a duration curve (of demand) can on average show a good fit, but from a different perspective (split based on PV) systematic deviations may occur. In summary, results of individual scenarios allow a better interpretation of more complex scenarios. Deviations of the modeling results can be traced back to specific time series (parameters) and, together with the knowledge gained from the identification of relevant time series parameters, effective extensions of the aggregation methods can be derived.

The proposed profiling method improves the representation of relevant time series parameters for aggregated time series leading to low average deviation of modeling results. Moreover, the standard deviation and outliers are significantly decreased. The performance of profiling is independent of the aggregated time series or its aggregation approach. Thus, it is possible to form a representative time series independent of the selected days and profiling can be seen as an extension to current aggregation methods to improve their reliability. However, the performance of profiling still varies between different years, although showing smaller deviations and outliers overall. It can be assumed that not all relevant time series parameters are identified or covered, yet. Analyzing the modeling results according to the number of days does not show a clear tendency that more days lead to a smaller mean deviation of installed capacities and thus the profiling works better. Based on this, it can be concluded that, depending on the scenario, there is a range of days, where the performance of profiling is independent of the number of days (from six to 20).

Besides the comparison to aggregation methods, a general assessment of the results based on the original time series is useful to evaluate the performance of profiling in absolute terms. The standard deviation of modeling results using original time series is on average 3 GW higher (from 1.2 GW for FPP and up to 10.1 GW for wind power) than those resulting from profiled time series (from 0.2 GW to 2.6 GW). Together with the low mean deviation (between -0.4 GW and 0.3 GW), the performance of profiling in terms of mean deviation and standard deviation of the modeling results can be interpreted as acceptable.

Comparing the profiling results to the CNR results, we observe a high reduction of the *MAE*, especially in the wind scenario. Thus, the difference between PV and wind power or the respective scenarios, that are significant in the CNR analysis, are eliminated. This suggests that profiling is able to transfer relevant characteristics of the original time series to the aggregated time series and, in addition, to represent further information not (explicitly) included in the ten parameter models of CNR or profiling.

The computing time of the energy system models can be significantly reduced by aggregating and profiling time series. This enables practitioners of complex energy system models - in addition to detailed energy system models - to carry out analyses with time series from several years. By additionally calculating a large number of profiled time series, for example, by applying different aggregation methods as done within this thesis, the sensitivity of the model with regard to different time series can be evaluated. In other words, a sensitivity analysis can be carried out, as it is usually done for other model parameters, such as investment costs or emission limits. Instead of only interpreting absolute results, these could be extended by a deviation that allows statements to be made about the sensitivity or stability of the results.

5.2 Application and classification

In addition to the aggregation methods described in the literature, there are python based modules that implement these aggregation methods. For example, *tsam* includes different aggregation methods such as k-means, k-medoids and hierarchical and provides additional features like tuning of aggregation parameters and extended representation methods to keep statistical attributes [29]. Two aggregation methods, k-mean and hierarchical, with default parameterization (except from noTypicalPeriods, hoursPerPeriod, clusterMethod) and advanced parameterization (including sortValues and evalSumPeriods) are used for a comparison with the profiling method. As in previous analysis, time series are aggregated to six to 20 representative days. The results are shown in Figure 5.1 including the modeling results from the original time series. As in the previous comparison of aggregated and profiled time series, we can observe high deviations (*ME* and *MAE*) for aggregated time series. Aggregated time series using default settings show deviations between -3.6 GW (FFP) and 11.0 GW (wind power). The *MAE* is between 2.6 GW (inert power plant) and 11.1 GW (wind power). The deviations of aggregated time series derived from advanced settings are slightly smaller and range between -3.3 GW (FPP) and 8.2 GW (wind power). The *MAE* is between -2.1 GW (IPP) and 8.7 GW (wind power). The deviations are significantly smaller for profiled time series that are -0.1 GW (IPP) to -0.4 GW (FPP) in terms of the *ME* and 0.5 GW (IPP) to 2 GW (wind power) for the *MAE*. Thus, the implemented aggregation methods within this thesis correspond to the latest state of externally implemented aggregation algorithms.

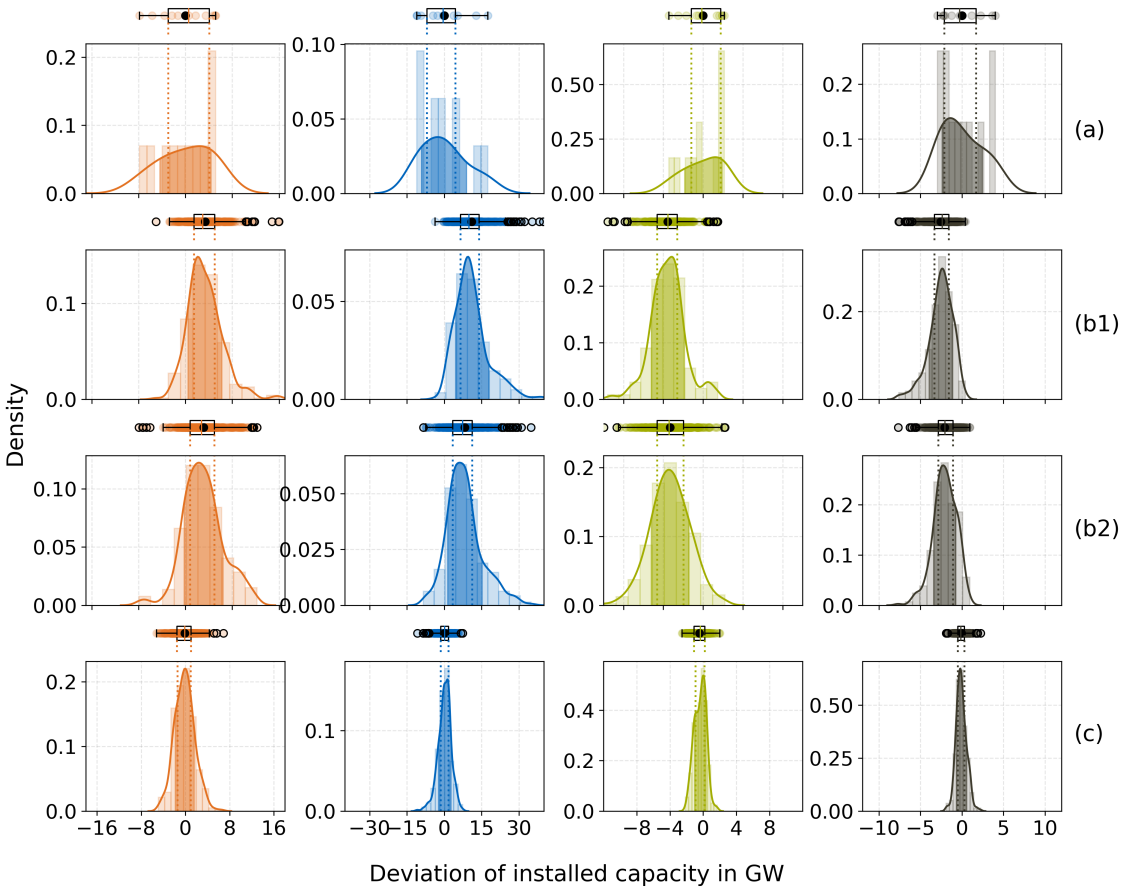

Figure 5.1: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The modeling results of (a) original time series, (b1) aggregated time series calculated with default *tsam*, (b2) aggregated time series calculated with advanced *tsam*, and (c) profiled time series are shown for the technologies PV, wind power, FPP, and IPP of the PV+WIND scenario.

