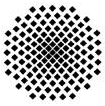

Universität Stuttgart
Institut für Energiespeicherung
Professor Dr. André Thess


Deutsches Zentrum
für Luft- und Raumfahrt
Institut für Technische Thermodynamik

Master Thesis

Development and implementation of a spatial clustering approach using a transmission grid energy system model

submitted by Johannes Metzdorf

Stuttgart, September 2016

Supervisors:
Prof. Dr. André Thess
M.Sc. Karl-Kiên Cao

Declaration

I, Johannes Metzdorf, born on 02.06.1987 in Ostfildern declare that I have developed and written the enclosed Master Thesis completely by myself, except the suggestions of my supervisors M.Sc. Karl-Kiên Cao and Dr. Hans Christian Gils, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Master Thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Place	Date	Signature

Zusammenfassung

Die Modellierung von Stromnetzen erfordert eine geographische Berücksichtigung aller Netzkomponenten. Bei der Anwendung von Stromsystemmodellen führt eine hohe räumliche Auflösung jedoch zu sehr hohen Rechenzeiten. Eine Möglichkeit hochaufgelöste Modelle zu beschleunigen, bietet eine Aggregation durch Clusterbildung. Durch die Clusterbildung sollen andererseits Informationen, die das Resultat stark beeinflussen, nicht verloren gehen. Um dieses Dilemma zu umgehen, wird in dieser Arbeit ein Ansatz gewählt, bei dem auf der einen Seite ähnliche Netzknoten zusammengefasst werden und auf der anderen Seite Netzengpässe weiterhin Bestand haben. Um dies zu erreichen, wird ein Spectral Clustering Algorithmus mit der Grenzkostentheorie (Locational marginal pricing) kombiniert. Weiterhin wird dieser kombinierte Algorithmus auf ein Fallbeispiel angewandt, welches das deutsche Übertragungsnetz in hoher räumlicher Auflösung abbildet. Die Anwendung des Algorithmus' auf das Fallbeispiel führt zu einer Aggregation des hochaufgelösten Modells. Durch die Aggregation werden Netzbeschränkungen abgebaut, was wiederum Abweichungen im Kraftwerkseinsatz (Dispatch) verursacht. Bezogen auf die gesamten Systemkosten, führt eine Aggregation des Modells zu Abweichungen von über 20 %, verglichen mit der höchsten Auflösung. Auf der anderen Seite ermöglicht eine Aggregation eine Verkürzung der Lösungszeit um 95 %. Aus diesem Grund muss ein Kompromiss zwischen der Genauigkeit der Ergebnisse und der Beschleunigung des Modells gefunden werden. Die in dieser Arbeit entwickelte Methodik, bietet eine Möglichkeit, die durch die Aggregation entstehenden Abweichungen abzuschätzen.

Abstract

Modeling power grids requires a geographically explicit consideration of all the grid components. However, a high spatial resolution leads to high computation times of fundamental power generation dispatch optimization models. An opportunity to accelerate highly resolved models, is an aggregation of the model by finding clusters. In doing so, information affecting the result too strongly should not be lost by the aggregation. To overcome this dilemma, an approach to find clusters containing similar grid vertices by keeping grid congestion is developed in this work. For this purpose, a spectral clustering algorithm is combined with the theory of locational marginal pricing. Furthermore, this combined algorithm is applied to a case study representing the German transmission grid in high spatial resolution. In the case study, the highly resolved model is aggregated, which leads to decreasing grid restrictions, and in turn causes a deviation of the dispatch. With regard to total system costs, aggregating the model accounts for deviations of up to 20 % compared to the highest resolution. On the other hand, aggregating the model enables accelerations of up to 95 % in terms of solving time. As a consequence, a compromise between accuracy of the results and acceleration of the model has to be found. The methodology developed in this work provides an opportunity to estimate the deviations arising from the aggregation.

Contents

List of Figures

Formula symbols

Symbol	Description	Unit
λ	Longitude	°
σ	Standard deviation	–
ϕ	Latitude	°
A_i	Vertex attribute of vertex i	–
$assoc(A, V)$	Total connection from vertices of A to the whole graph	–
B_j	Postcode area	–
c_j	Centroid of k-means algorithm	–
d	Distance of postcode area to vertex	km
D	Diagonal matrix	–
$\tilde{D}$	Distance between data point and centroid	–
$E_{agg,ou}$	Aggregated, overutilized edges	–
E_{ij}	Edge between vertex i and j	–
$E_{ou,tot}$	Total overutilized edges	–
$E_{percapita,a}$	Power demand per inhabitant per year	GWh
$E_{tot,a}$	Total power demand of Germany per year	GWh
$E_{v_i,a}$	Power demand at vertex i per year	GWh
I_{tot}	Total inhabitants of Germany	–
I_{v_i}	Inhabitants at vertex i	–
k	Number of clusters	–
l	Link length	km
L	Laplacian matrix	–
L_{rw}	Laplacian matrix according to Shi	–
L_{sym}	Laplacian matrix according to Ng	–
L_x	Link impendance	Ω
m	Maximum edge weight	–
M_i	Marginal costs at vertex i	EUR
$NCut$	Normalized cut	–
r	Radius of the earth	km
$r(k)$	Ratio of overutilized links	–
$RCut$	Ratio cut	–

v_i	Vertex i	–
w_{ij}	Edge weight on edge ij	–
W	Adjacency matrix	–
x	Reactance factor	$\frac{\Omega}{km}$
X_{ij}	Impedance between vertex i and j	Ω
y_i	Data point i	–
Y_{ij}	Admittance between vertex i and j	S
z_i	Data point i in transformed coordinate system	–

1. Introduction

This introduction Chapter provides the underlying motivation for this work. Furthermore, the state of research is examined and based on that, research questions are derived. Scope and structure of this work are mentioned at the end of this chapter.

1.1. Motivation

In order to reduce climate gas emissions and to mitigate climate change, a switch to a carbon neutral and sustainable energy supply is required. Due to the climate targets decided at the 2015 United Nations Climate Changes Conference (less than a 2% increase in global warming compared to the pre-industrial era) [1], sustainability goes along with the decarbonization of the power supply. Based on the energy concept decided by the German government in 2010 [2], a high share of the fossil power supply has to be substituted by renewable energy sources. In such a power system, the majority of the electricity generation is related to wind turbines and photovoltaic plants. Intrinsically, the power supply structure is getting more fragmented, decentralized and additionally less adjustable.

For analyzing power systems in terms of economic efficiency, environmental sustainability and supply reliability, so-called fundamental electricity market models are applied. Under the circumstances of high shares of renewable power generation, these models are growing more and more complex as analyzing power systems requires a holistic view of the technologies. These include the demand side, conventional and renewable power plants, storage technology and also the electrical grid.

In former days, the electrical grid was basically a one-way street, transferring the electricity from big scale power plants to the consumers in the lower voltage levels over short distances. In power systems with high shares of renewable energy this principle is partially reversed. Using Germany as an example, it can be illustrated how the demand and supply is getting spatially uncoupled. The power-intensive industry and the densely populated regions are located in the west and south of Germany, whereas the best potentials of wind energy exist in the north of Germany. Since in former times there was no need for transferring high amounts of electricity from the North to the South (or to deal with high supply of RE in distribution grids), the transmission capacities of the grid will have to be adjusted to the changing generation structure [3]. To achieve the goals of the German government in terms of a sustainable energy system and also to guarantee supply reliability, the electrical grid will play a key role in this context. By identifying grid congestion, it can be figured out where the transmission capacities have to be adjusted.

As already stated, power systems and thus power system models, are getting more and more complex. The growing complexity leads to increasing computation times and in some cases executing the analysis is not possible due to missing main memory capacity. To overcome this issue, different strategies regarding reduction and acceleration have been developed. One of these strategies is aggregating the model. In this context, aggregation means to aggregate the given information based on a defined detail level. For instance, aggregating the European power system on a country level, leads to one vertex in every country, which combines all parameters like power plant stock, demand or grid information. Obviously, by a simple spatial aggregation of the power system components, in particular grid information, is getting lost. An aggregated power system (aggregated on one vertex) can be interpreted as a 'copper plate' in terms of power transmission. Regarding power grids, 'copper plate' is characterized by unlimited transmission capacity of the grid. Consequently, every spatial power demand or every pumped storage can be supplied by any arbitrary power plant. It is therefore likely that the share of the power supply by power plants with low marginal cost, for instance lignite power plants (with today's commodity prices), is overestimated, compared to a system accounting for all grid restrictions. As a follow-up from this shortcoming, requirements for the aggregation can be derived. On the one hand the aggregation must lead to a complexity reduction of the power system (and thus to an acceleration in terms of computing), and on the other hand information affecting the result strongly may not be lost.

A basic aggregation process of a power system is depicted in Figure 1.1. Based on a highly resolved power system (upper level) a methodology for systematic aggregation is needed, in order to fulfill the requirements stated above (complexity reduction while maintaining information).

A proven approach in terms of systematic aggregation methods are clustering algorithms. By finding similar patterns in a data set, homogeneous groups of things can be build. Ordinary clustering algorithms like k-means aggregate elements of a (complex) system based on their numerical attributes. However, when aggregating power systems (with consideration of a power grid), also the topology of the grid has to be taken into account. The spectral clustering approach covers both requirements (clustering on numerical attributes, considering topology) and has already been used for the clustering of power grids by Sanchez et al. [4].

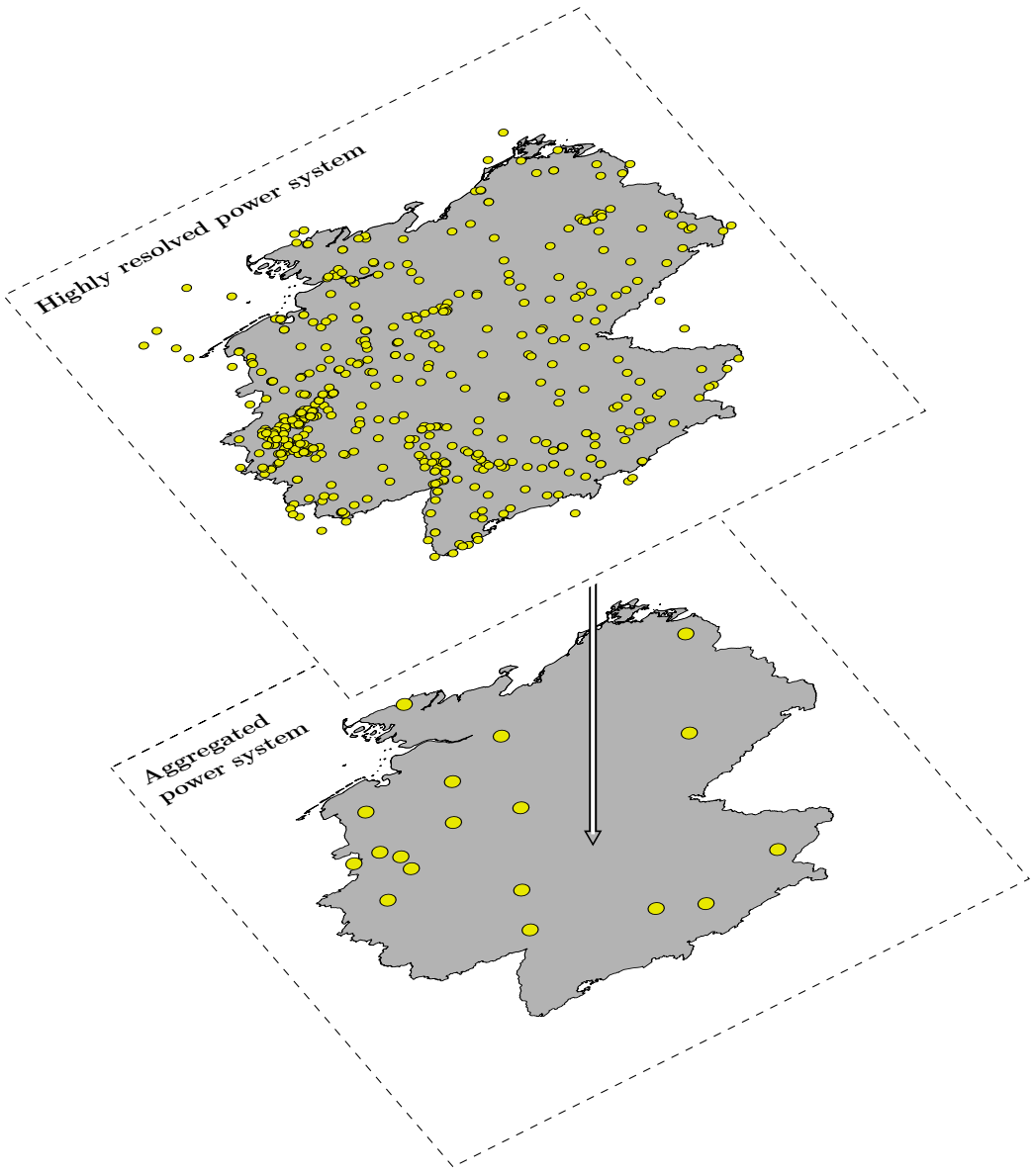


Figure 1.1.: Two levels of the clustering process: highly resolved data set on the upper level as input, aggregated data set at the bottom level

1.2. State of research

This section introduces the theoretical background of the study. In addition to the optimizing energy system model REMix, partition methods, clustering and the nodal pricing approach are described.

1.2.1. The optimizing energy system model REMix

REMix (Renewable Energy Mix) was developed by the German Aerospace Center (DLR) as a tool for bottom-up modeling of energy systems. The main goal of the tool is to determine minimal system costs of energy scenarios (primarily based on high shares of renewable energy resources). The objective function of the linear optimization problem represents minimal system costs in terms of dispatch and capacity expansion. For the formulation of the mathematical problem, REMix uses GAMS (General Algebraic Modeling System) [5]. The solution of the resulting linear optimization is done by CPLEX [6]. On the input side, modules for spatially and hourly resolved parametrization in terms of electricity, heat, storage and transportation are provided to set up energy system scenarios. REMix determines hourly system operation, capacity and grid expansion, system costs and CO_2 emissions. The whole process is schematically shown in Figure 1.2. The different components of REMix are described in [7], [8], [9], [10].

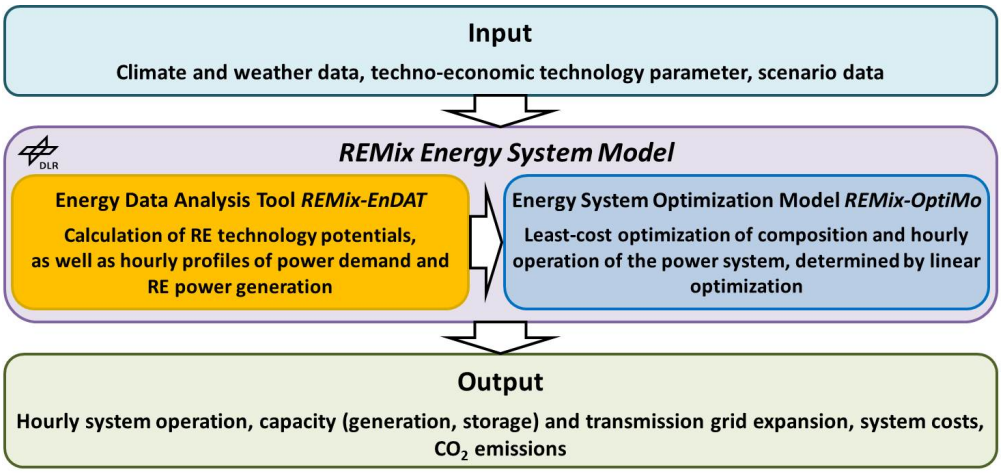


Figure 1.2.: Schematic layout of the REMix energy system model [7]

1.2.2. Identification of grid subsections

Basically, a grid consists of vertices (active vertices: in- and output of electricity, passive vertices: crossings) and links between vertices. By analyzing the grid, calculating electrical current, voltage or transferred power on the links is meant. Due to the growing share of renewable energy power plants and hence growing complexity (decentralization, fluctuating supply, decoupling of supply and demand) analyzing each element in the grid is getting more difficult [11]. Finding related subsections in electrical grids and thus reducing the complexity is a contribution to improve the analysis efficiency in power

system modelling, since one main goal of the grid analysis is to identify critical grid conditions. Links where critical grid conditions occur can indicate a need of expansion. A possible criterion for forming a grid subsection can then be the absence of critical grid conditions within a subsection.

Contemplating the German transmission grid, there is no transparent aggregation method available. An aggregation for the German power system applied in different studies is the *18-Regionen-Modell* depicted in Figure 1.3 [3]. The applied aggregation divides the German power system into 20 clusters. The basis for the division seems to be the geometry of the transmission grids.

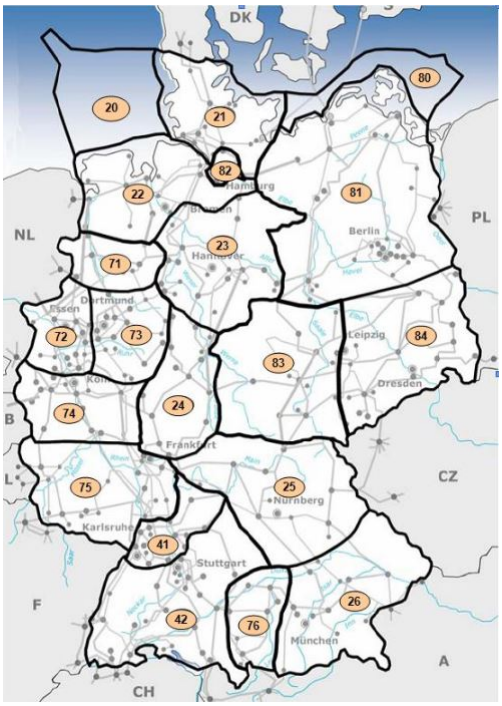


Figure 1.3.: *18-Regionen-Modell* of the German transmission system operators; naming of the regions according to ENTSO-E [3]

The methodology behind the aggregation is not published and thus the aggregation cannot be applied to other power systems such as to other countries or in a European context.

Other present established aggregations of power systems are a result of long-term work experience and usually cannot be applied to other grids or to derive general rules for partitioning as there is no comprehensible methodology underlying [12]. In addition, with the integration of renewable energy power plants, power systems are changing over time [12], which also complicates applying the same partitions for a long period. Furthermore, Gang et al. [12] mention different ways to identify transmission subsections like the *Girvan Newman algorithm* (GN) (developed by Michelle Girvan and Mark Newman) based on complex network theory. GN focuses on transmission betweenness which is linked to identify key transmission links. Key transmission links are determined by their total contribution of power transmission compared to the total power injection in the grid.

Another approach is given by Anderski et al. [13]. The method of the partition is to build clusters where small regions with similar properties are aggregated to a cluster and analogously to put regions with distinct properties in different clusters. Moreover, a reduction of the grid is considered as well by applying the method of *Equivalent impedances* where an estimated power flow is approximated to the real power flow of the system.

An alternative strategy to the mentioned methods are clustering algorithms using the spectrum of a graph. The term 'spectrum' refers to the sized Eigenvalues of the graph's Laplacian matrix or Adjacency matrix. In the following, this type of clustering is examined in detail.

1.2.3. Clustering analysis

Data analysis using clustering algorithms can be considered as a process to reveal data structures. Clustering analysis is an essential component in the fields of data mining, image segmentation or pattern classification. Inherently, clustering is grouping objects based on similarity criteria. A group or a cluster is hence a set of similar patterns [14]. Two of the most important clustering algorithms are k-means and spectral clustering. They are briefly introduced in the following. In the context of this work, clustering algorithms are used for reducing power grid data sets in terms of energy system model acceleration strategies.

