COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

MII EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

DEVELOPMENT OF A GLOBALLY APPLICABLE METHOD TO ANALYZE REGIONAL DIFFERENCES IN COUNTRY ENERGY SYSTEMS

Autor: Christian Perau

Director: Martin Küppers

Co-Director: Marco Franken

Madrid

Marzo de 2020

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Development of a globally aplicable method to analyze regional differences of country energy systems

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2019/2020 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

C. Perau

Fdo.: Christian Perau Fecha: 13/06/2020

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Martin Küppers Fecha: 17/06/2020

DEVOLOPMENT OF A GLOBALLY APPLICABLE METHOD TO ANALYZE REGIONAL DIFFERENCES OF COUNTRY ENERGY SYSTEMS

Author: Perau, Christian.

Supervisor: Küppers, Martin.

Collaborating Entity: Siemens AG, RWTH Aachen University

ABSTRACT

A methodology based on spatial clustering has been developed to identify different regions within country energy systems. Those identified regions are used in an energy system model to consider spatial fluctuations within a country´s energy system and better predict future transition pathways of decarbonization.

Keywords: Spatial clustering, energy system modeling, geographic information systems

1. Introduction

Meeting the Paris agreement [1], limiting the increase of average temperature to less than 2°C, preferably 1.5°C, requires decarbonization of different sectors, mainly electricity, mobility, heat, industry, buildings, and agriculture [2]. For example, in the electricity sector renewable energy technologies, like wind, solar, hydro, and biomass have to be installed and the generation by conventional power plant technologies using coal or lignite which are the technologies with the highest emissions must be reduced. Therefore, one characteristic of future energy systems is their high share of renewable energy and the volatile generation that they entail. Because of this, the transformation of current energy systems towards the future is under investigation.

Costs are an important factor in those transitions. Energy system modeling, especially optimization models, is used to calculate the cost-optimal pathway towards a decarbonized future energy system, i.e. which capacities of given technologies are installed or shut down and when.

In energy system modeling one important factor that has to be considered is the spatial resolution of the model [3]. Modeling each consumer and generator as one point interconnected by the grid is not feasible. Another approach is to model the energy system of a country as one point. However, this might lead to the neglect of important factors regarding the spatial differences of load, generation, and especially transmission and distribution of energy. Because of that, an in-between-approach of the two first mentioned approaches is to model the energy system based on regions by modeling their respective energy systems and the interactions between them [4]. Considering this approach, the question emerges on how to define those regions. Intuitively, one might consider the different states of a country as regions. However, their regional distribution can be very inhomogeneous regarding renewable potentials, load, or other factors that are important for energy system models. Therefore, the need for a so-called regionalization method rises. In this context, regionalization can be defined as the identification of a small number of groups of internally homogeneous and spatially contiguous regions out of a large set of spatial objects [5]. This way the validity of energy system modeling might be further improved in comparison to single-region models. Furthermore, the significance of the results improves in comparison to the modeling of administrative levels due to the consideration of regional occurrences.

2. State of the art

In order to classify different regions within a country depending on spatial and aggregated attributes Clustering is one mainly used option. Regarding clustering there are many different approaches and algorithms available. Since the suitability of a clustering algorithm strongly depends on the problem there is no optimal clustering technique. Therefore, all have to be taken into account and a profound evaluation has to be done to make a justified choice. The most common are listed and shorty described below:

- K-means: Initially it sets K centroids of clusters within the dataset and iteratively assigns the different data objects to the centroids. Even though the method is fast, the disadvantages make it less suitable for the use of this project. It is very dependent on the choice of initial cluster centroids and, therefore, the algorithm has to be executed several times in order to find a local optimum. Another problem is the tendency of K-means to find spherical clusters of same size which in this project is not desired. Furthermore, the spatial contiguity that shall be achieved by the developed method can hardly be implemented into K-means, because using the distances between regions as attribute of the sample results in the reduction of importance of its actual attributes and makes the spatial contiguity the most important factor.