The next example shows an application of profiled time series. As already introduced, not only the aggregation of time series is relevant to practitioners of energy system models but also the selection of representative annual time series (or year). As shown by the CNR analysis, there are multiple time series parameters that need to be taken into account when selecting a representative year. Alternatively, the selection of a representative annual time series can be included in the aggregation and profiling. Instead of annual time series, time series covering multiple years are aggregated and profiled according to the overall time series characteristics. The Figures 5.2 and 5.3 show the modeling results of original, aggregated and profiled time series when two or five years of data are considered. The aggregated and profiled time series include six to twenty days and are applied in the PV+WIND scenario. The deviation (*ME*) of aggregated time series derived from original time series of two years is on average between -2.9 GW and 12.5 GW, whereas profiling has an average deviation of -0.6 GW to 0.0 GW. The deviations of both, aggregated and profiled time series are slightly higher when they are representing time series of five years. The aggregated time series have an average

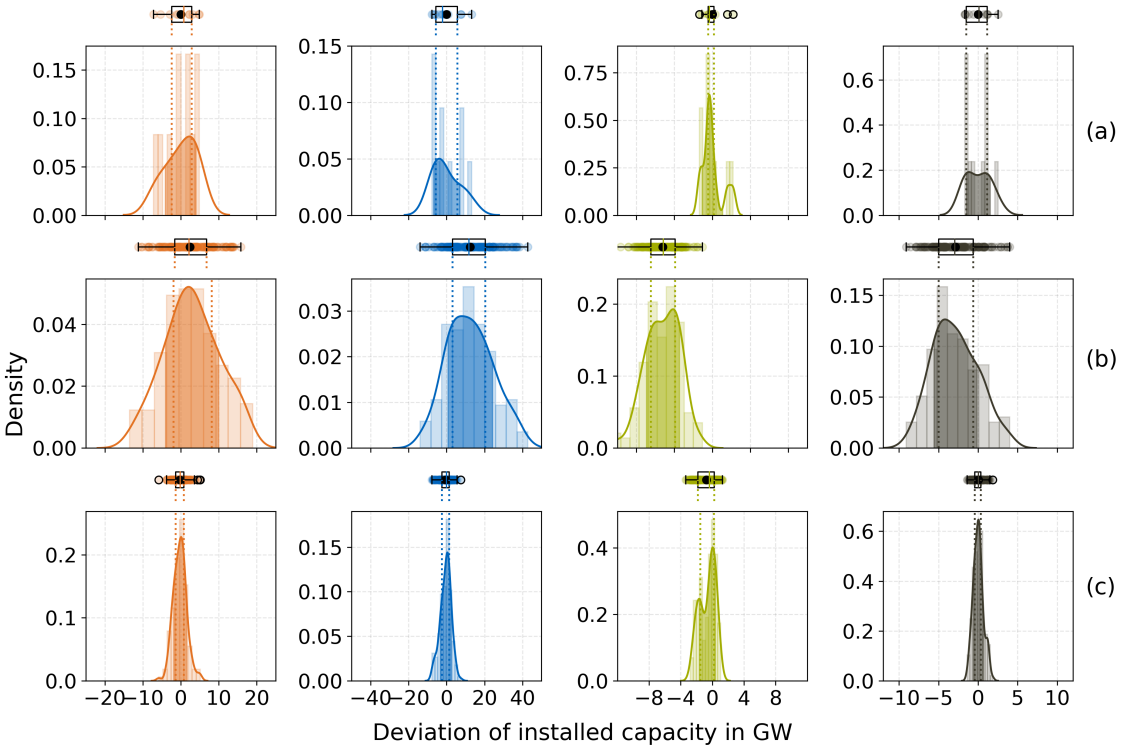

Figure 5.2: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The original time series compromises two years. The modeling results of (a) original time series, (b) aggregated time series calculated with default *tsam*, and (c) profiled time series are shown for the technologies PV, wind power, FPP, and IPP of the PV+WIND scenario.

deviation of -7.0 GW to 15.6 GW and profiling reaches an average deviation between -2.7 GW and 0.5 GW. Thus, multiple years with up to 1825 days can be aggregated and profiled by a maximum of 20 days (1 % of the original data) without losing relevant characteristics of the original time series.

5.3 Outlook

Not all relevant time series characteristics have been identified yet and further research is needed to understand and eliminate the remaining deviations. So far, profiling has only been applied to three scenarios consisting of four power generation technologies represented in a single-node model. A general transferability of the findings from simple systems (scenario PV and WIND) to a more complex system (scenario PV+WIND) has been demonstrated in this thesis. In the next step, an extension and transfer of the profiling algorithm especially to energy systems with multiple sites should be made. The application of CNR and profiling is not limited to energy systems and technologies selected in this thesis. Both methods can be transferred to other energy system configurations with further iRES and storage technologies in a next step (e.g., energy building model or extended energy system including hydro power) to further

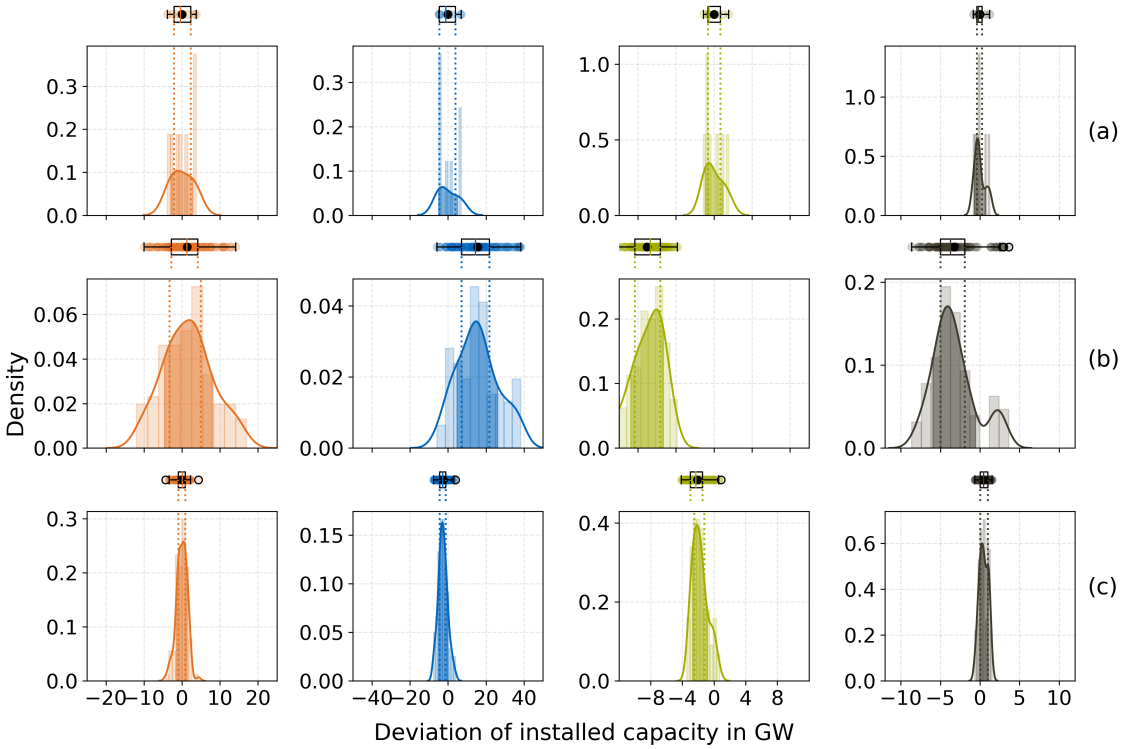

Figure 5.3: Distribution of the modeling results as deviation from the original results or average result shown as a box plot and histogram. The original time series compromises five years. The modeling results of (a) original time series, (b) aggregated time series calculated with default *tsam*, and (c) profiled time series are shown for the technologies PV, wind power, FPP, and IPP of the PV+WIND scenario.

demonstrate the transferability of CNR and profiling. Besides, the profiling algorithm can be improved, for example, by a day- or data-point-specific weighting to better represent specific parameters similar to extreme values. The CNR algorithm is computationally extensive and can be made more efficient, for example, by reversing the calculation from a one-parameter model to a multi-parameter model. In addition, further research can be carried out to determine the extent to which time series can be aggregated and profiled across multiple years or to use profiling for converting historic aggregated time series into future time series considering expected climate conditions of extreme weather events or changed averages (e.g., [32, 66]).

Chapter 6

Conclusion[1]

In conclusion, this thesis demonstrates that an information based time series aggregation approach improves the representation of relevant characteristics of the original time series and, thus, leads to a higher reliability of the resulting modeling results. A comprehensive analysis of time series parameters and modeling results as well as the evaluation of aggregation methods contributes to a better understanding of the interaction between time series and energy system models. In particular, the proposed profiling method complements existing time series aggregation methods. The CNR and profiling algorithms are generic and not limited to the defined energy system and selected time series of this thesis, but can be extended to additional time series. A general transferability of the findings from simple to complex energy systems has been shown. Further, CNR and profiling can make a valuable contribution to the selection of representative time series or years. In addition, modeling with several (profiled) time series of one or more years is recommended, so that a sensitivity analysis of the model with regard to the time series is possible and statements regarding the stability of the model and its results can be made. At best, this thesis motivates and inspires other researchers and discussed improvements for further research are taken up.

