Estimating costs of extending electricity distribution networks in Germany

Master's thesis

for obtaining the degree of
Master of Science (M.Sc.)

in Economics

at the School of Business and Economics
of Humboldt-Universität zu Berlin

submitted by
Manuel Linsenmeier
Student No. 574577

Examiner: Prof. Dr. Franz Hubert

Berlin, 27.11.2017

Abstract

The German power system is undergoing a fundamental transformation substituting electricity generated from fossil fuels with electricity generated from renewable resources. This transformation - the Energiewende - incurs substantial costs for extending and reinforcing electricity distribution networks in Germany. These costs have been estimated to be up to 40 billion EUR by the year 2030 (dena, 2012) but are relatively uncertain because no complete dataset of distribution networks exists. Previous studies thus relied on small samples of real networks and employed cluster methods to estimate the total costs of network expansion for Germany. However, this methodology has so far not been examined. In this thesis, a dataset of synthetic networks is used to examine different cluster models and cluster estimation methods. These models and methods include the methodology used by previous studies. In addition, alternative models and methods are proposed and examined. To this aim, first a theoretical framework is developed to identify and categorise 14 network attributes that are expected to determine the costs of network expansion. Each of these attributes is then examined regarding its effect on the performance of a cluster model if it is included in that model. Furthermore, the cluster models that result in the lowest within-cluster dispersion of costs are analysed. Based on these results, 57 cluster models are selected and assessed both in terms of the within-cluster dispersion of costs and the relative deviation of estimated total costs from calculated total costs. Finally, two cluster models are analysed in more detail including the geographic occurrence of clusters. Throughout the analysis, K-Mean and regression trees are used as two alternative cluster estimation methods. Furthermore, the number of clusters K is varied from 5 to 300. For the costs of network expansion, a worst-case scenario and a scenario with curtailment are constructed. Overall, the cluster model used in previous studies performs better than most of the proposed alternative models. The results show, however, that for all values of K at least one of the alternative cluster models performs better. Furthermore, regression trees as cluster estimation method generally result in clusters with lower within-cluster dispersion of costs and lower relative deviation of total costs than K-Mean estimation, which was used in previous studies.

Zusammenfassung

Im Rahmen der Energiewende in Deutschland wird die Elektrizitätsversorgung auf Erneuerbare Energien umgestellt. Damit sind substanzielle Kosten für den Netzausbau auf Verteilnetzebene verbunden. Diese Kosten wurden auf bis zu 40 Milliarden Euro bis zum Jahr 2030 geschätzt (dena, 2012), sind jedoch mit größeren Unsicherheiten verbunden, da keine vollständigen und öffentlich zugänglichen Daten über die Verteilnetze bestehen. Bisherige Studien waren daher auf kleine Stichproben realer Netze angewiesen und verwendeten Clustermethoden, um die Gesamtkosten für Deutschland zu schätzen. Diese Methodologie wurde bisher jedoch nicht untersucht. In dieser Arbeit wird ein Datensatz synthetischer Verteilnetze verwendet um alternative Clustermodelle und alternative statistische Methoden zur Schätzung von Clustern zu untersuchen. Diese Modelle und Methoden beinhalten die zuvor angewendete Methodologie. Darüber hinaus werden weitere Modelle und Methoden vorgeschlagen und analysiert. Um dies zu erreichen, wird zuerst ein theoretischer Rahmen entwickelt, in dem 14 Netzattribute identifiziert und kategorisiert werden, von denen vermutet wird, dass sie die Netzausbaukosten bestimmen. Jedes dieser Attribute wird dann daraufhin untersucht, wie sehr seine Berücksichtigung in einem Clustermodell zur Güte dieses Clustermodells beiträgt. Außerdem werden die Clustermodelle identifziert, die in der geringsten Streuung von Kosten innerhalb der Cluster resultieren. Basierend auf diesen Ergebnissen werden 57 Clustermodelle ausgesucht und hinsichtlich der Streuung der Kosten und hinsichtlich der relativen Abweichung der geschätzten Gesamtkosten von den tatsächlichen Gesamtkosten untersucht. Zudem werden zwei Clustermodelle detailliert ausgewertet und die geographische Verteilung der Cluster analysiert. In der Analyse werden durchgehend zwei Schätzmethoden, K-Mean und Regressionsbäume, angewendet und die Ergebnisse miteinander verglichen. Außerdem wird die Zahl der Cluster K von 5 bis 300 variiert. Für die Ausbaukosten werden zwei Szenarien, ein worst-case Szenario und ein Szenario mit Abregelung, entwickelt. Insgesamt zeigen die Ergebnisse, dass das in vorherigen Studien eingesetzte Clustermodell besser als die meisten vorgeschlagenen alternativen Clustermodelle abschneidet. Die Ergebnisse zeigen aber auch, dass für jede Zahl an Clustern K mindestens ein Clustermodell besser abschneidet. Außerdem weisen die Cluster, die mit Regressionsbäumen geschätzt werden, im Allgemeinen eine geringere Streuung der Kosten innerhalb der Cluster und eine geringere relative Abweichung der geschätzten Gesamtkosten von den berechneten Gesamtkosten auf als die Cluster, die mit K-Mean geschätzt werden. Vorherige Studien hatten ausschließlich K-Mean als Schätzmethode verwendet.

Table of contents

List of Figures

List of Tables

List of acronyms, abbreviations, and selected symbols

BMWi	Bundesministerium für Wirtschaft und Energie
DENA	... and other labels of cluster models: Figure 35 (Appendix)
DNO	Distribution network operator
EUR	Euro
GW	Gigawatt
HV	High-voltage
kW	Kilowatt
K	Number of clusters
kEUR	1,000 Euro
KM	K-Mean cluster estimation
LV	Low-voltage
MV	Medium-voltage
MVGD	Medium-voltage grid district
MW	Megawatt
OSM	OpenStreetMap
R	Relative deviation of estimated total costs from calculated total costs (Equation 5.12)
RT	Regression tree cluster estimation
W	Within-cluster dispersion of costs (Equation 5.11)
WIND-2015	... and other labels referring to network attributes: Table 2
WWT	... and other labels of cluster models: Figure 35 (Appendix)

1. Introduction

By 2050, the greenhouse gas emissions of Germany shall be reduced by at least 80 % compared to 1990 (Bundesregierung der Bundesrepublik Deutschland, 2010). In order to achieve this reduction, the German government has set several political targets: the use of electricity for heating and transportation shall be increased and 80 % of electricity shall be generated from renewable resources (Bundesregierung der Bundesrepublik Deutschland, 2010). If these targets shall be met, additional renewable energy power plants need to be integrated into the German power system.

Between 2000 and 2015, the installed generation capacity of renewable energy power plants in Germany has grown from 11.7 GW to 96.9 GW (BMWi, 2017). Most of this increase can be attributed to onshore wind (+ 35.2 GW) and solar photovoltaic power plants (+ 39.2 GW). Projections of Germany's transmission network operators predict that between 2015 and 2035 about + 20.3 GW of additional onshore wind and + 35.9 GW of additional solar photovoltaic generation capacity is going to be installed (NEP, 2017). This represents 80 % of the total additional generation capacity by 2035 (NEP, 2017). The two technologies are hence expected to dominate the expansion of renewable energy power plants in Germany also in the next two decades.

When the electrical grid of Germany was designed and constructed, most electricity was generated by conventional power plants, transmitted by the transmission system, transformed to lower voltage, and then distributed by distribution networks. In consequence, the transmission capacity of the grid was relatively large on the transmission level and relatively small on the distribution level. Because onshore wind and solar photovoltaic power plants are often located in relatively remote locations and tend to have a lower generation capacity than conventional power plants, they are typically connected to the grid on the distribution level. If the expansion of renewable energy power plants results in demand for additional transmission capacity, a substantial share of this demand will therefore concern distribution networks (dena, 2012).

The total costs of expanding the German electrical grid on the distribution level until

2030 are estimated at 23 - 40 billion EUR (dena, 2012; BMWi, 2014). This relatively large range of projected costs signals that these costs are relatively uncertain. The uncertainty concerns, for example, the extent to which energy storage facilities and smart grid technologies can reduce the demand for expanding the grid (dena, 2012). The uncertainty can however also be attributed to the lack of a dataset of distribution networks in Germany. All estimates of costs are thus based on relatively small samples of real networks. For example, the costs estimated by dena (2012) are based on networks that together contain only $0.5\,\%$ of the estimated total length of lines and cables of the distribution grid in Germany (Bundesnetzagentur, 2017; dena, 2012).

The lack of a complete dataset of distribution networks in Germany can partly be explained with the relatively large number of more than 800 distribution network operators. Each of these private companies operates one or more distribution networks. The number of networks is therefore even larger. On the low-voltage level, for example, there are more than 500,000 distribution networks in Germany (Amme et al., 2017). In order to estimate the total costs of expanding distribution networks, previous studies used statistical methods developed for cluster analysis (dena, 2012; Rehtanz et al., 2017; Ackermann et al., 2014). The authors then focused on a relatively small number of networks which were considered as representative for all networks. The costs of network expansion were then computed only for these networks. Although this methodology has already been applied in several studies, its accuracy and robustness has so far not been examined.

In this thesis, a dataset of synthetic distribution networks is used representing about $84\,\%$ of the total length of lines and cables of the national electrical grid on the corresponding voltage level. Based on this dataset, the performance of previously used and the performance of alternative cluster models is examined. Furthermore, two methods for estimating the cluster models are implemented and their performance is analysed and compared.

The thesis is structured as follows. First, the hierarchical structure of the German power system is described and the typical topology of distribution networks are explained. Furthermore, some fundamentals of electrical power transmission are introduced (Chapter 2). Then, the synthetic dataset is described and the methods and two alternative scenarios of network expansion are introduced (Chapter 3). Next, a theoretical framework is developed to identify and categorise attributes of distribution networks that are expected to determine the costs of network expansion. Moreover, descriptive statistics of the dataset of synthetic networks are given (Chapter 4). Then, two statistical methods to estimate cluster models, K-Mean and regression trees, are explained. Moreover, two metrics for the evaluation of cluster models are developed. The two methods and the two metrics are illustrated with an example (Chapter 5).

In Chapter 6, each of the 14 network attributes is examined regarding its effect on the within-cluster dispersion of expansion costs if the attribute is included or excluded in a cluster model. Then, the ten best cluster models are examined and based on these results, 57 cluster models are selected for further analysis. These cluster models are analysed with respect to the two metrics. The performance of some of the cluster models is then directly compared. Finally, two cluster models are analysed in more detail and the representative networks and the geographic distribution of clusters are examined.

The results are then discussed in Chapter 7. First, the total costs of network expansion of the synthetic networks is compared with the total costs that were estimated in previous studies. Furthermore, the selected 14 network attributes are discussed in light of previous studies. Then, the relative performance of the K-Mean and the regression tree method is discussed. Finally, the performance of the cluster model used by previous studies relative to the performance of alternative cluster models is discussed. In Chapter 8 conclusions are drawn and opportunities for future research are pointed out.

2. Fundamentals of distribution networks

In this Chapter, some fundamentals on distribution networks are presented. First, the hierarchy of the German power system is described and the functions and typical topologies of networks on the medium-voltage and low-voltage level are presented (Section 2.1). Then, the physical background of thermal limits and voltage limits of networks is explained and the conditions that determine whether these limits are violated are described (Section 2.2). Lastly, the context in which distribution network operators decide on expanding distribution networks is briefly addressed (Section 2.3).

2.1 Functions and typical topologies

Electric power systems can generally be considered to consist of three subsystems: the generation system, the transmission system, and the distribution system (Gönen, 1986, p. 1). In the generation system, electric power is generated. The generation system therefore includes all conventional power plants and renewable energy power plants. In the transmission and in the distribution system, power is transported from generation units to customers. The transmission and the distribution system can be distinguished based on different system properties (Biggar and Hesamzadeh, 2014, p. 53). In general, the transmission system carries a larger amount of power and carries power over larger distances than the distribution system. In order to reduce transmission losses, the voltage level is therefore higher in the transmission system than in the distribution system (Kirtley, 2010, p. 2).

The transmission and the distribution system in Germany can therefore be distinguished based on their voltage level. The transmission system features extra-high voltage ($\leq$ 220 kV). The distribution system features high voltage (110 kV), medium voltage (1-35 kV) and low voltage (< 1 kV).

These three voltage levels of the distribution system serve different purposes (Figure 1). On the extra-high and high-voltage level, power is transported over relatively long distances and generation units with a relatively large generation capacity such as nuclear power plants or large wind farms are connected to the grid. Furthermore, very large industrial consumers are connected. On the medium-voltage level, power is transported over smaller distances and smaller generation units such as wind power plants as well as smaller industrial consumers are connected. On the low-voltage level, power is transported to small consumers. Furthermore, small generation units such as roof-mounted photovoltaic installations are connected.

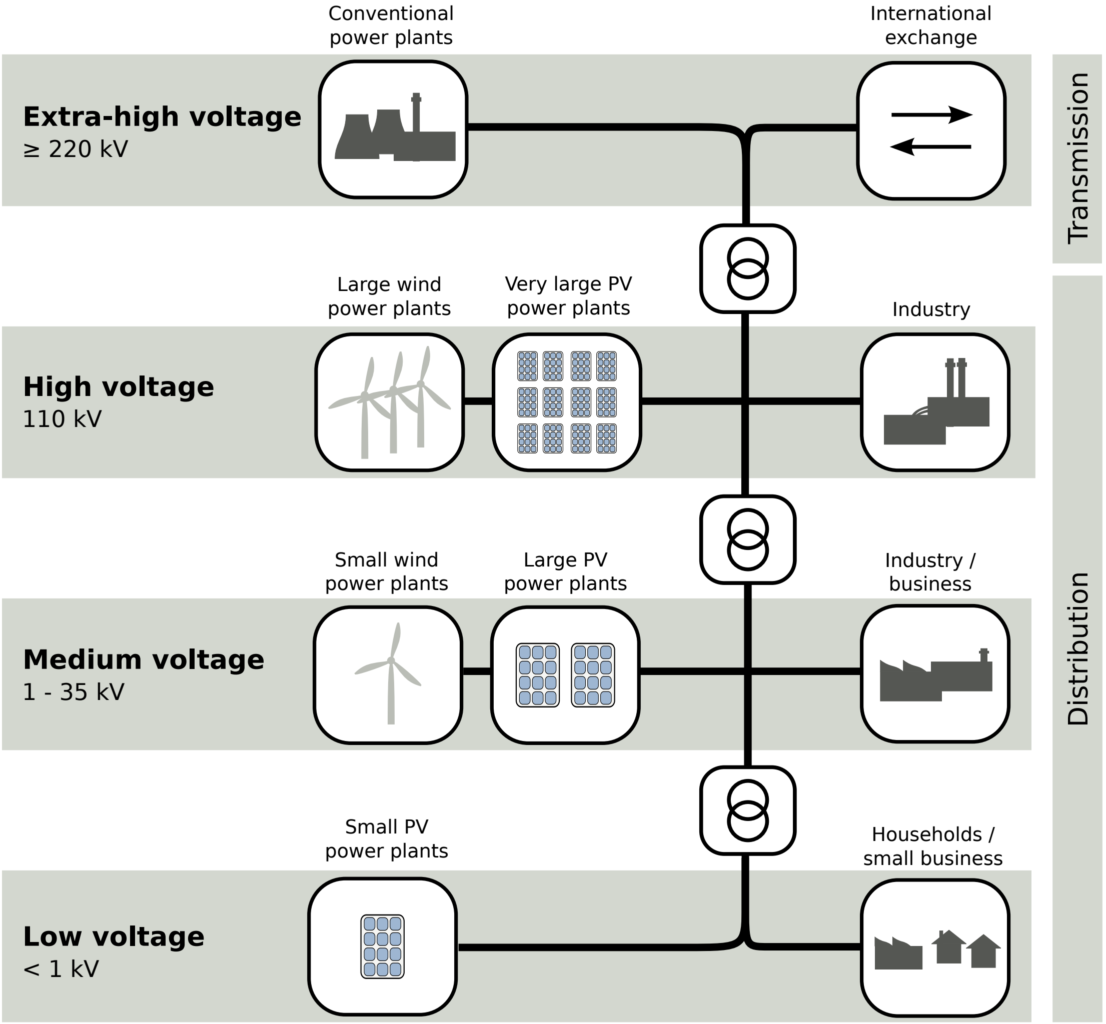


Figure 1: The German power system can be divided into the transmission and the distribution system, each with different voltage levels. For simplification, only two technologies for generating electricity from renewable resources are shown. Figure adapted from Gust (2014).

The three voltage levels feature different network topologies in the German electrical grid (Figure 2). On the high-voltage level, networks typically feature a meshed structure. On the medium-voltage level, most networks feature a ring structure. These networks are usually equipped with a switch disconnector and therefore operated as radial networks. In case of one line or one equipment failure, the open segment can then be closed in order to maintain power supply. Meshed networks and ring networks therefore comply with the n-1 criterion. On the low-voltage level, most networks are radial networks (Amme et al., 2017).

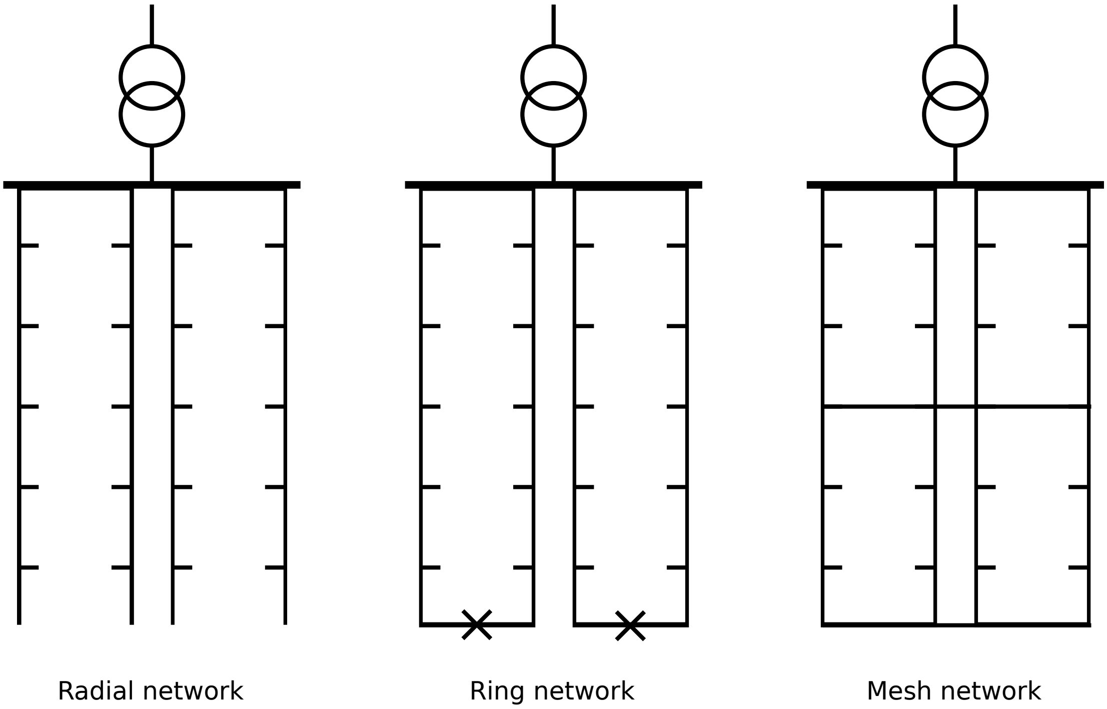


Figure 2: Typical simplified topologies of electricity networks in Germany. Figure adapted from dena (2012).

In this thesis, the medium-voltage and the low-voltage levels of the electrical grid are analysed. The main reason is that in Germany the availability of data is much better for networks on the high-voltage level than for networks on the medium- and on the low-voltage level. Much less is therefore known about the costs of network expansion on the medium- and low-voltage level. Furthermore, the focus on the medium-voltage and low-voltage level is supported by the fact that in 2012 about 95 % of renewable energy power plants were connected to these two voltage levels (dena, 2012) and that previous research has attributed about 50 % of total expansion costs to the expansion of medium-voltage and low-voltage networks (dena, 2012; BMWi, 2014; NEP, 2017).

2.2 Thermal limits and voltage limits

The amount of power that can be transported in a network while maintaining safety of operation is determined by certain properties of the network. This amount of power can be referred to as the transmission capacity of the network. If this amount is exceeded, technical equipment of the network becomes too warm. For this reason, the transmission capacity of a network can also be referred to as the thermal limit of the network. Furthermore, in order to ensure stability of the network and functioning of all connected devices, the voltage of a network needs to be kept within certain limits. The thermal limit and the voltage limits of a network are determined by certain properties of its technical equipment, such as the type and length of a cable or line.

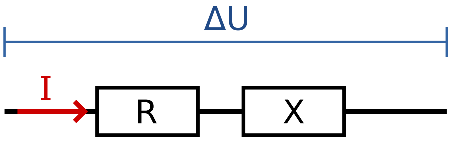


Figure 3: Simple model of an electric line with resistance R, reactance X, current I and drop of voltage ΔU.

Figure 3 shows a model of an electric line. The line features a resistance R and a reactance X, which are the real and the imaginary part of the impedance Z, respectively:

$$Z = R + j \cdot X \tag{2.1}$$

If the current I is transported along the line, the voltage U between the two ends of the line differs. This difference ΔU can be related to the current and the reactance of the line as follows:

$$\Delta U = I \cdot Z \tag{2.2}$$

In the model in Figure 3 the line consists of only one line segment. Equations 2.1 and 2.2 can also be formulated for the more general case of a line consisting of n segments. In this case the total voltage drop ΔU_{total} can be calculated from the sum of the impedance of each segment multiplied by the corresponding current:

$$\Delta U_{total} = \sum_{j=1}^{n} I_j \cdot Z_j \tag{2.3}$$

Whether a network is capable of hosting an additional renewable energy power plant with a certain generation capacity at a certain node within the network depends also on the thermal limit. This thermal limit is determined by the maximal allowed current I_{max} of the technical equipment. If the expected current exceeds I_{max}, the transmission capacity needs to be enhanced. If several additional power plants are connected to a network, it is most likely that thermal limits are exceeded at the line segment closest to the transformer station. This is because the current fed into the network by each of the additional power plants sums up to a total additional current there (Figure 4).

An additional renewable energy power plant can also result in a total voltage drop along a line that violates a voltage limit. In Germany, the maximum allowed deviation of actual voltage from the nominal voltage of a line ΔU_{max} on the low-voltage level is 10 % (DIN, 2011). Voltage limits are typically exceeded at the end of a line because the drop of voltage increases along the line from the HV/MV transformer station to the terminal node (Equation 2.3 and Figure 4).

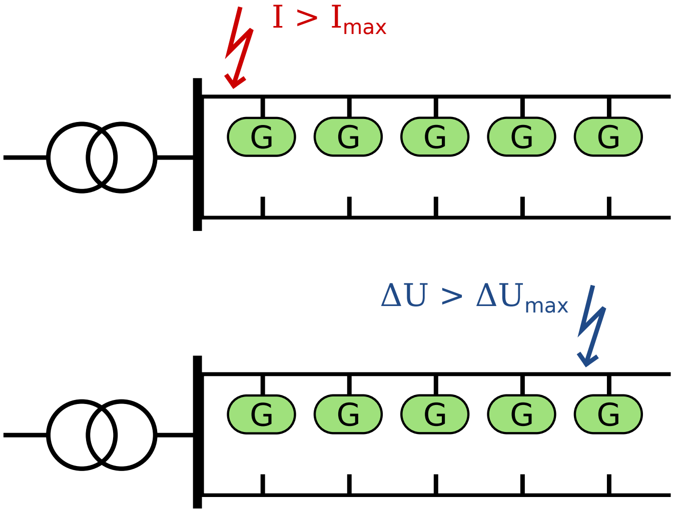


Figure 4: Locations of electricity networks at which thermal limits (top) and and voltage limits (bottom) are typically violated. The letter G denotes generation units such as renewable energy power plants. Figure adapted from dena (2012).