Toy example

To visualize the following clustering algorithms, a very simple toy example is introduced. The sample consists of 11 vertices $v_1, \ldots, v_{11}$ with attributes $A_i(v_i)$ (bold) and edges E_{ij} between vertex i and j. The Graph $G = (V, E)$ with vertex set V and edge set E is shown in Figure 1.4. In the sections, where the clustering algorithms are introduced, this toy example will be used to clarify how the algorithms work. The toy example can be interpreted as an electrical grid with distinct attributes at the vertices (e.g. installed power plants, see Section 1.2.4).

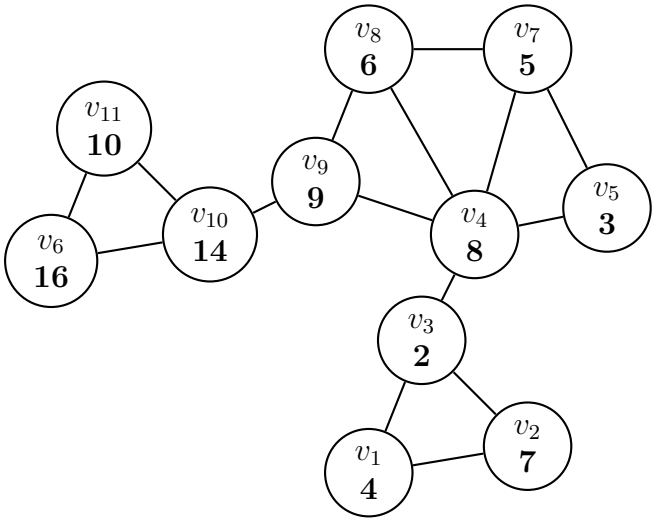


Figure 1.4.: Toy example: graph consisting of $i = 1, \ldots, 11$ vertices v_i labeled with differing numerical attributes $A_i(v_i)$ (bold)

The k-means clustering algorithm

The k-means clustering algorithm is one of the most popular methods to make a k-partition of a data set, whereas k is the number of requested clusters. For applying a k-means algorithm on a data set, two input parameters are needed: the data set itself $y_1, \ldots, y_n$ (points of p-dimensional vector space), which can be interpreted as attributes of the objects and a desired number of clusters k. Since the attributes represent coordinates in $\mathbb{R}^p$, the attributes have to be numerical. The algorithm proceeds as follows [15] (also seen in Figure 1.5):

1. Place centroids $c_1, \ldots, c_k$ at random locations in the p-dimensional vector space
2. Assign each data point y_i to its nearest centroid c_j (for instance with Euclidean distance - length of the line segment connecting y_i and c_j)
3. Update the centroids c_j by recomputing the centers of inertia of the data points y_{ij} in the clusters
4. Repeat step 2 and 3 until convergence (convergence means no further change in assignment of data y_i to c_j)

Due to the randomized placing of initializing centroids, k-means is not deterministic. That means k-means will not provide one specific solution for a data set. The output has always to be referred to a specific initialization of the k-means algorithm. Applications using k-means (e.g. Python's scikit-learn library) therefore perform a given number of initializations and return the 'best' result. In this context 'best' means the clustering with the most compact clusters, respectively the clustering with minimum inertia among the computed results [16].

If the toy example is converted into k-means' perspective the graph modifies to a data set in $\mathbb{R}^1$, since the attributes of the vertices are one-dimensional. The modification is depicted in Figure 1.6.

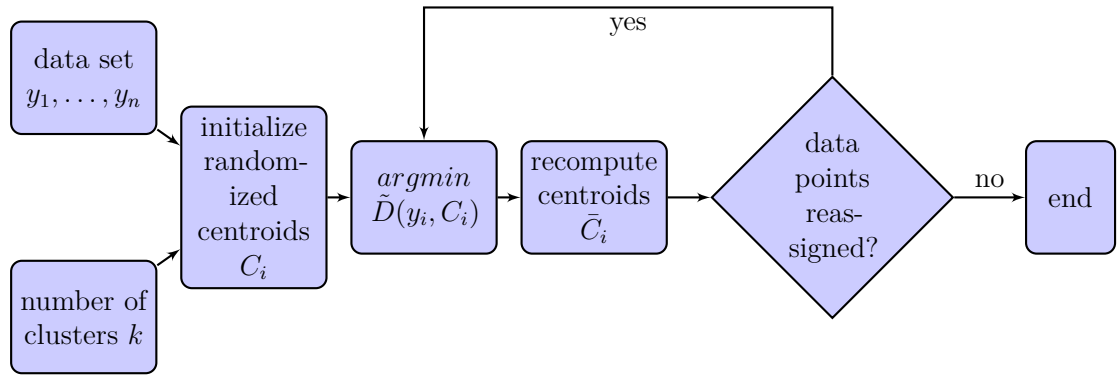


Figure 1.5.: The k-means clustering algorithm

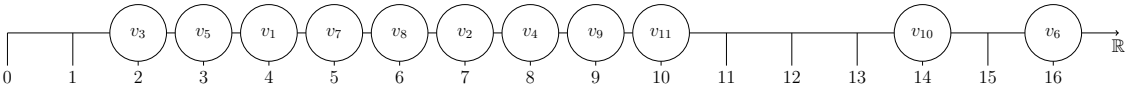


Figure 1.6.: Modified data set of the toy example into k-means' perspective

Applying the k-means algorithm on the toy example's data set with number of clusters $k = 3$ and centroids c_j set leads to the following result where each cluster is represented by a different color in Figure 1.7:

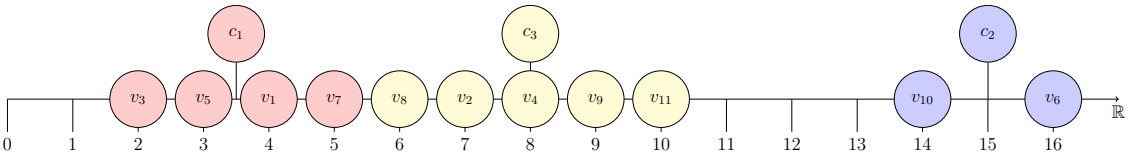


Figure 1.7.: Data set of the toy example clustered by the k-means algorithm with centroids c_j

Considering both the k-means clustering result and the topology of the graph is shown in Figure 1.8. In order to get a meaningful subclassification of a graph, the problem has to be contemplated from different perspectives. Besides the vertex attributes, the topology of the graph has to be considered as well. This means that classical clustering algorithms like k-means (which cannot take topological aspects into account) are not sufficient for creating a meaningful partition. Consequently, clustering of the graph only by the attributes does not lead to the desired result. As already mentioned, both the vertex attributes and the topology of the graph have to be considered for an appropriate clustering. This difficulty can be overcome by using spectral clustering. The spectral clustering algorithm is examined in detail in the following.

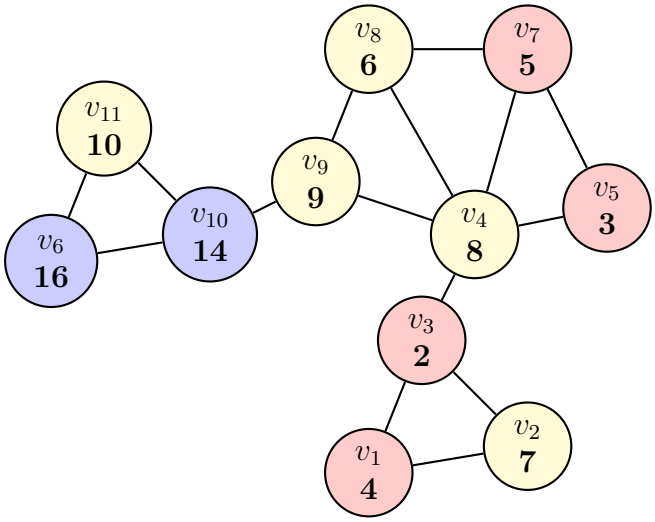


Figure 1.8.: Toy example clustered by vertex attributes using k-means ($k = 3$, cluster membership is indicated by color)

Spectral Clustering

Spectral clustering includes all clustering algorithms using the eigenvalues and eigenvectors of the Laplacian matrix of a graph [17]. Simplified, the Laplacian matrix visualizes the relationship between the different vertices v_i in a graph G. According to von Luxburg [17], there exists a whole field of studying different ways of Laplacian matrices. This section focuses on the creation of the unnormalized Laplacian matrix L which is fundamental for creating other Laplacian matrices, the normalized Laplacian matrix L_{sym} discussed by Ng. et al [18] and the Laplacian matrix L_{rw} according to Shi and Malik [19], because the characteristics of these Laplacian matrices are explained in detail in [17]. Contemplating not only the unnormalized and in a way 'easiest' Laplacian matrix L is due to the advice given by [17] that normalized Laplacian matrices are always to prefer against unnormalized Laplacian matrices. The difference between unnormalized and normalized spectral clustering is the way they serve the minimization of the cut problem. By the cut problem the methodology of removing edges from the graph is meant, as 'unimportant' edges have to be removed in order cluster the graph. Unnormalized spectral clustering is linked to the minimization of Ratio Cut (RCut), normalized spectral clustering refers to the Normalized Cut (NCut) problem. The RCut takes only the 'importance' of an edge into account, whereas the NCut also tends to create balanced clusters by additionally considering the connectedness of vertices to the rest of the graph (this is elaborated in detail in Section 2.5). Furthermore, Luxburg et al. advocate to use L_{rw}, as L_{sym} could lead to undesired artifacts due to its algorithm (see also Section 2.2) [17].

For creating the Laplacian matrix L, edge weights w_{ij} are introduced that take into account how similar two vertices v_i and v_j are. The interpretation of edge weights is a penalty for cutting the edge during the clustering process and can also be seen as a connection strength, as vertices with strong connection (edges) are more likely to get clustered [4]. Since two vertices v_i and v_j are considered similar if their vertex attributes A_i and A_j are similar, a function to weight the edges $f(A_i, A_j)$ has to be implemented.

For the toy example the function $f(A_i, A_j)$ is set according to Equation 1.1. For further information how to determine edge weights see Section 2.1.

$$\begin{aligned} m &= max|A_i - A_j| \\ f(A_i, A_j) = w_{ij} &= m - |A_i - A_j| + 1 \end{aligned} \tag{1.1}$$

The Graph $G = (V, E)$ with vertex set V and **weighted** edge set E (weighted according to Equation 1.1) is shown in Figure 1.9.

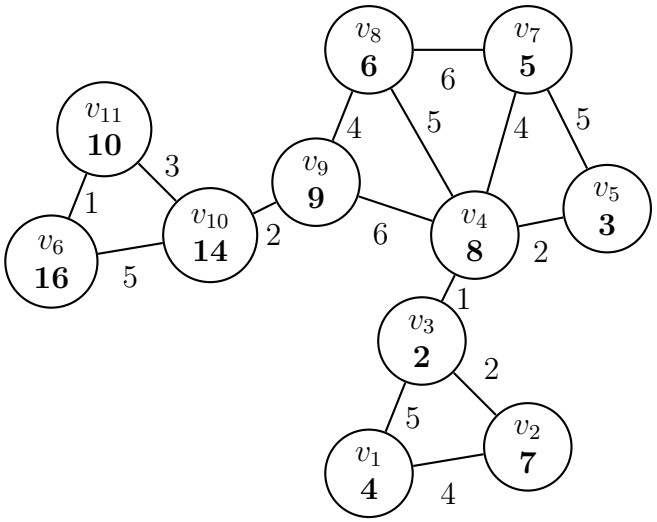


Figure 1.9.: Toy example with weighted edges between vertices

If the graph consists of n-vertices, the Laplacian matrix is an $n \times n$, real symmetric matrix with degree d_i of vertex v_i on the diagonal and negative values outside. By the degree of a vertex d_i, the sum of linked (weighted) edges E_{ij} on the vertex v_i is meant. The sum of the columns and the sum of the rows are zero [4].
Creating the unnormalized Laplacian matrix requires the degree matrix D and the weighted adjacency matrix W. For the unnormalizied Laplacian matrix L the following relationship can be stated:

$$L = D - W \tag{1.2}$$

The degree matrix D has the vertex degrees on the diagonal. The rest of the entries are zero. Furthermore, W represents the weighted adjecency matrix (also known as neighboorhood matrix). The adjacency matrix carries the information which vertices v_i of the graph are linked through an edge E_{ij}. The weighted adjacency matrix also considers the weight of the edges. For a better understanding the matrices L, D and W of the toy example are set up in the following:

$$\begin{bmatrix} 9 & -4 & -5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -4 & 6 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -5 & -2 & 8 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 18 & -2 & 0 & -4 & -5 & -6 & 0 & 0 \\ 0 & 0 & 0 & -2 & 7 & 0 & -5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & -5 & -1 \\ 0 & 0 & 0 & -4 & -5 & 0 & 15 & -6 & 0 & 0 & 0 \\ 0 & 0 & 0 & -5 & 0 & 0 & -6 & 15 & -4 & 0 & 0 \\ 0 & 0 & 0 & -6 & 0 & 0 & 0 & -4 & 12 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -5 & 0 & 0 & -2 & 10 & -3 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 18 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 15 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix} - \begin{bmatrix} 0 & -4 & -5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -4 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -5 & -2 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & -2 & 0 & -4 & -5 & -6 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & -5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5 & -1 \\ 0 & 0 & 0 & -4 & -5 & 0 & 0 & -6 & 0 & 0 & 0 \\ 0 & 0 & 0 & -5 & 0 & 0 & -6 & 0 & -4 & 0 & 0 \\ 0 & 0 & 0 & -6 & 0 & 0 & 0 & -4 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & -5 & 0 & 0 & -2 & 0 & -3 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & -3 & 0 \end{bmatrix}$$

For a spectral clustering approach using L, the following steps have to be applied:

1. Determining the first k eigenvectors $u_1, \ldots, u_k$ of L (first in terms of the corresponding k-smallest eigenvalues)
2. Creating new matrix $U \in \mathbb{R}^{n \times k}$ with vectors $u_1, \ldots, u_k$ as columns
3. Labeling the i-th row of U $(i = 1, \ldots, n)$ as $y_i \in \mathbb{R}^k$
4. Applying a standard algorithm like k-means to the points y_i leads to clusters $C_1, \ldots, C_k$

The calculation of the unnormalized Laplacian matrix L can be interpreted as preliminary work due to its basic property for other Laplacian matrices. The Laplacian matrices contemplated in this work are set up in the following. The relation between L and L_{rw} is described in Equation 1.3:

$$L_{rw} = D^{-1}L \tag{1.3}$$

The algorithm by [19] consists of the following workflow:

1. Determining the first k eigenvectors $u_1, \ldots, u_k$ of L (first in terms of the corresponding k-smallest eigenvalues) of the generalized eigenvalue problem $Lu = \lambda Du$
2. Creating new matrix $U \in \mathbb{R}^{n \times k}$ with vectors $u_1, \ldots, u_k$ as columns
3. Labeling the i-th row of U $(i = 1, \ldots, n)$ as $y_i \in \mathbb{R}^k$
4. Applying a standard algorithm like k-means to the points y_i leads to clusters $C_1, \ldots, C_k$

Also L_{sym} can be created out of the unnormalized Laplacian matrix L:

$$L_{sym} = D^{-\frac{1}{2}}LD^{-\frac{1}{2}} \tag{1.4}$$

After creating L_{sym} the spectral clustering algorithm by [18] can be executed. The different steps are shown below:

1. Determining the first k eigenvectors $u_1, \ldots, u_k$ of L_{sym} (first in terms of the corresponding k-smallest eigenvalues)
2. Creating new matrix $U \in \mathbb{R}^{n \times k}$ with vectors $u_1, \ldots, u_k$ as columns
3. Normalization of the rows of U to norm 1

4. Labeling the i-th row of U $(i = 1, \ldots, n)$ as $y_i \in \mathbb{R}^k$
5. Applying a standard algorithm like k-means to the points y_i leads to clusters $C_1, \ldots, C_k$

The following explanations are based on working with L_{sym}. The choice of which Laplacian matrix is used for the clustering in this work refers to Section 2.2.

What is implicitly done by forming a Laplace matrix, is to give geometric coordinates to the vertices in $\mathbb{R}^k$. The n rows of the normalized matrix U with k eigenvectors as columns contain these coordinates [4]. Actually, the vertices already have coordinates y_i given by the Laplacian matrix operations (rows and columns of the Laplacian matrices). Due to the characteristics of the Laplacian matrices a change of representation (from y_i to $z_i \in \mathbb{R}^k$) is advantageous. In principle, the coordinates are transformed to a new vector space $\mathbb{R}^k$ spanned by the eigenvectors of the Laplacian matrix. This transformation increases the cluster properties in the data and cluster information can be traced by simple clustering algorithms like k-means [17].

For visualization, the algorithm is applied to the toy example ($k = 3$). As the Laplacian matrix L_{sym} is already generated, the matrix $U_{norm} \in \mathbb{R}^{10\times 3}$ can be computed.

$$U_{norm} = \begin{bmatrix} -0.446 & -0.874 & 0.188 \\ -0.440 & -0.876 & 0.193 \\ -0.486 & -0.859 & 0.158 \\ -0.805 & 0.246 & -0.538 \\ -0.708 & 0.269 & -0.651 \\ -0.409 & 0.387 & 0.825 \\ -0.721 & 0.277 & -0.633 \\ -0.765 & 0.302 & -0.568 \\ -0.872 & 0.396 & -0.285 \\ -0.454 & 0.405 & 0.793 \\ -0.406 & 0.386 & 0.827 \end{bmatrix}$$

As already stated, U_{norm} carries the cluster information. For deriving the membership of a vertex to a cluster the eigenvectors $2, \ldots, k$ have to be contemplated [17]. The eigenvectors 2 and 3 of the toy example's L_{sym} are displayed in Figure 1.10.

Thresholding the second eigenvector at 0 (dashed line) [17] shows that vertex 1, 2 and 3 belong to one cluster and the rest of the vertices to another cluster. This leads to the red cut in Figure 1.11. Eigenvector 3 returns the second cut. By thresholding the third eigenvector at 0 the cut between the third and fourth vertex is already known. However, the vector contains all the missing information needed for the clustering. As vertex 6 is not connected with vertex 5 and 7, the graph has to be cut a second time between vertex 9 and 10. This is displayed by the blue line in Figure 1.11

Finally, the spectral clustering algorithm leads to the clustering in Figure 1.11 on the right side. While clustering simple networks - such as presented in the toy example - can be done manually, practical applications which include more vertices and edges require algorithms like k-means.