[1] This chapter is based on the *profiling paper* - Section 5 [40].

Appendix A

Extended theoretical background: Time series analysis

According to Box et al. "*[a] time series is a sequence of observations taken sequentially in time*" [9], p.1. Adjacent observations are dependent and the focus of time series analysis is on developing and applying models describing these inter-dependencies. In this thesis, time series or data determined from time series are used, so that the application of time series analysis seems obvious. However, the focus is not on modeling time series but on analyzing and understanding the interaction between time series and energy system models. Parameters from the time series analysis that can be used to describe the time series have not been identified to be relevant. For completeness, the time series analysis is summarized below based on [9], [10] and [59] that are recommended for further reading.

Models of univariate analysis, for example auto-regressive (AR), moving average (MA) or combined ARMA models require *stationary* time series. A time series as a *stochastic process* can be characterized by the mean μ (Equation 2.1), the variance σ^2 (Equation 2.8), the covariance γ as well as the auto-correlation ρ. The latter two are similar to the covariance and correlation defined in Equation 2.14 and Equation 2.15, but these parameters are related to the time series itself and not to another time series:

$$\gamma_k = cov[z_t, z_{t+k}] = E[(z_t - \mu)(z_{t+k} - \mu)]] \rightarrow \widehat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k}(z_t - \overline{z})(z_{t+k} - \overline{z}) \tag{A.1}$$

$$\rho_k = \frac{cov[z_t, z_{t+k}]}{\sigma_z^2} = \frac{\gamma_k}{\gamma_0} \rightarrow \widehat{\rho}_k = \frac{\sum_{t=1}^{n-k}(z_t - \overline{z})(z_{t+k} - \overline{z})}{\sum_{t=1}^{n-k}(z_t - \overline{z})^2} \tag{A.2}$$

where time series **z** compromise $t \in T$ observations, lag k is a shifting operator and $\widehat{\cdot}$ is an estimate[1]. A stochastic process is (strictly) stationary if its properties are time independent. Thus, regardless of the set of observations (by selecting different lag k to shift the timestamps of the realizations backward or forward), the properties remain unchanged. The *white noise* is one particular process that is also included in many models. It is strictly stationary with a mean of zero and uncorrelated observations ($\gamma_k = 0, k \neq 0$).

The *auto-correlation function (ACF)* is used to visualize the auto-correlation by applying different lag values k with $0 \leq k \leq l$ and $l \ll n$, for example, $l < n/4$. The *partial ACF (PACF)*

[1] Note, that the observations in time series are time stamps, thus i is replaced by t to make it more explicit

represents an adjusted ACF by removing the linear dependencies between z_t and z_{t+k} that are $\{z_{t+1}, ...z_{t+k-1}\}$:

$$\Phi_{kk} = corr(z_{t+k} - \hat{z}_{t+k}, z_t - \hat{z}_t) \tag{A.3}$$

with $\hat{}$ as regression of z_t and z_{t+k}, for example, $\hat{z}_t = \beta_1 z_{t+1} + ... + \beta_{k-1} z_{t+k-1}$ (see Equation 2.20). ACF and PACF are applied to parameterize AR(p), MA(q) or ARMA(p,q) models by deriving the order p and q as specific lags.
The ACF and PACF can also be applied to identify patterns of a non-stationary time series. For example, for a demand or PV time series with daily patterns, high auto-correlation for multiples of 24 hours ($k = [24, 48, ...]$) can be found. As already indicated by this example, time series are not stationary but often include trend and seasonal components. Using decomposition, both components can be identified and removed so that only a random noise component remains. For example, the *classical decomposition model* is additive:

$$z_t = m_t + s_t + v_t \tag{A.4}$$

where m_t is the trend, s_t is the seasonal component and v_t the random noise. In a first step, the trend is estimated using a moving average filter for a defined period $d = 2q$. In a second step, the seasonal component is estimated after adjusting the time series by its trend. In a third step, the trend is re-estimated from the adjusted time series according to its seasonal component before the random noise component is calculated using both, the recalculated trend and derived seasonal component.

Appendix B

Feature overview

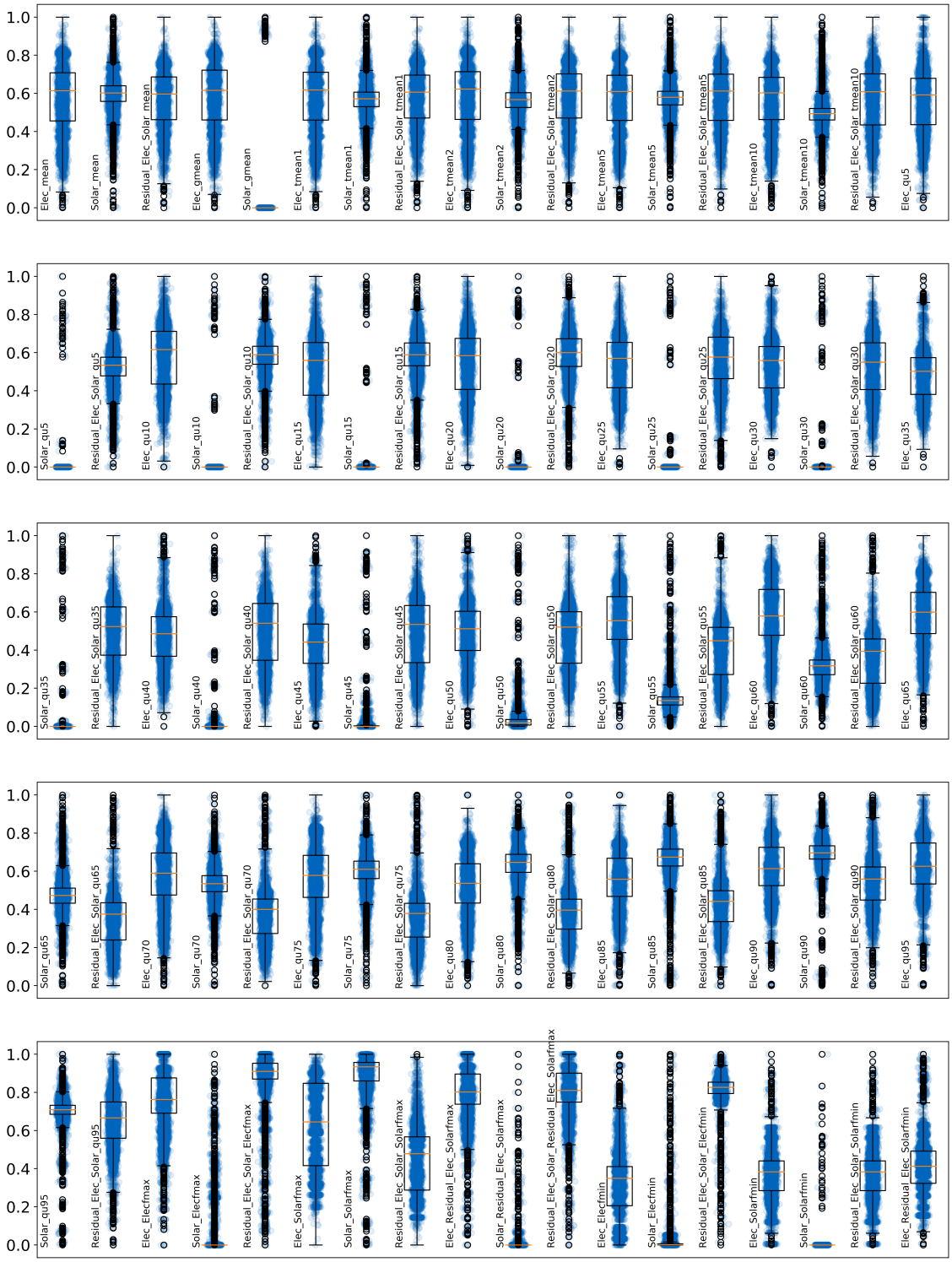

Figure B.1: Overview of normalized features of the PV scenario included in CNR - part I.