2.3 Network expansion

Each distribution network operator (DNO) is responsible for the safe operation of its network in Germany. This means that each DNO is also responsible for avoiding that thermal limits and voltage limits are exceeded. If the expected or the actual network status indicates that this is going to happen, the DNO can resolve the situation using one of several measures. The feasibility of these measures depends on the remaining time until the exceedance occurs. Furthermore, the DNO can base his decision on cost-benefit considerations.

If the DNO aims to prevent future exceedances of thermal or voltage limits by an expansion of the transmission capacity of the network, it can either install additional equipment such as overhead lines, underground cables or transformers, exchange existing equipment for equipment with larger transmission capacity, or change the topology of the network. These measures and the criteria for choosing among them have been described, for example, by Ackermann et al. (2014), dena (2012) and Rehtanz et al. (2017). The expansion of networks of the dataset used in this thesis is described in Section 3.4.

3. Dataset of synthetic networks

The dataset of synthetic electricity networks at medium- and low-voltage level in Germany that is analysed in this thesis was produced as part of the project open_ego. For this purpose, the software ding0 and the software eDisGo were written and used.[1] The networks are constructed based on the spatial distribution of load areas and generation units in Germany. In this Chapter, first the nature of load areas and their spatial distribution are briefly described (Section 3.1). Then, the spatial distribution of generation units is introduced (Section 3.2). Next, the construction of the synthetic networks is briefly explained (Section 3.3). Finally, the computation of the costs of network expansion for two scenarios is described (Section 3.4). More details on the dataset and the construction of the networks can be found in Hülk et al. (2017) and Amme et al. (2017).

For the work of this thesis, the author was provided with the final dataset. This dataset includes information on the medium-voltage network of each district, the associated low-voltage networks, the installed capacity of current and future generation units in that district, and the costs of network expansion for the two scenarios of electricity generation. The construction of the dataset was thus not part of the thesis but is explained here to anticipate the interpretation of the results in Chapter 6. The definition and implementation of the network attributes described in Chapter 4 was done by the author as part of this thesis.

3.1 Load areas and grid districts

Load areas are geographical polygons that aggregate individual points at which electricity is consumed. Load areas are identified based on data of land use and data of industrial infrastructure from the OpenStreetMap database (OSM) and data of population from the census 2011 for Germany (Hülk et al., 2017). Each load area

[1] The software of the project open_ego can be accessed online and downloaded from GitHub: https://github.com/openego.

is assigned four sector-specific electricity demand curves. The four sectors are the residential, retail, industrial and agricultural sector. The demand cures are derived from aggregated sectoral demand curves for Germany and allocated to each load area according to the approximate share of the sectoral gross value added (retail, industry, agriculture) and according to the approximate share of the total population of Germany (households).

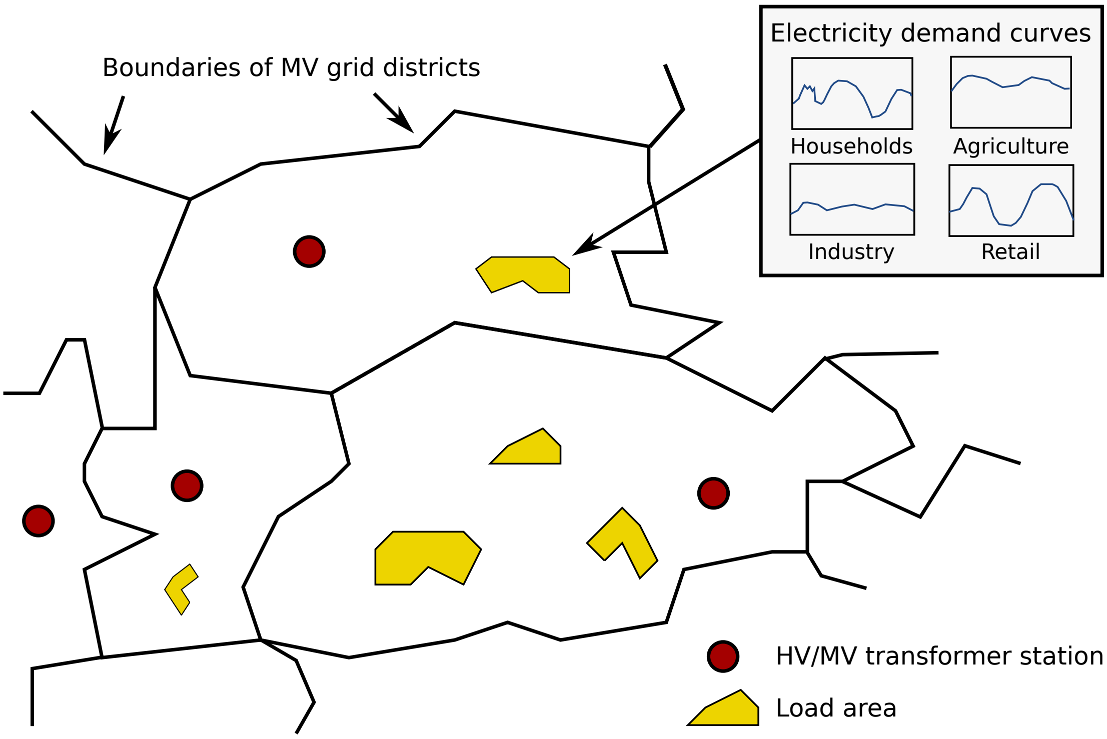

Boundaries of MV grid districts
Electricity demand curves
Households
Agriculture
Industry
Retail
HV/MV transformer station
Load area

Figure 5: Schematic illustration of medium-voltage network districts and load areas. Figure based on Amme et al. (2017).

For the construction of the dataset Germany is divided into about 3600 medium-voltage grid districts. These districts are based on locations and voltage levels of transformer stations, which are in turn taken from the OSM database. First, for each HV/MV transformer station the administrative boundaries of its municipality are used to define a geographical polygon (Figure 5). If there is more than one transformer station in one municipality, the municipality is split into further polygons using Voronoi partitioning (Hülk et al., 2017). It is then assumed that all load areas and generation units within one polygon are connected to the transformer station of that polygon. Each of the polygons therefore represents one medium-voltage grid district (Hülk et al., 2017).

The final dataset consists of 3606 medium-voltage grid districts covering the entire area of Germany. They are on average 99 km^2 large. In these grid districts overall 208,486 load areas are located. Each load area is assigned to one district. The typical size of load areas is about 5 ha (Hülk et al., 2017).

3.2 Existing and future power plants

For existing power plants in Germany, their locations are taken from public datasets (Hülk et al., 2017). Because this location is only approximately known, their allocation to grid districts and their exact location within these districts is further determined by an algorithm described in Hülk et al. (2017). The installed generation capacity of future power plants are taken from the scenario 2035B of the Netzentwicklungsplan 2030 (Table 1).

Table 1: Total installed capacities of different electricity technologies in the scenario NEP 2035B.

	Installed capacity [GW]		Changes 2035 vs. 2015	
Technology	2015	2035	GW	%
Natural gas	27.9	33.5	5.6	20
Hard coal	31.5	11	-20.5	-65
Oil	4.5	0.5	-4	-89
Waste	1.7	0.5	-1.2	-71
Biomass	7.2	8.4	1.2	17
Lignite	22.9	9.1	-13.8	-60
Uranium	12	0.5	-11.5	-96
Mixed fuels	2.6	2.4	-0.2	-8
Wind onshore	41.3	88.8	47.5	115
Wind offshore	5.6	18.5	12.9	230
Solar	38.5	59.9	21.4	56
Run off river	3.9	4.2	0.3	8
Reservoir	1.4	0	-1.4	-100
Pumped hydro	9.3	12.7	3.4	37

For future power plants, the total installed generation capacity of each technology is first allocated to medium-voltage grid districts. Within these districts, the installed capacity is then used to define a set of future power plants with certain generation capacities. These plants are then assigned to either the medium-voltage level or the low-voltage level, depending on their technology and nominal generation capacity (Amme et al., 2017). If a power plant is allocated to the medium-voltage level, the algorithm chooses a location in the medium-voltage grid district and the power plant is connected to the medium-voltage network (Section 3.3). If it is assigned to the low-voltage level, it is allocated to one of the low-voltage districts and there connected to the low-voltage network (Section 3.3).

The geographical distributions of installed generation capacity in the years 2015 and 2035 for onshore wind and solar photovoltaic are shown in Figure 6 and 7, respectively.

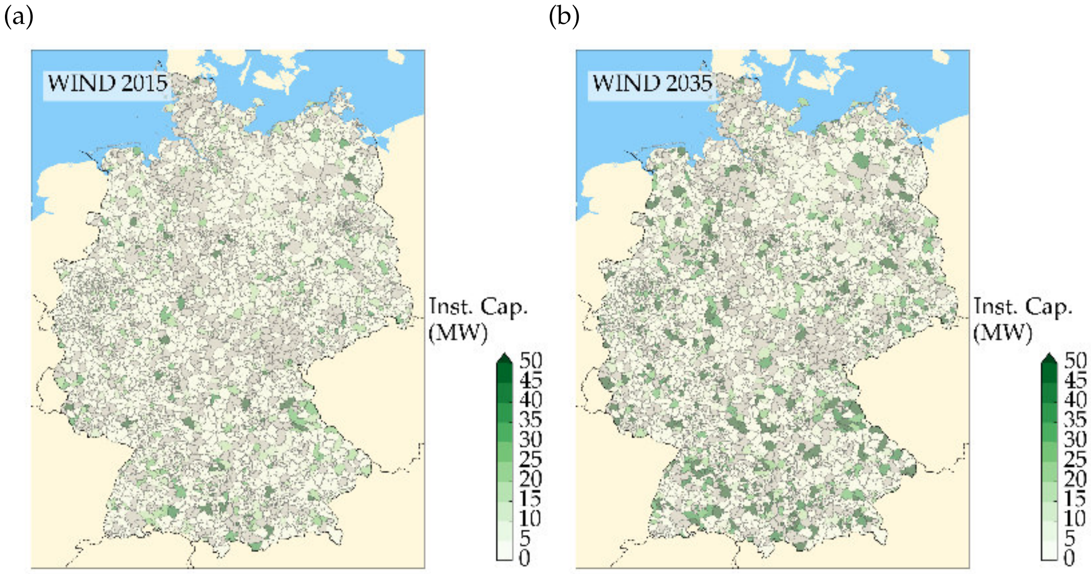


Figure 6: Geographic map of total installed generation capacity of onshore wind power plants in medium-voltage network districts in the dataset for the year (a) 2015 and (b) 2035.

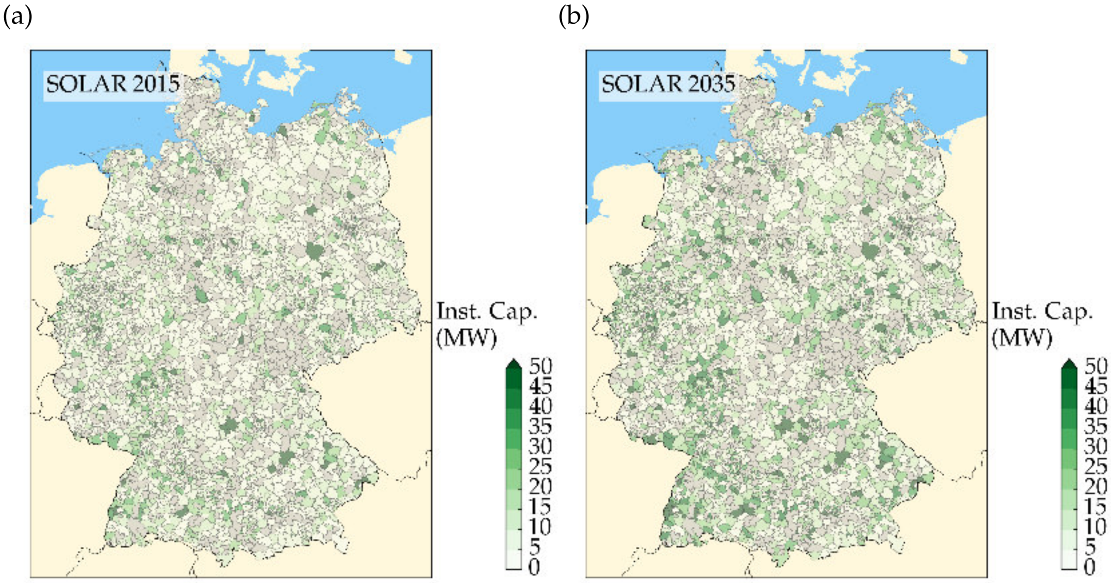


Figure 7: Geographic map of total installed generation capacity of solar photovoltaic power plants in medium-voltage network districts in the dataset for the year (a) 2015 and (b) 2035.

3.3 Medium- and low-voltage networks

Based on the spatial distribution of load areas and medium-voltage grid districts (Section 3.1) and the spatial distribution of generation units (Section 3.2), the software ding0 constructs a network topology for each of the medium-voltage grid districts connecting load areas and generation units to each other and to the transformer station. In the following, the algorithm is briefly described. More details can be found in Amme et al. (2017).

For the connection of load areas to the network, each load area is first categorised as either regular load area, satellite load area, or aggregated load area. Regular load areas feature a peak load $\geq 100\,\mathrm{kVA}$. The centre of each regular load area is always integrated into one of the rings of the medium-voltage network. Furthermore, each regular load area features regularly spaced MV/LV transformer stations (Figure 8). Each MV/LV transformer station is then assigned one of the 196 low-voltage network, which is randomly chosen from a sample of idealised low-voltage networks with simple radial topology (Figure 8).

The second category of load areas are satellite load areas. These are load areas with a peak load $\leq 100\,\mathrm{kVA}$. Satellite load areas are also connected to the medium-voltage grid and also feature several low-voltage grid districts. However, they are not necessarily integrated into one of the rings on the medium-voltage network and can also be connected to a branch of the network (Figure 8).

The third category of load areas are aggregated load areas. They are defined as load areas that require a cable with transmission capacity $\geq 1\,\mathrm{kVA\,km}^{-1}$. Aggregated load areas represent urban areas. Loads and generators in this area are treated as one aggregated production and consumption unit. Aggregate load areas are directly connected to the HV/MV transformer station. Furthermore, in contrast to regular load areas generation capacity and load within aggregated load areas is directly connected to the HV/MV substation's bus bar. Aggregated load areas are therefore not decomposed into low-voltage districts and the low-voltage networks are not explicitly modelled (Figure 8). This is because networks in urban areas are assumed to feature a high enough transmission capacity and local demand of electricity that they can host future renewable energy plants without network reinforcements. Their network topology at the low-voltage level is therefore considered to be irrelevant for total reinforcement costs (Amme et al., 2017).

Generator units are allocated to medium-voltage grid district and assigned to either the medium-voltage or the low-voltage level depending on their generation capacity (Section 3.2). If they are assigned to the medium-voltage level, they are placed somewhere in the medium-voltage grid district and can be either integrated into one of the rings or connected to a branch of the network. This is determined by the algorithm. If they are assigned to the low-voltage level, they are allocated to one of the corresponding low-voltage grid districts and there connected to the low-voltage network (Figure 8).

In sum, all synthetic networks on the medium-voltage level feature a topology consisting of rings and branches (Figure 8). This topology is constructed using a sophisticated algorithm (Amme et al., 2017). Networks on the low-voltage level feature a radial structure (Figure 8). Their topology is determined by a random choice of one of 196 idealised radial low-voltage networks. The topology of these idealised networks is based on Scheffler (2002) and Kerber (2010).

The final dataset consists of medium-voltage districts of the electricity distribution system. These districts can be considered as distinct parts of the distribution system because each district is connected to the transmission system only by one HV/MV transformer station and because the districts are connected to each other only via the transmission system. Each observation of the dataset hence represents one medium-voltage district and the corresponding distribution network. This distribution network consists of one medium-voltage network and potentially several low-voltage networks, depending on the number and category of load areas in the district.

When the topology of all networks has been determined, all load areas and generation units have been connected to the distribution network. The technical equipment required to safely operate the network (cables, cable distributors, transformers, switch disconnectors) is then determined using a power flow simulation. For this purpose, two situations are assumed. First, a situation with maximum load and minimum generation in the distribution network. Second, a situation with minimum load and maximum generation. The technical equipment is chosen such that neither voltage nor thermal limits are exceeded in any of the two situations. If necessary, also the topology of the

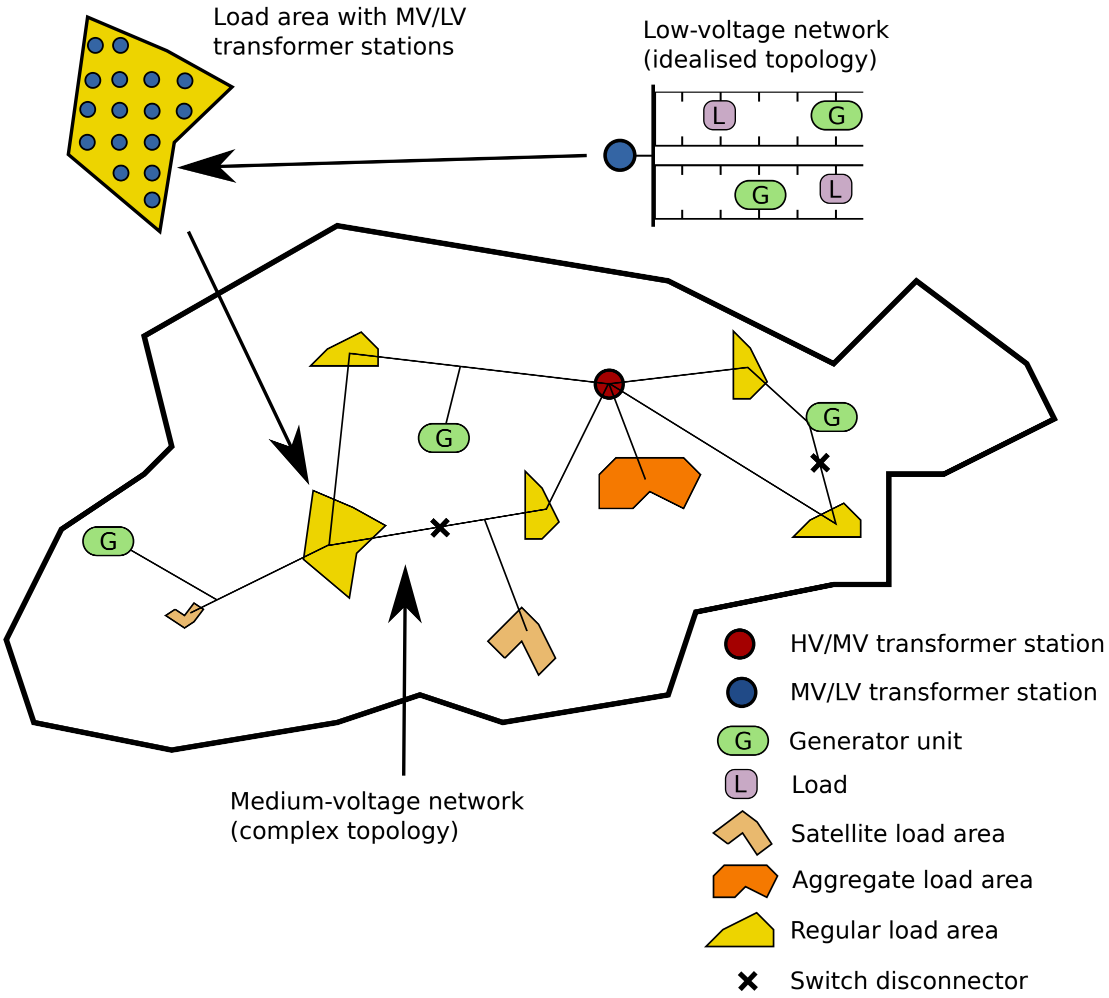

Load area with MV/LV transformer stations
Low-voltage network (idealised topology)
L
G
G
L
G
G
G
Medium-voltage network (complex topology)
HV/MV transformer station
MV/LV transformer station
Generator unit
Load
Satellite load area
Aggregate load area
Regular load area
Switch disconnector

Figure 8: Schematic illustration of medium-voltage network district with load areas, low-voltage network districts and low-voltage networks. Figure based on Amme et al. (2017).

distribution network is step-wise adjusted until a technically viable state of the network is obtained (Amme et al., 2017). This state considers only power plants existing in 2015. The algorithm used to determine the equipment and the costs of network expansion for scenarios with additional future generation units (Section 3.2) is explained in Section 3.4.

3.4 Scenarios of network expansion

For the scenario with additional future generation units (Section 3.2), the generation units are first allocated to a medium-voltage grid district and assigned to either the medium-voltage or the low-voltage level as described in Amme et al. (2017). They are then connected to the corresponding medium-voltage or low-voltage network using the algorithm described in Section 3.3.

For the connection of the generation units to the network a standard technical equipment can be used. This equipment does however not yet ensure that the network can operate in a safe state and that technical limits are maintained. To determine the additional equipment needed to ensure safe operation, the power flow in the network is simulated. Because the networks are already designed to cope with a situation with high load and low generation (Section 3.3), at this step only a situation with low load and high generation is simulated.

For this thesis, two scenarios of high generation are considered. The first scenario is a worst-case scenario. In that scenario, all generation units except solar photovoltaic power plants feed in 100 % of their nominal installed generation capacity. Solar photovoltaic power plants feed in 85 % of their nominal capacity. The load is set to 20 % of peak load.

The second scenario is a scenario with curtailment. For this scenario, the locations of onshore wind and solar photovoltaic power plants are combined with re-analysis weather time-series for the year 2011. From this, for each power plant a time-series of generation is constructed. Then, for each medium-voltage grid district the hour of the year 2011 with the maximum sum of generation from solar photovoltaic and onshore wind power plants is identified. For this hour, it is assumed that both photovoltaic power plants and onshore wind power plants are curtailed at 70 % of their nominal installed capacity. Furthermore, it is assumed that all other power plants feed in 100 % of their nominal capacity. The load is again set to 20 % of peak load. For the curtailment scenario, each medium-voltage network is therefore expanded according to the demand for transmission capacity in a potentially different hour of the year. However, the scenario is consistent in the sense that for each network the hour of the year is used in which, assuming a curtailment at 70 %, the demand for transmission capacity is maximum.

Both scenarios therefore represent situations with high generation and low load. Both scenarios are from an economic perspective unrealistic because in a situation with high generation from power plants using fluctuating renewable energy sources, electricity prices tend to be relatively low and the generation from conventional power plants is therefore likely to be low or even zero. However, this assumption is consistent with grid codes for distribution network operators in Germany and therefore current practice of network expansion (BDEW, 2008; VDE, 2011).