- Max-p Regions: In this method the clustering is written as a mixed-integer program where the number of clusters is minimized while a constraint has to be met, e.g. the number of solar generation capacity hast to be higher than a predefined threshold. Since in this project no such constraint exists as the algorithm has to decide on different attributes this method is not suited for the problem. As well the algorithm is not able to process a high number of samples, which is also a requirement that has to be met in this project. It should be mentioned that the constraint could also be chosen to limit the inner variance inside a cluster below a predefined value, but since defining such a highest variance is rather subjective this method is also not applicable to the problem of this project.
- Density-based & grid-based methods: Generally, these algorithms receive as an input a number of neighbors that a data sample has to have within a predefined distance in order to become a *core object*. If two core objects are within the distance defined before, they form a conjoint cluster. Every sample that is not within the distance form a core object is treated as noise. This would mean in this project to treat some regions that are highly different from all other data samples as outliers and, therefore, treat those as single clusters. Even though the last is not necessarily a disadvantage but could also be desired, the number of neighbors and the distance are rather subjective and make the method not suitable for the project. Furthermore, the algorithm is not suitable for high amounts of attributes of the samples. In case time series of wind, solar, load, etc. are used as attribute density-based methods are not suitable. Grid-based methods can handle high numbers of attributes better but have the other disadvantages of density-based methods and, therefore, are not suited either.
- Hierarchical Clustering: Output of hierarchical clustering is a hierarchy of how to combine the different data objects. The agglomerative clustering starts with each sample as an own cluster and combines them depending of a distance criterion, called linkage, until all samples are within one cluster. Afterwards, the optimal number of clusters has to be determined. This method is sensitive to outliers which is wanted for the project and a connectivity condition cannot only be easily implemented but also reduces the computing complexity of initially $O(n^2)$. The

drawbacks of this method are that once two data sample have been assigned to one cluster there is reassignment of the data objects to other clusters. Nevertheless, there are some methods that can be used to overcome this disadvantage.

Because of those advantages and disadvantages the hierarchical clustering is chosen for this project since it appears to be the most suited for the given problem.

3. Definition of the project

The goal of this work is to develop a methodology to automatically identify a suitable number of regions inside a country´s energy system which can be used for region-based energy system modeling. On the one hand, the loss of data inside those regions should be as low as possible. On the other hand, the number of regions should be in a reasonable range and reflect the complexity of the country. Another important factor is the electrical grid. Since the regions themselves are modeled as single-region models, they should have a considerable grid to justify this assumption. The methodology shall improve the significance of the results of an already existing energy system model, capable of identifying the cost-optimal decarbonization path by considering the regional differences. However, computational time and complexity should be kept low.

4. Description of the methodology

The developed methodology requires the following spatially resolved input data:

1) The smallest administrative regions of a country (municipal level)
2) Load data divided into private households and commerce, trade, service, and industry
3) Generation capacities based on different technologies
4) Wind and photovoltaic generation profiles
5) Grid data derived from OpenStreetMap data

Using the smallest administrative regions of a country it is assured that one city does not end up being divided, i.e. in two clusters, since this might lead to political issues. However, some smallest regions are not spatially contiguous. Since spatial contiguity of the final regions is required to model energy flows between them, those regions that consist of multiple not spatially contiguous regions are split. This also leads to the appearance of islands in the data set. Those are naturally not connected to other regions and, therefore, not spatial contiguous to other regions. Since those islands should not all be seen as an own cluster, the algorithm identifies spatially contiguous landmasses within the region set and considers only those with an area greater than a specified threshold. All other islands are connected to the closest regions if the distance is less than 50km. This assumes that islands might be connected to the mainland if they are not too far away which would make this unreasonable. The load, generation, and profiles are then aggregated at their corresponding smallest regions. Since bigger regions tend to accumulate more data points, the load and generation data are also normalized by the area of the corresponding region.