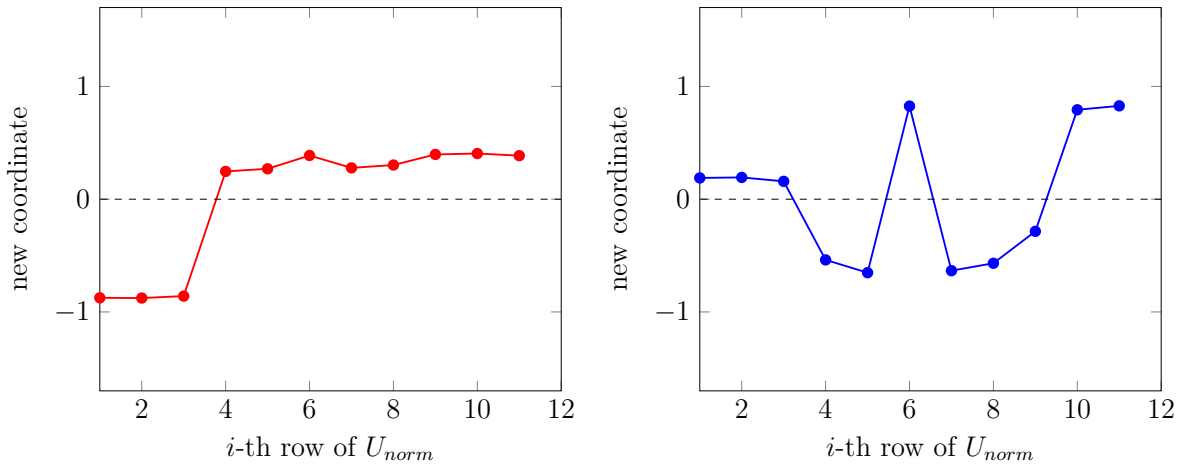


Figure 1.10.: Second and third eigenvector of the toy example

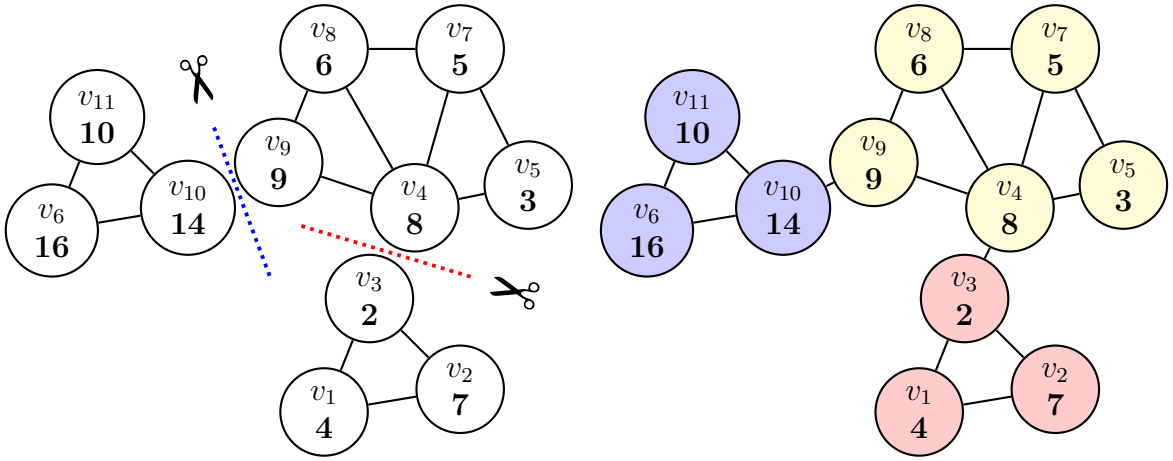


Figure 1.11.: Cut derived by second (red) and third (blue) eigenvector; on the right side: final result of the spectral clustering of the toy example

It can be stated, that spectral clustering is an appropriate approach for clustering based on vertex attributes and also considering the topology of the grid. Thus, spectral clustering is chosen for the clustering process in Chapter 2.

1.2.4. Choice of the edge weights

Choosing the edge weights, respectively choosing vertex attributes in the literal sense (as edge weights are determined by vertex attributes, see Section 1.2.3), has a major influence on the results of spectral clustering. As described in the following section, edge weights may take various shapes, however they should be chosen regarding the desired result of the clustering.

Unweighted graph

In some cases, it can be sufficient only to reveal the pure topology of the graph. This can be achieved by weighting all edges with the same value, for instance 1 (as the values are all equal it does not matter which value is selected). The mathematical expression is shown in Equation 1.5:

$$w_{ij} = 1, \ \forall (ij) \in E \tag{1.5}$$

Electrotechnical characteristics

A natural choice of edge weights can be based on electrotechnical parameters. With the intention of revealing the connection strength in the grid line, admittance is a reasonable choice of edge weights. The admittance of a line indicates the 'electrical distance' between vertices as a high admittance refers to low losses. A possible function for edge weights based on admittance criteria is set up in Equation 1.6 [4]:

$$w_{ij} = Y_{ij} = \frac{1}{|R_{ij} + jX_{ij}|}, \ \forall (ij) \in E \tag{1.6}$$

However, choosing admittance as edge weights returns no information about the grid utilization which is an important indicator in the context of clustering a power grid.

A further parameter in terms of electrotechnology is the power flow P_{ij} between vertex i and j. In contrast to admittance, the power flow implies information about the grid utilization. With regard to grid congestion, edge weights based on power flow do not make a reliable statement either. This results from the fact that grid utilization is contemplated, but it cannot be stated how critical the congestion is in terms of additional transmission capacity anyway.

Locational marginal pricing

Another approach of weighting edges is weighting according to marginal generation costs at the vertices of a grid. The theory behind this idea is explained in this section.

Marginal costs are defined by the change of economic costs linked to a specific increase of the output (mostly one unit increase of the output). Mathematically, marginal values

are associated with the first derivation. However, in practical applications the per unit change is more meaningful due to the indivisibility of, for instance, power plants [20].

In liberalized electricity markets, electricity has to be considered as a good which can be bought, sold and traded with temporal and spatial varying values. The time-dependent price of a unit of electricity takes the operating and capital costs of generating and transmitting the electricity into account [21]. The intention of locational marginal pricing (LMP) is to encourage an economic use of electrical energy with regard to interdependencies between generation and transmission. The most detailed form of LMP is called nodal pricing. By applying nodal pricing to an electrical grid, the spot prices (current price for electricity) for each vertex of the grid have to be calculated. In this context, vertices can be interpreted as individual markets (linked by the grid), where ad hoc market prices are determined simultaneously. If the vertices of the grid are taken in isolation, the locational value of the electricity is a function of operational costs. Consequently, the locational value of the electricity is dependent on the type of installed generators at the vertex. Thus, it is very likely that vertex prices within a power grid differ. As already stated, vertices can be interpreted as individual markets. By this interpretation, and based on the hypothesis that there are no transmission capacity limitations, the conclusion can be drawn that vertices tend to trade electricity until their marginal costs reach harmonization [22].

On the contrary, if a real grid is assumed (with capacity limitation between vertices), and marginal costs of linked vertices are still unbalanced, a strong indication of a line capacity limitation between those vertices is given. Consequently, differences in marginal costs can be grasped as an incentive to increase the line capacities.

With regard to this work, marginal costs seem to be a possible vertex attribute which reflect important vertex properties like power plant stock and also describes the transfer ability of the connected links.

1.3. Derivation of the research questions, objective and organization of this work

Based on the motivation, and on the state of research, two central questions for this work are derived.

- Which configuration of spectral clustering is suitable for aggregating a power grid?
- What impact does an aggregation based on marginal costs of a power grid have on the accuracy of the result compared to a highly resolved power grid?

The objective of this work is to develop a methodology for a spatial clustering of power grid systems. Besides the development, the methodology is implemented in an optimizing energy system model and also applied to a case study based on the German transmission grid.

Chapter 2 reveals the underlying methodology for the applied clustering. As spectral clustering provides many different configurations in terms of the choice of its parameters, the evaluation of the best choice for the underlying power system is described in detail. Another part of Chapter 2 outlines the parametrization of the German transmission

grid. Besides the data collection, the data processing from the raw data to REMix-interpretable data formats is also covered.

In the following Chapter 3, REMix outputs like electricity mix, network utilization or system costs, always comparing a varying numbers of clusters with a reference scenario, are investigated. Furthermore, the computation time for different levels of aggregation is evaluated to see how aggregation can lead to an acceleration of REMix computations.

In the last Chapter 4 the findings of this work are summarized. Moreover, shortcomings of the applied methodology are pointed out. Based on the shortcomings possible further developments are listed.

2. Methodology

This chapter gives an overview on how spectral clustering can be applied in terms of clustering a power system. By implementing spectral clustering, there exist many different variations. These variations relate to the choice of the edge weight function and the Laplacian matrices. To make an appropriate choice, case-related criteria as decision support have to be set up. Figure 2.1 shows the structure of the entire clustering process.

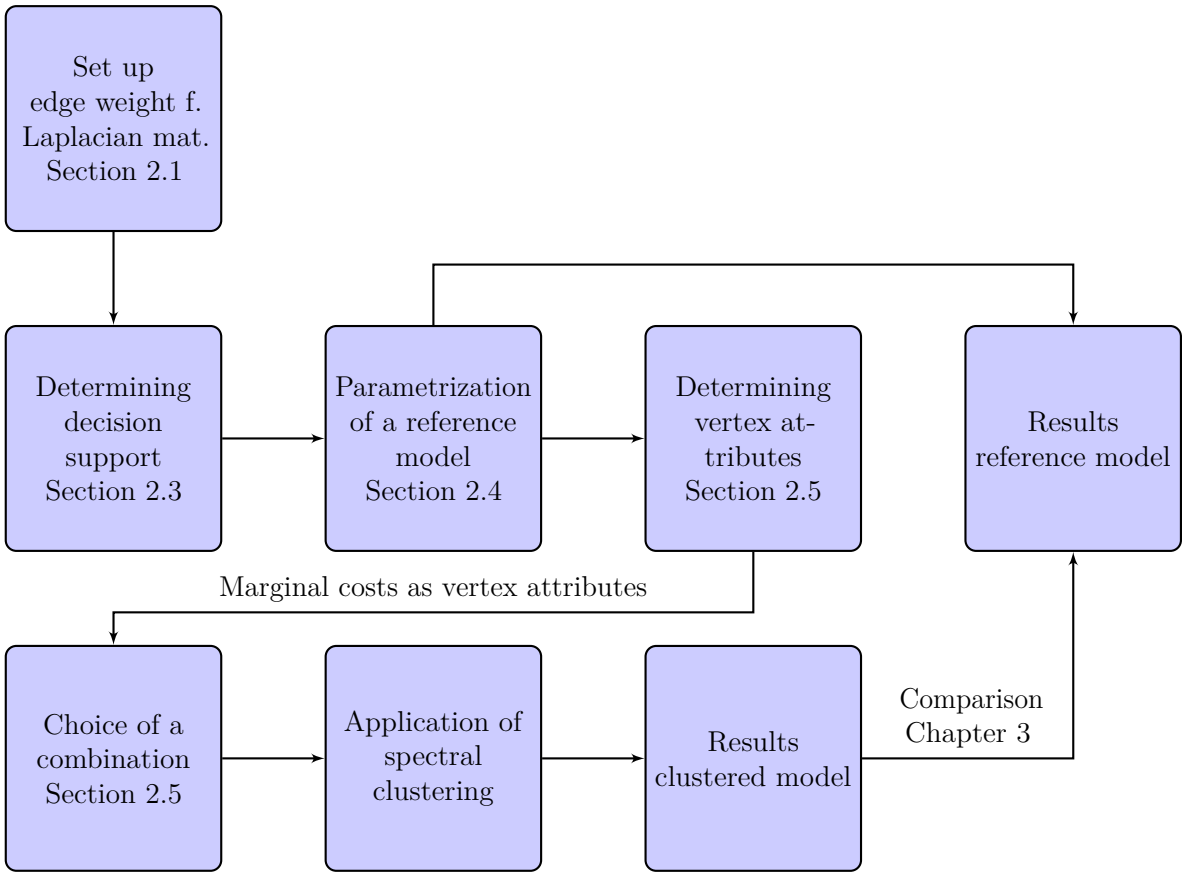


Figure 2.1.: Schematical structure of the clustering process

The choice of the edge weight function plays an important role for the clustering. By defining the edge weight function, it has to be considered, which information are desired to gain out of the clustering, as low weighted edges are likely to be removed. How an edge weight function can be determined that takes into account grid congestion, is elaborated in Section 2.1.

The choice of the edge weight function is associated to the desired results, whereas the choice of the Laplacian matrix is not that clear, as there exists no reliable advice in the

literature which Laplacian matrix has to be taken. In other words, the clustering has to be executed with different edge weight functions and Laplacian matrices in combination in order to make a choice. The different combinations are shown in Section 2.2. To support the choice, a parameter has to be defined to validate whether a combination (edge weight function/Laplacian matrix) provides a good clustering. The definition of this parameter refers to Section 2.3.

As Laplacian matrices are dependent on the used edge weight functions, and edge weight functions themselves are based on vertex attributes, attributes for each vertex have to be calculated in order to execute the validation of the possible combinations. For this reason, a spatially high resolved reference model is needed. The parametrization of the reference model is a further part of this chapter (see Section 2.4). Besides basic information about the different participants of a power system, data sources and the underlying data process for transferring the parameter data into REMix are described.

Based on the computed vertex attributes, the spectral clustering algorithm can be applied to the different combinations of edge weight functions and Laplacian matrices to validate which combination fits best in terms of the defined parameter of the decision support. This validation process is evaluated in Section 2.5.

After the evaluation, the actual spectral clustering process can be executed, as then the choice of the edge weight function and also the Laplacian matrix is done. The algorithm is executed with different number of clusters k which returns different levels of aggregation of the power system. The result of the REMix computations of the aggregated power systems are furthermore compared with the results of the reference model in Chapter 3. With results, different outputs of the REMix computations are meant. For instance, the power plant utilization, temporal resolved power plant dispatch and also system costs are investigated.

2.1. Determination of an edge weight function

As mentioned previously in Section 1.2.4, the choice of the edge weights has to be made with respect to the desired information, as the topological structure of the graph does not capture any functional indicators about the power grid. Basically, the goal of the clustering is the aggregation of the model and hence the reduction of the complexity, while preserving the research object. Since this work focuses on the aggregation of similar grid substructures to detect grid congestion, an edge weight function which takes into account the grid related vertex similarity has to be defined. In this case, grid related vertex similarity means how similar two vertices are in terms of the utilization of their connecting link. More precisely, a low utilization refers to a high similarity and a high utilization, especially an overutilization, leads to a low similarity. As already stated in Section 1.2.4, vertex marginal costs illustrate this similarity behavior.

Obviously, the relative power flow on a link (relative to its maximal capacity) could also be taken. The main advantage of marginal costs compared to relative power flow is that marginal costs also indicate how weighty the overutilization is, whereas the relative power flow returns just a binary information in terms of overutilization.

By allocating marginal costs on vertices, an obvious choice of the edge weight function is the difference of marginal costs between two vertices i and j. This can be expressed according to Equation 2.1:

$$w_{ij} = \Delta M_{ij} = |M_i - M_j|, \ \forall (i,j) \in E \tag{2.1}$$

In principle, Equation 2.1 takes into account a similarity measure based on marginal costs. However, to map the behavior described above, Equation 2.1 has to be modified, as at this point a low utilization (a low difference of marginal costs) would lead to a low connection strength respectively to a low edge weight w_{ij} (see also Section 1.2.3). The desired mapping can be expressed as the reciprocal of the difference of marginal costs (Equation 2.2):

$$w_{ij} = \frac{1}{_{ij}} = \frac{1}{|M_i - M_j|}, \ \forall (i,j) \in E \tag{2.2}$$

Considering the marginal costs difference, the desired similarity behavior can be achieved, as low differences of marginal costs lead to an high edge weight w_{ij}. In the case two marginal costs are equal, Equation 2.2 would lead to division by zero, which is probably for most of the programming environments a problem. This case is intercepted by a try/except statement where in case of a 'ZeroDivisionError' the belonging edge gets the maximum value of edge weights where marginal costs are not equal.

Another modification of the edge weight function with regard to the mapping of low price differences to high edge weights is set in Equation 2.3:

$$\begin{aligned} m &= max|M_i - M_j| \\ w_{ij} &= m - |M_i - M_j| + 1 \\ \forall (i,j) &\in E \end{aligned} \tag{2.3}$$

In the first step all the price differences are calculated. The variable m refers to the maximum price difference of all the links. In the second step all the origin price differences are subtracted from the maximum value and then added with 1 as edge weights with weight 0 refer to no existing link between two vertices. Consequently, the edge with the lowest price difference is mapped with the highest edge weight and vice versa.

2.2. Combinations of edge weight functions and Laplacian matrices

In Luxburg et al. [17] it is stated that there exists a whole field of studying different ways of creating Laplacian matrices. It is furthermore pointed out that a general statement which Laplacian matrix to use cannot be made. Due to the advice given by [17] (see also Section 1.2.3) using unnormalized spectral clustering should be avoided. Furthermore, they advocate to use L_{rw}, as L_{sym} can lead to undesired artifacts. In contrast, Sanchez et al. [4] use L_{sym} for their approach, however, they do not cluster by k-means afterwards,

but apply a hierarchical clustering using a dendrogram. In principle, the computational expense (in terms of coding) is quite equal concerning L, L_{rw} and L_{sym}.

In combination with the edge weight function introduced in the previous Section there exist $2 * 3 = 6$ possibilities of combining edge weight functions and Laplacian matrices:

- $w_{ij} = \frac{1}{|M_i - M_j|} \longleftrightarrow \mathrm{L}_{rw}$
- $w_{ij} = \frac{1}{|M_i - M_j|} \longleftrightarrow \mathrm{L}_{sym}$
- $w_{ij} = \frac{1}{|M_i - M_j|} \longleftrightarrow \mathrm{L}$
- $w_{ij} = m - |M_i - M_j| + 1 \longleftrightarrow \mathrm{L}_{rw}$
- $w_{ij} = m - |M_i - M_j| + 1 \longleftrightarrow \mathrm{L}_{sym}$
- $w_{ij} = m - |M_i - M_j| + 1 \longleftrightarrow \mathrm{L}$

Based on the underlying graph the possible combinations are shown in a decision tree in Figure 2.2.

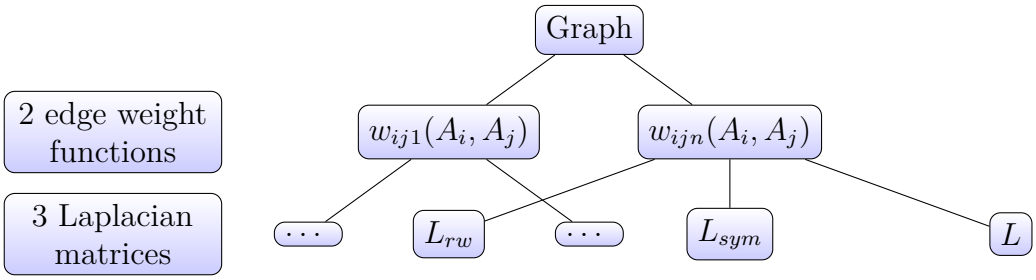


Figure 2.2.: Decision tree for the choice of the edge weight function and the Laplacian matrix

To determine which combination fits best to the clustering desired in the context of this work, a parameter for the decision support has to be set up. This is elaborated in the subsequent Section.

2.3. Decision support

Two criteria to support the decision of which edge weight function and Laplacian matrix combination has to be chosen are set up. The aggregation criterion is a sufficient criterion (the clustering would also work with bad aggregation results), whereas the consistency of the clustering results can be considered as a necessary criterion (the clustering would not work with inconsistent clusters).

2.3.1. Aggregation of overloaded links

To measure how good a combination of an edge weight function and a Laplacian matrix provides a clustering in the desired way, an indicator has to be introduced. As in the context of this work one goal is to identify grid congestion the applied indicator is derived from the number of overutilized links. The method can be described as follows:

- Identification of all overutilized links in the period under review
- Identification which of the overutilized links are lost by aggregation inside a cluster
- Determination of the ratio between aggregated overutilized links and overall overutilized links

Transferred to a mathematical expression the method leads to Equation 2.4:

$$r(k) = \frac{E_{(agg,ou)_k}}{E_{ou_{tot}}} \tag{2.4}$$

The ratio r and also the number of aggregated overutilized links is indexed by k, as the equation must always be seen in combination with a specific clustering. A low number of aggregated overutilized links indicates a good clustering, as it indicates that a low number of overutilized links vanish within the clusters obtained.