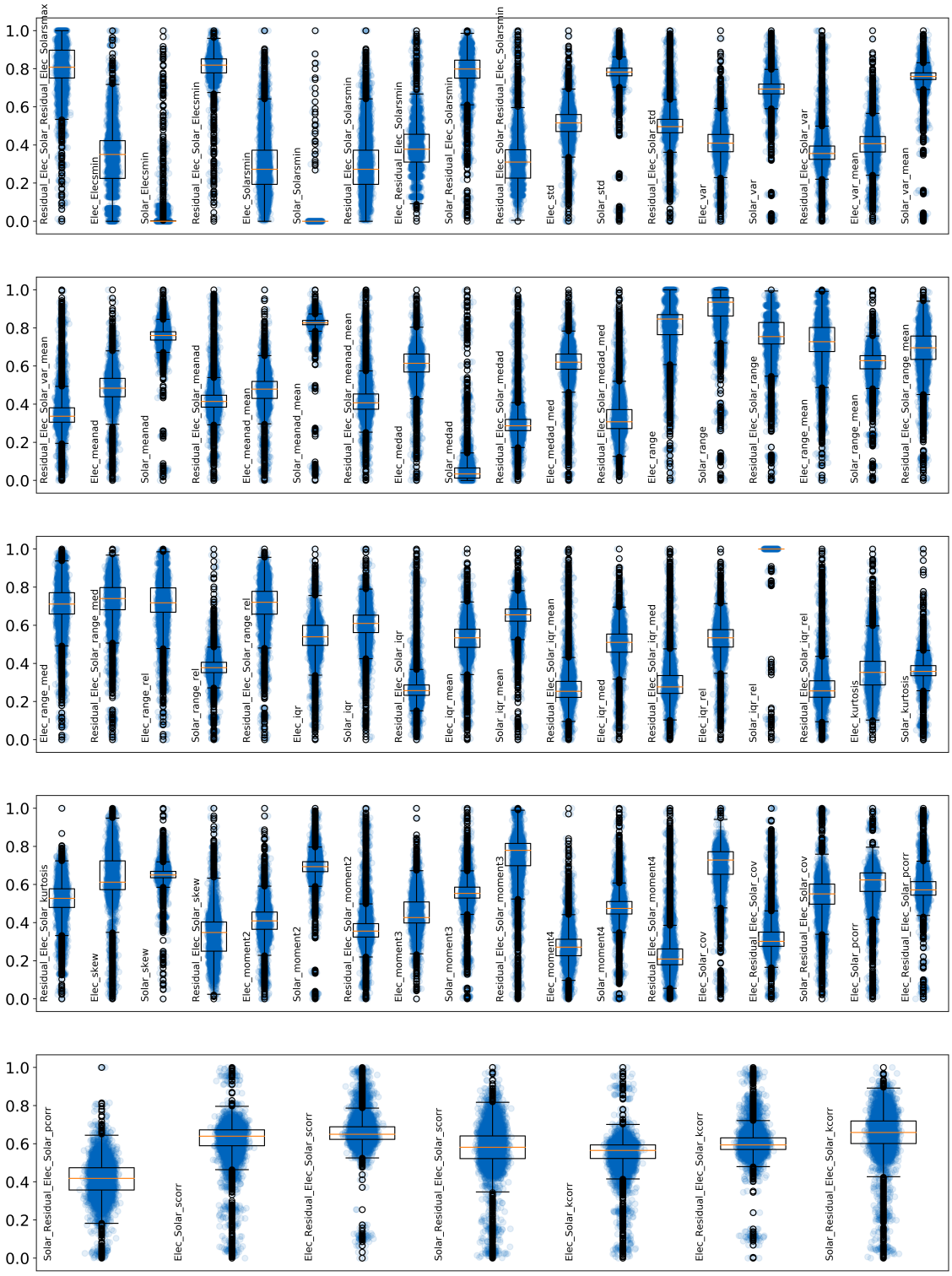

Figure B.2: Overview of normalized features of the PV scenario included in CNR - part II.

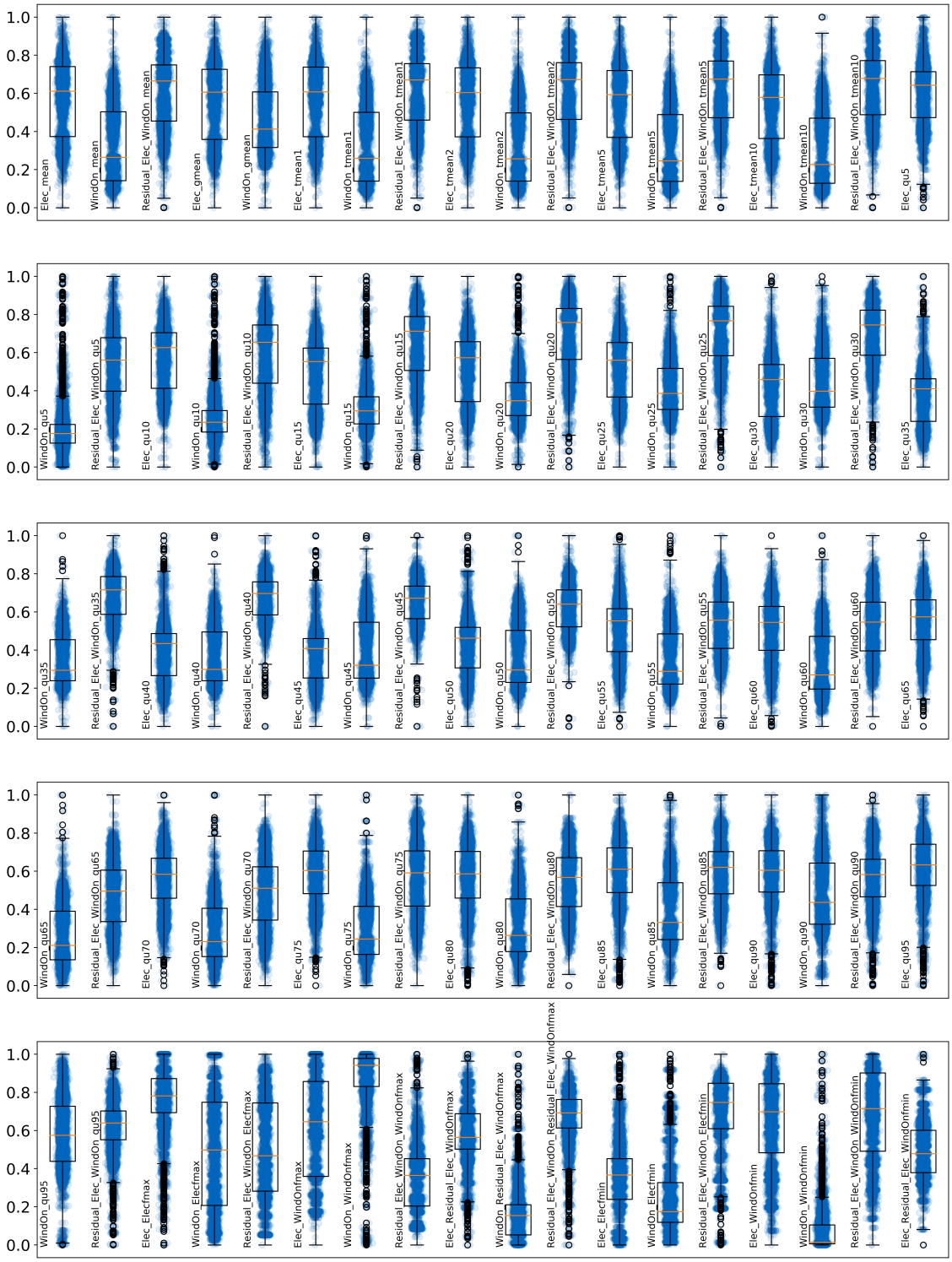

Figure B.3: Overview of normalized features of the WIND scenario included in CNR - part I.