To compare the different regions in a clustering the data have to be normalized. However, since there are more data corresponding to the wind and photovoltaic profiles, the data cannot be normalized to the range of zero to one. Doing so would lead the algorithm to consider each time step in the time series as important as the total load for one year. Therefore, the normalized values are also divided by their corresponding number of elements, i.e. 8760 for yearly time series.

Using a spatial hierarchical clustering the smallest base regions are grouped together. Using the elbow method [6] the optimal number of clusters is found. However, the hierarchical clustering alone tends to create a few very big clusters and some single region clusters. To improve the results a post-processing algorithm is applied [7]. Considering shifting each region into another spatially contiguous cluster the algorithm identifies changes that improve the result of the elbow curve. This process, however, has the drawback that only one region is shifted at a time. Therefore, this work proposes a method to shift multiple regions into spatially contiguous clusters if a path in one cluster can be identified that connects regions that are not spatially connected to more than one other region in the cluster. In doing so, the results of the clustering are further improved.

Using the identified regions and the OpenStreetMap grid data the electrical grid of the country is synthesized. Due to the occurrences of inconsistency of those data, a preprocessing is done applying standard rules. For example, lines with frequencies different than 50 or 60 Hertz are dropped from the data set, or the number of electrical circuits that one line carries is calculated to be a multiple of three since some user might also count the earthing cable. The transmission capacities between the different regions are then calculated using standard values depending on the voltage levels [8], [9].

Using the identified regions and their corresponding transmission capacities the energy system modeling is done using an already existing optimizer [10]. The complete workflow is depicted in the figure 1.

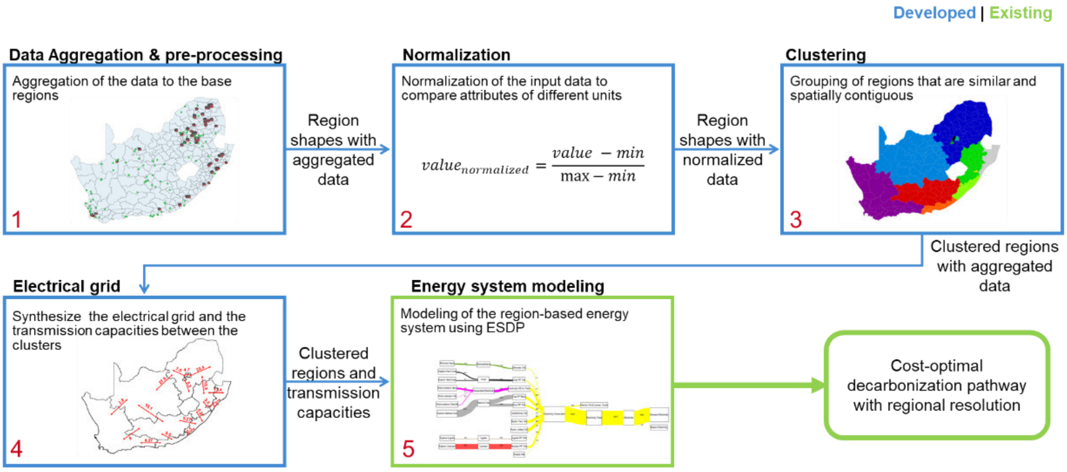
Developed | Existing
Data Aggregation & pre-processing
Aggregation of the data to the base regions
1
Region shapes with aggregated data
Normalization
Normalization of the input data to compare attributes of different units
$value_{normalized} = \frac{value - min}{max - min}$
2
Region shapes with normalized data
Clustering
Grouping of regions that are similar and spatially contiguous
3
Clustered regions with aggregated data
Electrical grid
Synthesize the electrical grid and the transmission capacities between the clusters
4
Clustered regions and transmission capacities
Energy system modeling
Modeling of the region-based energy system using ESDP
5
Cost-optimal decarbonization pathway with regional resolution

Figure 1: Workflow of the developed methodology

5. Results

Applying the developed methodology to the use case of South Africa twelve clusters are identified and presented in figure 2.