2.3.2. Consistency of the clustered graph

Besides the number of aggregated overutilized links, the consistency of the clustering has also to be contemplated by the choice of an edge weight function/Laplacian matrix combination. A shortcoming of spectral clustering in combination with k-means is the specification of the desired number of clusters k (see also Section 1.2.3). This property may lead to undesired artifacts as for a specified number of clusters the situation can occur that there exists no partition regarding to the solution of the cut problem (see Section 2.5). This means that there can exist clusters containing vertices, which are not connected among themselves. This connectedness is meant by consistency in terms of this work. As already mentioned, consistency refers to the inner connection of the clusters generated by the clustering. To put it in simpler terms: is there any vertex in a cluster which is not connected to the rest of the cluster, respectively is it possible to start at any vertex of a cluster and reach every other vertex of the cluster without moving through another cluster? Thus, it can be stated that a clustering is consistent if all the clusters of a clustering consist of only one connected component.

From a mathematical point of view, the number of connected components of a graph refers to the multiplicity of the eigenvalue $\lambda_i = 0$ of its Laplacian matrix. To execute the consistency check, the Laplacian matrices of all the clusters of a clustering are set up and in a further step all their eigenvalues are computed. If for a given clustering all the eigenvalues with value 0 of a cluster have multiplicity 1 (refers to 1 connected component) the clustering can be considered as consistent in terms of this criterion.

2.4. Parametrization of a reference model based on the German transmission grid

As already stated in Section 1.2.4, marginal costs seem to be a reasonable choice for vertex attributes. In order to determine marginal costs with REMix, a parametrized power system model has to be set up. The power system referred to this work consists of the following components:

- Vertices
- Links (connection lines between vertices)
- Loads
- Conventional power plants
- RE power plants
- Pumped storage

In the following subsections the parametrization of each of the components is described. The year of study is 2014. This means that loads, power plants and pumped storage are based on the state of the year of 2014. Apart from the choice of the underlying data (the whole data for the parametrization can be seen in Appendix A.1), the preparation and the processing of REMix input files is covered. In the following the term 'link' is used instead of 'edge', as in the context of power systems the term 'link' is more common then 'edge'. The term 'edge' refers to a mathematical consideration of graphs.

2.4.1. Transmission grid vertices and links

Vertices and edges are the basis for the parametrization of the reference model. For the first time, spatially resolved data (vertices and edges) of the German transmission grid is provided by SciGRID [23]. SciGRID represents a research project of the German energy supplier EWE. The overarching goal of the project is to provide an 'Open Source Reference Model of European Transmission Networks for Scientific Analysis'. For the moment, only the data of the German transmission grid is available and therefore the parametrization of the case study is based on that. The data itself consists of two csv-files (one for vertices, one for links) with characteristics like latitude and longitude of the vertices or transmission properties like capacity P or linked vertices of the links. The spatial and topographical information given by SciGRID is visualized in Figure 2.3.

The 'onshore' vertices are almost complete, however there are 'offshore' vertices missing. 'Offshore' vertices are not about the converter stations, but concerning the wind farms only, as in this case, offshore wind farms are linked directly to the landing points and not looped through the converter stations. In Figure 2.3, the augmentation of the SciGRID data set is displayed in red. The augmentation includes offshore wind farms in the North Sea (BARD 1, e.g.), wind farms in the Baltic Sea (Baltic 1 and 2) and a new landing point near Emden as well. The augmented data is extracted from *4C Offshore* [24].

Additionally to the augmentation, the SciGRID data set is reduced either. A very small number of vertices not linked to the grid are removed as well as the DC link between Herrenwyk (Germany) and Kruseberg (Sweden) as they have no real impact relative to the investigated power system. A further manipulation of the SciGRID data set in terms of reduction is the aggregation of multiple links between vertices. In this case aggregation means that the transmission capacities of the single links are summed up, and also the single links themselves are summarized to a resulting link. This reduction

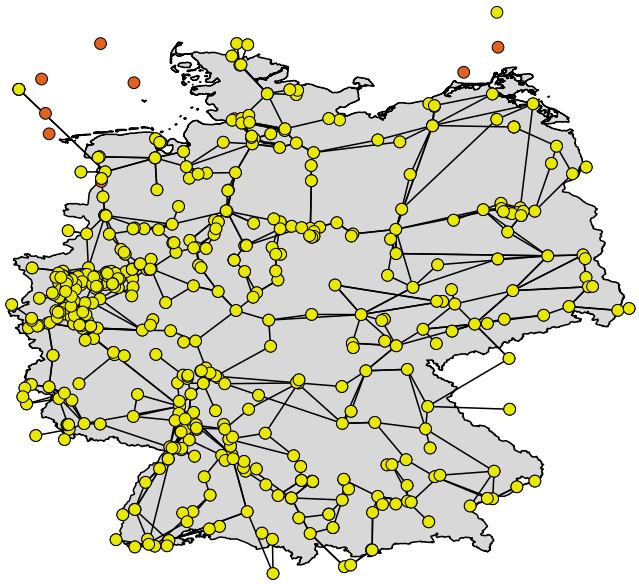


Figure 2.3.: The SciGRID data set of the German transmission grid plotted by QGIS (yellow), augmented offshore vertices in red

is necessary, as by using spectral clustering an unreduced graph would lead to erroneous results.

Besides the topological information of the grid (which vertex is connected to which vertex), the parametrization of the REMix grid module requires the link capacities P and also a pseudo link length L_x. This pseudo link length is defined in Equation 2.5

$$L_x = x * l \tag{2.5}$$

where x is the reactance factor of the link in $\frac{\Omega}{km}$ and l is the real link length in km. Thus, the 'link length' insert in REMix is actually the total line reactance L_x. For further information concerning the calculation of L_x, the SciGRID manual [23] is referred.

Apart from the AC links, the DC links have to be parametrized as well. [24] provides information about the length of the DC links from the offshore wind farms to the landing points. Due to the lack of information concerning the transmission capacity of the DC lines, the assumption is made that the DC capacities are infinite. The idea behind this assumption is that these lines are exclusively built to transfer the generated power of the wind farms and are therefore dimensioned adequately.

2.4.2. Mapping of loads, conventional generators, fluctuating generators and storage on vertices

As set out in Section 2.4.1, vertices and links are the basis for the parametrization. For a spatially resolved power system, the regional allocation of other components (power plants, demand) is needed as well. As the spatial information of the loads, generators and storage do not contain any information linkable to the SciGRID data set, geo-referencing is not straightforward. The approach made in this work is derived from the available data loads, generators and storage all have: postcodes. The linking process is schematically shown in Figure 2.4.

Figure 2.4.: Mapping of the loads, generators and storage on SciGRID data set

The linking pin between the SciGRID data set and the underlying data of the other participants is a postcode data base of Germany [25]. This data base includes latitude and longitude of its geographical center for every postcode area in Germany. In a first process step, every postcode area B_j is mapped on the geographical closest vertex. The distances between the centers of the postcode areas and the coordinates of the vertices are calculated by the Haversine formula. The formula is depicted in Equation 2.6

$$d = 2 * r * arcsin(\sqrt{\sin^2(\frac{\phi_2 - \phi_1}{2}) + \cos(\phi_1) * \cos(\phi_2) * \sin^2(\frac{\lambda_2 - \lambda_1}{2})}) \tag{2.6}$$

where ϕ_1 and ϕ_2 are the latitudes of point 1 and 2 in radians, λ_1 and λ_2 the longitudes of point 1 and 2 in radians and r the radius of the earth. In this case point 1 stands for the centers of the postcode areas and point 2 for the vertex coordinates. The mapping hence follows the algorithm displayed in Equation 2.7:

$$\begin{aligned} & B_j \longmapsto v_i \\ s.t. \quad & \min_{v_1 \dots v_i}(d(lat(B_j), lon(B_j), lat(v_i), lon(v_i))), \qquad \forall B_j \end{aligned} \tag{2.7}$$

For the better understanding what has been mathematically formulated in Equation 2.7, the mapping is verbalized: the distance from a postcode area center to each vertex in the SciGRID data set is calculated and based on that the postcode area is mapped on the vertex with shortest distance. Finally, every postcode area is mapped on a vertex. By this the geo-referencing of the loads, generators and storage can be executed as the postcode information given by their data sets are linked to SciGRID's data set.

2.4.3. Power demand

Contrary to the parametrization of generators and pumped hydro storage, there exists no detailed data base for spatially resolved loads. However, the total load curve of Germany created by Paul-Frederik Bach [26] is available. According to Bach, the values in this load curve cover around 91% of the total supply. Industrial on-site power supply is excluded, as well as some parts of the German railway system.

Furthermore, the number of inhabitants I of the postcode areas and the total number of inhabitants of Germany are obtainable. Therefore it is possible to determine a per capita power demand (Equation 2.8):

$$E_{percapita,a} = \frac{E_{tot,a}}{I_{tot}} \tag{2.8}$$

As the postcode areas are already mapped on the SciGRID data set, the number of inhabitants can also be mapped. In a further step the yearly power demand per vertex is calculated (Equation 2.9):

$$E_{v_i,a} = E_{percapita,a} * I_{v_i} \tag{2.9}$$

With this approach, the spatially resolved power demand of the power system can be approximated. The shortcoming of this method is that the vertex power demand is based on average per capita power demand. In addition, industrial and service sectors are not contemplated (but approximated by inhabitants mapping). So in detail the vertex demand can differ from real values if in reality the relation between inhabitants and industrial or service sectors at a vertex is particularly low or high.

Besides the spatial allocation of the power demand, a detailed parametrization also requires temporal resolved load. As this is even more difficult to determine on spatial aspects, the same load curve (ENTSO-E load curve Germany 2015) [26] is allocated to each of the vertices.

2.4.4. Conventional power plants

The choice of the data base for conventional generators is highly linked with the availability of geo-information (rated power values are presumed). The more detailed *Platts* data base [27] does not contain continuous postcode information for the individual power plants. The moment this gap is closed the *Platts* data base is more than a considerable alternative as it covers power plants all over Europe.

The data base used for this work (*Kraftwerksliste der Bundesnetzagentur*) [28] does not cover power plants all over Europe, but provides steady geo-information for conventional power plants in Germany. The data set contains power plants with rated power above 10 MW. Besides coal power plants, lignite, nuclear and natural gas power plants can be extracted from the data set. The information if a power plant refers to combined cycle technology is yet missing. To overcome this issue, the data set of the Bundesnetzagentur has to be linked with a data set provided by the Umweltbundesamt [29] which contains

information if a power plant is using combined cycle technology. Furthermore, it has to be considered that in many cases power plants are listed as single power plant units which have to be aggregated in a further process step.

Due to the preliminary work described in Section 2.4.2, the spatial resolving of the conventional generators is straightforward. As both the data sets of vertices and power plants have postcode information, the merging can be done easily.

As for a very low number of vertices there is not enough installed capacity to cover the demand (occurred by $k = 499$ computations in REMix), one run is executed with enabled capacity expansion of combined cycle gas power plants at these vertices. That run allows REMix to add capacity at the vertices with deficient covering. After the missing capacities are determined, they are added manually at the concerning vertices. A possible reason for the missing capacities are the allocation of the power demand according to inhabitants and also missing generation capacities in the *Kraftwerksliste*. The process is schematically shown in Figure 2.5.

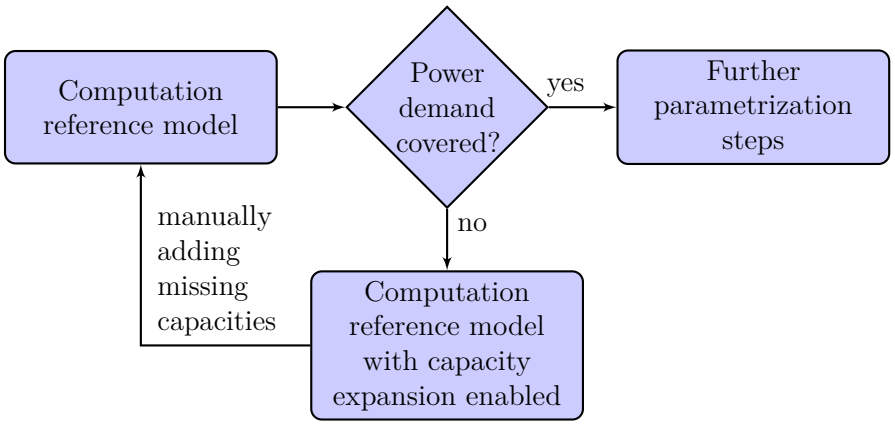


Figure 2.5.: Process of adding missing capacities

After adding the missing capacities, the capacity expansion is disabled. The reference model is computed again (with disabled capacity expansion) in order to check if uncovered power demand still occurs. If this is not the case, the actual process can be continued.

2.4.5. RE and biomass power plants

For the previous parametrization steps, manipulating and processing the data sets with *MS-Excel* is absolutely sufficient. Due to the high number of residential photovoltaic systems, the complete data set of RE power plants is way larger than the data set of conventional power plants. The chosen data set is provided by *energymap.info* [30] and is 363 MB in size with 1.5 million rows (corresponding to the number of installed generators). To deal with such huge data sets *Python* [31] is an appropriate environment. This is elaborated in detail in Section 2.4.9.

The data set itself covers wind onshore generators, photovoltaic modules, hydro power plants and biomass facilities with postcode information and rated power attribute. It has to be stated that this data set is not fully consistent as recent investigations (for

instance by [32]) point out that the data set contains duplicates and also stopped logging in 2013. Nevertheless, the *energymap.info* data set is the best available.

To gain data concerning offshore wind farms, the data set is less suitable because this part is not worked out properly yet. As already mentioned in Section 2.4.1, the data for offshore wind farms comes from *4C Offshore* [24]. The mapping on the vertices can obviously not be done by merging postcodes. In this case, the data sets (SciGRID and offshore wind farms) have to be merged manually, which is due to the small number of wind farms easy to handle.

Analog to Section 2.4.4, the processing of the 'onshore' wind farms input is also straightforward.

2.4.6. Storage

For the parametrization of storage, only pumped storage hydro power plants are contemplated. Since the number of pumped storage in Germany is lower than 40, a data base with postcode information is not necessarily needed. The geo-referencing can be done manually by researching on the internet. On the technical side, two main parameters are needed to integrate the pumped storage into the scenario: on the one hand, the rated power of the converter and on the other hand the storable amount of energy of the reservoir. The power of the converter can also be extracted from the *Kraftwerksliste* while the reservoir data comes from Marcos et al. [33]. The underlying data is listed in Appendix A.1.

2.4.7. Time series for fluctuating energy

In Section 2.4.5, the rated power of fluctuating generators and biomass power plants are determined. It is clear that a further parameter is needed to draw conclusions concerning generated power by RE power plants (for technical parameter see Section 2.4.8 and Appendix A). Fluctuating generators need spatial resolved time series in terms of solar radiation, wind and hydro flow to determine temporal and spatial resolved generated power. The time series are processed in an upstream environment of REMix named EnDAT (see also Figure 1.2). EnDAT provides RE technology potentials and hourly profiles of RE power generation [8], [9]. Biomass potentials are assumed as infinite due to the small amount of installed capacities.

2.4.8. Technical and economic parameters

Finally, as the rated power and potentials are set up, REMix needs further technical parameters like efficiencies or availability of the different power plant types, and also economic indicators (for instance operational expenditure) for determining a cost optimal solution. The technical and the economic parameters are taken from the short study 'Kapazitätsentwicklung in Süddeutschland bis 2025 unter Berücksichtigung der Situation in Deutschland und den europäischen Nachbarstaaten' [34] (see Appendix A.1).

2.4.9. Underlying process structure

The previous sections present the chosen data for the scenario. All the data is avalaible as xls- or csv-file. It is obvious that the data has neither the right structure nor the right data format to be interpretable by REMix. One strategy to transfer the data to a REMix-interpretable format is to use an existing *MS-Excel* file where the data can be manipulated, and also written to a REMix-interpretable dat-file. The main advantage of this strategy is the intuitive handling of *MS-Excel*. On the other hand, the shortcoming of using *MS-Excel* in terms of parametrization is firstly, the lack of adaptability concerning parameter adjustment, and secondly, (and also with the bigger influence regarding this work) the file size limitation. The upper limit of rows in a csv-file openable with *MS-Excel* is a little above 1 million rows. Most of the data needed (Kraftwerkliste, postcode data set) do not approach this upper limit, but, as already mentioned in Section 2.4.5, the underlying data set of RE generators is about 1.5 million rows.

To overcome this issue, the data manipulation, and also the writing of the dat-files is done with *Python*. The complete data processing is pictured in Figure 2.6.

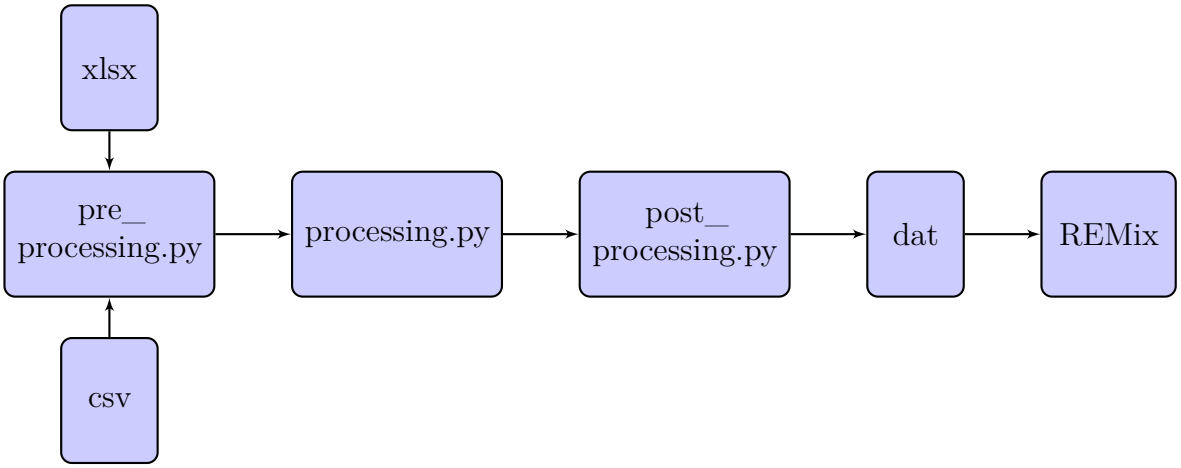


Figure 2.6.: Process structure of the parametrization

The data processing is divided into three parts, respectively *Python* scripts. The pre_processing.py script is fed with the xls- and csv-data containing the data provided by the *Bundesnetzagentur* or *energy-map.info*. In a first step the different technologies (coal, lignite, wind, photovoltaic) have to be separated. For this purpose, the *Python* environment provides a very powerful data analysis library named *Pandas* [35]. If through preliminary work, the corresponding technology terms of the data sets ('Steinkohle' →'coal', e.g.) are known, the data frames can be sliced easily with *Python's Pandas* library. Another task completed by the pre_processing.py script is the assignment of age classes to power plants to map appropriate efficiencies in a later process step. At the end of the pre_processing.py script the data is stored into a hdf-file (hierarchical data format) as it has excellent writing and reading access features, especially in terms of speed.

The hdf-file is then read by the consecutive processing.py script. At this part of the process, the main data manipulation is done. Besides the spatial mapping on the

SciGRID vertices (see Section 2.4.2), the various data frames are created and prepared into a format which is already close to the format REMix is able to read. Like in the previous process step, the data is stored into an hdf-file.

The last process step is done by the post_processing.py script. The script reads in the hdf-file from processing.py and writes out the dat-files interpretable by REMix.