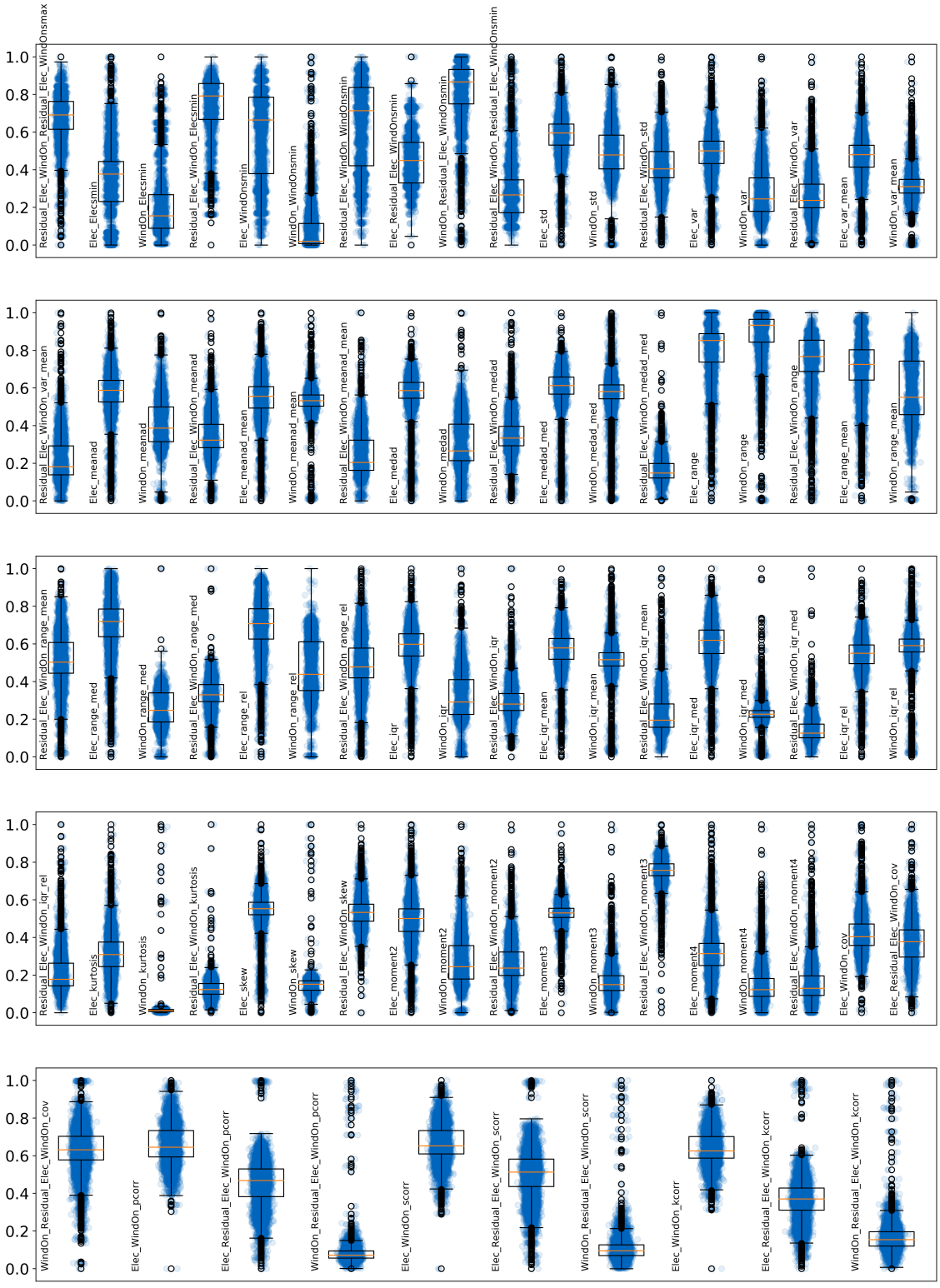

Figure B.4: Overview of normalized features of the WIND scenario included in CNR - part II.

Appendix C

Profiling derivation

Duration curve

For recalculating the duration curve, the Equation 3.25-3.29 can be simplified as follows: With $\tau = 1$ the quantiles q result in $q \in Q = \{\frac{1}{|\widehat{T}'|}, \frac{2}{|\widehat{T}'|}, ...\}$ and the step size $\delta = \frac{1}{|\widehat{T}'|}$ results. The set of considered time steps for each quantile is

$$\widehat{T}''_1 = \{1\}, \quad \widehat{T}''_2 = \{2\}, \quad ..., \widehat{T}''_q = \left\{|\widehat{T}'|\right\} \tag{C.1}$$

$$T''_1 = \left\{0 \leq t \leq \frac{|T'|}{|\widehat{T}'|}\right\}, T''_2 = \left\{\frac{|T'|}{|\widehat{T}'|} + 1 \leq t \leq 2\frac{|T'|}{|\widehat{T}'|}\right\}, ..., T''_q = \left\{|T'| - \frac{|T'|}{|\widehat{T}'|} + 1 \leq t \leq |T'|\right\} \tag{C.2}$$

For all $t \in \widehat{T}''_q \not\ni \widehat{B}'$, the weighting factor and the multiplier results in

$$\widehat{w}_q = \frac{|T''_q|}{|\widehat{T}''_q|} = \frac{|T'|}{|\widehat{T}'|} \tag{C.3}$$

$$\sigma_{i,q} = \frac{\sum_{t\in T''_q} \zeta_{i,t} - \widehat{w}_q \sum_{t\in \widehat{B}'\cap \widehat{T}''_q} \widehat{\zeta}_{i,t}}{\widehat{w}_q \sum_{t\in \widehat{T}''_q \setminus \widehat{B}'} \widehat{\zeta}_{i,t}} = \frac{\sum_{t\in T''_q} \zeta_{i,t}}{\widehat{w}_q \sum_{t\in \widehat{T}''_q} \widehat{\zeta}_{i,t}} = \frac{\widehat{T}'}{T'} \frac{\sum_{t\in T''_q} \zeta_{i,t}}{\sum_{t\in \widehat{T}''_q} \zeta_{i,t}} \tag{C.4}$$

Thus, the recalculation of the duration curve $\zeta_{i,t}$ can be expressed as

$$\widehat{\zeta}_{i,\widehat{t}'} = \frac{|\widehat{T}'|}{|T'|} \sum_{t\in T''_q} \zeta_{i,t} = \frac{|\widehat{T}'|}{|T'|} \sum_{t=1+\frac{|T'|}{|\widehat{T}'|}(t-1)}^{\frac{|T'|}{|\widehat{T}'|}t} \zeta_{i,t} \quad \forall \quad \widehat{t}' \in \widehat{T}''_q \setminus \widehat{B}' \tag{C.5}$$

Total average

For the alignment of the total average, Equation 3.25-3.29 can be simplified as follows: With the quantiles $q \in Q = 0.5$ and step size $delta = 0.5$, we get the time sets

$$\widehat{T}'' = \left\{1 \leq t \leq |\widehat{T}'|\right\} = \widehat{T}' \tag{C.6}$$

$$T'' = \{1 \leq t \leq |T'|\} = T' \tag{C.7}$$

The weighting factor and the multiplier results in

$$\widehat{w} = \frac{|T'|}{|\widehat{T}'|} \tag{C.8}$$

$$\sigma_{i,q} = \frac{\sum_{t \in T'} \zeta_{i,t} - \widehat{w} \sum_{t \in \widehat{B}'} \widehat{\zeta}_{i,t}}{\widehat{w} \sum_{t \in \widehat{T}''_q \setminus \widehat{B}'} \widehat{\zeta}_{i,t}} \tag{C.9}$$

Thus, the recalculation of the average can be expressed as

$$\widehat{\zeta}_{i,\widehat{t}'} = \frac{\sum_{t \in T'} \zeta_{i,t} - \widehat{w} \sum_{t \in \widehat{B}'} \widehat{\zeta}_{i,t}}{\widehat{w} \sum_{t \in \widehat{T}' \setminus \widehat{B}'} \widehat{\zeta}_{i,t}} \quad \widehat{\zeta}_{i,\widehat{t}'} \quad \forall \quad \widehat{t}' \in \widehat{T}' \setminus \widehat{B}' \tag{C.10}$$