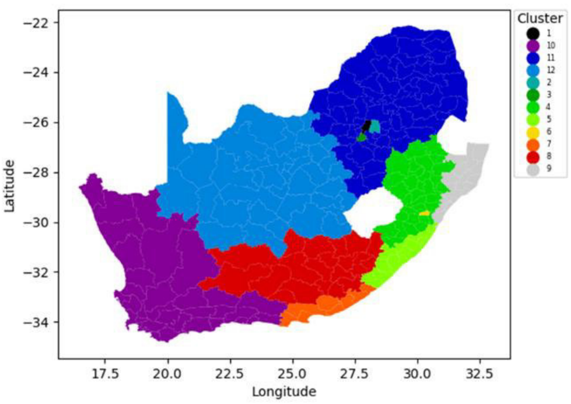

Figure 2: Result of the spatial clustering of South Africa

As can be seen, some clusters are single-regions around big cities of South Africa that are described by high load densities (clusters 1, 2, 3, and 6). Furthermore, different coastal areas are identified which entail high potentials for wind generation (clusters 5, 7, 9, and 10) but different load values. Cluster 12, on the other hand, is described by low load density but high photovoltaic generation capacities and potential.

Considering the electrical grid, results of the methodology are presented in figure 3. Where on the left the different lines are depicted and, on the right, the resulting transmission capacities between the regions.

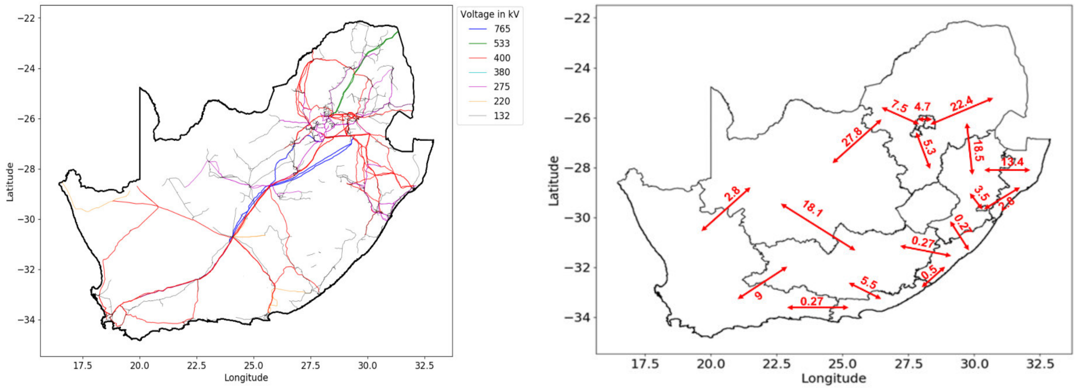

Figure 3: Results of the electrical grid synthesized by OpenStreetMap data. On the left the different voltage levels of the grid. On the right the resulting transmission capacities

As can be seen, the transmission capacities are located mainly from the North-East of the country – where high capacities of coal power plants are located – towards the South-West and South-East of the country. Due to different load centers being in the South-West (mainly Cape Town) this seems logical.

When modeling the energy system build on the identified regions and transmission capacities, the cost-optimal transformation pathway is calculated. The following figure compares the generation mix in 2015 and 2050, assuming a decarbonization of 80%.