2.5. Evaluation of the edge weight function - Laplacian matrix combinations related to the case study

Based on the parametrization of the high resolution model, it is possible to compute marginal costs for every time and every vertex of the grid. An obvious possibility for determining the marginal costs (and thus the vertex attributes A_i) is to execute the computation of the parametrized scenario for a period of one year to then determine the average marginal costs of the vertices. The problem is that by computing a whole year, the main memory of the available server is not sufficient (Intel Xeon CPU, 96 GB RAM). To overcome this issue, only a few hours of the year are considered. Therefore, eight hours of the year with representing grid usage are identified. They are characterized by their load level (L) and also photovoltaic (P) and wind (W) power generation level. The grid usages cases with associated hours of the parametrized model can be seen in Table 2.1. The minus (−) indicates a relatively low, plus (+) a relatively high value. The conditions low (−) and high (+) are related to the nominated load or nominated supply of photovoltaic and wind. 'Low' and 'high' have no fixed bounds as for some grid usage cases there exist no fitting hour. Instead, this must be considered on a case-by-case basis.

Table 2.1.: Grid usage cases with associated hours based on the scenario

Grid usage case	Load	PV	Wind	Hour
L- P- W-	low	low	low	5474
L- P+ W-	low	high	low	5486
L- P- W+	low	low	high	8619
L- P+ W+	low	high	high	6158
L+ P- W-	high	low	low	7115
L+ P+ W-	high	high	low	2077
L+ P- W+	high	low	high	65
L+ P+ W+	high	high	high	6155

The choice of eight grid usage cases further increases the options for the configuration of the clustering algorithm. By considering a combination of 2 edge weight functions, 3 Laplacian matrices and 8 grid usage cases a total of $2 * 3 * 8 = 48$ possibilities to implement the clustering are evaluated. The compared with Figure 2.2 expanded decision tree is depicted in Figure 2.7. Every level of the decision tree represents another level of the decision-making process.

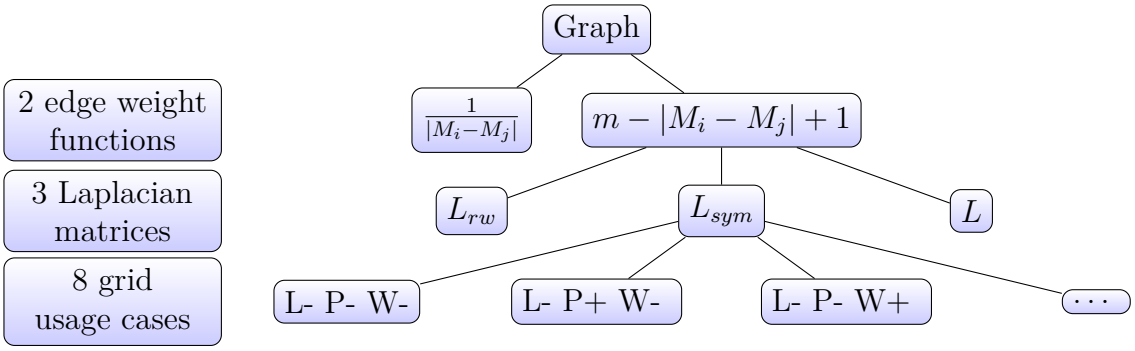


Figure 2.7.: Decision tree of edge weight function, Laplacian matrix and grid usage case - in total 48 cases are assessed.

For a better understanding of the evaluation, and also the clustering itself, the structure of the process is described in the following. The clustering is like the parametrization also processed in *Python*. The interaction of REMix and the clustering is schematically shown in Figure 2.8.

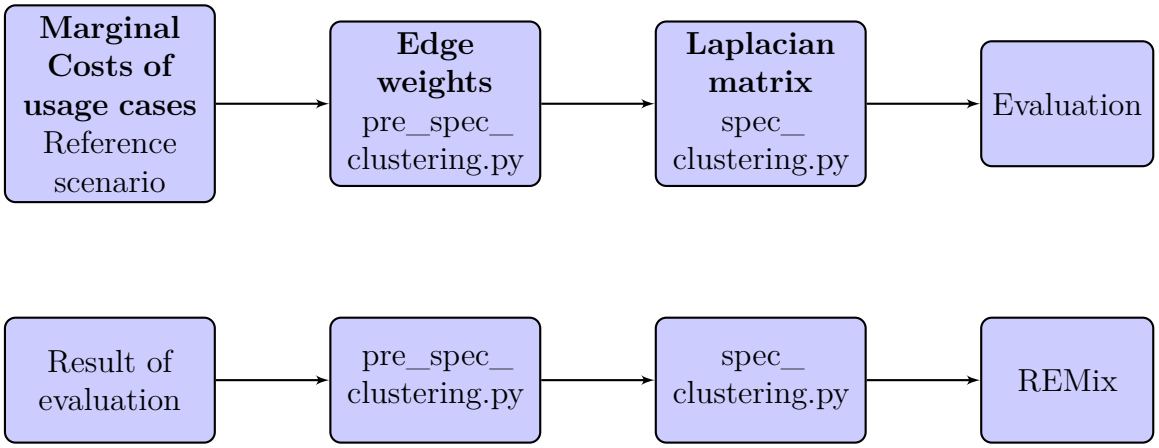


Figure 2.8.: Two levels of the clustering process: evaluation on the upper level, clustering process based on the result of the evaluation on the lower level

For evaluating which set of marginal costs (related to the usage case) are used for the clustering, the clustering is executed with all usage cases. The marginal costs of the REMix computations (for all eight 8 grid usage cases) are passed to the pre_spec_clustering.py where the preliminary work for the clustering takes place. In this step the different defined edge weights (see Section 2.1) of the grid are calculated and prepared for the main process. In the spec_clustering.py the actual spectral clustering is processed. At this part of the process, the different Laplacian matrices (see Section 2.2) are determined. In a further process step, these combinations of marginal costs, edge weight functions and Laplacian matrices are evaluated. As quality criterion for the evaluation the aggregation of overutilized links is contemplated (see Section 2.3). The marginal costs of the different usage cases contain only the overutilization information of their own usage case (see Section 1.2.4). Thus, the best approximation of a usage case to the global overutilization is searched.

Based on the result of the evaluation the clustering is processed with the 'best' (in terms

of aggregation) combination. The clustering is then executed for different number of clusters k. The result of the clustering process is written into a REMix-interpretable file (dat-file) and passed to REMix. In Chapter 3 the output of the computations of the clustered power systems are elaborated.

Considering the evaluation (upper level of Figure 2.8), Figure 2.9 shows an example of the different aggregation $r(k)$ of some selected parameter combinations (shown in the title of the subplots) as a function of the number of clusters k. The shown subplots represent a sample of parameter combinations.

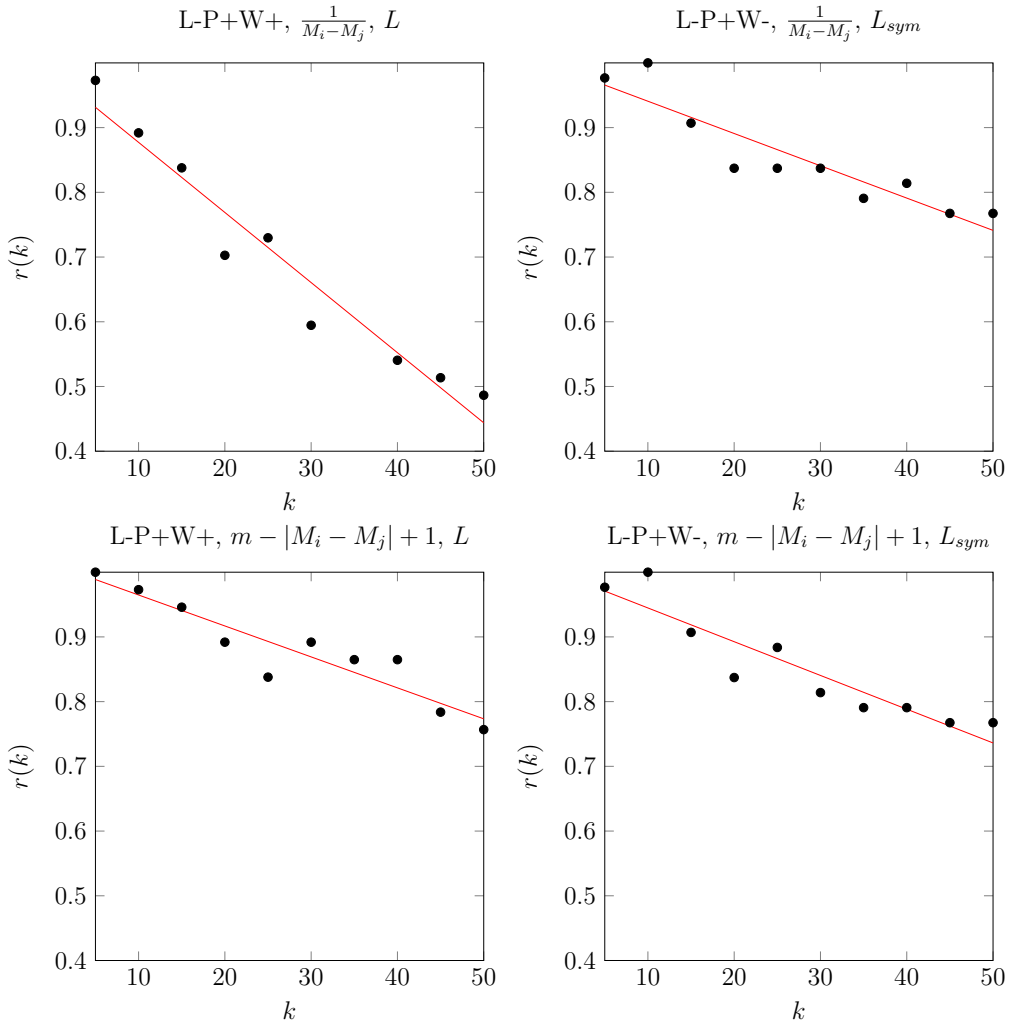


Figure 2.9.: Aggregation of overutilized links using different parameter combinations

Black dots represent the number of aggregated overutilized links in relation to the number of clusters. To visualize the behavior of the parameter combination, a regression curve (red) is included in the graphs. For a behavior to be good in the sense of aggregation, the gradient of the regression curve has to be high. A high gradient indicates that the number of overutilized links decreases considerably with only little increase of the number of clusters. A further indicator for the quality of the clustering is the cut with y-axis. The lower the intersection point with the y-axis, the better the clustering already for a little number of clusters.

The upper row is based on the use of the reciprocal edge weight function, the bottom row uses the difference function described in Equation 2.3. The first column consists of unnormalized spectral clustering results (using the unnormalised Laplacian matrix), whereas in the second column normalized spectral clustering by [18] using L_{sym} is executed. Furthermore, the underlying grid usage cases differ.

An observation that can be made is that the parameter combination of the upper left subplot best fits the aggregation criterion. Besides a low number of inconsistencies, this combination also leads to a high gradient of the regression curve. Hence, this combination (L-P+W+, $\frac{1}{M_i - M_j}$, L) will be used for the clustering. The choice of the usage case is comprehensive because L-P+W+ refers to the lowest residual load. In this usage case, the supply of RE is maximal, whereas the load is minimal, and this usage case has therefore the most overutilized links which is influencing the marginal costs difference between the vertices. Choosing the reciprocal function instead of the difference function is also plausible. By using reciprocals, the edge weights have a wider range of values, and so edges can be weighted with finer graduation.

For the right column (using L_{sym}), it does not make a huge difference which edge weight function is used. This is due to the underlying optimization problem (RCut vs. NCut). Also based on the optimization problem is the reason why L is used instead of L_{sym} or L_{rw}. This is explained in the following part of this section.

A further observation in the upper left subplot in Figure 2.9 reveals missing data, respectively a missing black dot (for $k = 35$). The missing data has its origin in missing consistency of the clustering with the underlying parameter combination. In general (but based on the SciGRID input graph), it can be stated that both L_{rw} and L_{sym} lead to much more inconsistent clustering results than clustering with L.

An assumption for the occurence of inconsistencies is based on the optimized cut problem. In contrast to L, L_{rw} and L_{sym} do not optimize the RatioCut problem but the NCut problem (a justification of the relation between the different cut problems and using the different Laplacian matrices is given by [17]). If the graph $G = (V, E)$ is cut into two disjoint subsets A and B (disjoint: sets with no element in common) the degree of dissimilarity between the two subsets is equal to the weight of the removed edges. Mathematically, this leads to Equation 2.10:

$$cut(A, B) = \sum_{x \in A, y \in B} w(x, y) \tag{2.10}$$

The optimal solution for a partition is the one that minimizes Equation 2.10. This optimization tends to create small sets with isolated vertices, as they only have a low number of edges connecting them to the rest of the graph.

In contrary, the NCut tries to create balanced clusters in terms of the number of vertices in a cluster. The NCut problem is depicted in Equation 2.11:

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, V)} + \frac{cut(A, B)}{assoc(B, V)} \tag{2.11}$$

Here $assoc(A, V)$ represents the total connection from vertices of A to the vertices of the whole graph. By finding the minimum of this equation it is very unlikely that small clusters occur as they will have small connectivity to the rest of the graph, and thus the denominator will get small which, on the other hand leads to huge NCut values [36].

It is assumed that the algorithms of L_{sym} and L_{rw} are tending stronger to create inconsistent clusters, as they have to fulfill the condition described in Equation 2.11. With another (more homogeneous) base graph the results are probably better in terms of consistency, as the German transmission grid consists of many chains, and also of nodes with only one connection to the rest of the graph. By using L, the consistency issue is solved, however this method has another shortcoming. As already mentioned, using L optimizes the RCut, and thus does not tend to create balanced clusters. This can lead to a very unequal allocation of vertices to clusters. Figure 2.10 shows the standard deviaton s of the cluster sizes for the two different parameter combinations in the first row of Figure 2.9.

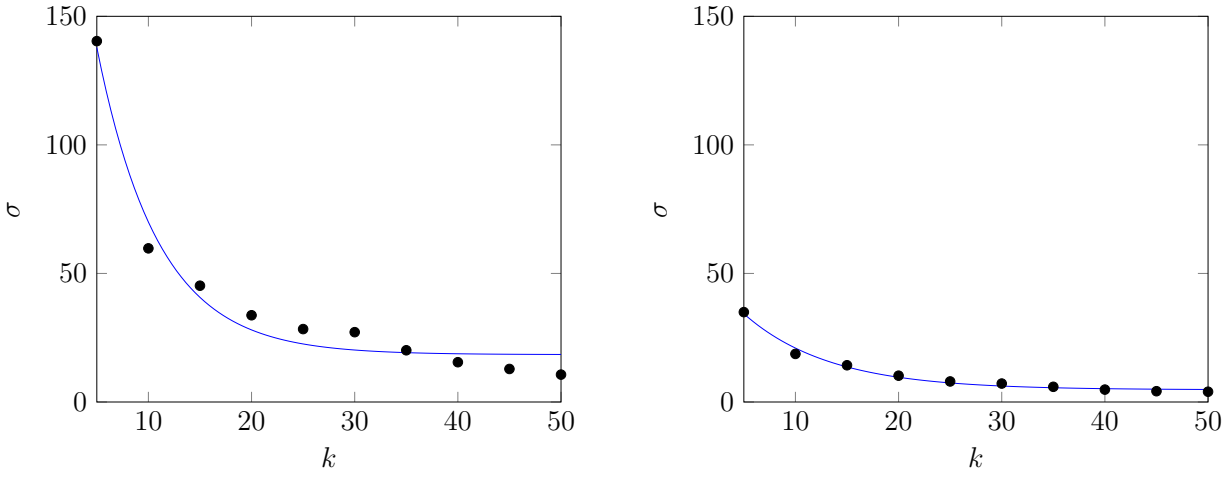


Figure 2.10.: Standard deviation σ of the parameter combinations of the upper row in Figure 2.9

It can be very clearly seen that the combinations using NCut exhibit to a lower standard deviation σ, and thus to more balanced cluster sizes. However, fulfilling the aggregation criterion and consistency is weighted more strongly than the balancing criterion, hence the algorithm based on unnormalized spectral clustering using L is preferred.

3. Results

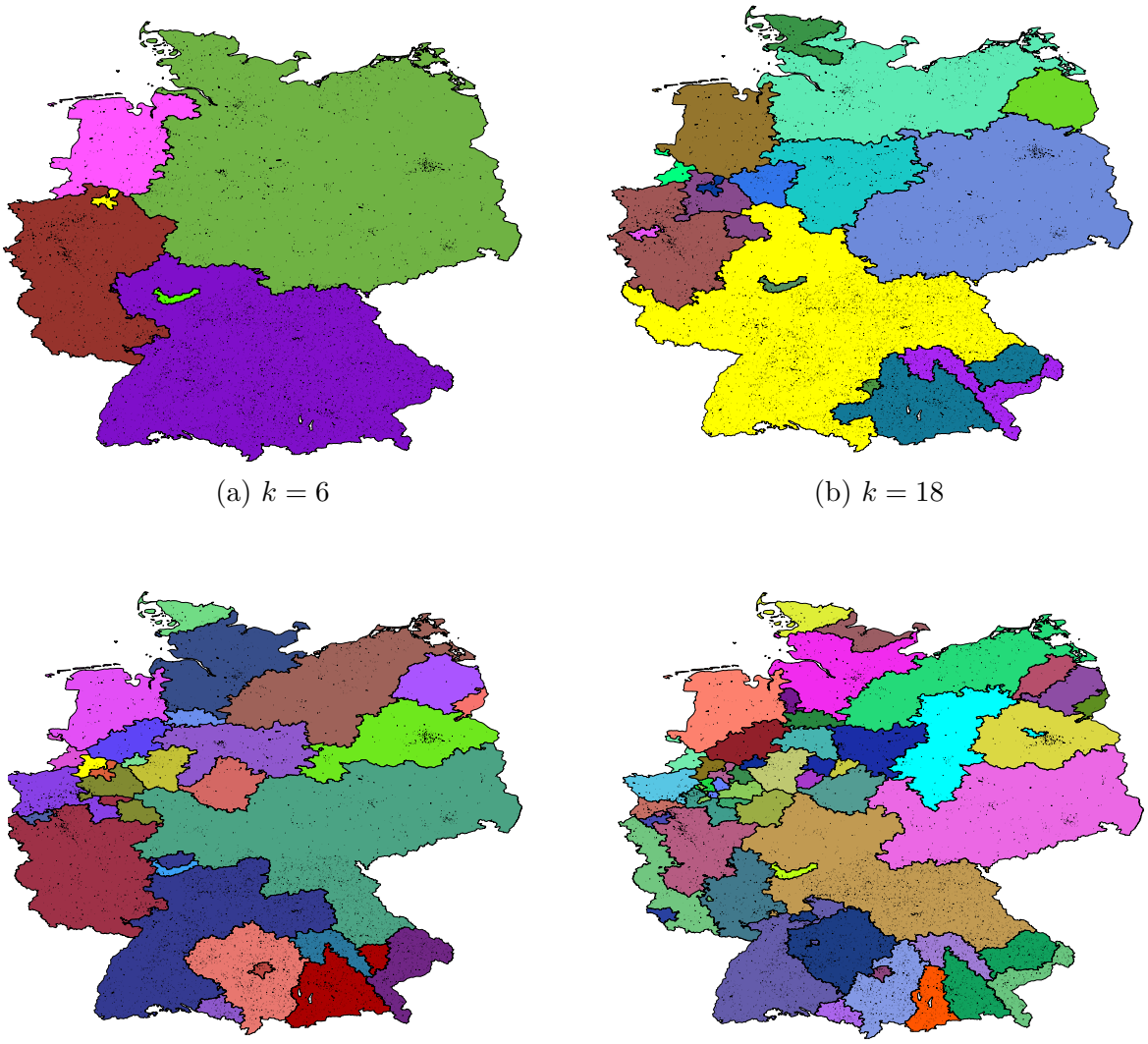
In this chapter, the developed clustering methodology is applied to the German transmission grid. In addition to the visual result of the spatial clustering (maps), the evaluation covers different power supply indicators like power plant and grid utilization or system costs, always comparing the results of different numbers of clusters k, the maximum resolution ($k = 499$) and a 'copper plate' scenario ($k = 1$). All output relies on computations in REMix of a fixed 30-day-period (from day 270 to day 300). This period is chosen because of the high wind supply during the autumn months. High wind supply is accompanied by critical grid situations, due to the high additional power supply. Besides the evaluation of the power supply indicators, the computation time is examined as well. At the end of this chapter a discussion of the applied methodology is examined.