Appendix D

Extended results

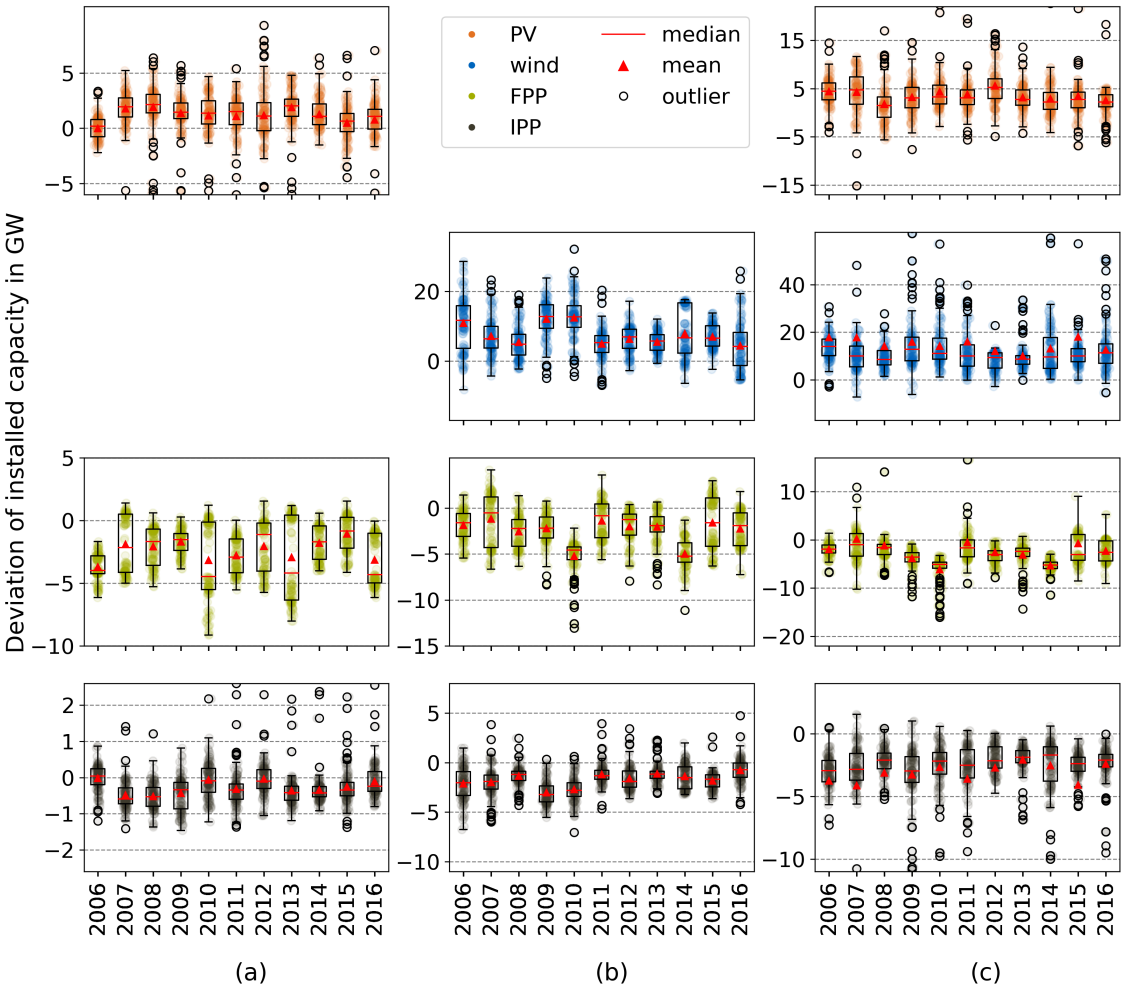

Figure D.1: Resulting installed capacities of aggregated time series as deviation from results of original time series shown for each year and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

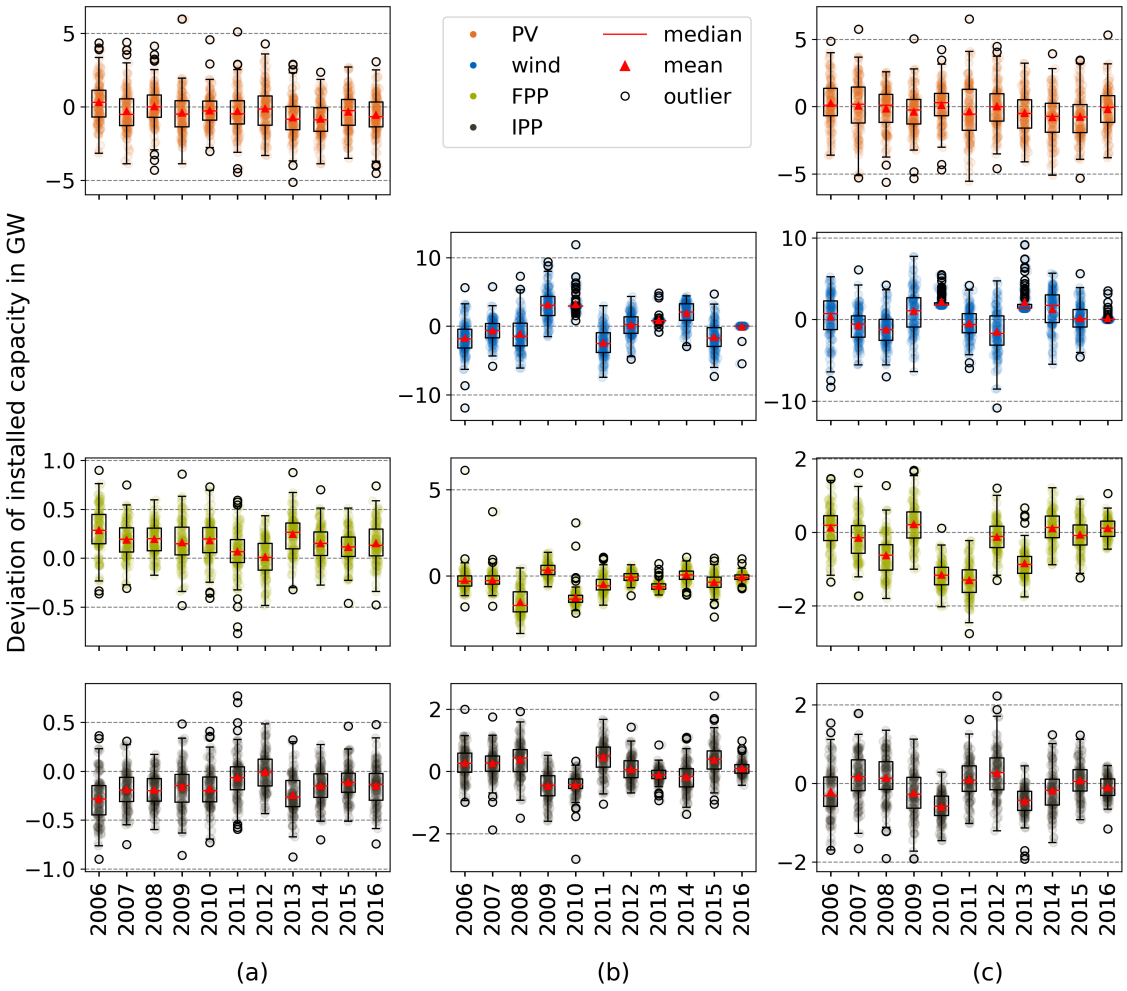

Figure D.2: Resulting installed capacities of profiled time series as deviation from results of original time series shown for each year and scenario: (a) PV, (b) WIND, and (c) PV+WIND.

Acronyms

cov Covariance
dem Electricity demand
gmean Geometric mean
hmean Harmonic mean
iRES Intermittend Renewable Energies Sources
kcorr Kendall's correlation
kurt Kurtosis
p p-value
pcorr Pearsons's correlation
res Residual load
scorr Spearman's correlation
skew Skewness
tmean Trimmed mean
var Varicance

AIC Akaike's Information Criterion

BIC Bayesian Information Criterion

CNR Clustering and Nested Based Regression

FPP Flexible Power Plant

GLS Generalized Least Squares

IPP Inert Power Plant
IQR Interquartile Distance

KDD Knowledge Discovery in Databases

LASSO Least Absolute Shrinkage and Selection Operator

MAD Mean Absolute Distance
MAE Mean Absolute Error
ME Mean Error

OLS Ordinary Least Squares

PI Performance Index

PV Photovoltaic

Q50 50 % quantile

R^2 Coefficient of determination

SI Siginificance Index
SSD Sum of Squared Distances
STD Standard Deviation

TSA Time Series Aggregation

Bibliography

[1] O. Abedinia, N. Amjady, and H. Zareipour. A new feature selection technique for load and price forecast of electrical power systems. *IEEE Transactions on Power Systems*, 32(1):62–74, 2017. doi:10.1109/TPWRS.2016.2556620. 34

[2] K. Abt. Descriptive data analysis: A concept between confirmatory and exploratory data analysis. *Methods of Information in Medicine*, 26(02):77–88, 1987. doi:10.1055/s-0038-1635488. 20

[3] W. Assenmacher. *Deskriptive Statistik*. Springer-Lehrbuch. Springer, Berlin, 3rd edition, 2003. doi:10.1007/978-3-662-06562-4. 19, 21, 22, 25

[4] A. Azzalini and B. Scarpa. *Data analysis and data mining: An introduction*. Oxford University Press USA, Oxford, 2012. 19, 20