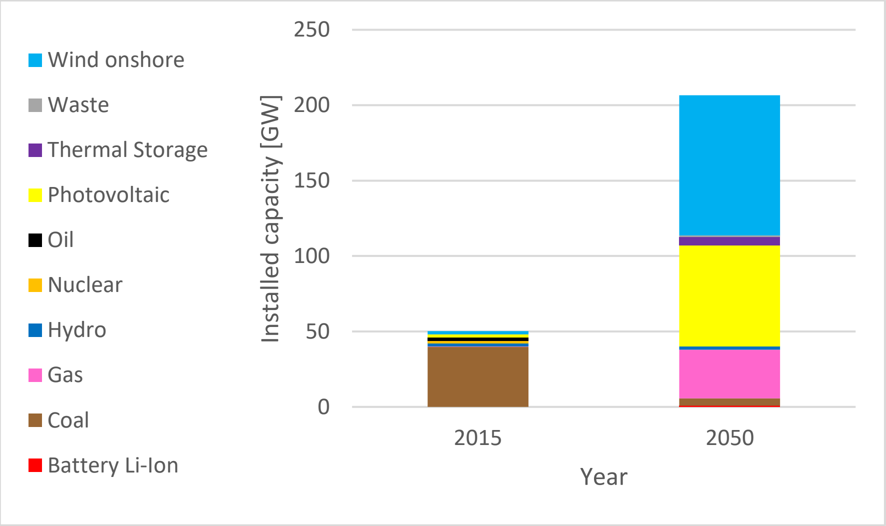

Figure 4: Generation capacities of South Africa in 2015 and 2050

As shown, the generation mix of primarily coal-based generation switches to a mix of mostly wind and photovoltaic based generation. In addition, a considerable amount of gas power plants is installed.

Comparing these results to a single-region model the following figure shows that a single-region model might evaluate poorly the potentials of renewables which leads to higher installed capacities. Figure 5 presents the comparison. On the other hand, comparing the total costs of the transformation pathways the single region model does not account for 16 billion € that are related to the electrical grid. The percentages on the right of the columns indicate the ratio of the costs of the respective technology to the total costs.

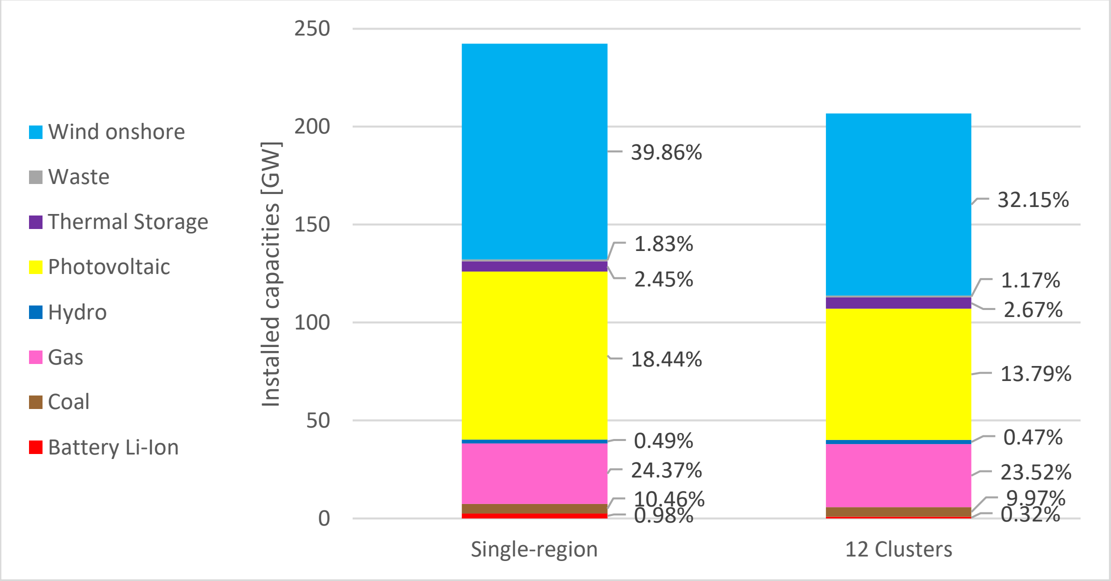

Figure 5: Comparison of the installed capacities in 2050 using either a single-region model or the 12 identified clusters. The percentages on the right indicate the ratio of the costs of each generation technology to the total costs