If the results of the power flow simulation of the respective scenario indicates that either a voltage or a thermal limit is exceeded, the transmission capacity of the network is expanded.[2] The expansion is done iteratively until no violation of voltage and thermal limits occurs any more with the following steps (Schachler, 2017):

First, all line segments at which thermal limits are exceeded are reinforced. This is done at the medium-voltage and at the low-voltage level. If a line segment can be sufficiently reinforced by adding one cable of the same type as the existing cable, a parallel line with this cable type is laid. Otherwise, the existing cable is replaced by a the standard cable type with sufficient transmission capacity. The reinforcement is likewise applied to transformers. If the current of a transformer station exceeds its technical limit, an additional transformer is added or the transformers are replaced. All reinforcements are applied simultaneously. This neglects that reinforcements in one line segment may affect the current in other line segments due to changes to impedance but has the advantage that the resulting network is independent of the order of reinforcements.

Second, violations of voltage limits are resolved. The order of reinforcements matters more for the violation of voltage limits than for thermal limits. For this reason, the following order was chosen. First, violations of voltage limits at the medium-voltage level are resolved. When all violations on the medium-voltage level are resolved, the MV/LV transformer stations are addressed. If the voltage of the low-voltage network close to the transformer station is close to the limit, transformers are added to the transformer station in order to reduce its impedance. Lastly, violations of voltage limits at the low-voltage level are resolved. The algorithm of the reinforcement is the same on the medium-voltage and on the low-voltage level and described in the following.

For each branch of the network, first the largest violation of a voltage limit is identified. This violation is resolved by laying a parallel line with the same transmission capacity over $\frac{2}{3}$ of the length of the line. If this does not yet sufficiently reduce the voltage deviation, another line is laid parallel to the second one to further reduce the line impedance. This is iteratively done until the voltage deviation is sufficiently small.

[2] The terms extension, expansion and reinforcement of networks are used synonymously in this thesis

When the largest violation of voltage limits has been resolved, the power flow in the updated network is simulated. Then the largest violation of the remaining violations of voltage limits is resolved. This is done iteratively: after each successful reinforcement, the power flow is simulated in order to determine whether the reinforcement already resolved other violations of voltage limits.

The costs of network expansion are calculated by multiplying the required equipment by its market prices. The prices are taken from the OpenEnergyPlatform[3] and originate from several published sources (Ackermann et al., 2014; Consentec et al., 2006; dena, 2012; Rehtanz et al., 2017).

Finally, the worst-case scenario and the curtailment scenario are therefore represented by costs for network expansion for each medium-voltage grid district. The geographical distributions of these costs are shown in Figure 9. There is no clear North-South or East-West gradient of costs. Furthermore, there are no clear regional hotspots of costs. Similar to the geographical distribution of installed capacity of renewable energy power plants (Figure 6 and 7) the costs do not follow any clear spatial patterns.

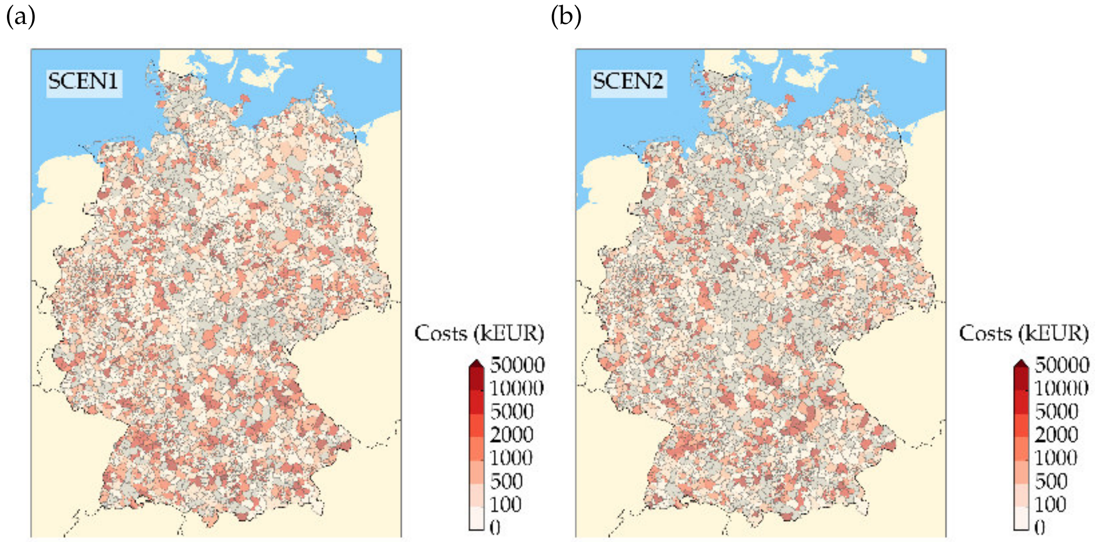


Figure 9: Geographic map of costs of network expansion of medium-voltage network districts in the dataset for the (a) worst-case scenario (SCEN1) and (b) curtailment scenario (SCEN2).

[3] The platform is accessible online at: https://oep.iks.cs.ovgu.de/

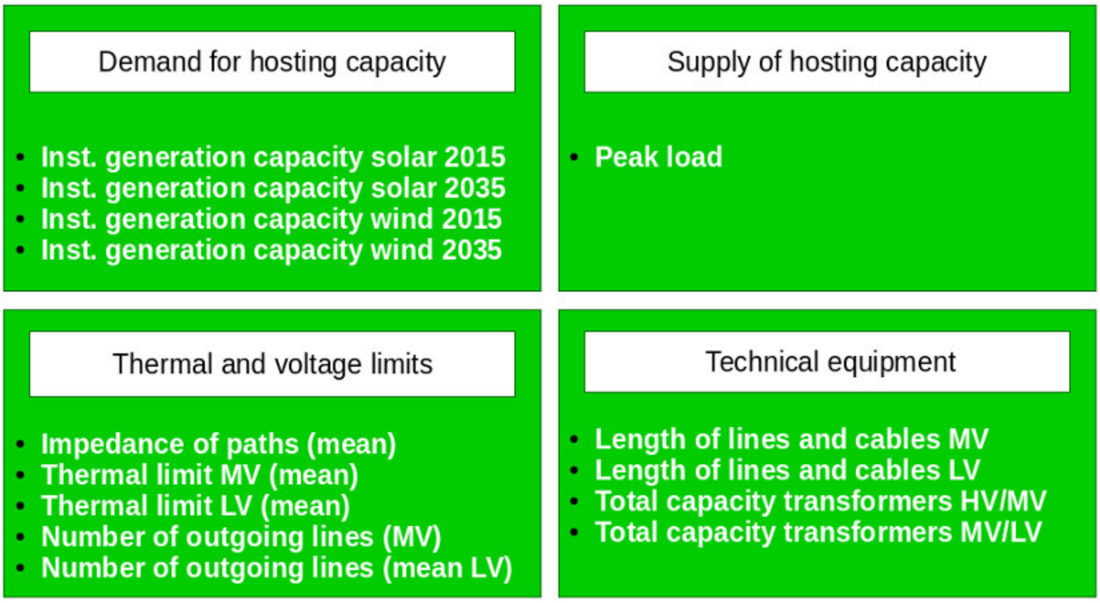
4. Network attributes

The demand for expansion of distribution networks in Germany is driven by a number of factors. When the distribution grid was planned and built, its capacity was oriented at the maximum amount of power that is consumed at the terminal end of a branch. Power was generated by few centrally located power plants and from these plants transported into regions and to consumers. Today, more and more power is generated locally. Whenever the amount of power that is locally generated exceeds what is locally consumed, the excess power needs to be transported away. This means that today in many places power flows in reverse direction and additional transmission capacity on the distribution level is required.

The approach taken in this thesis is to estimate the total costs of network expansion using a cluster analysis (Chapter 5). Clusters of networks, in turn, require network attributes that somehow determine the expansion costs of an individual network. The network attributes are selected based on theoretical considerations on the relationship between network attributes and expansion costs. Each attribute can be assigned to one of four categories (Figure 10). The first category contains attributes that are considered as drivers of the demand for hosting capacity (Section 4.1). In this context, hosting capacity describes the capability of a network to host additional power plants. The second category comprises attributes that describe the supply of hosting capacity (Section 4.2). The third category includes attributes that determine the occurrence of violations of voltage or thermal limits (Section 4.3). The fourth category contains attributes that determine the costs of the technical equipment required to resolve a potentially violated technical limit (Section 4.4).

Overall, 14 network attributes are identified and selected for the analysis in the following Chapters. Their selection was based on theoretical considerations and presented and discussed with experts from the Reiner-Lemoine-Institute Berlin. Finally, for these network attributes descriptive statistics are given (Section 4.5).

In addition to the network attributes described in the following, there may be other



Demand for hosting capacity

- Inst. generation capacity solar 2015
- Inst. generation capacity solar 2035
- Inst. generation capacity wind 2015
- Inst. generation capacity wind 2035

Supply of hosting capacity

- Peak load

Thermal and voltage limits

- Impedance of paths (mean)
- Thermal limit MV (mean)
- Thermal limit LV (mean)
- Number of outgoing lines (MV)
- Number of outgoing lines (mean LV)

Technical equipment

- Length of lines and cables MV
- Length of lines and cables LV
- Total capacity transformers HV/MV
- Total capacity transformers MV/LV

Figure 10: Theoretical framework to identify and categorise network attributes that potentially determine the costs of network expansion on the distribution level.

factors that determine the costs of network expansion. The authors of dena (2012) expect, for example, that there are peculiarities of states and regions in Germany which influence costs. Examples for these peculiarities are historically evolved typical network structures, the status of ongoing network development, and different principles applied in network development. However, these factors are not taken into account in the synthetic dataset and therefore not discussed here.

4.1 Demand for hosting capacity

The main driver of additional demand for transmission capacity in distribution networks are additional generation units (dena, 2012). In Germany, most additional future generation units are projected to be onshore wind and solar photovoltaic power plants (Table 1). Other drivers are, for example, changes in the use of power in other energy sectors (due to e.g. technological innovations such as heat pumps or electrically powered transport vehicles) and changes in the demand for power by existing power consumers. However, in the project open_ego and therefore also in this thesis it is assumed that the power demand does not change until 2035. This is assumption is not realistic but it is made to focus on the costs due to changes in generation technologies.

In this thesis the currently and the future total installed generation capacity of onshore wind and solar photovoltaic power plants are included as network attributes. All other technologies of electricity generation are neglected. Non-renewable technologies are excluded because according to the scenario B of the NEP, their total generation capacity will decrease until 2035 (Table 1). This means that these technologies are unlikely to incur significant network expansion costs. Furthermore, most power plants using non-renewable technologies are connected to networks on the extra-high and high-voltage level.

The only non-renewable technology that does not show a decreasing trend and whose plants are sometimes connected to the medium-voltage level is natural gas. However, increases in total installed capacity of natural gas are both in absolute and relative terms substantially smaller than changes for onshore wind and solar photovoltaic (Table 1). The same is true for other renewable energy technologies including biomass and pumped hydro (Table 1). Off-shore wind plants are excluded because they are generally connected to the extra-high- or high-voltage grid.

4.2 Supply of hosting capacity

Whether an additional demand for transmission capacity (Section 4.1) requires the expansion of the network depends on the supply of transmission capacity. Because historically transmission capacities were chosen to satisfy situations with high load and low generation, it is expected that the higher the aggregate load in a grid district, the larger the existing transmission capacity of the network (Chapter 2). Although the relatively large transmission capacity may have been intended to transport the amount of power required to cover the peak load, it can likewise be used in situations with minimum load and maximum generation to transport power away from generation units.

4.3 Thermal limits and voltage limits

Future additional demand of transmission capacity (Section 4.1) that exceeds the supply of hosting capacity of a network (Section 4.2) can result in a violation of technical limits. These technical limits are thermal limits and voltage limits. Thermal limits are violated if the current becomes too large. Thermal limits are typically violated close to the transformer station (Section 2.2, Figure 4). The transmission capacity of the first line segment of a branch is therefore included as network attribute, both on the medium-voltage and on the low-voltage level.

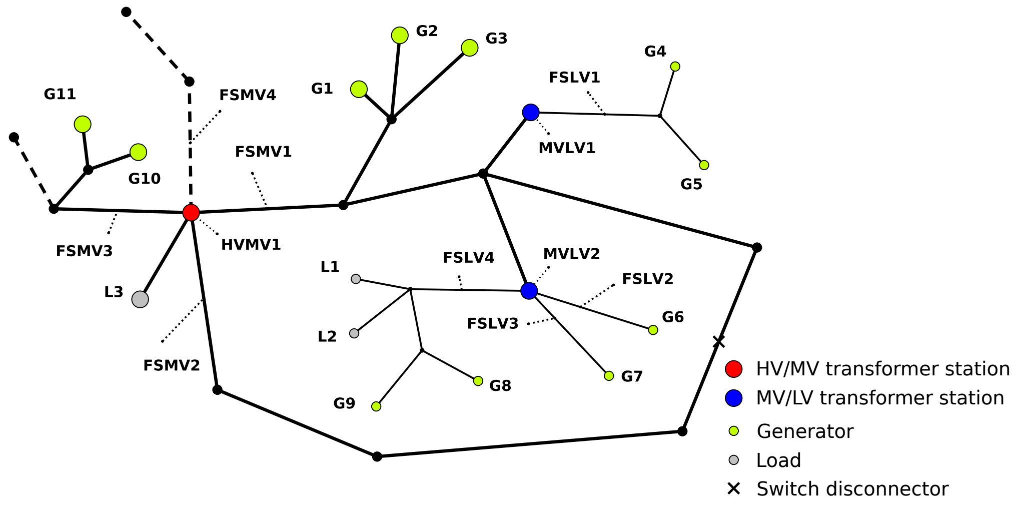


Figure 11: Schematic illustration of an electricity distribution network with medium-voltage and low-voltage level. Terminal nodes are either generators (G) or loads (L). First segments of network paths on medium-voltage level and low-voltage level denoted as FSMV and FSLV, respectively.

Without any knowledge about the position of future additional generation units in the grid district, it is assumed that new plants will be connected to one of the existing terminal nodes of the network. These terminal nodes are either generation units or loads. All terminal nodes are assumed to be equally likely. For this reason, the network attribute describing the likelihood of a violation of a thermal constraint is defined as the mean value of the transmission capacity of the first segments of all network paths to terminal nodes in the same district.

The network attribute is computed for both the medium-voltage and the low-voltage level. For the medium-voltage level, all terminal nodes irrespective their voltage level are used. Furthermore, the transmission capacity of the first segment of the path starting at the HV/MV station is used (FSMV in Figure 11). For the computation of the mean transmission capacity on the low-voltage level, only terminal nodes on the low-voltage level are used. Furthermore, the transmission capacity of the first segment of the path starting at the MV/LV station is used (FSLV in Figure 11). Furthermore, for the low-voltage level the mean transmission capacity is first computed for each MV/LV station and then averaged over all MV/LV stations in the network district.

For example, in Figure 11 each generator G1, G2, G3, ... and each load L1, L2, L3, ...represents one terminal node. Each of these terminal nodes is connected to the HV/MV transformer station with one unique network path (in standard operation, the switch disconnectors are open). In order to compute the network attribute based on the mean transmission capacity on the medium-voltage level, for each of these paths the transmission capacity of the first segment of the path on the respective voltage level (e.g. FSMV1 for G1 and L1 and FSMV3 for G10 on the medium-voltage level) is used.

Voltage limits are typically violated at the end of a line because the total impedance of the line has its maximum value there (Section 2.2, Figure 4). It is again assumed that each terminal node is equally likely to be chosen as position of a future generation unit. For this reason, the network attribute describing the likelihood of a violation of a voltage constraint is defined as the mean value of the total impedance of all network paths to terminal nodes. For the computation of the impedance, no distinction is made between the medium-voltage and the low-voltage level because the voltage drop along a line is determined by the total impedance across voltage levels. Furthermore, for terminal nodes on the low-voltage level the voltage drop over the transformer station is added.

For example, in Figure 11 the total impedances of all network paths to terminal nodes G1, G2, G3, ... and L1, L2, L3, ... irrespective their voltage levels are averaged.

Each of the two attributes is defined to describe the likelihood of a violation of either a thermal limit or a voltage limit. Because the costs of network expansion are determined by the number of violations that need to be resolved, two additional network attributes are included. These are the number of outgoing branches from the transformer station on the medium-voltage level and on the low-voltage level, respectively.

4.4 Technical equipment

If a technical limit of a network is violated (Section 4.3), the network can be expanded to resolve the violation. In this context, expansion can mean the replacement of existing equipment (e.g. exchange of a cable for a cable with larger capacity) and the use of additional equipment (e.g. laying a parallel cable; see also Section 3.4). The costs of these measures are mainly determined by the type of and the amount of the required equipment. This type of and amount of additional equipment is likely to be related to the type of and amount of existing equipment. For example, in a grid district with relatively long cables it is likely that also relatively long cables are required for network expansion.

In this thesis, the total length of lines and cables on the medium-voltage and on the voltage level are therefore included as network attributes. Furthermore, the total capacity of HV/MV transformers and of MV/LV transformers are included to account for the costs of their reinforcement.

4.5 Descriptive statistics

In this thesis, the estimation of expansion costs is based on a dataset of synthetic networks. The dataset was created using open-access data on demand and supply of electricity and covers the whole area of Germany with 3608 network districts. For 2928 (81 %) of these districts, network attributes and expansion costs were computed. The remaining 680 districts are neglected because expansion costs could not be computed. The total length of lines and cables of the synthetic networks on the medium-voltage level is 428 869 km. This corresponds to 84 % of the total length of lines and cables on the medium-voltage level in Germany (Bundesnetzagentur, 2017).

In order to facilitate the illustration, description and discussion of results, the selected network attributes are in the remainder of this thesis denoted by short labels. These labels are shown in Table 2. Furthermore, descriptive statistics of each of the 14 network attributes are presented in Table 3.

Table 2: Network attributes as defined in this Chapter and used throughout this thesis. Labels are used in text and illustrations in the following Chapters.

Label	Network attribute
WIND_2015	Installed generation capacity of onshore wind power plants in 2015
WIND_2035	Installed generation capacity of photovoltaic power plants in 2015
SOLAR_2015	Installed generation capacity of wind power plants in 2035
SOLAR_2035	Installed generation capacity of photovoltaic power plants in 2035
LOAD	Aggregated peak load
IMPEDANCE	Impedance of paths to terminal nodes (mean value)
IMAX_MV	Thermal limit of first segment of path from MV station to terminal node (mean value)
IMAX_LV	Thermal limit of first segment of path from LV station to terminal node (mean value)
NLINES_MV	Number of lines and cables going out from MV stations
NLINES_LV	Number of lines and cables going out from LV stations (mean value)
LENGTH_MV	Length of underground cables and overhead lines on MV level
LENGTH_LV	Length of underground cables on LV level
TRAFO_MV	Total capacity of transformers HV/MV
TRAFO_LV	Total capacity of transformers MV/LV

Table 3: Descriptive statistics of network attributes for the 2928 network districts in the dataset. Each observation in the dataset represents one medium-voltage network district including the corresponding networks on the low-voltage level.

Label	Unit	Minimum	Median	Mean	Maximum
WIND_2015	kW	0	0	5470	151943
WIND_2035	kW	0	3	13048	293816
SOLAR_2015	kW	0	6813	10901	156171
SOLAR_2035	kW	0	10630	16106	193939
LOAD	kW	156	19000	25606	551262
IMPEDANCE	Ohm	166	1468	1901	11657
IMAX_MV	A	210	362	377	609
IMAX_LV	A	218	294	297	419
NLINES_MV	-	2	6	7	65
NLINES_LV	-	1	232	306	2373
LENGTH_MV	km	2	107	147	1135
LENGTH_LV	km	0	110	141	824
TRAFO_MV	kVA	40000	80000	102007	1028000
TRAFO_LV	kVA	50	11730	16292	133820

5. Methods for cluster analysis

Groups of observations that are relatively similar can be referred to as clusters. These clusters can be identified with cluster analysis. There exist several alternative methods for the identification of clusters. In this thesis, the identification of clusters is also referred to as estimation of clusters. Methods from both supervised and unsupervised statistical learning can be used for this purpose. Clusters are typically estimated using one of several alternative algorithms based on a certain criterion of similarity between observations (Hastie et al., 2009).

Cluster estimation can be used to reduce the number of observations for a certain subsequent computation. Once the clusters have been identified, one or several observations can be selected for each cluster. These selected observation can then be used for the computations. Based on the fact that observations of the same cluster are more similar than observations of different clusters, it is then often assumed that the results of the computation for the selected observations are representative for all observations of the same cluster.

Because the computation of the costs of network expansion requires a lot of computing resources, it can only be done for relatively few selected distribution networks. The main aim of the cluster estimation in this thesis is therefore to identify representative networks, which can then be used to reduce the required time and resources for subsequent computations of the costs of network expansion, potentially for several alternative scenarios. Based on the assumption that the costs of the selected networks are representative for the costs of all networks of the same cluster, the representative networks can then be used to estimate the total costs of all networks.

An alternative method to estimate the total costs of all networks from the costs of few selected networks would be to train a model based on the sample of networks and use this model to estimate the total costs. This approach could be referred to as a prediction model. The main advantage of cluster analysis is that once the clusters and the representative networks have been identified, they can be used for several computations. For example, the representative networks can be used to compute expansion costs for alternative scenarios of future generation technologies. With prediction models one would need to estimate one model for each of the scenarios.

Furthermore, cluster analysis can be used to describe certain structures within a dataset. For example, once clusters of electricity networks have been identified, these clusters can be used to identify a set of typical networks. These networks can be used, for example, to derive classes of networks. Furthermore, the typical networks can be used to map their occurrence across Germany in order to visualise and analyse regional differences of distribution networks. Prediction models can allow one to identify typical observations, for example in the case of regression trees that are used as prediction models, but not all prediction models and estimation methods produce groups of observations.

In this thesis, two methods for cluster estimation are used. These are K-Mean and regression trees. Both methods provide first an assignment of observations to clusters and second one representative observation for each cluster. However, the way how observations are grouped into clusters and the way how the representative observations are identified differ between the two methods. The two methods, K-Mean (Section 5.1) and regression trees (Section 5.2) are described in the following. Then, two metrics are introduced and explained that can be used to evaluate the performance of a cluster model (Section 5.3). Finally, a simple example is given that illustrates the application of the two estimation methods, their differences and the two metrics for evaluation (Section 5.4).

5.1 K-Mean

One of the most popular methods for cluster analysis is K-Mean estimation (also referred to as K-Mean in the following) (Hastie et al., 2009). K-Mean requires that the number of clusters K is specified. Each observation $i = 1, ..., N$ with p attributes $X_i = (x_{i1}, ..., x_{ip})$ is then assigned to one of the K clusters based on its distance from the clusters' centroids. The centroids of clusters are points in the p-dimensional space of attributes. They can be determined by an algorithm that minimises the sum of distances of observations from the centroid of the cluster to which they are assigned.

The clusters are created in an iterative procedure (Hastie et al., 2009). The main result of the algorithm is an assignment of each observation $i = 1, ..., N$ to one cluster with index $k = 1, ..., K$, which can be written as an assignment function $C(i) = k$.

For K-Mean, the Euclidean distance $D(\cdot, \cdot)$ is typically used as measure of distance between observations. This means that the result of K-Mean is relatively sensitive to outliers. For this reason, Hastie et al. (2009) recommend to use the more robust K-Medoid if it is computationally feasible. K-Medoid was tested but the computation time turned out to be too large for this thesis.