[5] K. Bareiß, Schönleber Konrad, and T. Hamacher. The role of hydrogen, battery- electric vehicles and heat as flexibility option in future energy systems. *20th European Conference on Power Electronics and Applications, Riga, Latvia*, pages 1–10, 2018. 14

[6] N. Baumgärtner, B. Bahl, M. Hennen, and A. Bardow. Rises3: Rigorous synthesis of energy supply and storage systems via time-series relaxation and aggregation. *Computers & Chemical Engineering*, 127:127–139, 2019. doi:10.1016/j.compchemeng.2019.02.006. 14, 32

[7] K. A. Bollon and J. S. Long. *Testing structural equation models*, volume 154 of *Sage focus editions*. Sage, Newbury Park, CA, illustrated edition, 1993. 38

[8] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos. *Feature selection for high-dimensional data*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham, 2015. doi:10.1007/978-3-319-21858-8. 26

[9] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: Forecasting and control*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, 5th edition, 2016. 89

[10] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer, Cham, 3rd edition, 2016. doi:10.1007/978-3-319-29854-2. 89

[11] T. Burandt, K. Löffler, and K. Hainsch. GENeSYS-MOD v2.0 – Enhancing the global energy system model: Model improvements, framework changes, and European

data set. URL: https://www.diw.de/de/diw_01.c.594278.de/publikationen/data_documentation/2018_0094/genesys-mod_v2.0_____enhancing_the_global_energy_system_mode___model_improvements__framework_changes__and_european_data_set.html. 14

[12] C. Bussar, P. Stöcker, Z. Cai, L. Moraes Jr., D. Magnor, P. Wiernes, N. van Bracht, A. Moser, and D. U. Sauer. Large-scale integration of renewable energies and impact on storage demand in a European renewable power system of 2050 - sensitivity study. *Journal of Energy Storage*, 6:1–10, 2016. doi:10.1016/j.est.2016.02.004. 13

[13] S. Chen, Y. Ren, D. Friedrich, Z. Yu, and J. Yu. Sensitivity analysis to reduce duplicated features in ANN training for district heat demand prediction. *Energy and AI*, 2:100028, 2020. doi:10.1016/j.egyai.2020.100028. 34

[14] T. Cleff. *Deskriptive Statistik und moderne Datenanalyse: Eine computergestützte Einführung mit Excel, PASW (SPSS) und STATA*. Gabler, Wiesbaden, 2nd edition, 2012. doi:10.1007/978-3-8349-7071-8. 22, 25

[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi:10.1214/009053604000000067. 27

[16] ENTSO-E. Historical data. URL: https://www.entsoe.eu/data/dataportal/. 32

[17] A. T. Eseye, M. Lehtonen, T. Tukia, S. Uimonen, and R. John Millar. Machine learning based integrated feature selection approach for improved electricity demand forecasting in decentralized energy systems. *IEEE Access*, 7:91463–91475, 2019. doi:10.1109/ACCESS.2019.2924685. 34

[18] M. Ester and J. Sander. *Knowledge Discovery in Databases: Techniken und Anwendungen*. Springer, Berlin, 2000. doi:10.1007/978-3-642-58331-5. 19, 21, 28, 29

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 82—88. AAAI Press, 1996. 20

[20] C. Feng, M. Cui, B.-M. Hodge, and J. Zhang. A data-driven multi-model methodology with deep feature selection for short-term wind forecasting. *Applied Energy*, 190:1245–1257, 2017. doi:10.1016/j.apenergy.2017.01.043. 34

[21] B. A. Frew and M. Z. Jacobson. Temporal and spatial tradeoffs in power system modeling with assumptions about storage: An application of the power model. *Energy*, 117:198–213, 2016. doi:10.1016/j.energy.2016.10.074. 13, 14

[22] O. Grothe and J. Schnieders. Spatial dependence in wind and optimal wind power allocation: A copula based analysis, 2011. URL: https://www.ewi.uni-koeln.de/cms/wp-content/uploads/2015/12/EWI-WP-11-05-Spatial-Dependence-in-Wind-Power-Allocation.pdf. 13

[23] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. 26

[24] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002. doi:10.1023/A:1012487302797. 26

[25] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, nov 2009. doi:10.1145/1656274.1656278. 26

[26] J. Han, M. Kamber, and J. Pei. *Data mining: Concepts and techniques*. The Morgan Kaufmann series in data management systems. Elsevier/Morgan Kaufmann, Amsterdam, 3rd edition, 2012. 19, 20, 21, 22, 28

[27] T. Hastie, M. Wainwright, and R. Tibshirani. *Statistical learning with Sparsity: The lasso and generalizations*. Monographs on statistics and applied probability 143. CRC Press LLC, Boca Raton, 2015. 22, 27

[28] A. E. Hoerl and R. W. Kennard. Ridge regression. *In Encyclopedia of Statistical Sciences*, 8:129–136, 1988. 25

[29] M. Hoffmann, L. Kotzur, and D. Stolten. The pareto-optimal temporal aggregation of energy system models. *Applied Energy*, 315:119029, 2022. URL: https://www.sciencedirect.com/science/article/pii/S0306261922004342, doi:10.1016/j.apenergy.2022.119029. 82

[30] M. Hoffmann, J. Priesmann, L. Nolting, A. Praktiknjo, L. Kotzur, and D. Stolten. Typical periods or typical time steps? A multi-model analysis to determine the optimal temporal aggregation for energy system models. *Applied Energy*, 304:117825, 2021. doi:10.1016/j.apenergy.2021.117825. 14

[31] W. H. Jefferys and J. O. Berger. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72, 1992. 38

[32] S. Jerez, I. Tobin, R. Vautard, J. P. Montávez, J. M. López-Romero, F. Thais, B. Bartok, O. B. Christensen, A. Colette, M. Déqué, G. Nikulin, S. Kotlarski, E. van Meijgaard, C. Teichmann, and M. Wild. The impact of climate change on photovoltaic power generation in europe. *Nature communications*, 6:1–8, 2015. doi:10.1038/ncomms10014. 85

[33] H. Jiang and Y. Dong. A novel model based on square root elastic net and artificial neural network for forecasting global solar radiation. *Complexity*, 2018:1–19, 2018. doi:10.1155/2018/8135193. 26, 34

[34] Johannes Dorfner, Konrad Schönleber, Magdalena Dorfner, sonercandas, froehlie, smuellr, dogauzrek, WYAUDI, Leonhard-B, lodersky, yunusozsahin, adeeljsid, Thomas Zipperle, Simon Herzog, kais siala, and Okan Akca. tum-ens/urbs: urbs v1.0.1, 2019. doi:10.5281/zenodo.3265960. 14, 32

[35] C. Kath and F. Ziel. The value of forecasts: Quantifying the economic gains of accurate quarter-hourly electricity price forecasts. *Energy Economics*, 76:411–423, 2018. doi:10.1016/j.eneco.2018.10.005. 26, 27

[36] L. Kotzur, P. Markewitz, M. Robinius, and D. Stolten. Impact of different time series aggregation methods on optimal energy system design. *Renewable Energy*, 117:474–487, 2018. doi:10.1016/j.renene.2017.10.017. 14, 31

[37] N. Ludwig, S. Feuerriegel, and D. Neumann. Putting big data analytics to work: Feature selection for forecasting electricity prices using the lasso and random forests. *Journal of Decision Systems*, 24(1):19–36, 2015. doi:10.1080/12460125.2015.994290. 26, 34

[38] R. O. Mueller. *Basic principles of structural equation modeling*. Springer, New York, 1996. doi:10.1007/978-1-4612-3974-1. 38

[39] I. M. Müller. Feature selection for energy system modeling: Identification of relevant time series information. *Energy and AI*, 4:100057, 2021. doi:10.1016/j.egyai.2021.100057. 14, 17, 24, 26, 34, 41, 57, 79, 80

[40] I. M. Müller. Energy system modeling with aggregated time series: A profiling approach. *Applied Energy*, 322:119426, 2022. doi:10.1016/j.apenergy.2022.119426. 13, 14, 17, 28, 31, 42, 57, 62, 74, 79, 80, 87

[41] I. M. Müller, M. Reich, F. Warmer, H. Zohm, T. Hamacher, and S. Günter. Analysis of technical and economic parameters of fusion power plants in future power systems. *Fusion Engineering and Design*, 146(B):1820–1823, 2019. doi:10.1016/j.fusengdes.2019.03.043. 14

[42] P. Nahmmacher, E. Schmid, L. Hirth, and B. Knopf. Carpe diem: A novel approach to select representative days for long-term power system modeling. *Energy*, 112:430–442, 2016. doi:10.1016/j.energy.2016.06.081. 13, 14, 30, 31