6. Conclusions

In this work, a methodology has been developed that identifies a suitable number of regions inside a country´s energy system model. The corresponding electrical grid is synthesized using OpenStreetMap data and the corresponding transmission capacities are calculated using standard values depending on the voltage level of the respective line. Optimizing the decarbonization pathway towards a specified goal, the region-based modeling improves the insight that can be obtained and the significance of the calculated results. It also indicates which regions are of special importance to the energy system and where to install additional generation capacities in which amount. However, the number of regions might be arbitrary due to a set limit that has been implemented to keep the computational complexity and running time low. The developed methodology also might not apply to a high number of regions (multiple thousands) due to the high computation time of the clustering post-processing. Possible further developments are

1) The specifications of high voltage cables and high voltage direct current cables which are currently considered to be standard overhead lines
2) Considering the available space inside a region which might be limited

3) Connections to other countries and the possibility of exchange of electrical energy
4) Offshore wind generation, either including it in the clustering process or afterwards

DESARROLLO DE UNA METODOLOGÍA GLOBALMENTE APLICABLE PARA ANALIZAR DIFERENCIAS REGIONALES DE SISTEMAS ENERGÉTICOS DE PAÍSES.

Autor: Perau, Christian.

Supervisor: Küppers, Martin.

Entidad Colaborada: Siemens AG, RWTH Aachen University

RESUMEN DEL PROYECTO

Se ha desarrollado una metodología basada en el clustering territorial para identificar diferentes regiones dentro del sistema energético de un país. Estas regiones se utilizan en un modelo de sistema energético para considerar las fluctuaciones regionales y mejorar la predicción de futuras vías de transición de la descarbonización.

Palabras clave: Clustering territorial, modelización de sistemas energéticos, sistema de información geográfica (SIG)

1. Introducción

Para cumplir el acuerdo de París [1], limitando el aumento de la temperatura a menos de 2°C, preferiblemente 1,5°C, se requiere la descarbonización de sectores diferentes, principalmente los de la electricidad, la movilidad, el calor, la industria, los edificios y la agricultura [2]. Por ejemplo, en el sector de electricidad habría que instalar tecnologías de energía renovable como la eólica, la solar, la hidroeléctrica y la biomasa. También habría que reducir la generación convencional como el carbón y el lignito. Por lo tanto, una característica de los sistemas energéticos del futuro es su alta proporción de energía renovable y la generación volátil que éstas acarrean. Por ello, se está investigando la transformación de los sistemas energéticos actuales hacia el futuro.

Los costes son un factor muy importante en esa transición. La modelización de sistemas energéticos, especialmente los modelos de optimización, se utilizan para calcular la vía más barata hacia un sistema energético descarbonizado. Eso implica calcular qué tecnologías se instalan o se apagan y cuándo.

En la modelización de sistemas energéticos un factor importante que debe considerarse es la resolución espacial del modelo [3]. No es factible modelar cada consumidor y generador conectado a la red eléctrica. Otro enfoque es modelar el sistema energético como un punto. Sin embargo, esto puede conducir a la negligencia de factores importantes con respecto a las diferencias espaciales de demanda, generación y especialmente transmisión y distribución de energía. Por ello, un enfoque intermedio consiste en modelar el sistema energético basándose en regiones mediante la modelización de sus respectivos sistemas energéticos y las interacciones entre ellos [4]. Con respecto a este enfoque surge la pregunta de cómo definir dichas regiones. Intuitivamente se pueden considerar las provincias de un país como regiones. No obstante, su distribución regional puede ser poco homogénea en lo que respecta a potenciales para renovables, demanda u otros factores importantes para los modelos de sistemas energéticos. Por lo tanto, surge la necesidad de un método de regionalización de un país y su sistema energético. En este contexto, la regionalización se puede definir como la identificación de un número pequeño de grupos internamente homogéneos y adyacentes a partir de un gran conjunto de objetos espaciales [5]. De esa forma se puede mejorar la validez de la modelización con respecto al modelo de un solo punto.