The assignment of observations to clusters, represented by the assignment function $C(\cdot)$, can be evaluated based on a measure of the within-cluster dissimilarity $W_{C(\cdot)}$ (James et al., 2017):

$$W_{C(\cdot)} = \sum_{i=1}^{N} \sum_{k=1}^{K} D(X_i, M_k) I_k(X_i) \tag{5.1}$$

whereby M_k is the centroid of cluster k and $I_k(\cdot)$ is an indicator function:

$$I_k(X_i) = \begin{cases} 1 & \text{if} \qquad\qquad C(i) = k \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

K-Mean requires an initial assignment of observations to clusters $C(\cdot)^{\text{init}}$. The final clusters are sensitive to this assignment. In order to choose the initial assignment, the cluster estimation can be repeated several times with random initial assignments. One can then choose the initial assignment $C(\cdot)^{\text{init}}$ and its corresponding final assignment $C(\cdot)$ that results in the lowest value of $W_{C(\cdot)}$. The algorithm used in this thesis does this by itself.

The number of clusters K needs to be specified. In this thesis, cluster models are estimated for $K = 1, ..., 20$ to analyse how K influences the quality of the cluster model. The value of K can also be chosen based on the structure of the data. For example, Tibshirani et al. (2001) propose to generate a synthetic data set with observations uniformly distributed over a rectangle that includes all observations of the actual data set. They propose to compute $W_K \equiv W_{C(\cdot)|K=K}$ from the actual data set ($\log(W_K)$) and from the synthetic data set ($\log(W_K^u)$) for different K. The optimal value $K^\star$ can then be obtained as:

$$K^\star = \underset{K}{\operatorname{argmax}}\ \log(W_K) - \log(W_K^u) \tag{5.3}$$

The optimal value $K^\star$ was investigated but it was finally decided not to include the results in this thesis. The reason is that the main objective of this thesis is not to describe certain properties of the dataset of synthetic networks, such as the optimal number of clusters for certain network attributes, but rather to compare alternative cluster models for a fixed value of K in order to determine which value of K provides a good balance between model performance and the required computing resources. In this context, model performance refers to the error made by using only K representative observations to estimate the total costs of network expansion. Because this model performance generally tends to increase with K, it is the constraint of computational feasibility rather than the structure of the dataset which determines the "optimal" value of K.

The results of K-Mean are the final assignment of observations to clusters $C(\cdot)$ and the coordinates of the centroid of each cluster M_k. The observation with the minimum distance to the centroid of its cluster can then be chosen as representative observation of that cluster. The coordinates of the representative observation of cluster k can be written as $\tilde{X}_k$ and are determined by:

$$\tilde{X}_k = \underset{X_i|C(i)=k}{\operatorname{argmin}}\ D(X_i, M_k) \tag{5.4}$$

The Euclidean distance assigns equal weight to each attribute. This means that if the data is not normalised, the clustering assigns the same weight, for example, to one meter of cable length as to one kW installed capacity of wind generation plants. In order to normalise the data, z-scores are calculated from the expected value $\mathrm{E}(x)$ and the sample variance $\mathrm{Var}(x)$ of attribute x:

$$z = \frac{x - \mathrm{E}(x)}{\mathrm{Var}(x)} \tag{5.5}$$

5.2 Regression trees

Regression trees are based on the recursive partitioning of a sample. They are especially useful to detect non-linear effects and interactions between variables. Furthermore, regression trees are well suited for the analysis of data with many variables because they automatise the process of variable selection (Varian, 2014). If the dependent variable data is non-continuous and unordered, regression trees are referred to as classification trees (James et al., 2017, chap. 8).

Partitions of a tree are sometimes also referred to as *nodes*. A distinction can then be made between *internal nodes* and *terminal nodes*. Internal nodes are parents of other nodes. Terminal nodes do not have any children nodes and are also referred to as *leaves*.

Regression trees can be constructed using one of several algorithms that differ with respect to the splitting rule and the prediction model (Loh, 2011). For example, some algorithms conduct statistical test to pre-select some variables when partitioning the sample, whereas other algorithms consider all variables for splitting the sample. Furthermore, while some algorithms fit linear or quadratic models to the observations within one leaf, other algorithms use the mean value of observations as estimate of the leaf (Loh, 2011). There also exist algorithms based on likelihood and Bayesian estimation as well as for logistic and Poisson regression (Loh, 2011).

In this thesis only the CART (Classification And Regression Tree) algorithm is used. The CART algorithm is one of the most popular and widely used algorithms (Hastie et al., 2009). The algorithm is usually attributed to Breiman et al. (1984). The algorithm considers all possible split candidates when partitioning the sample and uses the mean value of observations as estimate of the leaf.

The mathematical representation of a tree built with the CART algorithm is relatively simple. The tree can be represented by K partitions of the sample $R_1, ..., R_K$. With the independent variable X, and the dependent variable y the tree can be written as:

$$y(X) = \sum_{k=1}^{K} c_k I_k \left(X\right) \tag{5.6}$$

where $I_k(.)$ is an indicator function (Hastie et al., 2009):

$$I_k(X) = \begin{cases} 1 & \text{if} \qquad X \in R_k \\ 0 & \text{otherwise} \end{cases} \tag{5.7}$$

For estimating the tree based on observations $i = 1, ..., N$, the constants c_k can be set to the arithmetic mean of all observations (y_i, X_i) with p covariates $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ falling into the corresponding region R_k:

$$\hat{c}_k = \frac{1}{\sum_{i=1}^{N} I_k\left(X_i\right)} \sum_{i=1}^{N} y_i I_k\left(X_i\right) \tag{5.8}$$

The estimated tree can then be written as:

$$\hat{y}(X) = \sum_{k=1}^{K} \hat{c}_k I_k\left(X\right) \tag{5.9}$$

The CART algorithm constructs trees as follows: for each terminal node and for each of the $j = 1, ..., p$ covariates (attributes), the algorithm considers all possible split-points s of the interval $s \in (\min_i\{x_{ij}\}, \max_i\{x_{ij}\})$ with some step-size Δs. For each split point, the algorithm first computes the predicted value (typically the arithmetic mean) and second some measure of the quality of a cluster model (typically the mean-squared-error) for each of the two potential children nodes. The algorithm then chooses the split that yields the best fit in the two children nodes. The children nodes are then added as new terminal nodes. The algorithm typically continues splitting terminal nodes until a stopping criterion is met.

This criterion determines the complexity of a tree, typically measured by the number of terminal nodes. The criterion can be based, for example, on the minimum number of observations in a terminal node or on the improvement of the fit of the tree stemming from a potential additional split (Hastie et al., 2009). In this thesis, the number of leafs is specified and equal to the number of clusters.

The result of a regression tree is an assignment of observations to leafs. These leafs can be considered as clusters. For each cluster k, the observation whose value of y is closest to the mean value of all observations in the same cluster is considered as representative observation of that cluster. With the costs of the representative observation of cluster k as $\tilde{y}_k$ one can write:

$$\tilde{y}_k = \underset{y_i|C(i)=k}{\operatorname{argmin}} \left(y_i - \overline{y}_k\right)^2 \tag{5.10}$$

whereby $\overline{y}_k$ is the arithmetic mean of y_i of all observations i for which $C(i) = k$.

5.3 Metrics for evaluation of cluster models

The result of a cluster model can be evaluated in different ways. In this thesis, two metrics that describe the performance of a cluster model are used. The first metric describes the average distance in terms of y between observations of the same cluster. This metric is in the following referred to as the within-cluster dispersion of costs and denoted as W:

$$W = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\left(\sum_{j=1}^{N} \sum_{k=1}^{K} I_k\left(X_i\right) I_k\left(X_j\right)\right) - 1} \sum_{j=1}^{N} \sum_{k=1}^{K} |y_i - y_j| \cdot I_k\left(X_i\right) I_k\left(X_j\right). \tag{5.11}$$

In Equation (5.11), $\tilde{y}_k$ is the value of the independent variable of the representative observation of cluster k (determined by Equation (5.4) for K-Mean and by Equation (5.10) for regression trees).

The second metric describes the relative deviation of estimated total costs from calculated total costs if a certain cluster model is used to estimate total costs. This metric is sometimes in the following referred to as the relative deviation of total costs and denoted as R:

$$R = \left| \sum_{i=1}^{N} \sum_{k=1}^{K} \tilde{y}_k I_k\left(X_i\right) - \sum_{i=1}^{N} y_i \right| / \sum_{i=1}^{N} y_i \tag{5.12}$$

In sum, R is a metric based on the distance of observations from the representative observation of their cluster whereas W is based on the average difference of costs for observations of the same cluster. In consequence, R is sensitive to the choice of the representative observations whereas R is not. Both W and R are used in the analysis of the results in Chapter 6. The differences between W and R are also illustrated in the following in Section 5.4.

5.4 Simple example

In Section 5.1 and Section 5.2, two methods to estimate cluster models were described, K-Mean and regression trees respectively. In the following a simple example is developed that illustrates the two methods and some differences between them. For this purpose, a dataset is generated with an independent variable y, two attributes $X = (X_1, X_2)$ and $N = 100$. In this example, the association between X_1 and y is much weaker than the association between X_2 and y. Furthermore, the relationship between X_2 and y is non-linear (Figure 12).

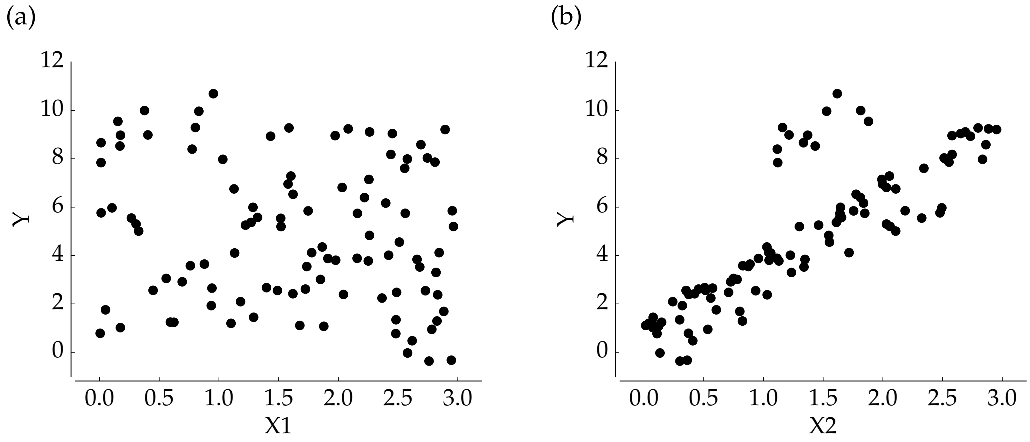


Figure 12: Scatter plot of attributes (X_1, X_2) and dependent variable (y) for simple example with two attributes: (a) X_1 and y and (b) X_2 and y.

For this dataset, a cluster model with the two attributes X_1 and X_2 is estimated using K-Mean and a regression tree. The number of clusters $K = 4$. The results of the cluster estimation are shown in Figure 13. The assignment of observations to clusters is indicated by symbols. Furthermore, the representative observations as determined by Equation (5.4) and Equation (5.10) are marked with X.

When assigning observations to clusters, K-Mean assigns equal weight to each attribute and does not take y into account (Section 5.2). In Figure 13(a), each of the 4 clusters therefore covers one of the four corners of the two-dimensional space. By contrast, the

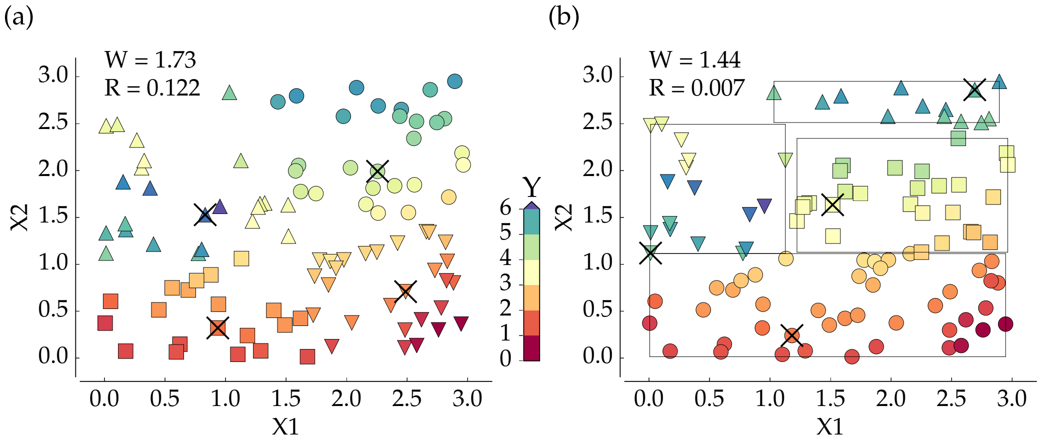


Figure 13: Clusters and representative observations for simple example with two attributes and $K = 4$ for (a) K-Mean and (b) regression tree estimation. Clusters are indicated by symbols, values of y are indicated by colour, and representative observations are marked with X.

regression tree also considers the association between attributes and y when assigning observations to clusters. In Figure 13(b), the two-dimensional space is hence partitioned in three segments with respect to X_2 and in at most two segments with respect to X_1. In addition, while K-Mean uses the observation closest to the geometric centroid as representative observation, the regression tree uses the observation with the value of y closest to the arithmetic mean of y within the cluster as representative observation. Figure 13 reveals that in this example the observation closest to the mean value of y is often located at the boundary of the cluster.

One consequence of the different partitioning of the two-dimensional space is that observations are more similar with respect to y within clusters for the clusters estimated with a regression tree than for the clusters estimated with K-Mean. This is also indicated by the values of the metrics W and R shown in Figure 13. Both the within-cluster dispersion of costs W and the relative deviation of estimated total costs from calculated total costs R are larger for the clusters estimated with K-Mean than for the clusters estimated with a regression tree.

The results of the same analysis with $K = 6$ are shown in Figure 14. As K increases from 4 to 6, the clusters estimated by K-Mean become smaller. For the assignment of observations to the new clusters, both attributes X_1 and X_2 carry the same weight. By contrast, as K increases to 6 the regression tree partitions the (X_1, X_2) space with additional segments in the direction of X_2. As indicated by the colour of symbols in Figure 14, this means that observations of the same cluster feature generally more similar costs than for the clusters estimated with K-Mean. The lower values of W and R for regression tree than for K-Mean estimation confirm this result from visual inspection.

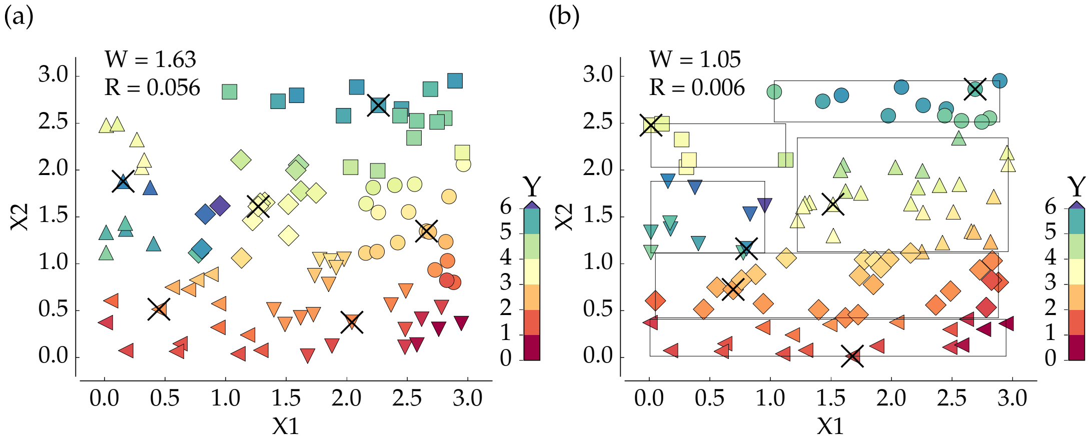


Figure 14: Clusters and representative observations for simple example with two attributes and $K = 6$ for (a) K-Mean and (b) regression tree estimation. Clusters are indicated by symbols, values of y are indicated by colour, and representative observations are marked with X.

In sum, the simple example illustrates how K-Mean and regression tree estimation assign observations to clusters and identify representative observations. In both steps, regression tree estimation takes the dependent variable y, in the remainder of this thesis the costs of network expansion for the worst-case scenario, into account whereas K-Mean does not. As a result, the clusters estimated with regression trees generally feature lower values of the metrics W and R and therefore tend to perform better. In this example, however, regression trees were trained with one dataset and then applied to the same dataset. In general, two different datasets or two different samples of the same dataset are used for training and applying a regression tree. In order to evaluate the performance of K-Mean and regression tree for an alternative scenario, in Chapter 6 both methods are applied to both the worst-case and the curtailment scenario. Because the computation of the costs of network expansion requires less computation for the worst-case scenario than for other scenarios, the worst-case scenario is generally used to train regression trees.

6. Results

This Chapter contains results of the cluster analysis. The Chapter is structured as follows. First, the average effect of including an attribute on the within-cluster dispersion of costs W (Section 5.3) of a cluster model is investigated. To this aim, cluster models that exclude the attribute are compared with models that include it (Section 6.1). Second, all cluster models are ranked according to W. This ranking is used to identify and examine the ten cluster models with the lowest value of W (Section 6.2).

Based on the results of these two steps of the analysis and the theoretical framework of network attributes (Chapter 4), 57 cluster models are selected. These cluster models are then assessed in terms of both the within-cluster dispersion of costs W and the relative deviation of estimated total costs from calculated total costs R (Section 5.3) for the worst-case scenario (Section 6.3) and the curtailment scenario (Section 6.4). Furthermore, some selected cluster model are analysed and compared in more detail (Section 6.5).

Finally, for three selected cluster models the resulting clusters themselves are analysed in more detail. For this, the representative networks of each cluster are identified and their coordinates in terms of network attributes are analysed. Furthermore, geographical maps of the occurrence of clusters are drawn and examined (Section 6.6). Throughout this Chapter, all steps of the analysis are done in parallel for both K-Mean and regression tree cluster estimation and all results are analysed regarding differences between the two methods.

6.1 Effect of attributes on within-cluster dispersion of costs

In order to conduct a cluster analysis, one first needs to choose a combination of attributes. Each of these attributes is then potentially used to assign observations to clusters. Regression trees consider each of the attributes and split the sample based on those attributes that improve the cluster model with respect to the target variable y

the strongest (Section 5.2). By contrast, K-mean assigns observations to clusters based on their distance from cluster centres. For the calculation of distances, each attribute carries equal weight. This means that for regression tree estimation a cluster model with more attributes yields at least the same performance (in terms of e.g. W, R) as a cluster model that includes only a subset of the same attributes. In the case of K-Mean a larger number of attributes does not necessarily yield a better cluster model (Section 5.1).

In order to account for interactions between attributes, in the following the average effect of including an attribute in a cluster model with other attributes on the performance of that cluster model is examined. For this purpose, the performance of a cluster model is assessed in terms of the within-cluster dispersion of costs W. The reason is that W does not account for differences between K-Mean and regression tree cluster estimation in how these methods choose representative networks (Section 5.3).

In order to compute the average effect of including an attribute in a cluster model, all possible cluster models with a certain number of attributes were estimated. Then, for each attribute A two cluster models are identified and compared: the performance of model with n-1 attributes excluding A and the performance of the corresponding models with the same n-1 attributes plus A. For these models, the difference $\Delta W^{A,c} = W_n^{A,c} - W_{n-1}^{A,c}$ is calculated. This is done for all possible models with n-1 attributes that exclude A. From the results, the average effect $\overline{\Delta W^A} = \frac{1}{C}\sum_{c=1}^{C} \Delta W^{A,c}$ is calculated.

The results show that the order of importance of the 14 network attributes according to $\overline{\Delta W^A}$ is generally similar for K-Mean and regression tree estimation (Figure 15). For the six attributes with the strongest effect on W, the order is indeed exactly the same for both estimation methods (in decreasing order of importance): WIND-2035, WIND-2015, IMAX-MV, SOLAR-2035, LENGTH-MV, and SOLAR-2015 (Figure 15).

In Figure 15(a) the result for $n = 3$ and $K = 100$ is shown. In general, the larger the number of included attributes n the more effects from interactions between variables are included in the calculation of $\overline{\Delta W^A}$. To quantify this effect, Figure 15(b) shows the the results of the same computations but with $n = 5$. For K-Mean, fewer attributes tend to reduce W for $n = 5$ than for $n = 2$. This confirms that including more attributes does not necessarily improve the performance of a cluster model that is estimated using K-Mean. Furthermore, the marginal effect is for all attributes smaller for $n = 5$ than for $n = 2$. That is also the case for regression tree estimation. However, the order of importance of the 14 network attributes does not change with n, neither for K-Mean nor for regression tree estimation.

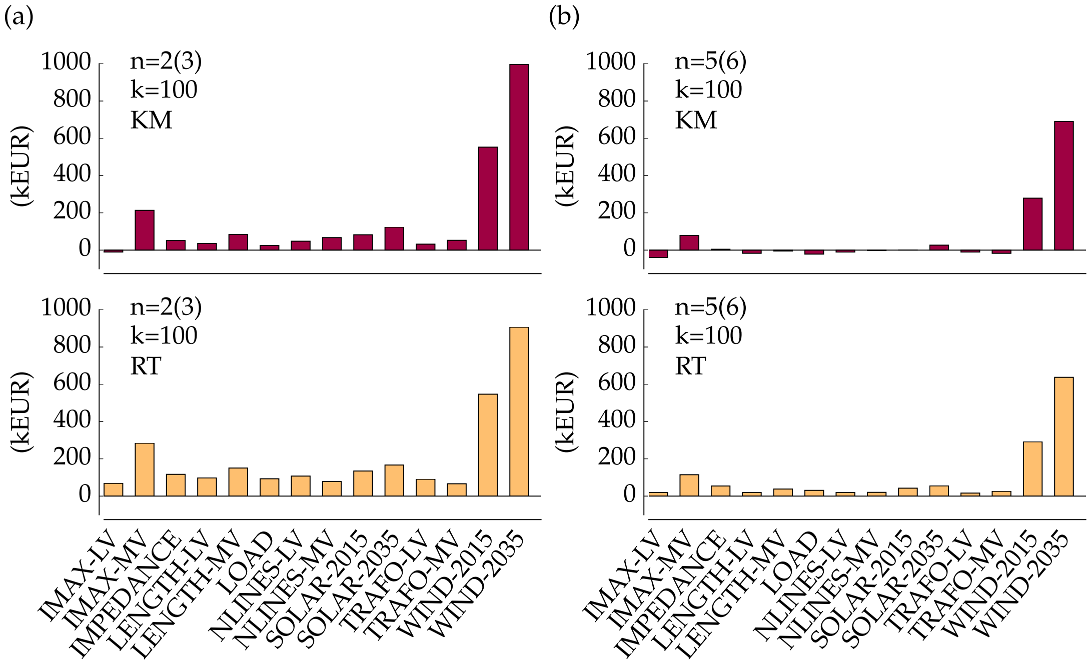


Figure 15: Effect of including an attribute in cluster models with (a) $n = 2$ and (b) $n = 5$ on within-cluster dispersion of costs, as measured by $\overline{\Delta W^A}$ (see text).

The average effect of including an attribute in a cluster model $\overline{\Delta W^A}$ takes interaction effects of several attributes into account. An alternative metric to quantify the potential effect of including attributes in a cluster model are correlation coefficients between attributes and y. One can generally assume that the larger the correlation coefficient, the more the attribute tends to improve existing cluster models. Correlation coefficients neglect however interactions between attributes.