3.1. The clustered German transmission grid

As elaborated in detail in Chapter 2, a parametrized power system based on the SciGRID data set is clustered. For a detailed investigation, the results of different clustering computations (with different number of clusters k) are depicted in Figure 3.1.

In general, due to the model size cause-effect relationships are difficult to comprehend. By investigating the $k = 6$ plot, a partition of the graph in 4 four big clusters and two very small clusters can be stated. Furthermore, the graph is divided into a big north cluster (green) and a south cluster (purple). Since the clustering is based on obtaining transmission capacity limitations, this indicates a grid congestion between the northern and southern part of Germany. This is congruent to the often made assertion that there is not enough transmission capacity to supply the structurally strong demand regions in the South with sustainable wind energy from the North. Another partition, which seems plausible, is the pink north western cluster. This cluster is influenced by low marginal costs due to the installed wind offshore capacities. This cluster also indicates grid congestion between this cluster and the power-intensive industry in Nordrhein-Westfalen (brown cluster).

In the $k = 18$ plot the biggest cluster is located in the south west and includes Baden-Württemberg in total and also parts of Saarland, Rheinland-Pfalz, Nordrhein-Westfalen, Hessen and Bavaria. In the north of Germany there exist two clusters that contain the landing points of the offshore wind farms (brown and mint green). These clusters are influenced by low marginal costs of the offshore wind farms and also by the high amount of installed onshore wind farms. As already in the clustering with $k = 6$, the small cluster containing the Frankfurt a.M. metropolitan area still exists. Comparing the $k = 18$ plot with the *18-Regionen-Modell* it can be stated that some of the clusters are similar.



(a) $k = 6$ (b) $k = 18$

(c) $k = 30$ (d) $k = 50$

Figure 3.1.: Clustering results for different k, different colors refer to different cluster membership, clusters mapped on postcode areas

For instance, the blue cluster in the middle of Germany is similar to the cluster with number 23 of the *18-Regionen-Modell*. Another similarity can be observed in the west of Nordrhein-Westfalen. The result of the clustering as well as the *18-Regionen-Modell* splits this area into clusters. In addition, the big sized clusters in the east of Germany are similar.

Contemplating the $k = 30$ and $k = 50$ plots still shows two clusters including the landing points of the offshore wind farms. This suggests a high influence of the offshore wind farms on the marginal costs in these clusters. Furthermore, the Frankfurt a.M. metropolitan area is still a single cluster like in the $k = 6$ and $k = 18$ plots.

In general, the cluster sizes vary a lot. This behavior is probably linked to the underlying optimization problem (NCut vs. RatioCut). A detailed investigation of the cut problem is examined in Section 2.5. In terms of consistency, some of the clusters seem to be

inconsistent (for instance in the $k = 50$ plot: the 'Berlin cluster' belongs to the big cyan cluster in the west). In fact, they are consistent in terms of their underlying transmission grid vertices. However, the presentation in Figure 3.1 is based on a mapping of postcode areas to transmission grid vertices and there exist links that go over entire postcode ares. Furthermore, it may be assumed that by an increasing number of clusters k, existing clusters (referring to a lower k) would split into a higher number of new clusters. This split-up sometimes occurs, but a general statement about the splitting behavior cannot be made. However, the cluster sizes are more balanced the higher number of clusters k are.

3.2. Power plant utilization

One of the central questions of a clustered power system is the impact of the clustering on the power supply. Since in an aggregated grid the grid restrictions decrease, the power system has more options to cover the demand.

In Figure 3.2, the power plant utilization for different number of clusters k is plotted. On the ordinate the supplied power in TWh is displayed, in each case summed over all clusters for the different number of clusters. The abscissa refers to different number of clusters.

Firstly, the increasing amount of total supplied electricity by increasing number of clusters can be stated. This increase is due to the grid losses. In REMix, grid losses are taken into account by percentage losses per length values. Since through the clustering the number of links and thus the total grid length (cumulated link lengths) change, also the amount of supplied electricity has to change for different number of clusters k, as the grid losses differ as well. In the reference scenario ($k = 499$), the grid losses are highest. Consequently, in the reference scenario the amount of supplied electricity is also highest. By contemplating decreasing number of clusters k, the amount of supplied electricity decreases. The lower the number of clusters, the lower the grid losses, as highly aggregated grids have less links and hence a lower total grid length. Quantitatively, the losses increase from 0.3 % of the total demand at $k = 5$ to 3 % of the total demand at $k = 499$. However, only the transmission grid losses are taken into account, whereas distribution grids are not considered.

Another observation reveals the different shares of the used technologies for the supply. The shares of hydro, nuclear and non-adjustable RE power plants (photovoltaic, wind) are not changing noticeably, whereas the shares of lignite, coal, natural gas and biomass power plants are varying for different number of clusters k. The not changing technologies have the lowest variable costs and thus they are the last power plants to be curtailed. Furthermore, especially the RE power plants consist of small units and are distributed more evenly. This means that increasing grid restrictions have a lower influence on them. The curtailment of the RE technologies is very low for all the scenarios and tends to increase with increasing number of clusters.

The decreasing share of electricity supplied by lignite power plants for increasing number of clusters has its origin in the low variable costs of this technology. For instance in the 'copper plate' scenario ($k = 1$), where no grid restrictions exist, the share of lignite power

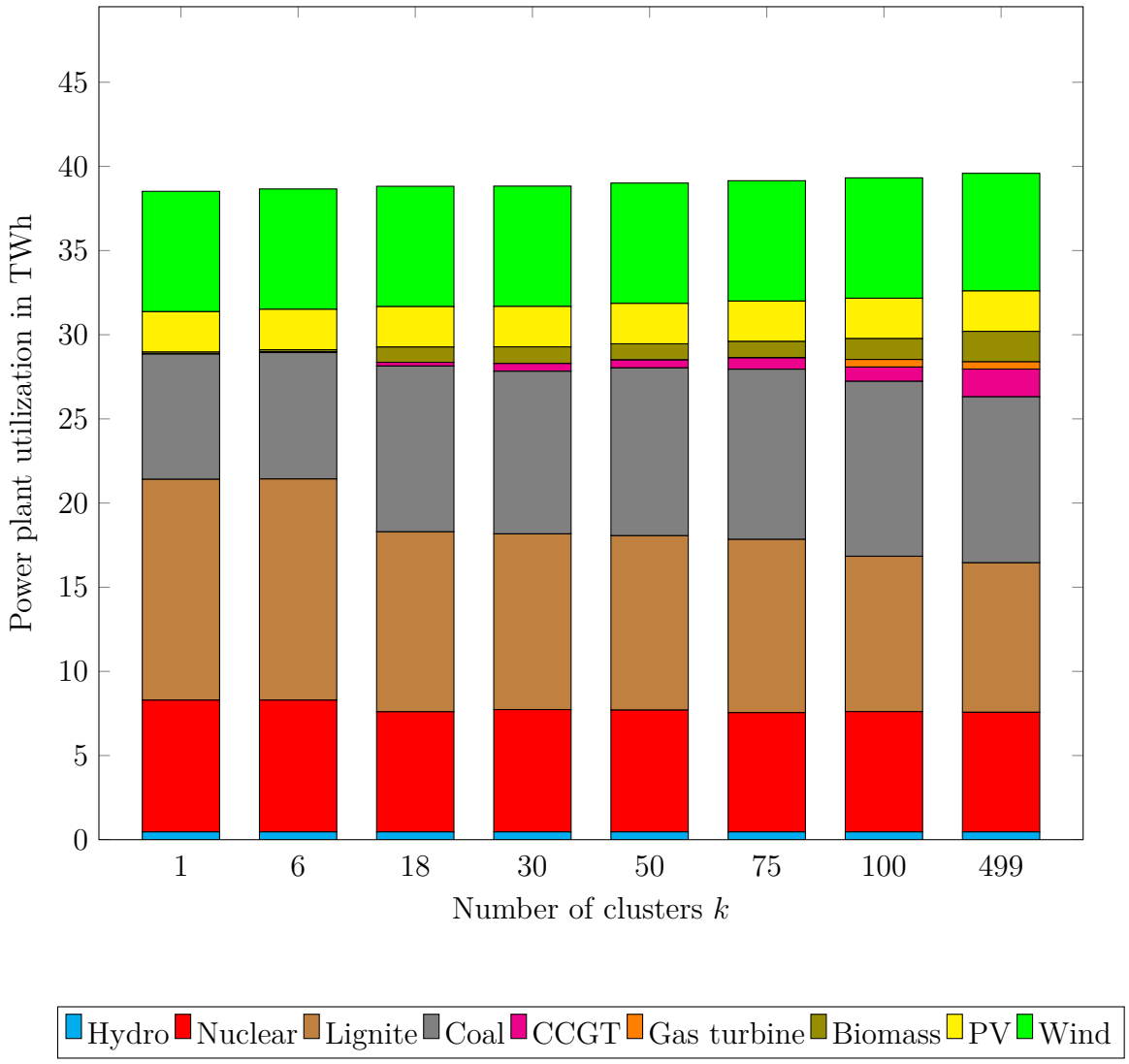


Figure 3.2.: Power plant utilization in TWh for the power supply of different numbers of clusters for the investigated 30-day-period

supply is maximal. This is logical, since REMix determines a cost optimal solution. The reason for the decrease of the lignite power supply by increasing number of clusters are rising grid restrictions and strongly linked to that, the uneven distribution of lignite power plants in Germany. The behavior of the power supply by lignite power plants is also valid for nuclear power plants. Lignite and nuclear power supply both decrease remarkable when increasing the number of clusters k from 6 to 18. The reason for this behavior is a significant decrease of the cluster sizes from $k = 6$ to $k = 18$ (see Figure 3.1).

In contrary to lignite and nuclear power plants, the power supply by coal plants increases with increasing number of clusters k. Due to the better distribution and the higher number of coal power plants (compared to lignite and nuclear power plants), coal power plants cover a high share of the decrease of power supply by lignite and nuclear power plants. The reason for this is that increasing grid restrictions less influence better distributed technologies.

Looking at smaller units and better distributed technologies like natural gas, a contrary behavior can be noticed. As the share of lignite power supply decreases due to the rising grid restrictions, the demand has to be covered with other technologies. As a result, the share of technologies with high variable costs (natural gas, biomass) increases and has its maximum, consequently, in the reference scenario with strongest grid restrictions.

3.3. Dispatch comparison

The previous evaluation is based on summed (over all vertices and all time steps) amounts of power supply. By analyzing the dispatch of different number of clusters k, the so far gained knowledge can be contemplated with time reference. Figure 3.3 shows a dispatch comparison between clustered scenarios and the reference scenario. On the x-axes the time in h, on the y-axes the generated power in GW is plotted. For better visibility only two representative days are plotted.

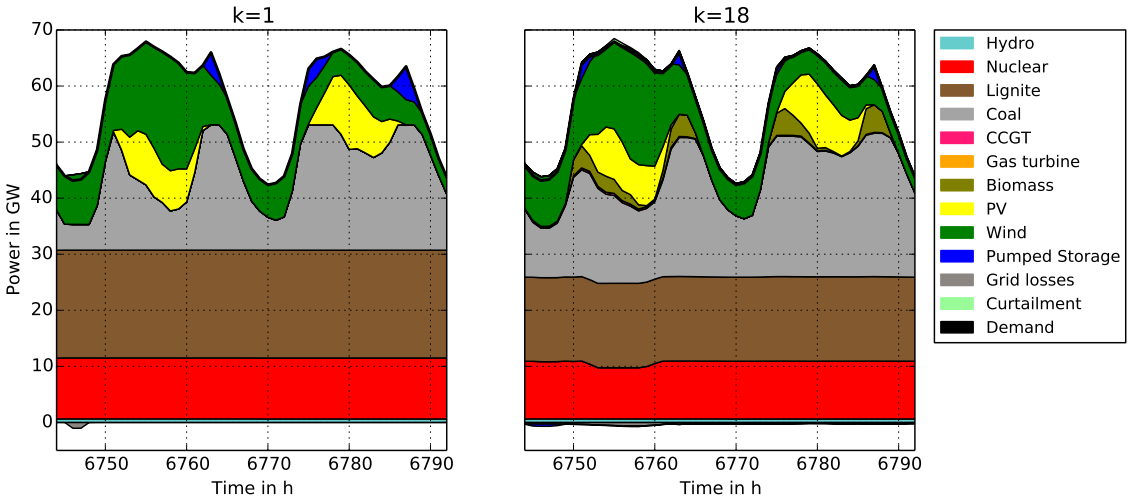
Differences in hourly power plant dispatch between aggregated and high resolved systems can be illustrated most clearly by comparing the reference and the 'copper plate' scenario. Furthermore, the approximation to the high resolved system by increasing the resolution can be seen in the plots with $k = 18$ and $k = 50$.

As already stated in Section 3.2, the amount of power supplied by lignite power plants is significantly reduced if all transmission restrictions are accounted for. Another aspect which is detectable only in a time dependent plot, is the profile of the supply. In a power system with high grid restrictions, lignite and nuclear power supply have to be adjusted to the demand and the supply of RE, whereas without any grid restrictions ($k = 1$) this is not the case. The lignite and the nuclear power supply remain constant over the time, as any spatial demand can be covered with any power plant.

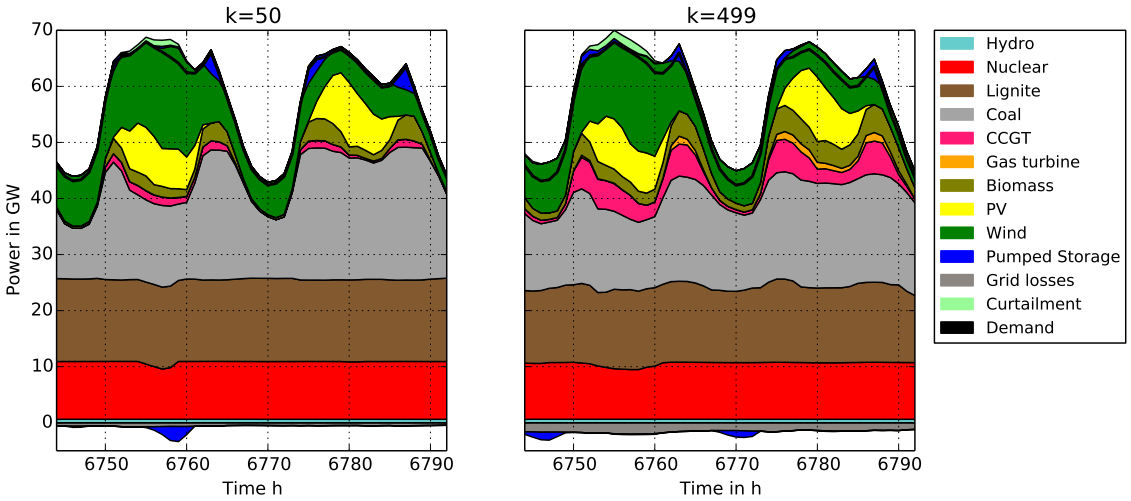
Contemplating the supply by coal power plants reveals that in a maximal aggregated system, coal power plants are used for peak load covering. This leads to a challenging ramping behavior of the coal power supply. Actually, coal power plants do not have start-up characteristics fitting this ramping behavior and are normally used for base load covering. With increasing number of clusters k, the profile of the coal power supply is getting more flat, as biomass power plants are used for peak load covering. Furthermore, with increasing grid restrictions, the utilization of pumped storage hydro power plants are partially substituted by biomass power plants.

Due to the decrease of coal power peak load covering by increasing grid restrictions, also combined cycle gas power plants are used for peak load covering ($k = 50$). The amount of coal power supply is still higher compared to the reference scenario ($k = 499$), whereas the amount of combined cycle gas power supply is lower. Gas turbines are not used at all in the aggregated power systems.

The curtailment of the RE power generation tends to increase with increasing number of clusters k. In the fully aggregated power system, there is almost no curtailment. The reason are missing grid restrictions and thus the power can be used for covering any spatial demand. With increasing grid restrictions and also increasing spatial allocation of the demand, it is not always possible to transfer the RE power to the demand locations.



(a) $k = 1$ vs. $k = 18$



(b) $k = 50$ vs. $k = 499$ (reference scenario)

Figure 3.3.: Comparison of power plant dispatch during 2 October days for different number of clusters k

In the case, the grid restrictions do not allow a transfer, the RE power plants have to be curtailed.

3.4. Power plant ramping

In this section the power plant utilization is analyzed by contemplating ramping behavior. The ramping behavior shows the power plant control. Considering conventional power plants, ramping is associated with costs that are higher for coal, lignite and nuclear power plants then for gas turbines.

The summed ramp-up and -down power for the different number of clusters is depicted in Figure 3.4. The summed ramping in GW is depicted on the y-axis, different number of clusters on the x-axis.

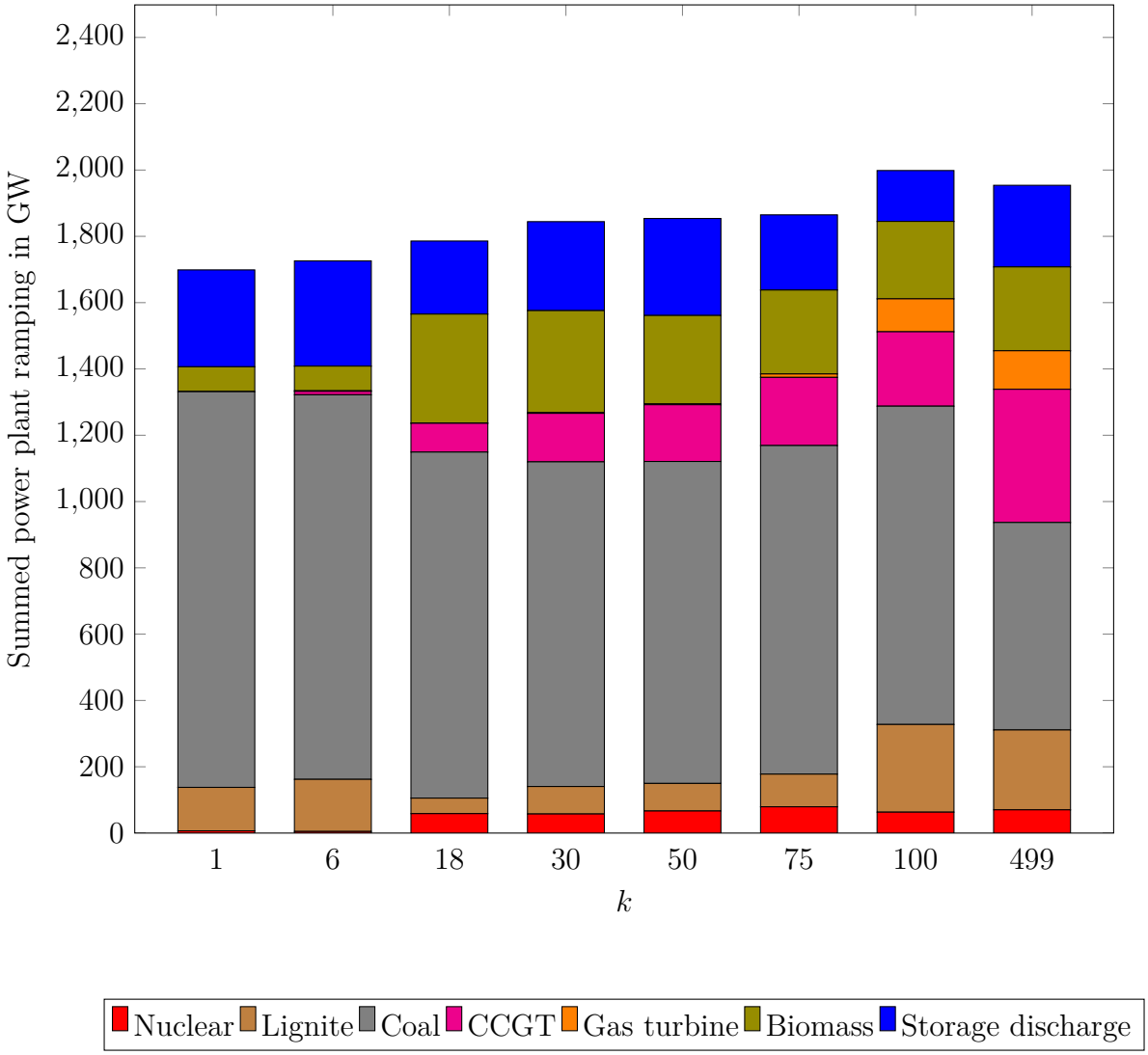


Figure 3.4.: Power plant ramping in GW during the investigated 30-day-period

The knowledge gained through the dispatch (Figure 3.3) can be drawn by investigating the ramping behavior. It can be stated that the ramping tends to increase with increasing number of clusters k, with an outlier at $k = 100$. Contemplating the stronger aggregation (low k) the less ramping is due to the constant operation mode of few, cost-efficient technologies. With increasing number of clusters k the ramping also increases. The reason for this behavior are increasing grid restrictions. In addition, the amount of the

actual peak load covering gas power plants is increasing. By increasing the number of clusters k, the increasing ramping of gas power plants is plausible as well. The reason for this behavior is elaborated in Section 3.3.