2. Estado del arte

Para clasificar regiones distintas dentro de un sistema energético se puede usar el clustering territorial. Existen diferentes tipos de clustering. La adecuación de un clustering depende del problema. Por ello, no hay una técnica de clustering que sea superior a todas las demás. Por lo tanto, todos los enfoques han de ser considerados y evaluados con respecto al problema de este proyecto. Los más comunes son:

- K-means: Inicialmente se establecen K centroides en el conjunto de datos. Después se asignan iterativamente los diferentes objetos a los centroides. Aunque este método es rápido, las desventajas que presenta lo hacen poco adecuado para su aplicación en este proyecto. El resultado depende de la selección de los centroides iniciales y por ello se debe ejecutar el algoritmo varias veces para identificar un óptimo local del problema. Otro problema es la tendencia del algoritmo K-means a identificar grupos esféricos, lo que no se desea en este proyecto. Sin embargo, la desventaja más problemática es la inclusión de la contigüidad espacial de los clústeres, dado que ésta no puede ser forzada en K-means. Por ello, se utilizan las coordinados como atributos en el clustering con un peso muy alto. No obstante, esto resulta en la reducción de la importancia de otros atributos que se quiere usar.

- Max-p Regions: En este enfoque el clustering está programado como programa mixed-integer donde el número óptimo de los clústeres se calcula sujeto a una restricción. Esa restricción puede ser, por ejemplo, que la capacidad de generación solar sea más alta que un valor especificado. No obstante, no existe una restricción adecuada para el problema de este proyecto. Se podría usar la varianza de los clústeres, pero definir un valor adecuado puede resultar subjetivo. Además, el algoritmo no es apto para un número muy elevado de regiones, necesarias en este proyecto.
- Metodologías basadas en la densidad y las mallas: Generalmente, la metodología basada en la densidad recibe dos valores: el número de vecinos que debe tener un objeto y la distancia máxima a la cual este número mínimo de vecinos tiene que estar. Si un objeto tiene suficientes vecinos se trata como un objeto central. Si la distancia entre dos objetos centrales es menor que la distancia especificada, ambos forman un único clúster. Un objeto que se encuentra demasiado lejos de un objeto central se trata como ruido y por ello se convierte en un clúster individual. Aunque esto no es necesariamente una desventaja el número de vecinos y la distancia son valores subjetivos y hacen que el método no sea adecuado para el proyecto. Además, algoritmo no es adecuado para grandes cantidades de atributos. Por ejemplo, en el caso de las series temporales eólicas la metodología basada en densidad no es aplicable. Metodologías basadas en las mallas puede utilizar un mayor número de atributos, pero tienen las mismas desventajas que la metodología basada en la densidad y por tanto no son adecuadas.
- Clustering jerárquico: El resultado de este enfoque es una jerarquía indicando cómo combinar los diferentes objetos. Se empieza con un clúster para cada objeto y en cada etapa se une dos clústeres basándose en la similitud de los clústeres y de un criterio de conexión. Eso se hace hasta que todos los objetos estén en un único clúster. El número óptimo de clústeres se debe identificar en otro proceso. Este método es sensible a objetos atípicos, lo que se desea en este proyecto y permite incluir una restricción de relativa a la conexión espacial de los clústeres. Usando dicha restricción se consigue reducir el tiempo de cálculo de este método. La desventaja de este enfoque es que una vez que forman un

clúster, los objetos no pueden cambiar de clúster en otro momento. Es decir, no hay una reasignación de los objetos. No obstante, existen formas de vencer esa desventaja.

Debido a las características de los distintos enfoques el clustering jerárquico es seleccionado porque aparece como el más adecuado con respecto al problema de este proyecto.