Pearson correlation coefficients between network attributes and y are shown in Figure 16. Overall, the order of relative importance differs for many pairs of attributes between the result for $\overline{\Delta W^A}$ and the correlation coefficients. Furthermore, attributes which on average have a negative effect on W of a cluster model if they are included in the cluster model and if the cluster model is estimated with K-Mean feature correlation coefficients comparable to the correlation coefficients of attributes with a positive effect on W. Overall, the results thus suggest that correlation coefficients are not suitable for identifying those attributes which improve the performance of existing cluster models the most.

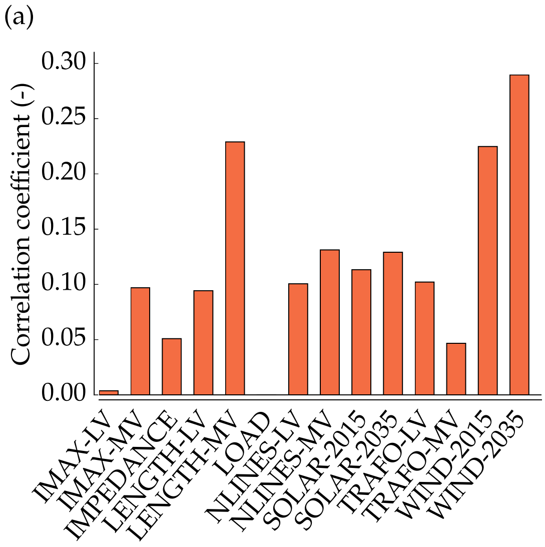


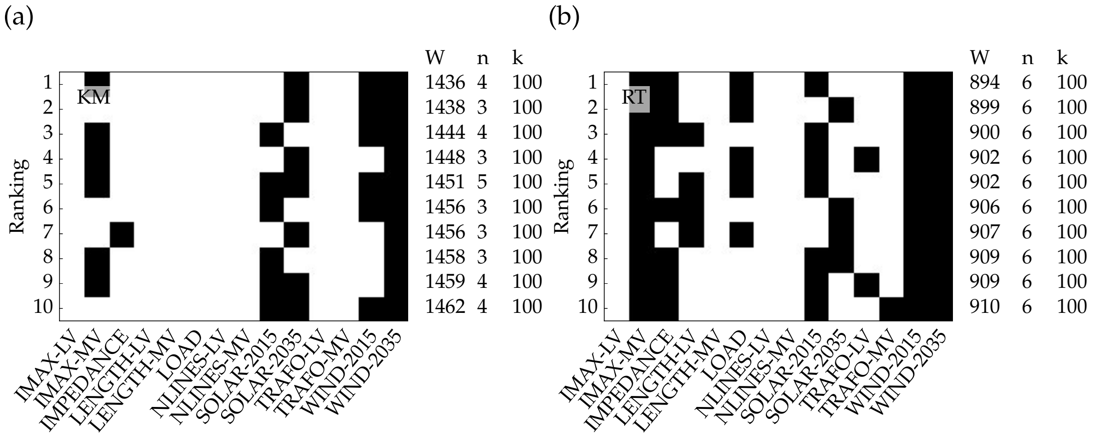
Figure 16: Pearson correlation coefficients between network attributes and costs of network expansion for worst-case scenario.

6.2 Ranking of cluster models

The results of the previous Section indicate which network attributes improve the performance of a cluster model the strongest if they are included in that model. However, the results do not reveal which are the best cluster models, which here refers to the combinations of attributes that perform best according to W.

For this purpose, all cluster models with $n = 2, 3, ..., 6$ and $K = 100$ were estimated. The ten cluster models with the lowest value of W are shown in Figure 17. The result reveals that some attributes appear in more than one of the ten best combinations. If one ranks attributes according to their occurrence in these combinations, the order of the top six attributes is: WIND-2035, SOLAR-2035, WIND-2015 and SOLAR-2015, IMAX-MV, IMPEDANCE for K-Mean and WIND-2035 and WIND-2015 and IMAX-MV, SOLAR-2015 and IMPEDANCE, LOAD for regression trees (Figure 17).

These are almost the same attributes as the attributes with the largest average effect on W in Section 6.1. The only difference is that in the best cluster models, LOAD replaces LENGTH-MV. In this analysis, only the ten best performing cluster models according to W were analysed. The reason is that only a limited number of cluster models can be



Ranking	W	n	k
1	1436	4	100
2	1438	3	100
3	1444	4	100
4	1448	3	100
5	1451	5	100
6	1456	3	100
7	1456	3	100
8	1458	3	100
9	1459	4	100
10	1462	4	100

Ranking	W	n	k
1	894	6	100
2	899	6	100
3	900	6	100
4	902	6	100
5	902	6	100
6	906	6	100
7	907	6	100
8	909	6	100
9	909	6	100
10	910	6	100

Figure 17: The ten best cluster models with $K = 100$ according to W for (a) K-Mean and (b) regression tree estimation.

analysed at this level of detail. For this thesis, those cluster models are considered most relevant that yield the best performance.

Figure 18 shows how the performance of cluster models as indicated by W declines if more cluster models are considered. Overall, W increases with a decreasing slope. From the first to the tenth cluster model, W increases most strongly. From there on, the slope declines more gradually. This can be regarded as support for the decision to focus on the top ten cluster models.

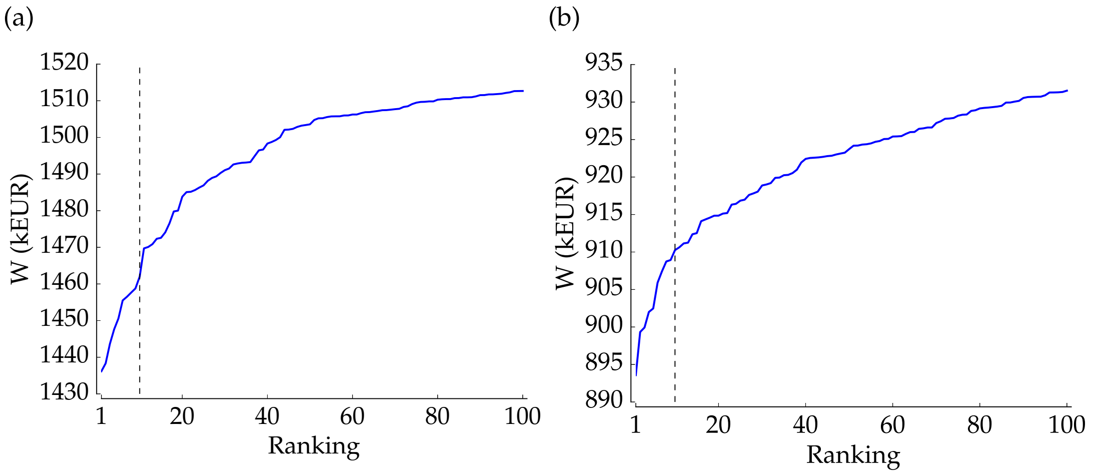


Figure 18: Within-cluster dispersion of costs W for 100 best cluster model with $K = 100$ according to W for (a) K-Mean and (b) regression tree estimation.

Moreover, the results in Figure 18 show that for regression trees, the performance of models is largest if all possible attributes are included. This is because regression trees consider only those variables that improve the performance of the cluster model when building the model. All other variables are neglected. In contrast to this, for K-Mean clustering all attributes carry equal weight for assigning observations to clusters. The optimal number of attributes is hence smaller than for regression trees. The results indicate that cluster models with 3 or 4 attributes have the lowest value of W (Figure 18).

This raises the question whether starting with the optimal cluster model for a certain value of n and then increasing n stepwise by including additional attributes can yield the optimal cluster model of attributes for subsequent values of n. The optimal cluster models for $n = 2, 3, 4, 5, 6$ and $K = 100$ are shown in Figure 19. The results give a mixed picture. For K-Mean estimation, the best cluster model of all $n \leq 5$ can indeed be obtained by adding one attribute to the best cluster model for $n - 1$. For $n = 6$, the attributes of the best cluster model change. For regression trees, there is two exceptions for $n = 3$ and $n = 4$. Apart from these two exceptions, the best cluster model can be obtained by adding one attribute to the best cluster with $n - 1$ attributes.

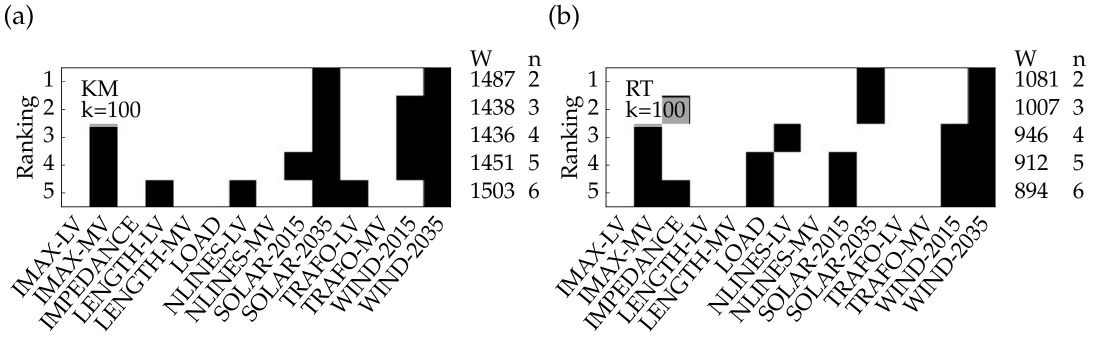


Figure 19: The best cluster model with $K = 100$ and $n = 2, 3, ..., 6$ according to W for (a) K-Mean and (b) regression tree estimation.

For the further analysis, a set of cluster models needs to be selected. Based on the theoretical considerations in Section 6.1 and the previous results, these combinations are mostly made up of the following network attributes: WIND-2035, WIND-2015, SOLAR-2035, SOLAR-2015, IMAX-MV, IMPEDANCE, LOAD, and LENGTH-MV. These are the attributes with the strongest average effect on W (Section 6.1). Furthermore, interactions that provide a plausible model for expansion costs according to the considerations in Section 6.1 are included. For example, the mean impedance (IMPEDANCE) is for some models combined with the number of branches NLINES-MV and NLINES-LV. Also the combination used in dena (2012) is further considered. Furthermore, the combination of all 14 attributes is included in the analysis. An overview over all 57 selected cluster models and the labels used to refer to them is given in Figure 35 (Appendix).

6.3 Evaluation of cluster models for worst-case scenario

A heuristic method to estimate total costs consists of drawing a random sample of networks with size K, computing expansion costs for these networks, and aggregating these costs by assuming that each of the networks in the sample is representative for $\frac{1}{K}$ of all networks. This heuristic method is in the following used as to derive a benchmark for the relative deviation of estimated costs from cluster models R (Section 5.3).

In order to compute this benchmark, for each value of K 10,000 random samples of size K are drawn. For each of these samples, the total costs are estimated by a simple aggregation in which each network in the sample is assigned equal weight. From these 10,000 estimates, the mean value of the relative deviation R is shown in Figure 20. Furthermore, the 25 percentile and the 75 percentile of the relative deviations are shown.

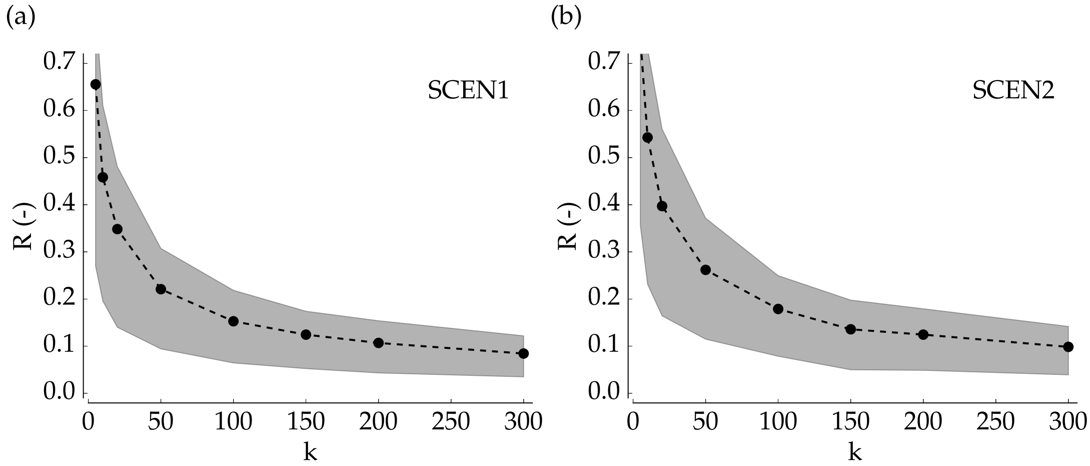


Figure 20: Relative deviation of estimated total costs from total costs (R in Equation 5.12) for a random selection of K networks. Black line indicates arithmetic mean from 10,000 random samples of size K, grey shaded area ranges from 25 percentile to 75 percentile. (a) Worst-case scenario (SCEN1) and (b) curtailment scenario (SCEN2).

This benchmark is used to filter the 57 selected cluster models. Each cluster model is only included in the following analysis if for all K it performs better in terms of the relative deviation then the mean relative deviation of the random samples. This reduces the selected 57 cluster models to 32 cluster models.

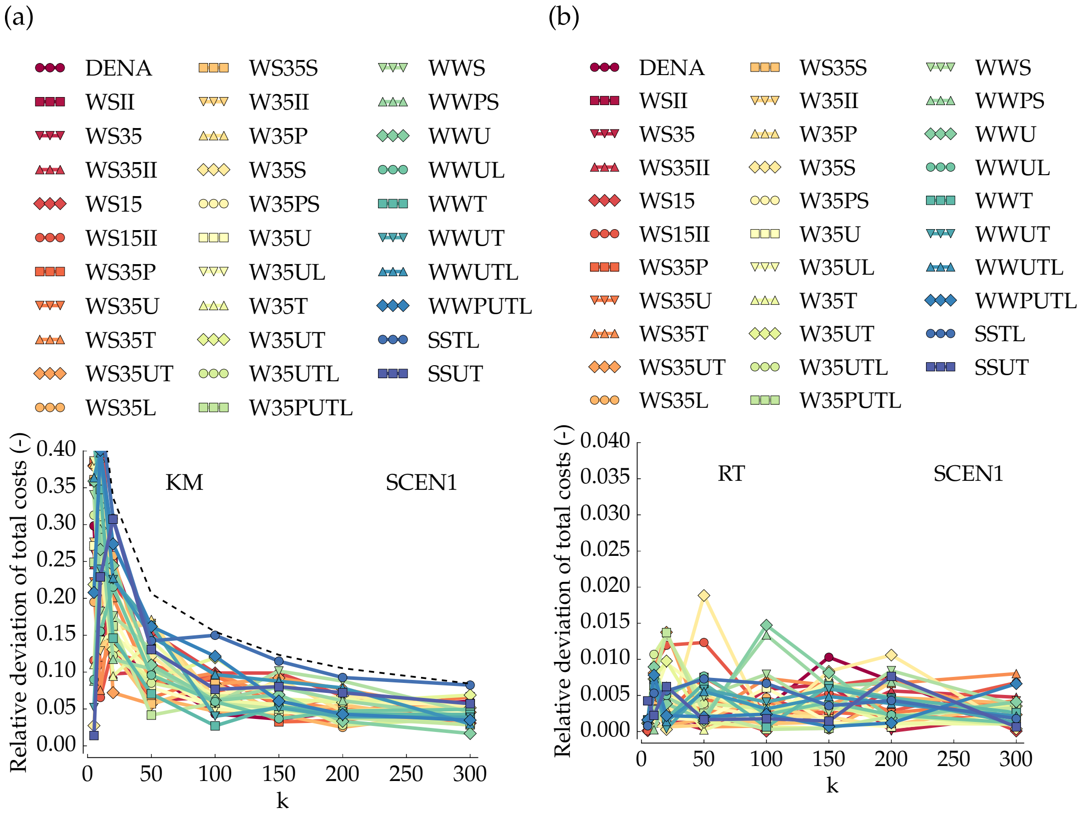


Figure 21: Relative deviation of estimated total costs from total costs for 32 cluster models and different number of clusters K for the worst-case scenario and (a) K-Mean and (b) regression tree estimation. The black dashed line indicates the mean relative deviation for random samples of size K (see Figure 20).

For each of the remaining 32 cluster models, the relative deviation R is computed for $K = 5, 10, 20, 50, 100, 150, 200, 300$ (Figure 21). The results reveal that for K-Mean estimation, the relative deviation of estimated total costs from calculated total costs tends to decrease as K increases. For regression tree estimation, the relative deviations are generally lower by about one order of magnitude than for K-Mean estimation. This can be explained as follows. For K-Mean estimation, the network closest to the geometric centre of the cluster is considered as representative. For regression tree estimation, the network with costs closest to the mean value of the costs of all networks in the cluster is considered representative (Section 5.1 and Section 5.2). This implies that the total costs of a cluster are estimated by multiplying a value close to the mean value of the costs of the cluster by the number of networks in the cluster. If the mean value of the costs coincides with one of the networks of the cluster, the estimate of total costs will therefore coincide with the calculated total costs. Most of the difference between K-Mean and regression trees shown in Figure 21 can therefore be attributed to the method by which the representative networks are selected.

For regression tree estimation, the relative deviation is relatively constant with K. In contrast to the results for K-Mean estimation, there is no clear decreasing trend of the relative deviations as K increases. Two mechanisms with opposite effect may affect the results here. First, a larger value of K means that clusters tends to consist of fewer networks. This means that the probability that the costs of the network with costs closest to the mean value of costs and the actual mean value coincide decreases. At the same time, the larger number of clusters means that costs within one cluster tend to become more similar because the tree can partition the sample into more segments based on differences in costs.

The changes of W with K for the remaining 32 cluster models are shown in Figure 22. For all models, W declines with K. The slope of the curve is generally steeper for smaller values of K. Furthermore, the level of W of a model is generally lower for regression trees than for K-Mean clustering. In other words, observations of the same cluster of a model that was estimated as a regression tree tend to be more similar with respect to costs than observations of the same cluster of that model estimated with K-Mean.

Based on the assessment of W four models stand out because of their relatively large value of W as compared to all other cluster models. Furthermore, the value of W of these models is relatively large for both K-Mean and regression tree estimation (Figure 22). These are the models that account for the installed generation capacity of solar photovoltaic but not of onshore wind power plants (SS and S35II). Because of their substantially larger value of W, these four models are also excluded in the following, which results in 28 remaining cluster models.

6.4 Evaluation of cluster models for curtailment scenario

One of the main advantages of estimating a cluster model as compared to the estimation of a prediction model is that once the representative observations have been identified, they can be used for any future computation and estimation. The total costs can then be

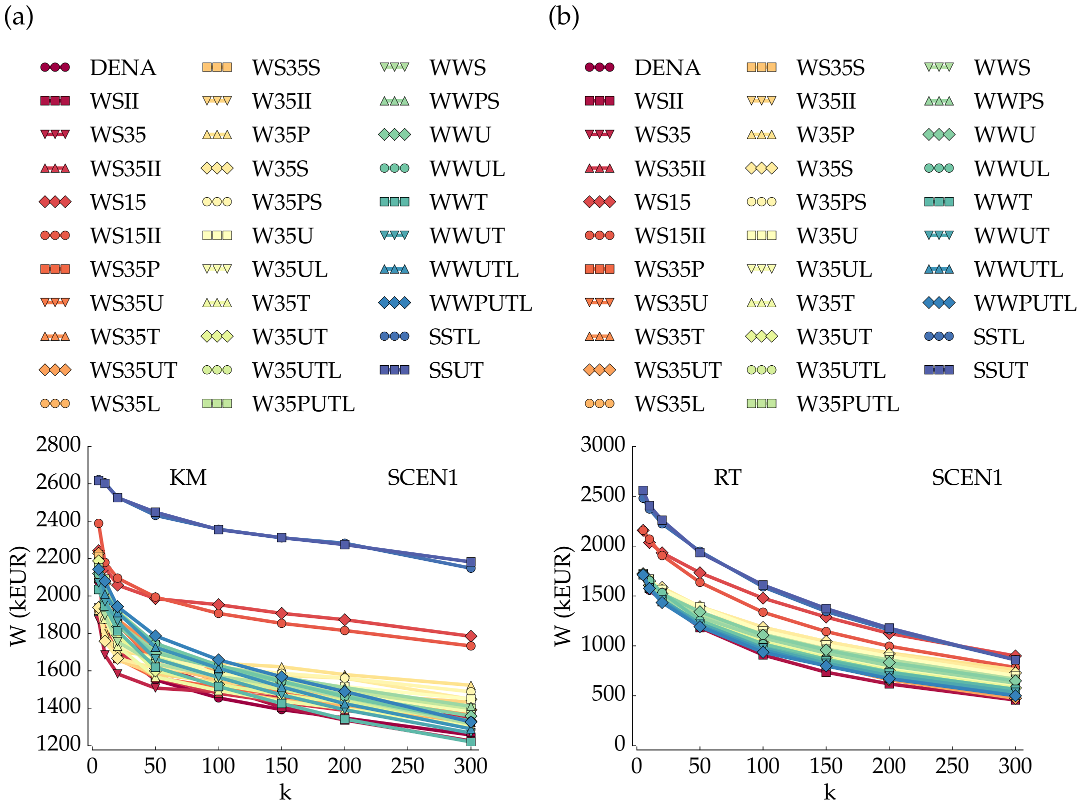


Figure 22: Within-cluster dispersion of costs for 32 cluster models and different number of clusters K for the worst-case scenario and (a) K-Mean and (b) regression tree estimation.

estimated with a simple multiplication of the costs with the size of the corresponding cluster. This means that no additional model on the relationship between attributes and costs of network expansion needs to be estimated. In the context of this thesis, for example, one can use the representative networks to estimate total costs of network expansion for scenarios with different flexibility options. This can be achieved, for example, by applying a power flow model to each of the representative networks. With a prediction model, one can likewise apply the power flow model to a sample of networks. However, on would then need to estimate a new model for the association between network attributes and expansion costs for each of the scenarios.

The results in Section 6.3 indicate that an estimation of clusters using regression trees can yield a better estimate of total costs than an estimation using K-Mean. The clusters of the regression tree estimation were however identified using costs of the worst-case scenario. For this reason, the performance of cluster models estimated with regression trees may be more specific to this scenario than the performance of cluster models estimated with K-Mean. In order to test the robustness of the relative performance of cluster models estimated with these two alternative methods, the cluster models are applied to the second scenario of costs of network expansion, the curtailment scenario (Section 3.4). For this step, the representative networks are the same as estimated for the worst-case scenario but expansion costs are replaced by the costs computed for the curtailment scenario.