3.5. System costs

All in the computation incurring costs are aggregated under system costs. The system costs include the annuities of the overnight investment costs of capacity expansion as well the operational costs of the utility dispatch. The latter consists of fuel, implicit grid usage (implicit due to grid losses), emission certificate as well as operation and maintenance costs (OM). As in the context of this work, dispatch and not capacity expansion is optimized, only the operational costs of the utility dispatch are considered.

Figure 3.5 shows the development of the system costs for different levels of aggregation. On the x-axis the different number of clusters are plotted, the y-axis refers to system costs in Million Euro.

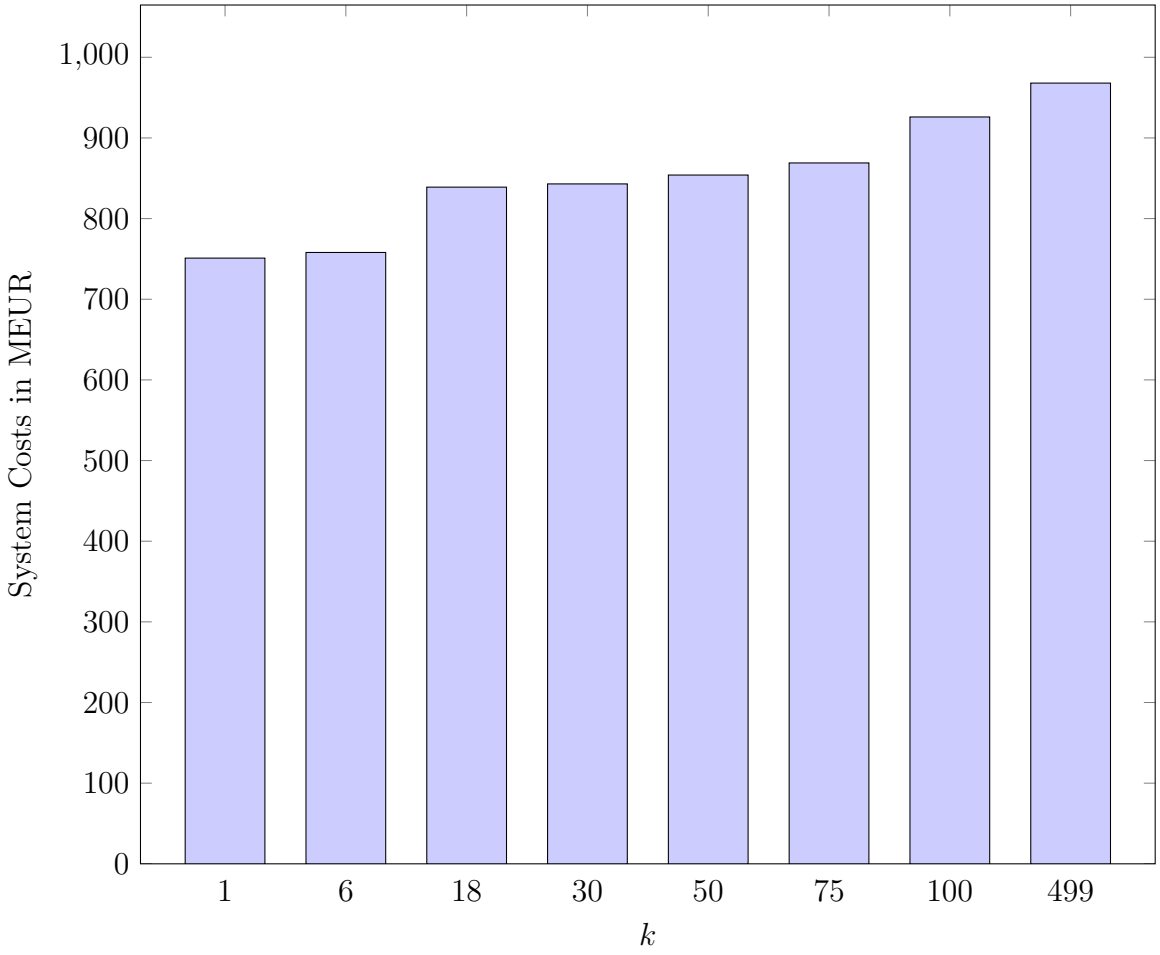


Figure 3.5.: Total system costs in MEUR for different number of clusters k

A first observation reveals the cost trend by contemplating different number of clusters. System costs are increasing by increasing number of clusters. There are various reasons

for this behavior. The most important reason is that the power plant utilization varies for different number of clusters (see also Figure 3.2), due to the different grid restrictions. By low grid restrictions (refer to a low number of clusters) more electricity generated from cheaper technologies can be used. For instance, the share of lignite is partially substituted by more expensive technologies like natural gas or biomass power plants by growing grid restrictions. This leads to the main share of cost difference between different numbers of clusters.

A further reason for increasing system costs by increasing number of clusters are grid losses. As a higher number of clusters is accompanied by higher grid utilization, the grid losses increase the higher the utilization is. Since only the transmission grid and no distribution grids are considered, which would lead to higher total losses, the impact of grid losses on the costs is fairly low.

In summary it can be stated that the 'copper plate' scenario maps only about 80 % of the costs of the reference scenario. By contemplating future scenarios including capacity expansion optimization the results would differ more, as grid and storage expansion plays a key role in this context, in particular on low aggregated levels. For the 'copper-plate' scenario, grids do not play a role at all and storage is also used less than in the reference scenario.

3.6. Grid utilization

The evaluation of the grid utilization is linked with a series of challenges. By executing a clustering, links are removed or aggregated. Hence, the number and also the names of the links vary by contemplating different clustering results. The convention for the naming is based on 'startvertexID___endvertexID'. By changing the number of clusters k, the IDs of the start vertices or end vertices of the links of the clustered grid are changing as well. Firstly, the number of vertices decreases (and so the set of possible names does) and secondly the naming of the resulting vertices (in this context clusters) is randomized because of the randomized initializing of the k-means algorithm (see also Section 1.2.3). Hence, it is hardly possible to investigate the utilization of a specific link in different clustering results. As an alternative, the summed power imports of the different clustering results are analyzed. The term power imports denotes the amount of electricity which is not generated in the cluster where it is used.

In Figure 3.6 the total imports (blue bars) in TWh, summed over all vertices respectively clusters, for a different number of clusters k are shown on the left y-axis. On the right y-axis the relation between electricity imports and total electricity demand is plotted (red dots).

Since in the scenario with only 1 cluster ($k = 1$) all the electricity demand is covered with generation from this cluster, there is no import. With increasing number of clusters the import is increasing. As the demand is distributed by the inhabitants of the postcode areas (see Section 2.4.3), the distribution of the demand can be assumed as different compared to the real demand. However, the distribution of the power plants is partially coupled to the real demand. Furthermore, by increasing number of clusters, the clusters themselves are getting smaller (on average). It can thus be concluded that

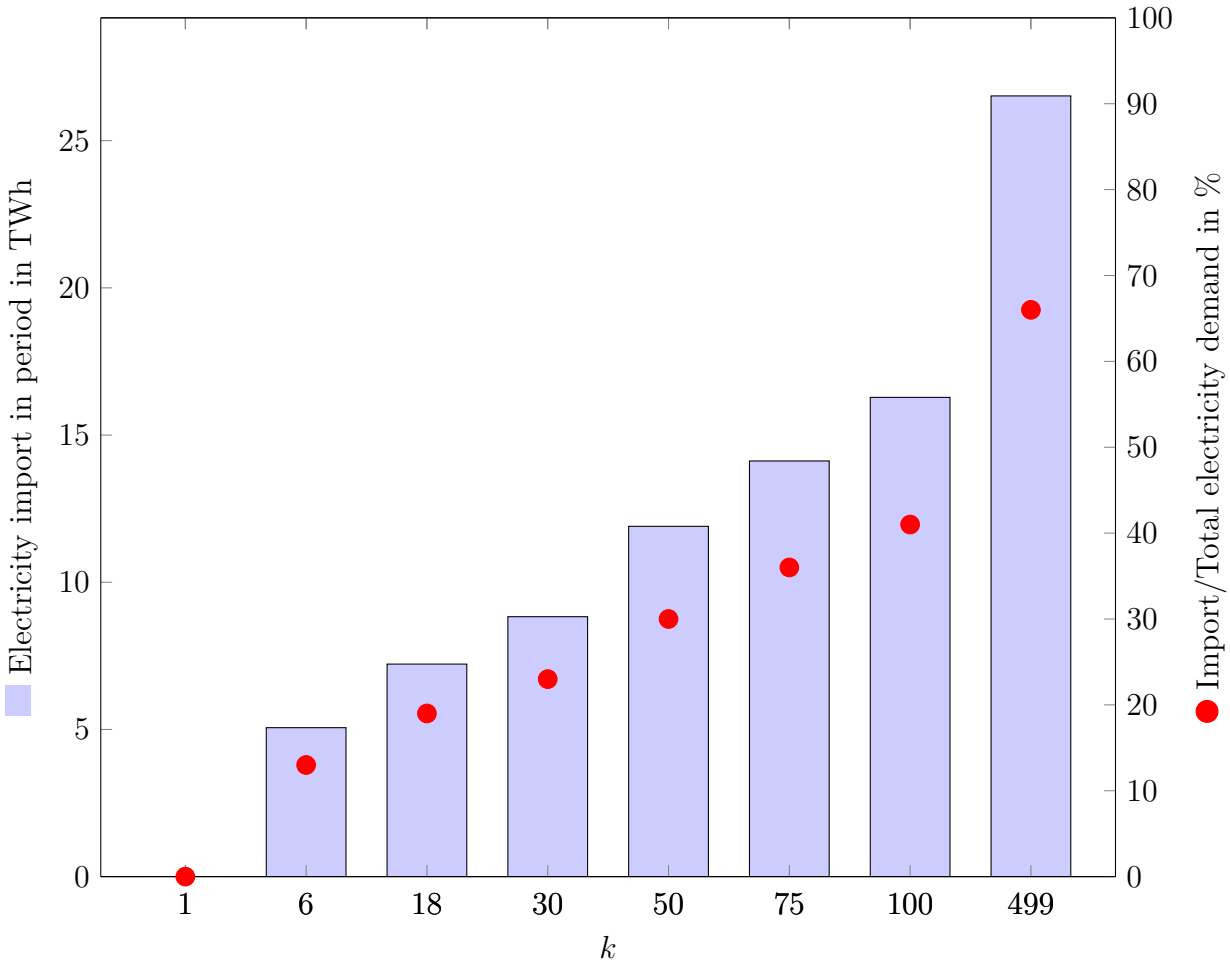


Figure 3.6.: Electricity imports in TWh summed over all clusters for the different scenarios with different number of clusters

the probability that a cluster contains enough installed capacity to cover its demand is decreasing by increasing number of clusters. As a result, the amount of imported electricity and also the relation between imported and total electricity demand rises. The import ratio rises from approximately 15 % ($k = 6$) to 65 % ($k = 499$). Hence, 35 % of the demand in the reference scenario ($k = 499$) can be covered by local power generation.

3.7. Computing time

As mentioned in Section 1.2.1, REMix uses the CPLEX algorithm to solve the optimization problem. Figure 3.7 shows the computation time of the CPLEX algorithm for different levels of aggregation k. This computation time excludes the preparation and follow-up processes executed by REMix and only takes into account the pure solving time for the linear system of equations. The computations are performed on a Intel Xeon CPU with 96 GB RAM.

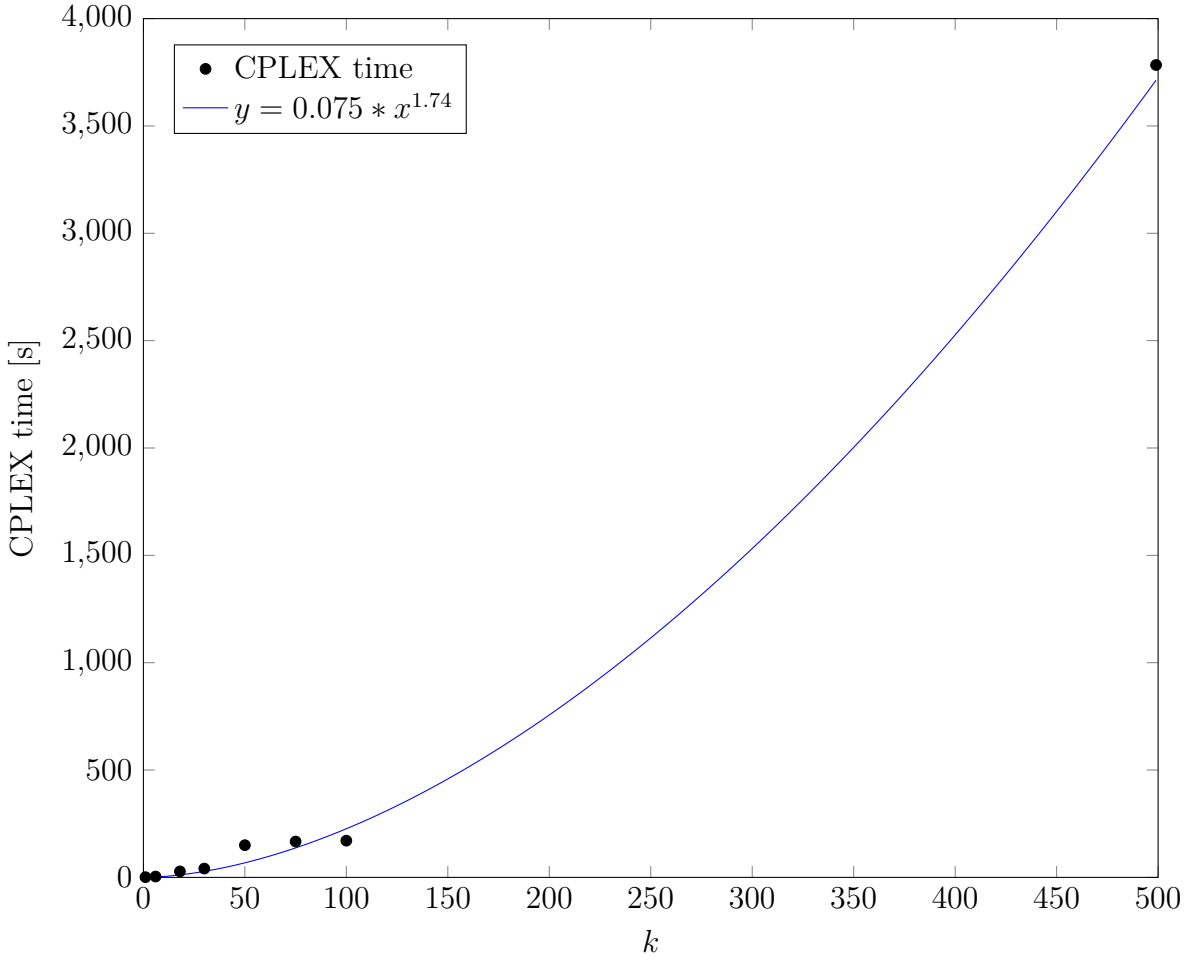


Figure 3.7.: CPLEX time [s] for different levels of aggregation for the investigated time period in logarithmic representation

The plot shows that the CPLEX time can be approximated by Equation 3.1:

$$y = 0.075 * x^{1.74} \tag{3.1}$$

what indicates an almost square behavior (the curve can be assumed as going through (0,0) as $k = 0$ is an empty set and does not refer to computation time). A better approximation might be achieved with further data points, like grid points at $k = 200$ and $k = 300$.

In general, in an aggregated power system the equation matrix is smaller and thus the CPLEX algorithm needs less time to solve. Derived from the almost square behavior of the computation time, it can be stated that with a slight change in aggregation, the computation time can be reduced noticeable. By comparing the $k = 499$ scenario with the $k = 100$ scenario, the computation time can be reduced by 95 % (from 3784 seconds to 171 seconds). However, it has to be considered that the testing conditions can differ from computation to computation as other users could also perform computations and occupy main memory. Nonetheless, a trend is recognizable.

3.8. Discussion of the results

In Section 2.5, the evaluation of the choice of the edge weight function, of the best fitting grid usage case and also of the Laplacian matrix is examined. During the evaluation inconsistencies for different parameter combinations (edge weight function, grid usage case, Laplacian matrix) occurred. With inconsistencies unassociated clusters are meant (for more detailed information see Section 2.5). As there is no hint in the literature concerning this behavior, it is necessary to validate the model. For the validation of the model developed in this work, the spectral clustering algorithm has been executed by Marco Schindler from the TU Clausthal on the underlying graph. His spectral clustering algorithm (based on Matlab) leaded to inconsistencies by clustering the graph either. This implies that these inconsistencies are not related to errors in the spectral clustering algorithm used in this work, but possibly on the topology of the graph, as the German transmission grid consists of many chains and also vertices with only one connection to the rest of the graph (see Section 2.5).

For a better classification of the results the shortcomings concerning this work are explained in the following. The shortcomings can be divided into two main parts: method related and data related. Method related shortcomings can be overcome by developing REMix or the clustering process itself. Data related shortcomings are more difficult to improve, as generating grid data, for instance, is linked to enormous effort.

REMix endogenous limitations of this work are some characteristics of the power plant scheduling. Due to the high computation times, all the scenarios are executed using linear optimization, which leads to a simplification in terms of start-up costs, minimum online and offline times or part load behavior. A more realistic reproduction of the power plant scheduling can be done by mixed integer optimization. On the other side, mixed integer optimization is accompanied by higher computation times, which is critical due to the model size. Another aspect of the power plant scheduling with regard to the REMix configuration used in this work, is the non-consideration of the heating sector. In general, REMix computes a most cost-efficient dispatch based on the variable costs of the generation technologies. However, in reality some of the power plants are cogeneration power plants. This means that they generate electrical power and heat. Thus, the situation can occur that a less cost-efficient power plant actually provides a lower-cost alternative, as the generated heat has also to be taken into account.