[43] R. Nisbet, J. F. Elder, and G. Miner. *Handbook of statistical analysis and data mining applications*. Academic Press/Elsevier, Amsterdam and Boston, 2009. 19, 20, 26

[44] Observ'ER, TNO Energy Transition, RENAC, Frankfurt School of Finance and Management, Fraunhofer ISI and Statistics Netherlands. The state of renewable energies in europe: Edition 2019 - 19th EurObserv'ER Report. URL: https://www.isi.fraunhofer.de/content/dam/isi/dokumente/ccx/2020/The-state-of-renewable-energies-in-Europe-2019.pdf. 31

[45] S. K. Paramasivan and D. Lopez. Forecasting of wind speed using feature selection and neural networks. *International Journal of Renewable Energy Research (IJRER)*, 6(3):833–837, 2016. doi:10.20508/ijrer.v6i3.3855.g6866. 34

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 27, 28, 43

[47] S. Pfenninger. Dealing with multiple decades of hourly wind and pv time series in energy models: A comparison of methods to reduce time resolution and the planning implications of inter-annual variability. *Applied Energy*, 197:1–13, 2017. doi:10.1016/j.apenergy.2017.03.051. 13, 14, 31, 32, 42

[48] S. Pfenninger and B. Pickering. Calliope: A multi-scale energy systems modelling framework. *Journal of Open Source Software*, 3(29):825–826, 2018. doi:10.21105/joss.00825. 14

[49] S. Pfenninger and I. Staffell. Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data. *Energy*, 114:1251–1265, 2016. doi:10.1016/j.energy.2016.08.060. 32

[50] M. Pirhooshyaran, K. Scheinberg, and L. V. Snyder. Feature engineering and forecasting via derivative-free optimization and ensemble of sequence-to-sequence networks with applications in renewable energy. *Energy*, 196:117136, 2020. doi:10.1016/j.energy.2020.117136. 34

[51] K. Poncelet, H. Höschle, E. Delarue, A. Virag, and W. D'haeseleer. Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems. *IEEE TRANSACTIONS ON POWER SYSTEMS,*, 32(3):1936–1948, 2017. doi:10.1109/TPWRS.2016.2596803. 14, 31, 32, 44

[52] L. A. Roberto, J.-W. Arnulf, V. Marika, S. Bergur, M. Davide, J. Mindaugas, P. F. M. Del Mar, L. Stavros, G. Jacopo, W. R. Eveline, et al. ETRI 2014-energy technology reference indicator projections for 2010-2050, 2014. URL: https://op.europa.eu/en/publication-detail/-/publication/79a2ddbd-5ba1-4380-93af-2ce274a840f0/language-en. 11, 32, 33

[53] T. A. Runkler. *Data analytics: Models and algorithms for intelligent data analysis*. Springer Vieweg, Wiesbaden, 2nd edition, 2016. doi:10.1007/978-3-658-14075-5. 19, 20, 22, 28, 30

[54] G. Savvidis, K. Siala, C. Weissbart, L. Schmidt, F. Borggrefe, S. Kumar, K. Pittel, R. Madlener, and K. Hufendiek. The gap between energy policy challenges and model capabilities. *Energy Policy*, 125:503–520, 2019. doi:10.1016/j.enpol.2018.10.033. 14

[55] N. Scarlat, J.-F. Dallemand, F. Monforti-Ferrario, M. Banja, and V. Motola. Renewable energy policy framework and bioenergy contribution in the European Union – an overview from national renewable energy action plans and progress reports. *Renewable and Sustainable Energy Reviews*, 51:969–985, 2015. doi:10.1016/j.rser.2015.06.062. 13

[56] K. Schermelleh-Engel, H. Moosburger, and H. Müller. Evaluating the fit of structural equation models:tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online 2003*, 8(2):23–74, 2003. 38, 40

[57] S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pages 92–96, 2010. doi:10.25080/Majora-92bf1922-011. 25, 38

[58] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x. 34

[59] R. H. Shumway and D. S. Stoffer. *Time series analysis and its applications: With R examples*. Springer Texts in Statistics. Springer, Cham, 4th edition, 2017. doi:10.1007/978-3-319-52452-8. 25, 89

[60] K. Siala and M. Y. Mahfouz. Impact of the choice of regions on energy system models. *Energy Strategy Reviews*, 25:75–85, 2019. doi:10.1016/j.esr.2019.100362. 13

[61] H. Son and C. Kim. Forecasting short-term electricity demand in residential sector based on support vector regression and fuzzy-rough feature selection with particle swarm optimization. *Procedia Engineering*, 118:1162–1168, 2015. doi:10.1016/j.proeng.2015.08.459. 34

[62] A. Steland. *Basiswissen Statistik: Kompaktkurs für Anwender aus Wirtschaft, Informatik und Technik*. Springer-Lehrbuch. Springer Spektrum, Berlin, 4th edition, 2016. doi:10.1007/978-3-662-49948-1. 22

[63] P. Stenzel, J. Linssen, J. Fleer, and F. Busch. Impact of temporal resolution of supply and demand profiles on the design of photovoltaic battery systems for increased self-consumption. *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6, 2016. doi:10.1109/ENERGYCON.2016.7514010. 32

[64] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL: http://jmlr.org/papers/v21/20-091.html. 36

[65] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x. 26, 27

[66] I. Tobin, S. Jerez, R. Vautard, F. Thais, E. van Meijgaard, A. Prein, M. Déqué, S. Kotlarski, C. F. Maule, G. Nikulin, T. Noël, and C. Teichmann. Climate change impacts on the power generation potential of a European mid-century wind farms scenario. *Environmental Research Letters*, 11(3):034013, 2016. doi:10.1088/1748-9326/11/3/034013. 85

[67] J. W. Tukey. We need both exploratory and confirmatory. *The American Statistican*, 34(1):23–25, 1980. 20

[68] B. Uniejewski, G. Marcjasz, and R. Weron. Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. *International Journal of Forecasting*, 35(4):1533–1547, 2019. doi:10.1016/j.ijforecast.2019.02.001. 27, 34

[69] B. Uniejewski, J. Nowotarski, and R. Weron. Automated variable selection and shrinkage for day-ahead electricity price forecasting. *Energies*, 9(8):621, 2016. doi:10.3390/en9080621. 28

[70] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi:10.1038/s41592-019-0686-2. 43

[71] F. vom Scheidt, H. Medinová, N. Ludwig, B. Richter, P. Staudt, and C. Weinhardt. Data analytics in the electricity sector – a quantitative and qualitative literature review. *Energy and AI*, 1:100009, 2020. doi:10.1016/j.egyai.2020.100009. 34

[72] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845, doi:10.1080/01621459.1963.10500845. 30

[73] I. H. Witten, E. Frank, and M. A. Hall. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann series in data management systems. Elsevier/Morgan Kaufmann, Amsterdam, 3rd edition, 2011. 19, 20, 21, 22

[74] W. Zappa, M. Junginger, and M. van den Broek. Is a 100% renewable European power system feasible by 2050? *Applied Energy*, 233-234:1027–1050, 2019. doi:10.1016/j.apenergy.2018.08.109. 13

[75] C. Zhang, H. Wei, J. Zhao, T. Liu, T. Zhu, and K. Zhang. Short-term wind speed forecasting using empirical mode decomposition and feature selection. *Renewable Energy*, 96:727–737, 2016. doi:10.1016/j.renene.2016.05.023. 34

[76] L. Zhang and J. Wen. A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings*, 183:428–442, 2019. doi:10.1016/j.enbuild.2018.11.010. 34

[77] Z. A. Zhao and H. Liu. *Spectral feature selection for data mining*. Chapman & Hall/CRC, Boca Raton, 2012. 26

[78] F. Ziel. Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. *IEEE Transactions on Power Systems*, 31(6):4977–4987, 2016. doi:10.1109/TPWRS.2016.2521545. 26, 34

[79] F. Ziel and R. Weron. Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks. *Energy Economics*, 70:396–420, 2018. doi:10.1016/j.eneco.2017.12.016. 26

[80] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 03 2005. URL: https://doi.org/10.1111/j.1467-9868.2005.00503.x, arXiv:https://academic.oup.com/jrsssb/article-pdf/67/2/301/49795094/jrsssb_67_2_301.pdf, doi:10.1111/j.1467-9868.2005.00503.x. 26, 27