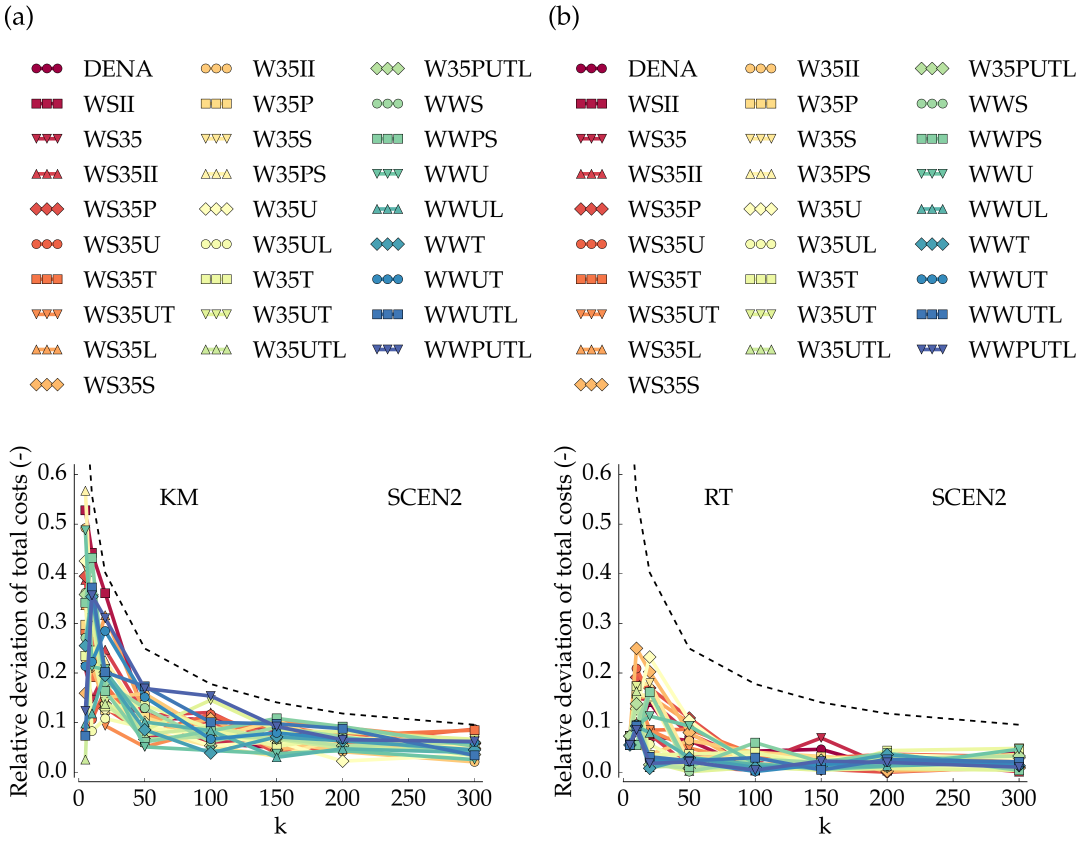


Figure 23: Relative deviation of estimated total costs from total costs for 28 cluster models and different number of clusters K for the curtailment scenario and (a) K-Mean and (b) regression tree estimation. The black dashed line indicates the mean relative deviation for random samples of size K (see Figure 20).

The results for the curtailment scenario are shown in Figure 24. The results are qualitatively similar to the results for the worst-case scenario (Figure 22). The costs are generally lower for the curtailment scenario, which is reflected in the generally lower value of W. As for the worst-case scenario, the regression tree estimation yields cluster models with a lower value of W than the K-Mean estimation.

The relative deviation of the estimated total costs from the calculated total costs R is shown in Figure 23. All cluster models yield a similar relative deviation for K-Mean estimation as for regression tree estimation. The cluster models estimated using regression trees still perform slightly better with respect to R than the cluster models estimated using K-Mean but the differences between the two methods are much smaller than for the worst-case scenario. In general, the relative deviation of cluster models estimated using K-Mean is similar for both scenarios (Figure 21 and Figure 23). In contrast to this, the relative deviation of cluster models estimated using regression trees is much larger for the curtailment scenario than for the worst-case scenario. This supports the hypothesis that the cluster models estimated using regression trees are generally more specific to the worst-case scenario than the cluster models estimated using K-Mean.

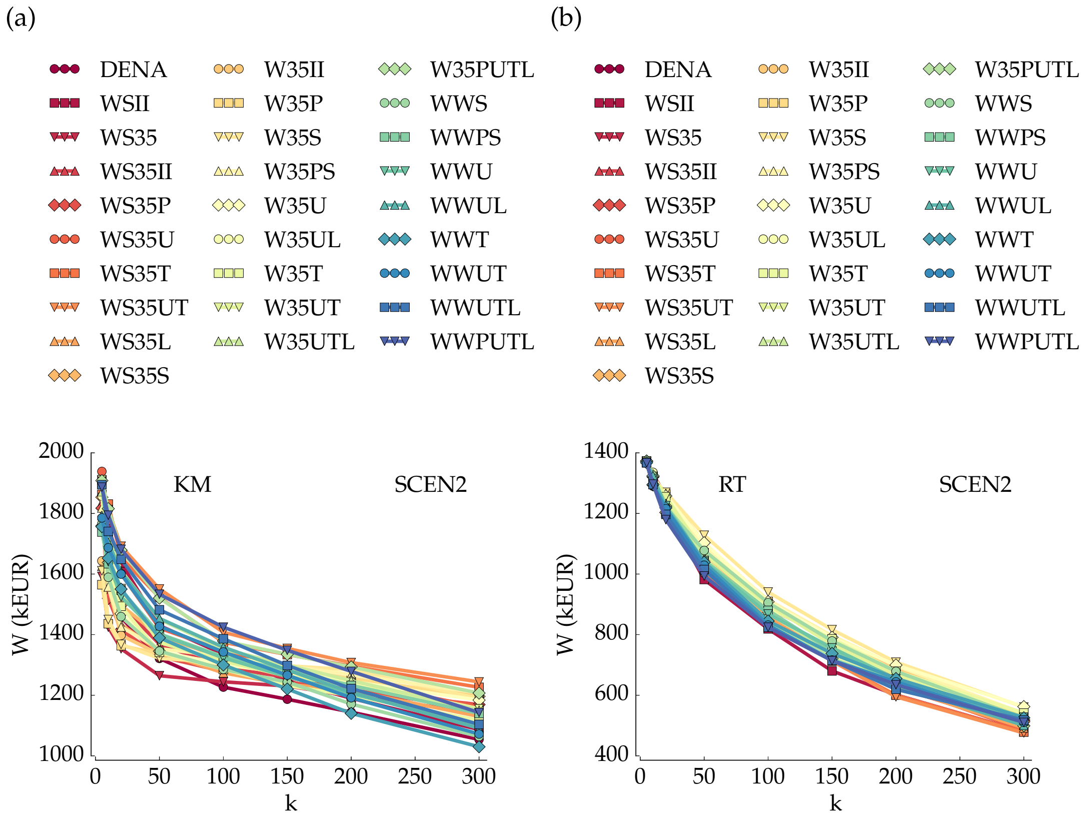


Figure 24: Within-cluster dispersion of costs for 28 cluster models and different number of clusters K for the curtailment scenario and (a) K-Mean and (b) regression tree estimation.

6.5 Evaluation of selected cluster models

The results in Section 6.3 and Section 6.4 reveal that cluster models estimated with K-Mean can be better compared in terms of W than in terms of R. The reasons is that W is not sensitive to the choice of the representative networks.

The cluster models selected for the analysis in this Section are hence assessed only in terms of W. Furthermore, only the results from K-Mean estimation are examined here. One reason for this is that the performance of cluster models generally features a stronger variation for K-Mean than for regression tree estimation. Furthermore, previous studies relied on K-Mean as cluster estimation method which provides an additional motivation to keep and attempt to improve the same estimation method.

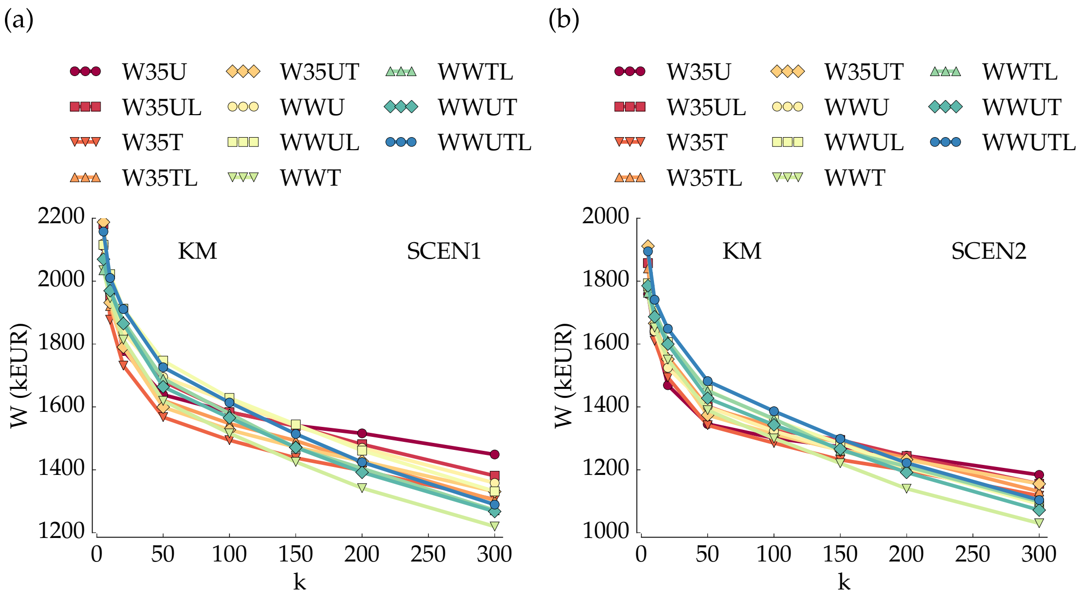


Figure 25: Within-cluster dispersion of costs for twelve selected cluster models and different number of clusters K using K-Mean cluster estimation for the (a) worst-case scenario and (b) curtailment scenario.

The values of W for selected cluster models and the two scenarios SCEN1 and SCEN2 are shown in Figure 25. There are several new insights that can be derived from this comparison of cluster models. First, for $K > 100$ the cluster model with the lowest within-cluster dispersion of costs W is WWT for both scenarios of network expansion costs. This cluster model includes four network attributes: WIND-2035, WIND-2015, IMAX-MV, and NLINES-MV. Second, this cluster model has a lower W than alternative cluster models that include additional attributes (WWTL, WWUT, WWUTL). Third, for $K \leq 100$ other cluster models that include WIND-2035 but not WIND-2015 have a lower value of W (e.g. W35T). This is the case for both scenarios.

The sensitivity of the relative performance of cluster models to the value of K is further examined in Figure 26. There, also cluster models that include both installed generation capacities of wind and photovoltaic are included. As before, WWT has the lowest value of W for large K, in this case for $K > 150$. However, for $K < 200$ cluster models that include both WIND-2035 and SOLAR-2035 perform better. Among these models is the model DENA that includes WIND-2015, WIND-2035, SOLAR-2015 and SOLAR-2035 (Figure 26).

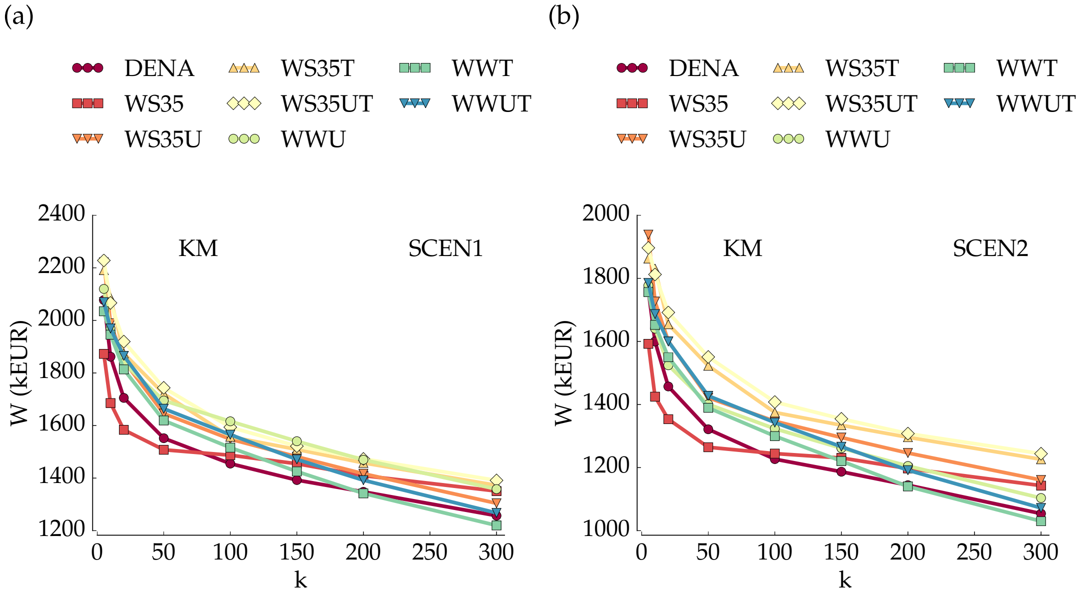


Figure 26: Within-cluster dispersion of costs for eight selected cluster models and different number of clusters K using K-Mean cluster estimation for the (a) worst-case scenario and (b) curtailment scenario.

In sum, for relatively low values of K ($K \leq 100$) cluster models including only the installed generation capacity in 2035 of both onshore wind and solar photovoltaic power plants perform best. For relatively large values of K, cluster models that include WIND-2015 and WIND-2035 tend to perform better. This is especially the case for the cluster model WWT which includes WIND-2015, WIND-2035, IMAX-MV and NLINES-MV. Furthermore, the model DENA is among those models performing best, especially for relatively low values of K.

6.6 Detailed analysis of two cluster models

In Section 6.5 the cluster model WWT performed relatively well with respect to both metrics W and R for several values of K. This cluster model is therefore analysed in more detail in the following. To this aim, the cluster model WWT with $K = 10$ is applied to the worst-case scenario and the ten clusters are analysed based on their representative network. Furthermore, because the cluster model DENA has been used in previous studies and also tends to perform relatively well according to the results of the previous Sections, it is also applied to the worst-case scenario and the resulting clusters are also analysed.

The representative networks of the cluster model DENA estimated with K-Mean and with a regression tree are shown in Table 4 and Table 5, respectively. For each representative network, the coordinates in terms of the attributes of the cluster model DENA are given. Furthermore, the additional future installed capacity for onshore wind (Δ WIND) and solar photovoltaic (Δ SOLAR) are calculated and shown. In addition, the costs of network expansion and the size of the clusters are given.

Table 4: Representative networks of the cluster model DENA estimated with K-Mean and K=10. ΔWIND = WIND-2035 - WIND-2015, ΔSOLAR = SOLAR-2035 - SOLAR-2015. n_k refers to the size of the corresponding cluster. Costs refer to the costs of network expansion for the worst-case scenario. See Table 3 for units of the attributes. $W = 1870$, $R = 0.15$.

k	WIND -2015	WIND -2035	SOLAR -2015	SOLAR -2035	Δ WIND	Δ SOLAR	Costs	n_k
1	1502	2748	8428	14158	1246	5730	529	656
2	19251	58453	26115	40739	39202	14624	5151	102
3	2617	5217	38570	50094	2600	11524	465	162
4	4	8	83797	106009	4	22212	4788	33
5	32696	68669	9601	14019	35973	4418	18103	130
6	62396	122351	86935	97299	59955	10364	15474	10
7	1809	3618	20450	26676	1809	6226	423	386
8	12000	30000	6206	11408	18000	5202	788	314
9	62700	123687	13335	19792	60987	6457	12968	51
10	750	1282	2141	2969	532	828	172	1084

For K-Mean, there are three clusters for which the costs of the representative network are larger than 10,000,000 EUR ($k = 5, 6, 9$). They represent together $6.5\,\%$ of all networks. The three representative networks all feature ΔWIND > 35 MW. If one compares these and the other representative networks in more detail, Table 4 also reveals a potential weakness of the cluster model DENA. The representative network of the cluster $k = 2$ features a larger ΔWIND and a larger ΔSOLAR than the representative network of the cluster $k = 5$. At the same time, its costs of network expansion are lower. This indicates that there are other network attributes which have a relatively strong influence on costs but are not taken into account by the cluster model DENA.

If one compares the results for K-Mean (Table 4) and regression tree estimation (Table 5), one can see that the regression tree identifies two relatively small clusters ($k = 1, 2$) whose representative networks feature relatively large costs. These two representative networks also feature relatively large ΔWIND. At the same time, two relatively large clusters ($k = 5, 6$) feature relatively low costs and their representative networks feature no additional wind power plants (ΔWIND = 0). These representative networks are more extreme with respect to costs and ΔWIND than the most extreme representative networks for K-Mean estimation. The identification and clustering of networks with relatively extreme costs can be considered one of the strengths of regression tree estimation as compared to K-Mean. Overall, the regression tree estimation of this example yields a lower within-cluster dispersion of costs W and a lower relative deviation of estimated total costs R than the K-Mean estimation (see caption of Table 4 and 5).

Table 5: Representative networks of the cluster model DENA estimated with a regression tree and K=10. ΔWIND = WIND-2035 - WIND-2015, ΔSOLAR = SOLAR-2035 - SOLAR-2015. n_k refers to the size of the corresponding cluster. Costs refer to the costs of network expansion for the worst-case scenario. See Table 3 for units of the attributes. $W = 1557$, $R = 0.0067$.

k	WIND -2015	WIND -2035	SOLAR -2015	SOLAR -2035	Δ WIND	Δ SOLAR	Costs	n_k
1	4500	124698	38449	54535	120198	16086	40091	2
2	67750	184693	24121	28197	116943	4076	20849	21
3	0	43200	4731	33551	43200	28820	11556	35
4	17100	49266	4903	8627	32166	3724	7127	181
5	0	0	7629	9738	0	2109	256	1540
6	0	0	20093	37939	0	17846	1195	695
7	3000	26500	7727	10630	23500	2903	3161	286
8	8100	14850	18450	32838	6750	14388	5460	102
9	19425	88320	14228	17805	68895	3577	16883	27
10	42250	94500	18931	27879	52250	8948	11007	39

Figure 27 shows the geographic distribution of the ten clusters from the K-Mean and the regression tree cluster estimation. There are no clear spatial patterns that can be recognised. Overall, few clusters dominate because of their larger cluster size, as also shown in Table and Table .

The representative networks of the model WWT estimated with K-Mean and their coordinates are shown in Table 6. From these coordinates, also the additional future installed capacity for onshore wind (ΔWIND) is calculated and shown. The result

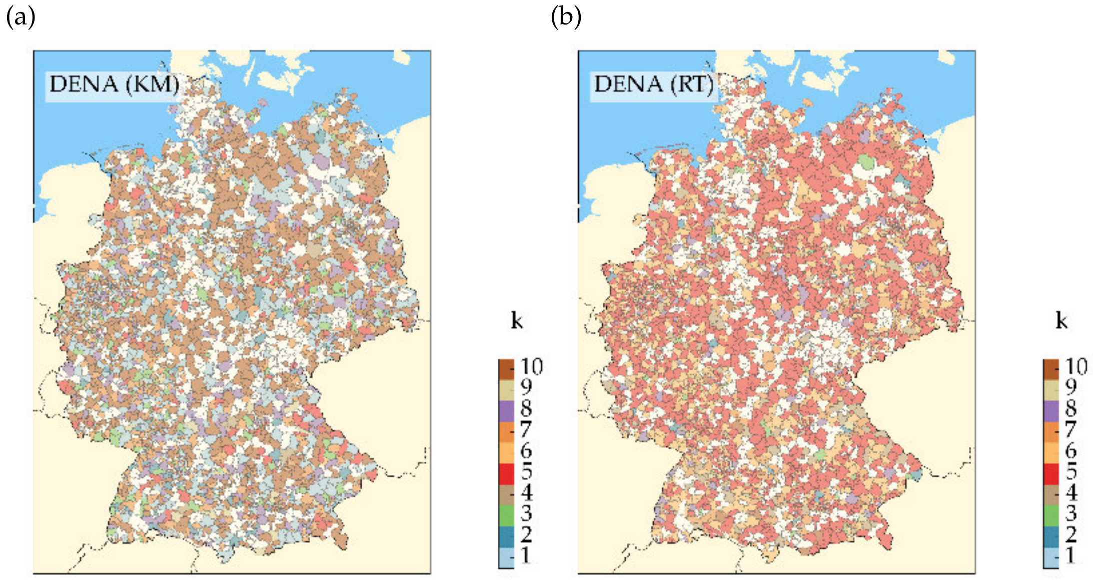


Figure 27: Geographical map of the distribution of clusters for the cluster model DENA with $K = 10$ using (a) K-Mean and (b) regression tree estimation.

illustrates one reason why the WWT cluster model tends to perform better than most cluster models that do not include network attributes describing thermal and voltage limits (Section 6.3). For example, the representative network of the cluster $k = 7$ features the largest ΔWIND of all representative networks. Its cost of network expansion are however smaller than the costs of four of the other representative networks ($k = 2, 3, 8, 9$). The attribute IMAX-MV indicates one potential reason for this as the representative network of the cluster $k = 7$ features a larger IMAX-MV than each of the four other representative networks.

Table 6: Representative networks of the cluster model WWT estimated with K-Mean and K=10. ΔWIND = WIND-2035 - WIND-2015. n_k refers to the size of the corresponding cluster. Costs refer to the costs of network expansion for the worst-case scenario. See Table 3 for units of the attributes. $W = 1957$, $R = 0.32$.

k	WIND -2015	WIND -2035	IMAX -MV	NLINES -MV	ΔWIND	Costs	n_k
1	1850	3700	382	9	1850	523	498
2	39400	86824	452	10	47424	10646	76
3	16800	36000	429	8	19200	5123	198
4	800	800	371	4	0	33	1135
5	2000	4000	518	5	2000	761	307
6	1300	1950	257	8	650	795	372
7	85304	161396	461	15	76092	4754	16
8	22960	65950	375	15	42989	5571	120
9	43380	83032	387	29	39652	8060	29
10	850	4250	342	15	3400	1682	177

The representative networks of the cluster model WWT estimated with a regression tree are shown in Table 7. As expected from Section 6.3, the cluster model WWT performs bettern in terms of both the within-cluster dispersion of costs W and the relative deviation of estimated total costs R it if is estimated with a regression tree than if it is estimated with K-Mean (see caption of Table 6 and 7). Furthermore, similar to the results for the model DENA the regression tree identifies two relatively small clusters ($k = 2, 5$) with relatively large costs and one relatively large cluster ($k = 7$) with relatively low costs. These clusters are more extreme than the most extreme clusters of the estimation with K-Mean in terms of costs (Table 6 and Table 7). This difference between K-Mean estimation and regression tree estimation is also similar to the result for the cluster model DENA discussed above.

Table 7: Representative networks of the cluster model WWT estimated with a regression tree and K=10. ΔWIND = WIND-2035 - WIND-2015. n_k refers to the size of the corresponding cluster. Costs refer to the costs of network expansion for the worst-case scenario. See Table 3 for units of the attributes. $W = 1580$, $R = 0.0063$.

k	WIND -2015	WIND -2035	IMAX -MV	NLINES -MV	ΔWIND	Costs	n_k
1	7870	40735	210	8	32865	3747	388
2	4500	124698	389	19	120198	40091	2
3	67750	184693	444	11	116943	20849	21
4	67140	125905	438	14	58765	12937	64
5	35620	120602	608	13	84982	42535	2
6	17100	49266	267	12	32166	7127	181
7	1025	1025	274	15	0	340	1865
8	6100	12199	422	7	6099	1576	370
9	850	54604	357	4	53754	8841	24
10	4100	44348	421	6	40248	16013	11

Figure 28 shows the geographic distribution of the ten clusters of the cluster model WWT. Similar to the results of the model DENA shown in Figure 27, the geographical distribution of the clusters does not follow any clear spatial patterns. For the results from both estimation methods, few relatively large clusters dominate. Furthermore, as illustration of the use of regression trees as cluster estimation method and to further visualise differences between the cluster models DENA and WWT, the regression trees of the two cluster models are shown in Figure 33 (Appendix) and Figure 34 (Appendix), respectively.