A further model endogenous shortcoming is the temporal resolution of the REMix computations. The maximum resolution is hourly, which means, that although hourly resolved results are computable, it is not possible to investigate what happens within the hours and leads to averaged and smoothed hourly profiles. The impact on the results of this work should be small as also the model exogenous profiles (load curve) are hourly resolved. If in future load curves with higher temporal resolution are available, a higher temporal resolution of the result will be necessary.

As there exist no reliable data for a spatial resolved electricity demand and spatial demand profiles, determining the demand at the vertices and also the vertex demand profiles are based on assumptions. As the mapping of the demand on the vertices is based on population density, the demand of regions with small population and high industrialization is underestimated, whereas the demand of so-called dormitory towns

(no industrialization, only residential areas) is overestimated. Furthermore, every vertex has the same load curve, derived from the total German load curve. This can also distort the results, since high industrialized regions tend to have a more balanced demand. Hence, the mapping of the demand and also the same load curve at each vertex could lead to deviations from real world clusters, if the used methodology leads to a grid congestion between vertices where no power transfer is needed in real world. On the other hand there could be transmission links which cannot provide enough capacity, but due to the mapping there is no demand in the model and thus the grid congestion would not occur.

Another simplification is the non-consideration of distribution grids. In the model, every power plant, every storage and every demand is connected to a transmission grid vertex. In a real power grid, the parts of the grid are connected to the grid level their installed capacity or demand refers to. As grid losses are reciprocal to the voltage level (higher losses in low voltage levels) and the transmission grid only represents a small part of the total grid (34,810 km of 1,800,000 km, respectively 2 %) [37], grid losses are underestimated in this work. In addition, grid congestion occurring in distribution grids cannot be revealed, which is in particular for the high number of domestic photovoltaic systems of interest. A method related shortcoming concerning the power transmission is the DC (direct current) approximation of the actual AC (alternate current) power flows. This simplification linearizes the AC power flow by ignoring reactive power and assuming all voltage angles as nominal. More detailed information about the DC approximation can be found in [9].

The SciGRID data set only covers the transmission grid of Germany and consequently no imports or exports of electrical energy are taken into account for the computations. Thus, the developed model can be interpreted as an island system with no exchange of electrical energy to other countries. This limitation has an influence on the power flows, since in a pan-European power system the power flow does not consider any state borders. This means that in a model contemplating an island system, flexibility options are getting lost and thus grid congestion can be overestimated. For instance, the high amount of fluctuating wind energy from the north of Germany must be transferred to the South and cannot be exported to the western or eastern neighboring states, which would lead to extra flexibility options.

As this work focuses on developing a clustering methodology, which covers a parametrization, and not on a detailed as possible simulation of the German power grid, this approximation is chosen (see also Section 4).

A classification of the results is exemplarily executed by contemplating lignite power supply. In the investigated period (autumn month) lignite power supply in the scenario with maximum resolution ($k = 499$) is about approximately 8.9 TWh. Projected to one year, the lignite power supply in the scenario is about 108 TWh. Considering the published power plant utilization data of the Bundesministerium für Wirtschaft und Energie of 2014 [38], lignite power supply accounts for 156 TWh. Compared to the reference scenario in this work the deviation amounts to 31 %. Possible explanations for this deviation are the high wind power supply in the investigated period and also the non-consideration of exports.

4. Conclusion and outlook

This work presents a methodology for aggregating power grids by using spectral clustering, considering both the grid topology and vertex attributes. In contrast to other works dealing with the identification of subsections in power grids, the spectral clustering algorithm used in this work is based on locational marginal pricing (marginal costs as vertex attributes) and not on electrical parameters, such as power flow or line admittance. The idea behind using marginal costs as vertex attributes is derived from the intention of locational marginal pricing: encouraging an economic use of electrical energy with regard to interdependencies between generation and transmission. This means that linked vertices tend to transfer electrical energy from vertices with lower marginal costs to vertices with higher marginal costs to reduce total system costs. In a power system with infinite transmission capacity this would lead to harmonized marginal costs at the vertices. Conversely, if linked vertices have differences in marginal costs a strong indicator for grid congestion is given. Furthermore, the degree of the grid congestion is determinable as high differences refer to high grid congestion, which is an advantage compared to only contemplating the power flows. This approach has only a binary behavior, utilized or underutilized. Insofar, spectral clustering based on locational marginal pricing provides a meaningful clustering seen from the perspective of the grid.

Due to a spatial aggregation of power grids, the computed results are influenced, as grid restrictions decrease and thus the power plant dispatch varies. Through the methodology developed in this work, it is possible to estimate that deviation of the results of an aggregated power system to the reference power system (maximum resolution). The application of this this methodology on the German transmission grid leads to two main conclusions. First, the aggregation of a power system is accompanied by a decrease of grid restrictions, since the clusters transform their inner grid to a 'copper plate'. This favors base load power plants with low marginal generation costs over peak load power plants, as the demand can be supplied without any grid limitations or losses. This leads to an overestimation of base load power plant operation times on the one hand and to an underestimation of peak load power plant operation times on the other. Second, due to the decreasing grid restrictions, the total system costs are underestimated in an aggregated power system. The degree of the underestimation depends on the degree of the aggregation. The more a system is aggregated, the more the system costs deviate from the reference scenario. However, increasing aggregation leads to decreasing computation times. Hence, a compromise between accuracy of the result and acceleration of the computations in a case-by-case contemplation has to be done. A general rule for an aggregation degree is not derivable.

The shortcomings explained at the end of the previous chapter may serve as a starting point for further analysis. By using a mixed integer optimization instead of linear

optimization, the computation time would increase, but the power plant behavior would be more realistic. This means that even in aggregated systems, the peak load covering is not provided by base load power plants anymore, as the ramping behavior can be better represented by mixed integer optimization. Hence, the dispatch could be less deviating in aggregated systems compared to the reference scenario. A possible approach for an analysis could be a comparison regarding accuracy of the results and computation time between an aggregated power system computed with mixed integer optimization and a highly resolved power system computed with linear optimization.

The mapping of the electrical demand on the vertices provides data related potential for improvements. An investigation of the power demand of industrial areas and also a spatial mapping of these areas would contribute to a finer parametrization which would lead to a more realistic power flow.

On the grid side, there are two main aspects which can get closer to a real power system. The implementation of distribution grids into the underlying case study would lead to a more precise mapping of the demand, small RE power plants and also the storage. This is accompanied by a more realistic representation of grid losses, more grid restrictions (in particular concerning the domestic PV supply), but also with higher computation times. Also here an investigation of the relation between accuracy and computing time must evaluate if a consideration of distribution grids would make sense. Another starting point for an analysis also considers the grid, but on a different voltage level. By contemplating the German transmission grid as an island system, flexibility options are underestimated. Through the parametrization of a pan-European transmission grid, flexibility options are augmented and also a more realistic power flow can be achieved, as also power imports and exports can be considered

Besides the elimination of the shortcomings, the developed method in general and spectral clustering in particular can be executed with a high number of different configurations. Concerning the method, the locational marginal pricing could be based on average marginals of a longer period instead of using the marginals of only one hour of the year. Furthermore, weighting the edges is linked with various opportunities. There exist a whole field of studying different types of Laplacian matrices and their influence on the results. The reason why both L_{sym} and L_{rw} are linked with a high probability of occurring inconsistencies is probably related to the topology of the transmission grid. The transmission grid consists of many chains and vertices with only one connection to the rest of the grid. By modifying the grid, L_{sym} or L_{rw} could lead to less inconsistencies. An analysis on how to modify the graph and not to influence the results too much, could lead to more balanced clusters.

In power systems with a high share of RE, flexibility options such as demand side management, energy storage or power-to-X technologies are playing a key role, due to the high amount of non adjustable power supply. In particular, such power systems require a long term investigation (for instance 1 year) to make an assessment that takes into account seasonal variations in wind, solar and hydro power availability. However, long term investigations cause high computation times. A reduction of the computation time can be achieved by aggregating the power system. Applying the developed aggregation methodology on such power systems requires a further analysis, in order to investigate the influence of the aggregation on flexibility options.

Bibliography

[1] Christoph Bals, Sönke Kreft, and Lutz Weischer. Wendepunkt auf dem Weg in eine neue Epoche der globalen Klima- und Energiepolitik. `https://germanwatch.org/de/download/13982.pdf`. Accessed: 2016-08-16.

[2] Bundesregierung. Energiewende, 2015. `http://www.bundesregierung.de/Content/DE/StatischeSeiten/Breg/Energiekonzept/0-Buehne/ma%C3%9Fnahmen-im-ueberblick.html` Accessed: 05.06.2015.

[3] Dena. dena-Netzstudie II. Integration erneuerbarer Energien in die deutsche Stromversorgung im Zeitraum 2015 - 2020 mit Ausblick 2025. Technical report, Deutsche Energieagentur (Dena), 2010.

[4] Ruben J. Sanchez-Garcia, Max Fennelly, Sean Norris, Nick Wright, Graham Niblo, Jacek Brodzki, and Janusz W. Bialek. Hierarchical spectral clustering of power grids. *IEEE Trans. Power Syst.*, 29(5):2229–2237, sep 2014.

[5] GAMS Development Corporation. General Algebraic Modeling System (GAMS) Release 24.2.1. Washington, DC, USA, 2013.

[6] IBM. *IBM ILOG CPLEX Optimization Studio CPLEX User's Manual*, 2013.

[7] Hans Christian Gils. *Balancing of Intermittent Renewable Power Generation by Demand Response and Thermal Energy Storage*. PhD thesis, Universität Stuttgart, 2015.

[8] Yvonne Scholz. *Renewable energy based electricity supply at low costs: development of the REMix model and application for Europe*. PhD thesis, Universität Stuttgart, 2012.

[9] Daniel Stetter. *Enhancement of the REMix energy system model: Global renewable energy potentials, optimized power plant siting and scenario validation*. PhD thesis, Universität Stuttgart, 2014.

[10] Diego Luca de Tena. *Large Scale Renewable Power Integration with Electric Vehicles*. PhD thesis, Universität Stuttgart, 2014.

[11] Department of Economic and Social Affairs of the United Nations. Multi dimensional issues in international electric power grid interconnections, 2006.

[12] Luo Gang, Chen Jinfu, Duan Xianzhong, and Shi Dongyuan. Automatic identification of transmission sections based on complex network theory. *IET Generation, Transmission & Distribution*, 8(7):1203–1210, jul 2014.

[13] Thomas Anderski, Yvonne Surmann, Simone Stemmer, Nathalie Grisey, Eric Momot, Anne-Claire Leger, Brahim Betraoui, and Peter van Roy. European cluster model of the pan-european transmission grid. Technical report, Seventh Framework Programme, 2012.

[14] Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta. A survey of kernel and spectral methods for clustering. *Pattern Recogn.*, 41(1):176–190, January 2008.

[15] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, July 2002.

[16] Brucher Cournapeau, Millmann. *Documentation of scikit-learn 0.17.*

[17] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007.

[18] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856. MIT Press, 2001.

[19] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[20] Ralph Turvey. What are marginal costs and how to estimate them? Technical report, The University of Bath, March 2000.

[21] Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, and Roger E. Bohn. *Spot Pricing of Electricity (Power Electronics and Power Systems)*. Springer, 2013.

[22] Anke Eßer-Frey. *Analyzing the regional long-term development of the German power system using a nodal pricing approach.* PhD thesis, Karlsruher Institut für Technologie (KIT), 2012.

[23] Wided Medjroubi, Carsten Matke, and David Kleinhans. SciGRID - An Open Source Reference Model for the European Transmission Network (v0.2), November 2015.

[24] 4C Offshore. Offshore wind farms. `http://www.4coffshore.com/windfarms/`. Accessed: 2016-07-22.

[25] FA Technik. German postcode list. `http://www.fa-technik.adfc.de/code/opengeodb/PLZ.tab`. Accessed: 2016-07-22.

[26] Paul-Frederik Bach. Entso-e load curve germany 2015. `http://pfbach.dk/firma_pfb/time_series/ts.php`. Accessed: 2016-07-22.

[27] PLATTS McGraw Hill Financial. World Electric Power Plants Database, 2015.

[28] Bundesnetzagentur. Kraftwerksliste. `http://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/Versorgungssicherheit/Erzeugungskapazitaeten/Kraftwerksliste/kraftwerksliste-node.html`. Accessed: 2016-07-22.

[29] Umweltbundesamt. Kraftwerke in Deutschland. `https://www.umweltbundesamt.de/dokument/datenbank-kraftwerke-in-deutschland`. Accessed: 2016-07-22.

[30] Deutsche Gesellschaft für Sonnenenergie e.V. EEG-Anlagenregister. `http://www.energymap.info/download.html`. Accessed: 2016-07-22.

[31] Drake van Rossum. *Python Reference Manual*, 2016.

[32] Dominique Pascal Heiken. Zubaudynamik der erneuerbaren Energien in Deutschland. Master's thesis, Hochschule für Wirtschaft und Umwelt Nürtingen-Geislingen, 2016.

[33] Gimeno-Gutiérrez Marcos and Lacal-Arántegui Roberto. Assessment of the European potential for pumped hydropower energy storage: A GIS-based assessment of pumped hydropower storage potential. Technical report, Joint Research Centre, 2013.

[34] F. Borggrefe, T. Pregger, H. C. Gils, K. K. Cao, M. Deissenroth, S. Bothor, M. Blesl, U. Fahl, M. Steurer, and M. Wiesmeth. Kurzstudie zur Kapazitätsentwicklung in Süddeutschland bis 2025 unter Berücksichtigung der Situation in Deutschland und den europäischen Nachbarstaaten. Technical report, Deutsches Zentrum fü Luft- und Raumfahrt, Institut f ür Energiewirtschaft und rationelle Energieanwendung der Universitä Stuttgart im Auftrag des Umweltministeriums Baden-Württemberg, 2014.

[35] Python Software Foundation. pandas - python data analysis library, version 18.1. `http://pandas.pydata.org/`. Accessed: 2016-07-28.

[36] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[37] Bundesverband Energie und Wasserwirtschaft e.V. Deutsches stromnetz ist 1,8 millionen kilometer lang. `https://www.bdew.de/internet.nsf/id/20130412-pi-deutsches-stromnetz-ist-18-millionen-kilometer-lang-de`. Accessed: 2016-08-26.

[38] Bundesministerium für Wirtschaft und Energie. Stromerzeugung in Deutschland. `http://www.bmwi.de/DE/Themen/Energie/Strommarkt-der-Zukunft/zahlen-fakten.html`. Accessed: 2016-09-01.

[39] William W Hogan. Contract networks for electric power transmission. *Journal of Regulatory Economics*, 4(3):211–42, 1992.

A. Annex

A.1. Parametrization of the REMix reference scenario

	Decade of construction						
	1959	1969	1979	1989	1999	2009	2019
Technology	Installed capacity in MW						
Nuclear	0	0	0	12068	0	0	0
Lignite	832	2885	5939	3851	3825	1773	2843
Coal	306	4553	6324	11216	2570	428	5452
CCGT	0	175	4121	431	1430	4910	5130
Gas turbine	156	290	5368	1055	2863	2482	1055
	Gross efficiency						
Nuclear	0.324	0.324	0.324	0.324	0.324	0.324	0.324
Lignite	0.355	0.355	0.38	0.395	0.425	0.4505	0.4611
Coal	0.387	0.387	0.407	0.417	0.467	0.4752	0.4968
CCGT	0.409	0.409	0.429	0.509	0.53795	0.53795	0.555
Gas turbine	0.234	0.234	0.274	0.274	0.3045	0.3248	0.355
	Netto efficiency						
Nuclear	0.309	0.309	0.309	0.309	0.309	0.309	0.309
Lignite	0.33	0.33	0.355	0.37	0.4	0.425	0.435
Coal	0.35	0.35	0.37	0.38	0.43	0.44	0.46
CCGT	0.4	0.4	0.42	0.5	0.53	0.53	0.5475
Gas turbine	0.23	0.23	0.27	0.27	0.3	0.32	0.33
	Availability						
Nuclear	0.945	0.945	0.945	0.945	0.945	0.945	0.945
Lignite	0.935	0.935	0.935	0.935	0.935	0.935	0.935
Coal	0.94	0.94	0.94	0.94	0.94	0.94	0.94
CCGT	0.977	0.977	0.977	0.977	0.977	0.977	0.977
Gas turbine	0.902	0.902	0.902	0.902	0.902	0.902	0.902
	O&M cost in kEUR/MWh						
Nuclear	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Lignite	0.0044	0.0044	0.0044	0.0044	0.0044	0.0044	0.0044
Coal	0.004	0.004	0.004	0.004	0.004	0.004	0.004
CCGT	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Gas turbine	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	Tear & wear cost in kEUR/MW						
Nuclear	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015
Lignite	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015
Coal	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015
CCGT	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
Gas turbine	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005

Technology	Inst. cap. in MW	Eff.	Avail.	O&M cost in kEUR/MWh
Hydro	1532	N/A	[1]	0
Photovoltaic	36876	N/A	[1]	0
Wind onshore	37222	N/A	[1]	0
Wind offshore	2922	N/A	[1]	0
Pumped storage	6301	0.8	0.98	0
Biomass	6836	0.2	0.9	0.002

Technology	Fuel Cost in kEUR/MWh	Certificate Cost in kEUR/tCO2
Nuclear	0.000286	0.01
Lignite	0.0064	0.01
Coal	0.0138	0.01
CCGT	0.03	0.01
Gas turbine	0.03	0.01
Biomass	0.01	0.01

[1] Already considered by EnDAT

A.2. Results

	Number of clusters k							
	1	6	18	30	50	75	100	499
Technology	Power-plant utilization in TWh							
Hydro	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46
Nuclear	7.83	7.83	7.14	7.27	7.24	7.09	7.15	7.11
Lignite	13.12	13.14	10.69	10.44	10.36	10.29	9.23	8.88
Coal	7.44	7.51	9.85	9.66	9.97	10.11	10.39	9.86
CCGT	0.0	0.05	0.22	0.46	0.47	0.67	0.85	1.65
Gas turbine	0	0	0	0	0	0.02	0.45	0.43
Biomass	0.12	0.12	0.91	0.99	0.96	0.96	1.24	1.81
PV	2.41	2.41	2.41	2.41	2.41	2.41	2.41	2.41
Wind	7.15	7.15	7.15	7.15	7.15	7.15	7.15	6.99
Technology	Power plant ramping in GW							
Nuclear	6.71	4.8	58.66	57.57	66.56	79.21	63.08	70.09
Lignite	131.06	157.28	46.76	82.88	83.22	98.57	264.51	240.97
Coal	1193.68	1160.06	1043.8	979.31	971.10	991.15	960.13	625.75
CCGT	0	10.24	86.87	146.3	171.16	205.65	224.45	401.88
Gas turbine	0	1.82	0.43	2.39	2.88	10.32	99.27	116.17
Biomass	75.03	74.3	328.92	307.52	266.41	253.12	233.62	252.9
Pumped storage	292.29	317.31	220.43	268.36	292.56	226.79	153.44	246.27
	System costs in MEUR							
	751	758	839	843	854	869	926	968
	Electricity import in TWh							
	0	5.06	7.22	8.83	11.9	14.12	16.28	26.52
	Import/total electricity demand in %							
	0	13	19	23	30	36	41	66
	CPLEX time in s							
	1	4	28	41	150	167	171	3784