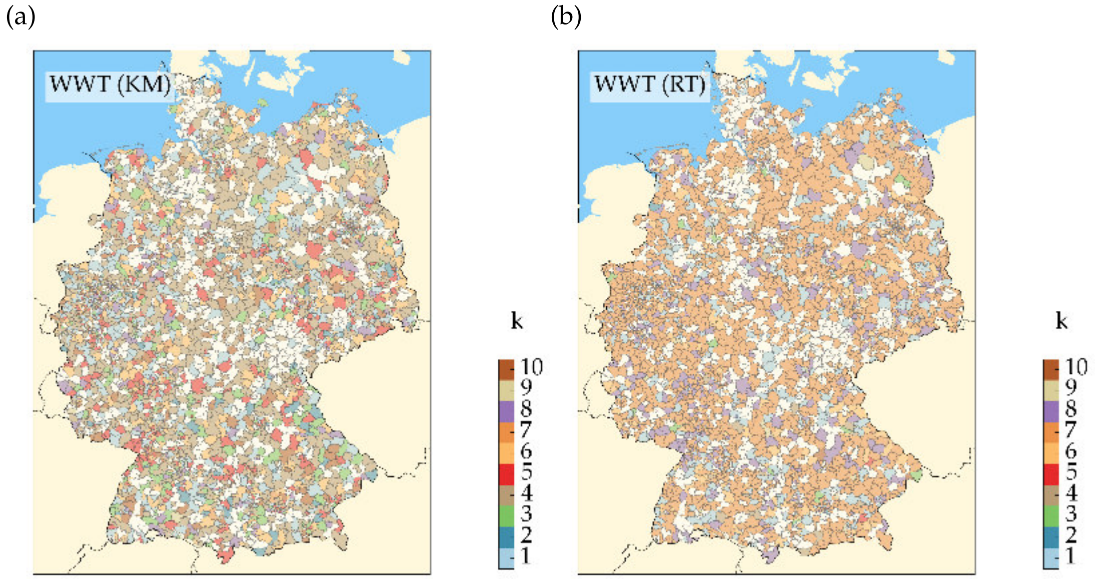


Figure 28: Geographical map of the distribution of clusters for the cluster model WWT with $K = 10$ estimated with (a) K-Mean and (b) a regression tree.

7. Discussion

In this Chapter, the results shown in the previous Chapters are discussed. Furthermore, some of the strengths and weaknesses of the dataset used in this thesis are addressed. In order to examine and discuss the quality of the dataset of synthetic networks, first the total costs of network expansion of the synthetic networks are compared with the total costs that were estimated in previous studies (Section 7.1). Then, the selected 14 network attributes and their relative importance are discussed in light of previous studies (Section 7.2). Next, the relative performance of the K-Mean and the regression tree method is discussed (Section 7.3). Finally, the performance of the cluster model used by previous studies relative to the performance of alternative cluster models is discussed (Section 7.4).

7.1 Data on distribution networks

The costs for expanding distribution networks were estimated from a dataset of synthetic networks. This dataset was constructed as part of an ongoing research project at the Reiner-Lemoine-Institut Berlin and used in this thesis because there is no complete dataset of real electricity distribution networks in Germany. Previously, the costs of network expansion were estimated from samples of real distribution networks (dena, 2012; BMWi, 2014). While the synthetic networks allowed the author to examine alternative cluster models and cluster methods in more detail than samples of real networks would have allowed because of the much larger size of the dataset, the use of a synthetic dataset introduces some uncertainty about the accuracy of the costs of network expansion.

One way to assess the accuracy of these costs is to compare the total costs with the total costs that were estimated in previous studies. For this thesis the costs of network expansion were calculated for two scenarios, one worst-case scenario and one scenario with curtailment. Table 8 shows the total costs for these two scenarios and the total costs estimated in dena (2012) and in BMWi (2014). Because these previous studies assumed different future installed capacities of onshore wind and solar photovoltaic power plants, the total costs are divided by the total installed capacity of these two technologies. Furthermore, the share of the total costs on the MV level of the total costs at both the MV and the LV level is calculated (Table 8).

Table 8: Total costs of network expansion of worst-case scenario (SCEN1) and curtailment scenario (SCEN2) and of two previous studies: BMWi (BMWi, 2014) and DENA (dena, 2012). Additional installed generation capacity of onshore wind (Δ W), solar photovoltaic (Δ PV), and other renewable energy technologies (Δ EE). Total costs denoted as TC and divided by sum of installed generation capacity of onshore wind and solar photovoltaic (TC per GW). Share of total costs on medium-voltage level of total costs on both medium-voltage and low-voltage level (Share MV).

Source	Year 1	Year 2	Δ W [GW]	Δ PV [GW]	Δ EE [GW]	TC [G EUR]	TC per GW	Share MV
DENA	2010	2030	34.3	44.9	3	11.4	0.14	0.68
BMWi	2012	2032	34.3	31.7	0	15.5	0.23	0.64
SCEN1	2015	2035	47.5	21.4	0	7.9	0.11	0.78
SCEN2	2015	2035	47.5	21.4	0	5.9	0.09	0.87

The scenarios of this thesis are based on installed capacities of wind power and photovoltaic in 2035 that differ greatly from previous studies (Table 8). This can partly be attributed to differences in the year of publication and the years used as reference and for the future. For example, dena (2012) projected a larger future expansion of solar photovoltaic than wind power, whereas the scenario used on this thesis expects a larger future expansion of wind power (Table 8).

In previous studies, the costs per additional Watt of installed generation capacity of solar photovoltaic and onshore wind power plants were about 0.14 EUR / W (dena, 2012) and 0.23 EUR / W (BMWi, 2014). In the scenarios of this thesis, the costs are estimated at 0.11 EUR / W (worst-case scenario) and 0.09 EUR / W (curtailment scenario). The worst-case scenario is more similar to the scenarios examined in the two previous studies than the curtailment scenario. Furthermore, the methodology of this thesis is more similar to dena (2012) than to BMWi (2014). For these two reasons, in the following the deviations of the worst-case scenario from dena (2012) are discussed.

There are several explanations why the total costs per Watt of installed generation capacity in this thesis might be lower than the total costs per Watt in dena (2012). For example, the scenario of dena (2012) is based on the assumption that solar photovoltaic expands more strongly than onshore wind. If the costs of network expansion tend to be higher for one additional Watt of solar photovoltaic installed capacity than for one additional Watt of onshore wind installed capacity, this can explain why the costs are lower in the worst-case scenario. One possible explanation is that most photovoltaic power plants are connected to networks on the low-voltage level. This means that one additional Watt of installed capacity can result in a demand for network expansion on both the low-voltage and the medium-voltage level if it causes an exceedance of thermal limits on both levels.

Furthermore, the total costs per Watt may be lower in the scenarios of this thesis than in dena (2012) because some real networks are only partly represented in the dataset of synthetic networks. This concerns network districts with aggregated load areas. These load areas do not contain any network on the low-voltage network. They represent urban areas and it is therefore assumed that the costs of network expansion would be relatively low there (Section 3.1). However, it is not clear how realistic this assumption is. Table 8 shows that the networks expansion on the low-voltage level has a generally lower share for the scenarios of this thesis than in dena (2012). This can partly be attributed to the larger expansion of photovoltaic in dena (2012) than in the scenarios of this thesis. However, it may also indicate that the costs of expanding networks on the low-voltage levels are underestimated in this thesis because aggregated load areas do not contain low-voltage networks.

The focus of this thesis is on associations between network attributes and the costs of network expansion. The associations are discussed in Chapter 4. They are then used to define the cluster models at the end of Section 6.2. Overall, the results in Chapter 6 support some of the conjectured associations more strongly than others. Even if the total costs of network expansion agree well with previous results based on samples of real networks, the associations between network attributes and expansion costs that are reflected in the results of this thesis are not necessarily realistic. These associations have not been analysed before. Also dena (2012) and BMWi (2014) do not examine these associations. The accuracy of the associations can therefore only be discussed qualitatively based on the strengths and weaknesses of the dataset that includes both network attributes and costs of network expansion. Some of the weaknesses of the dataset of synthetic networks are discussed in the following.

The synthetic networks were built from open-access geographical information about demand and supply of electricity in Germany. This means that systematic errors in the geographical information also affect the representativeness of the distribution networks (Hülk et al., 2017).

Furthermore, the networks were generated with an algorithm based on principles of distribution network operation. Many of these principles, such as the n-1 criterion, were taken from laws and regulation (Amme et al., 2017). However, not all options to implement these principles are considered by the algorithm that constructs the synthetic networks. For example, the n-1 criterion can sometimes be met by interconnections between neighbouring network districts on the MV level in case of an equipment failure. This option was not considered in the construction of the dataset.

The algorithm constructs the distribution networks based on geographical data on electricity supply and demand. This means that some additional factors that lead to variation among distribution networks in Germany are not taken into account. Examples are the historical evolution of typical network topologies or regional differences in how principles of network planning and development are implemented. These differences are not represented in the dataset.

Finally, the methods for network expansion were derived based on published literature (Section 3.4). This literature takes laws and regulation in Germany into account. However, one must assume that there are more options and criteria considered in practice than by the algorithm. If these options are correlated with network attributes, the actual costs of network expansion may systematically deviate from the costs of the dataset.

7.2 Network attributes

In this thesis, a list of possible network attributes was derived from theoretical considerations. For each of these attributes, the average effect in terms of the change in performance of a cluster model that excludes and then includes that attribute was examined. Based on the results, the most important network attributes were derived. The top six network attributes are: WIND-2035, WIND-2015, IMAX-MV, SOLAR-2035, LENGTH-MV, and SOLAR-2015 (Section 6.1).

To the knowledge of the author this is the first examination which network attributes should be included in a cluster analysis in order to identify electricity networks that are representative for an estimation of total costs of distribution network expansion. The work that comes closest to this thesis is dena (2012). In that study, each network district is first categorised as urban or rural based on population density. These categories are distinguished because the sample of real networks is not representative with respect to population density. Then, within each category a cluster model with four attributes is applied: WIND-2035, WIND-2015, SOLAR-2035, and SOLAR-2035. The relative importance of these or alternative attributes is however not addressed.

In some previous publications, authors have proposed attributes that can be used to characterise and distinguish electricity distribution networks. These authors did however not account for the association between network attributes and expansion costs. For example, Kerber (2010) classified low-voltage networks in Germany based on population density of their local administrative district. In his classification, network districts are classified as either urban, semi-urban or rural. Based on a sample of real low-voltage networks, Kerber (2010) shows how attributes such as the capacity of the transformer station or the distance between two consumers differ systematically between these three categories.

Gust (2014) examined which network attributes can be used to distinguish distribution networks in a sample of real networks from Switzerland. These attributes are however specific to low-voltage networks and, for example, do not include installed capacity of wind power plants. Furthermore, the examination is not specific to the estimation of expansion costs. In consequence, many of the attributes in Gust (2014) can either not be computed from the synthetic networks or do not fit into the theoretical framework developed in Chapter 4. Nevertheless, some of the attributes resemble attributes used in this thesis. For example, Gust (2014) includes the capacity of transformer stations, peak load, installed capacity of solar photovoltaic plants, impedance of lines and cables and their thermal capacity.

Walker et al. (2014) identify representative low-voltage networks for the area of one distribution network operator in Germany. To this aim, they conduct a cluster analysis. Before the cluster analysis, Walker et al. (2014) apply a factor analysis to identify which attributes are relatively strongly correlated. Finally they propose five attributes for the cluster analysis: total length of lines and cables on the LV level, the capacity of the transformer station, the impedance, and the installed capacity of solar photovoltaic plants. Furthermore, they propose the average age of the population as a socio-demographic variable. These attributes resemble attributes used in this thesis. However, the analysis of is also not Walker et al. (2014) specific to the estimation of expansion costs.

Before a cluster analysis one can also conduct a principal component analysis based on potential attributes and then include some of these components in the cluster model (Hastie et al., 2009). In this thesis, it was decided to include network attributes as they are defined in Chapter 4. The main reason is that the results of the cluster models in Chapter 6 are easier to interpret for these original attributes. Furthermore, the results are easier to reproduce and the cluster models are easier to adopt in future applications.

The development of a theoretical framework in Chapter 4 had the consequence that finally only 14 network attributes were examined in Chapter 6. In general, a much larger number of attributes could have been examined. For example, one could not only include the mean value of the impedance of all network paths to terminal nodes but also its maximum and minimum value and potentially some additional percentiles. Furthermore, one could include population density and the area of a network district. Especially the regression tree estimation would allow for a larger number of network attribute since the tree simply ignores less important attributes. However, the theoretical framework and the focus on attributes that are included in that framework allowed the author to interpret the cluster models based on the theoretical considerations in Chapter 4. This was done, for example, in the more detailed examination of alternative cluster models in Section 6.5.

For the examination of further scenarios of network expansion, the theoretical framework can integrate additional network attributes. For example, if options for storing energy were included in the networks and in the power flow simulation, one could add an attribute describing the use or the capacity of these storages in a network district as an attribute determining the supply of hosting capacity.

7.3 Cluster estimation methods

In this thesis two alternative cluster estimation methods were considered and compared, K-Mean and regression trees (Chapter 6). For these two methods, the error from aggregation was quantified. This was made possible by the dataset of synthetic networks. Figure 29 shows the average relative deviation of estimated total costs from calculated total costs for both K-Mean and regression tree estimation. For both costs scenarios, the relative deviation is on average lower for regression tree than for K-Mean estimation.

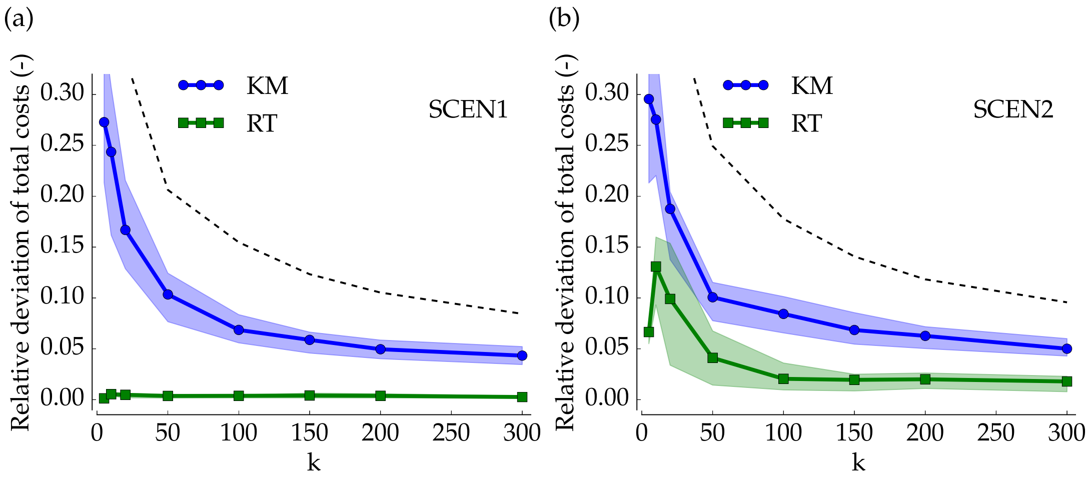


Figure 29: Relative deviation of estimated total costs from total costs (R in Equation 5.12) for different number of clusters K and for the (a) worst-case scenario and (b) curtailment scenario. Lines indicate arithmetic mean for the 28 cluster models in Section 6.4, shaded area ranges from 25 percentile to 75 percentile of these models. The black dashed line indicates the mean relative deviation for random samples of size K (see Figure 20). K-Mean (KM) and regression tree estimation (RT).

The regression trees were fitted to the costs of network expansion of the worst-case scenario. For this scenario, one would therefore expect that regression trees perform better than K-Mean estimation. Whether the fitted trees perform better for an alternative scenario is determined by the association between the costs of the two scenarios. Figure 31 shows the relationship between the costs of the worst-case scenario and the curtailment scenario. Overall, the costs are relatively similar with a correlation coefficient of 0.97 and a root-mean-squared-error (RMSE) of 1. As expected from the construction of the scenarios (Section 3.4), the costs tend to be lower for the curtailment scenario. Despite the overall good agreement for some networks costs are different by a factor of two or more.

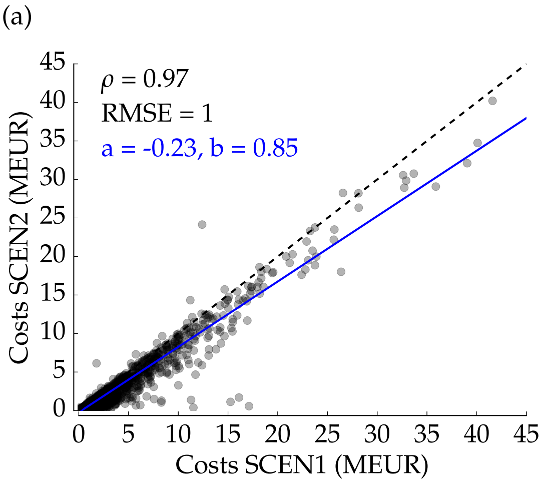


Figure 30: Scatter plot of costs of network expansion of worst-case scenario (SCEN1) and curtailment scenario (SCEN2). Pearson correlation coefficient $\rho = 0.97$. $a = -0.23$ and $b = 0.85$ denote intercept and coefficient of linear regression model, respectively.

The overall good agreement between the costs of the two scenarios can explain why regression trees performed generally better than K-Mean clustering also for the curtailment scenario. For a third scenario, this relative advantage could however change depending on how well the costs of the two scenarios are related. For example, for a scenario that includes battery storage, the costs could be relatively different. Whether regression trees that are fitted to the worst-case scenario still produce clusters that are better suited for an estimation of total costs than K-Mean estimation cannot be inferred from the results of this thesis.

For some applications, it may be desired that clusters of distribution networks have a certain minimum size. The regression trees used in Chapter 6 do not constrain the size of clusters. While some clusters include more than 1000 networks, other clusters consist of only one network (not shown). Figure 31 shows the relative deviation for regression trees of which each cluster consists of at least 10 networks. For the worst-case scenario, this additional constraint increases the relative deviation of estimated costs from actual costs. For the curtailment scenario, however, for $k < 300$ and especially for low values of K, the constraint tends to improve the performance of cluster models.

The results in Figure 31 illustrate how the minimum size of clusters can change the performance of regression tree estimation of clusters. There are further parameters of tree such as the maximum size of a cluster or the depth of the tree which could also be examined. Furthermore, one could examine how pruning of trees affects

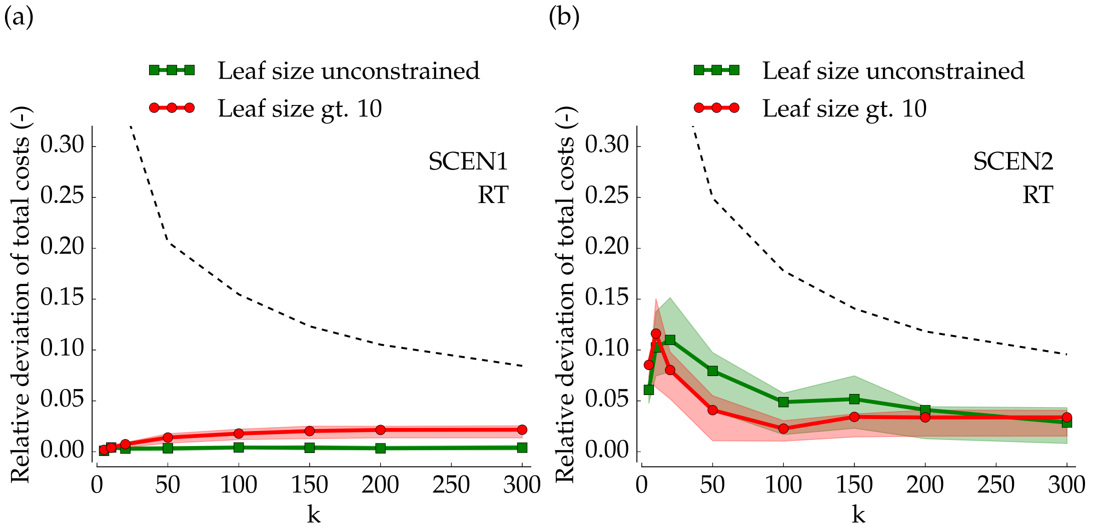


Figure 31: Relative deviation of estimated total costs from total costs (R in Equation 5.12) for different number of clusters K and for the (a) worst-case scenario and (b) curtailment scenario. Lines indicate arithmetic mean for a subset of the 28 cluster models in Section 6.4, shaded area ranges from 25 percentile to 75 percentile of these models. The black dashed line indicates the mean relative deviation for random samples of size K (see Figure 20). Regression tree estimation (RT) with size of leafs unconstrained (green color) and size of leafs $n_k > 10$ (red color).

their performance for the scenario of costs to which they are fitted and for alternative scenarios. Due to time and resource constraints of this thesis, this examination remains for future research.

It is important to note at this point that the association between costs of the scenario to which the trees are fitted and the scenarios to which they are finally applied is generally not known. This is because for the scenarios to which the trees are finally applied costs are only computed for few representative networks. This is the central idea behind the use of clusters to estimate total costs, the performance of which is examined in this thesis. This means that the relative deviation of estimated from calculated total costs for the two scenario shown here may be considered as only one of several criteria for the choice of a cluster estimation method. For a scenario more different from the worst-case scenario than the curtailment scenario, K-Mean may be preferred for example because one may consider it to be a less scenario-specific and therefore more robust cluster estimation method. More insights into the robustness of the two estimation methods could be gained by examining a larger set of possible cost scenarios in the same way as the two scenarios were examined here.

7.4 Cluster models

One of the main objectives of the analysis was to identify those cluster models that yield the lowest within-cluster dispersion of costs and the lowest relative deviation of estimated total costs from calculated total costs. This was also motivated by the fact that all previous studies adopted the cluster model proposed by dena (2012) although the performance of this model had not yet been assessed.

The performance of the cluster model DENA relative to the average performance of the other 27 cluster models selected for the detailed analysis in Section 6.3 and Section 6.4 is shown in Figure 32. For $20 \leq K < 200$ the model DENA performs better than the average of the alternative models. An exception is $K = 50$ for which the model DENA performs worse than the average for regression tree estimation. Overall, the model DENA tends to perform better than the average of the alternative models (Figure 32).

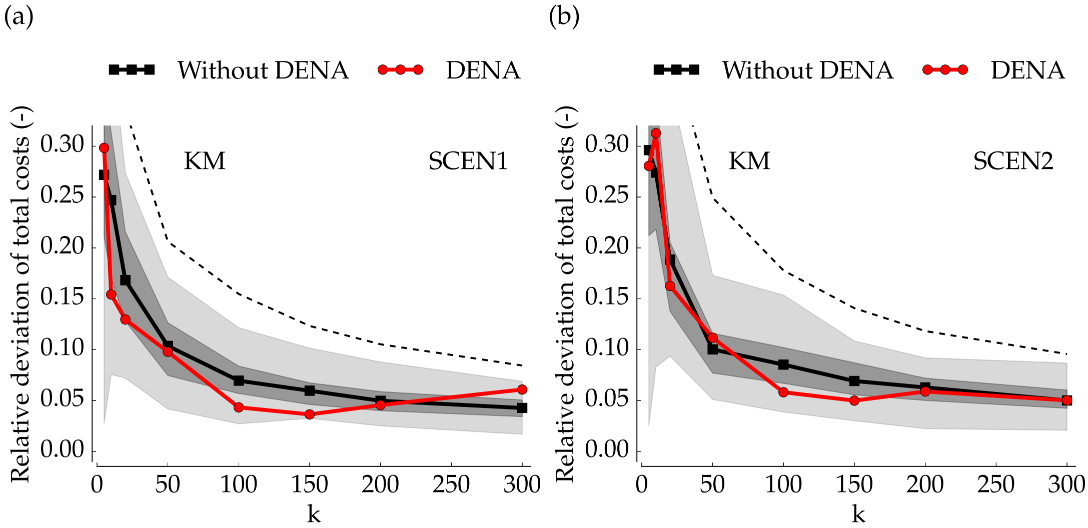


Figure 32: Relative deviation of estimated total costs from total costs (R in Equation 5.12) for different number of clusters K and for the (a) worst-case scenario and (b) curtailment scenario. Lines indicate arithmetic mean for the 27 cluster models in Section 6.4 excluding the model DENA (black colour) and for the cluster model DENA (red colour). Light grey area ranges from minimum to maximum value, dark grey area ranges from 25 percentile to 75 percentile of the 27 models. The black dashed line indicates the mean relative deviation for random samples of size K (see Figure 20). K-Mean (KM) estimation.

If one compares the model DENA with individual cluster models, some other models perform better for $K > 150$ (WWT) and for $K < 100$ (WS35) (Section 6.3 and Section 6.4). Furthermore, the detailed analysis revealed that for relatively small values of K it is better to include WIND-2035 and SOLAR-2035 rather than only one technology. At the same time, for relatively large values of K it is better to include WIND-2035 and WIND-2015. Furthermore, the results indicated that it is better to include relatively few attributes describing the electricity network itself. For example, the cluster model WWT that performed better than most other cluster models in this thesis included only network attributes describing the occurrence of an exceedance of thermal capacity of a line, IMAX-MV and NLINES-MV (Section 6.3).

This relative performance of cluster models is subject to the same uncertainties of the dataset as the relative importance of network attributes. These uncertainties were discussed in Section 7.1. Furthermore, some of the conclusions above are sensitive to the assumed locations of future onshore wind and solar photovoltaic power plants. The choice of their location was described in Section 3.2. If the algorithm that distributes installed capacity to network districts and chooses their location within these districts was different, also the relative performance of cluster models can be different. For example, the correlation coefficient between additional onshore wind (ΔWIND = WIND-2035 - WIND-2015) and additional solar photovoltaic installed capacity (ΔSOLAR = SOLAR-2035 - SOLAR-2015) across network districts in the dataset is 0.25. If this correlation was even lower, network attributes related to solar photovoltaic plants (SOLAR-2035, IMAX-LV) could be more important. This is because the higher this correlation, the less important it is to include this attributes if attributes related to wind power plants are already included.

The same holds true for the correlation between installed generation capacity of one technology in 2015 and 2035. The correlation coefficient between SOLAR-2015 and SOLAR-2035 is 0.97 in the dataset. For WIND-2015 and WIND-2035, the correlation coefficient is 0.88. If it was lower, the attributes describing the installed capacity in 2015 (WIND-2015, SOLAR-2015) would be relatively more important.

8. Conclusions

The German power system is undergoing a fundamental transformation substituting electricity generated from fossil fuels with electricity from renewable resources. This transformation - the Energiewende - poses challenges also for the German electricity distribution system. If the installed capacity of onshore wind and solar photovoltaic power plants increases as projected, the total costs of distribution network expansion by the year 2035 are estimated to be up to 40 billion EUR (dena, 2012).

These and other estimates of the costs of expanding distribution networks at the national (BMWi, 2014) and sub-national level (Ackermann et al., 2014; Rehtanz et al., 2017) were derived from samples of relatively few networks in Germany and aggregated to total costs using a cluster analysis. This methodology was developed and applied for the first time by dena (2012) and has since then be adopted by subsequent authors. Whether the methodology can produce accurate results and whether it is better suited than alternative methods has so far however not been examined.

In this thesis, a dataset of synthetic networks was used to examine different cluster models and cluster estimation methods as alternatives to the methodology developed and first applied in dena (2012). To this aim, first a theoretical framework was developed. This framework was used to define 14 network attributes. Second, each of these attributes was examined regarding its effect if it is included in an existing cluster model. Third, the cluster models that yield the lowest within-cluster dispersion of expansion costs were identified. Fourth, the previous two results were used to define 57 cluster models and compare their performance for two alternative scenarios of network expansion. The performance was assessed both in terms of the within-cluster dispersion of costs and the relative deviation of estimated total costs from calculated total costs. Finally, some of the cluster models were analysed in more detail including the geographic occurrence of clusters.

The six network attributes that reduce the within-cluster dispersion of costs of an existing cluster model the strongest were in this order: WIND-2035, WIND-2015, IMAX-MV, SOLAR-2035, LENGTH-MV, SOLAR-2015 (see Table 2 for their definition). The results show that these attributes are also most frequently included in those cluster models that feature the lowest within-cluster dispersion of expansion costs.

Overall, the results show that for the best cluster models, which include the model of dena (2012), the relative deviation of estimated total costs from calculated total costs decreases with the number of clusters K. This number was varied in fixed intervals from 5 to 300, whereby the total number of networks in the dataset is 2928. The relative deviation decreased relatively strongly for $K \leq 50$ and less strongly for $K > 50$. For this reason, $K = 50$ can represent a good choice when searching for a balance between the accuracy of the model and the resources required for computations. For $K \geq 50$ the relative deviation of estimated total costs was on average less than $10\,\%$ for the best 28 cluster models. These 28 cluster models feature a smaller relative deviation than the average deviation of random samples of networks of size K.

Furthermore, for $K \leq 100$ cluster models that include future installed generation capacity of onshore wind and solar photovoltaic power plants as network attributes results in the lowest relative deviation. For $K > 150$, the cluster model WWT, which includes the installed generation capacity of onshore wind in 2015 and in 2035, as well as attributes that describe thermal limits of the network, performed best. The model DENA, which includes the installed generation capacity of onshore wind and solar photovoltaic power plants in 2015 and in 2035, featured a relative deviation that was lower than the average of the best 28 models for $K < 200$ but not lower than the best of these models.

For all results, two scenarios, one worst-case scenario and one scenario with curtailment (see Section 3.4 for their definition) were examined. Furthermore, K-Mean and regression trees were compared as two alternative cluster estimation methods. Regression trees were fitted to the costs of the worst-case scenario and applied to the scenario with curtailment. Overall, the relative deviation of estimated total costs from calculated total costs was lower for regression tree than for K-Mean estimation also for the curtailment scenario.

In sum, the results indicate that the methodology used by previous studies can produce good results: the cluster model that was used previously performs better than the average of the alternative cluster models proposed in this thesis. At the same time, the results indicate two ways by which the performance of the methodology can be improved. The first is the choice of the cluster model. For all values of K, at least one of the alternative models performed better. The second way to improve the performance of the methodology is the use of regression trees rather than K-Mean to estimate clusters. For the scenarios of this thesis, regression trees tended to outperform K-Mean with respect to both metrics that were developed for the evaluation of cluster models.

The use of a dataset of synthetic electricity networks allowed the author the analysis in this thesis. At the same time, it introduced additional uncertainty about the accuracy of the costs of network expansion and the association between network attributes and these costs. The construction of the dataset was in detail described in Chapter 3. The dataset has also been successfully validated against statistics of the total length of lines and cables in Germany (Amme et al., 2017). In order to examine the costs that were computed for the dataset, the total costs were compared to the total costs in previous studies. Overall, the total costs per additional installed generation capacity are 0.11 EUR / W and hence close to the result of dena (2012) (0.14 EUR / W), especially if compared to the alternative estimate of total costs in BMWi (2014) (0.23 EUR / W, Table 8).

The results of this thesis and their discussion in Chapter 7 point to several avenues for future research. This includes a more detailed validation of the dataset of synthetic networks. For this purpose, statistics published by distribution network operators due to the German energy law could be used in order to compare statistics not only on the national level but also on the level of network districts. Furthermore, the performance of regression tree estimation relative to K-Mean estimation requires further examination. For this, the effect of additional parameters of regression trees could be analysed, such as the size of clusters and the depth of the tree. Furthermore, additional scenarios of costs of network expansion could be analysed in order to examine the robustness of the good performance of trees if a scenario differs more strongly from the worst-case scenario to which the tree has been fitted.

Acknowledgements

This thesis would not have been successful without the support from my supervisor, my colleagues at RLI, my family, my friends and my partner.

I would like to thank Prof. Dr. Franz Hubert for granting me the freedom to choose and develop the topic of this thesis and for the fruitful feedback to three seminar presentations. Furthermore I am very thankful to my colleagues at the Reiner-Lemoine-Institut (RLI) Berlin for offering me this topic and for guiding me from the start of this thesis down to the finish line. My special thanks goes to Guido Plessmann for supporting me in many ways. Furthermore, I would like to thank Birgit and Jonathan for sharing their expertise with me and for providing valuable feedback.

I would also like to thank my course mates with whom I collaborated in seminars and lectures and spent time inside and outside the classroom. You provided invaluable support and motivation throughout my Master's studies.

Finally, I would like to thank my close friends, my family and my partner Tania. Without your support, I would not have had the optimism and endurance for finishing this thesis.

Appendix

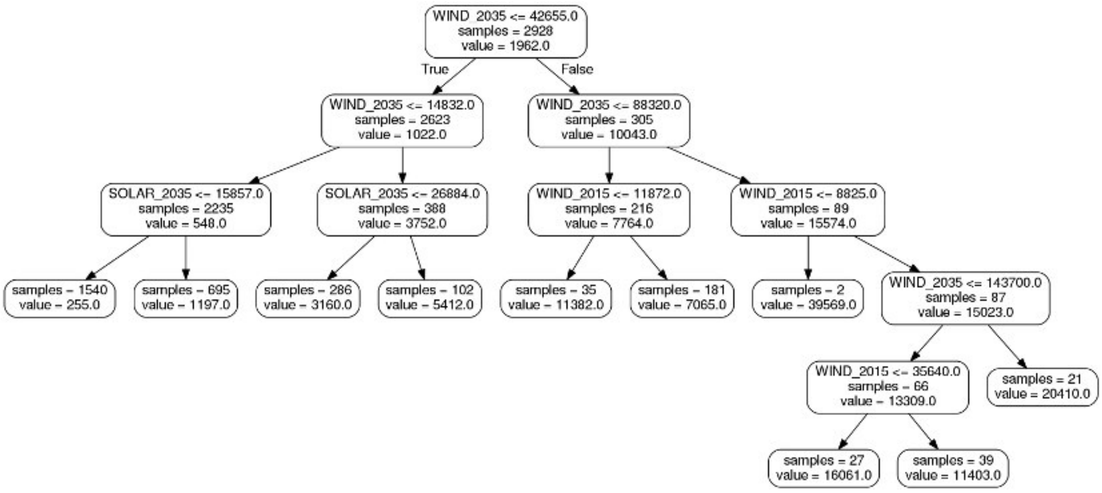


Figure 33: Regression tree for the cluster model DENA with $K = 10$.

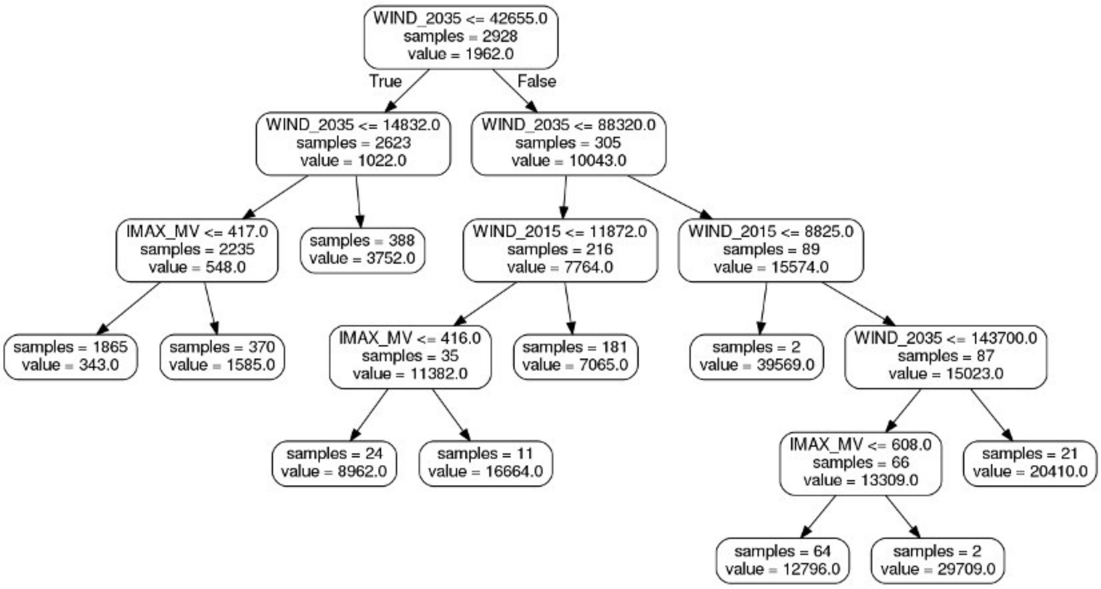


Figure 34: Regression tree for the cluster model WWT with $K = 10$.

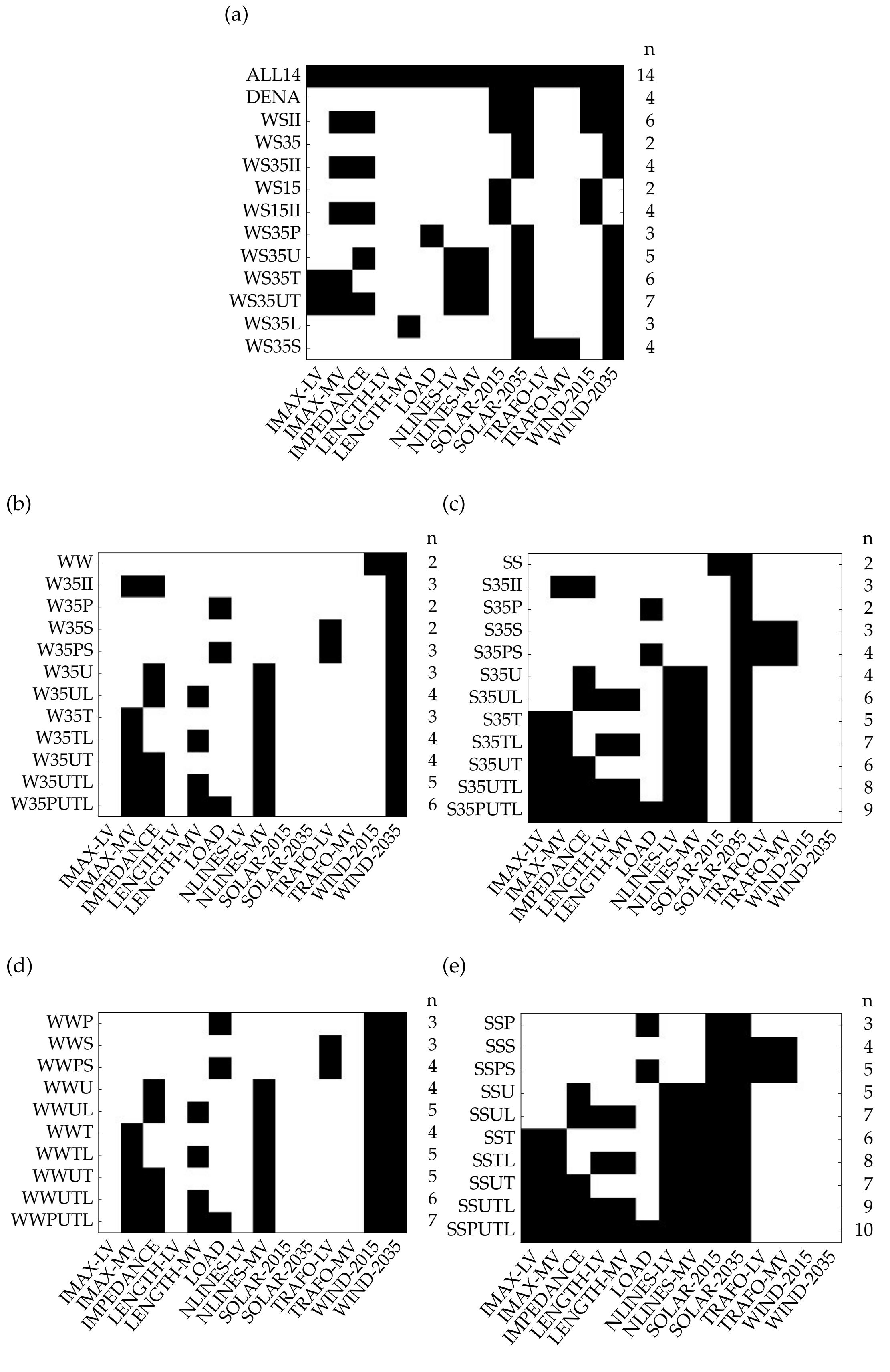


Figure 35: Cluster models for further analysis, e.g. in Section 6.3 and Section 6.4, and the labels used to refer to them throughout this thesis.

References

Ackermann, T., Martensen, L., Brown, T., Untsch, S., Tröster, E., and Geidel, S. (2014). Verteilnetzstudie rheinland-pfalz 2030. *Available online: https://mwkel.rlp.de/fileadmin/mwkel/Abteilung_6/Energie/Verteilnetzstudie_RLP.pdf [last accessed: 26.11.2017]*.

Amme, J., Plessman, G., Buehler, J., Huelk, L., Koetter, E., and Schwaegerl, P. (2017). The eGo grid model: An open-source and open-data based synthetic medium-voltage grid model for distribution power supply systems. *Working Paper (submission on request)*.

BDEW (2008). *Bundesverband der Energie- und Wasserwirtschaft e.V. (BDEW): Technische Richtlinie Erzeugungsanlagen am Mittelspannungsnetz - Richtlinie fuer Anschluss und Parallelbetrieb von Erzeugungsanlagen am Mittelspannungsnetz*.

Biggar, D. R. and Hesamzadeh, M. R. (2014). *The Economics of Electricity Markets*. IEEE Press and John Wiley & Sons Ltd.

BMWi (2014). Moderne Verteilernetze fuer Deutschland (Verteilernetzstudie). Studie im Auftrag des Bundesministeriums für Wirtschaft und Energie (BMWi). *Available online: https://www.bmwi.de/Redaktion/DE/Publikationen/Studien/verteilernetzstudie.pdf ?blob=publicationFile&v=5 [last accessed: 26.11.2017]*.

BMWi (2017). Erneuerbare energien zeitreihen. *Available online: http://www.erneuerbare-energien.de/EE/Navigation/DE/Service/Erneuerbare_Energien_in _Zahlen/Zeitreihen/zeitreihen.html [last accessed: 26.11.2017]*.

Breiman, L., Friedman, J., Olshen, R., and C.J.Stone (1984). *Classification and Regression Trees*. CRC Press.

Bundesnetzagentur (2017). Quartalsbericht zu netz- und systemsicherheitsmaßnahmen viertes quartal und gesamtjahr 2016. *Available online: https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Bundesnetzagentur/Publikationen/Berichte/2017/Quartalsbericht_Q4_Gesamt_2016.pdf?__blob=publicationFile&v=2 [last accessed: 26.11.2017].*

Bundesregierung der Bundesrepublik Deutschland (2010). Energiekonzept für eine umweltschonende, zuverlässige und bezahlbare Energieversorgung. *Available online: http://www.bundesregierung.de/ContentArchiv/DE/Archiv17/_Anlagen/2012/02/energiekonzept-final.pdf [last accessed: 26.11.2017].*

Consentec, RWTH Aachen, Rechenzentrum fuer Versorgungsnetze Hart/Wehr, and Frontier Economics Limited (2006). *Untersuchung der Voraussetzungen und möglicher Anwendung analytischer Kostenmodelle in der deutschen Energiewirtschaft. Untersuchung im Auftrag der Bundesnetzagentur für Elektrizität, Gas, Telekommunikation, Post und Eisenbahnen.*

dena (2012). dena-Verteilnetzstudie. Ausbau- und Innovationsbedarf der Stromverteilnetze in Deutschland bis 2030. *Available online: https://www.dena.de/themen-projekte/projekte/energiesysteme/dena-verteilnetzstudie/ [last accessed: 26.11.2017].*

DIN (2011). *Voltage characteristics of electricity supplied by public distribution networks. German version EN 50160:2010 + Cor.: 2010.* German Institute for Standardisation (DIN).

Gönen, T. (1986). *Electric Power Distribution System Engineering*. McGraw Hill Inc.

Gust, G. (2014). Analyse von niederspannungsnetzen und entwicklung von referenznetzen. *Masterarbeit an der Fakultaet für Wirtschaftswissenschaften des Karlsruher Institut fuer Technologie. Available online: https://publikationen.bibliothek.kit.edu/1000045150/3370439 [last accessed: 26.11.2017].*

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Hülk, L., Wienholt, L., Cußmann, I., Müller, U., Matke, C., and Kötter, E. (2017). Allocation of annual electricity consumption and power generation capacities across multiple voltage levels in a high spatial resolution. *International Journal of Sustainable*

Energy Planning and Management, 13(0):79–92.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning*. Springer.

Kerber, G. (2010). *Aufnahmefähigkeit von Niederspannungsverteilnetzen für die Einspeisung aus Photovoltaikkleinanlagen*. Dissertation an der TU Muenchen. Available online: https://mediatum.ub.tum.de/doc/998003/998003.pdf [last accessed: 26.11.2017].

Kirtley, J. L. (2010). *Electric Power Principles*. Wiley.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):14–23.

NEP (2017). Netzentwicklungsplan strom 2030, version 2017 (2. entwurf). *Available online on: https://www.netzentwicklungsplan.de/de/netzentwicklungsplaene/netzentwicklungsplaene-2030-2017 [last accessed: 26.11.2017].*

Rehtanz, C., Greve, M., Häger, U., Hagermann, Z., Kippelt, S., Kittl, C., Kloubert, M.-L., Pohl, O., Rewald, F., and Wagner, C. (2017). *Verteilnetzstudie für das Land Baden-Württemberg. Studie für das Ministerium für Umwelt, Klima und Energiewirtschaft Baden-Württemberg*. ef.Ruhr GmbH.

Schachler, B. (2017). *Network expansion in the software eDisGo. Personal communication on November 10th 2017. The software is available online on:https://github.com/openego/eDisGo*.

Scheffler, J. (2002). Bestimmung der maximal zulaeassigen netzanschlussleistung photovoltaischer energiewandlungsanlagen in wohnsiedlungsgebieten. *Ph.D. thesis TU Chemnitz. Available online: http://monarch.qucosa.de/fileadmin/data/qucosa/documents/4595/data/Dissertation_Scheffler.pdf [last accessed: 26.11.2017].*

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

VDE (2011). *Verband der Elektrotechnik Elektronik Informationstechnik e.V. (VDE):*

VDE-AR-N 4105: Erzeugungsanlagen am Niederspannungsnetz – Technische Mindestanforderungen fuer Anschluss und Parallelbetrieb von Erzeugungsanlagen am Niederspannungsnetz.

Walker, G., Krauss, A.-K., Eilenberger, S., Schweinfort, W., and Tenbohlen, S. (2014). *Design of a Standardized Approach for the Classification of Distribution Grids*. VDE-Kongress 2014, Frankfurt am Main.