3. Definición del proyecto

El objetivo de este Proyecto es el desarrollo de una metodología para identificar automáticamente el número adecuado de regiones dentro del sistema energético de un país. Estas regiones deben ser adecuadas para una modelización de un sistema energético. La pérdida de datos debe ser reducida al mínimo. Otro factor importante es la red eléctrica. Debido a la modelización de regiones como un punto, dentro de las regiones debería existir una red eléctrica considerable para justificar esa suposición. La metodología debe resaltar la importancia de los resultados de la modelización. Para ello se usa un modelo existente, capaz de calcular la vía más económica para la transición al sistema energético del futuro.

4. Descripción de la metodología

La metodología desarrollada requiere los siguientes datos espaciales:

1) Las regiones administrativas más pequeñas de un país (nivel municipal)
2) Los datos de demanda divididos en hogares privados y comercios, servicios y industria
3) Capacidades de generación basadas en las distintas tecnologías
4) Perfiles de generación eólica y fotovoltaica
5) Datos de la red de transmisión derivados de OpenStreetMap

Usando las regiones administrativas más pequeñas de un país se asegura que una ciudad no termine en dos clústeres diferentes, lo que podría causar en cuestiones políticas. No obstante, algunas de las regiones más pequeñas no son espacialmente contiguas.

Debido a la necesidad de clústeres espacialmente conectados para modelar el flujo de energía entre ellos, las regiones que consisten en múltiples regiones no contiguas espacialmente se separan. Esto también resulta en la aparición de islas. Éstas no están naturalmente conectadas a otras regiones. Cada isla no debería considerarse como un clúster único. Por ello, el algoritmo identifica las masas terrestres espacialmente contiguas dentro del conjunto de regiones y sólo considera las que tienen una superficie superior a un umbral especificado. Todas las demás islas se conectan a las regiones más cercanas si la distancia es inferior a 50km. Esto asume que las islas pueden estar conectadas al continente si no están demasiados lejos. Los datos de demanda, generación y los perfiles se agregan a las áreas correspondientes. Sin embargo, regiones más grandes tienden a acumular más datos. Por ello, los valores de demanda y generación se normalizan al área su región.

Para comparar las regiones en el proceso de clustering se deben normalizar los datos. Sin embargo, debido a la cantidad muy alta de valores que pertenecen a las perfiles eólicos y fotovoltaicos los datos no pueden ser normalizados únicamente al rango de cero a uno. Por ello, los valores normalizados se dividen por su correspondiente número de elementos. Por ejemplo, la serie temporal anual del perfil eólico se divide por 8760.

La agrupación de las regiones administrativas más pequeñas se realiza por el clustering jerárquico. Usando el método del codo [6] se identifica el número óptimo de clústeres. No obstante, el clustering jerárquico por sí sola tiene tiende a identificar pocos clústeres grandes – uno o dos – y muchos pequeños que consisten en una sola región. Por ello, se aplica algoritmo de post-procesamiento [7]. Se considera desplazar de cada región a otro clúster espacialmente contiguo. Así se identifican los cambios de regiones a otros clústeres que mejoran el resultado de la curva del codo. Este proceso tiene la desventaja de sólo considerar el cambio de una sola región a la vez. Por lo tanto, en este proyecto se ha desarrollado otro proceso para cambiar múltiples regiones a la vez, mejorando así el resultado del clustering.

Con las regiones identificadas y datos de la red eléctrica de OpenStreetMap, se sintetiza la red eléctrica del país. Debido a la inconsistencia de los datos, estos últimos deben ser procesados antes de su utilización. Para ello se aplican normas estándar. Por ejemplo, cada línea que tiene una frecuencia distinta de 50 o 60 Hz se elimina de conjunto de datos. Las capacidades de transmisión se calculan utilizando valores estándar dependiendo de los nivel de tensión [8], [9].

Utilizando las regiones identificadas y sus correspondientes capacidades de transmisión, se realiza la modelización del sistema energético [10]. La figura 1 presenta el proceso completo.

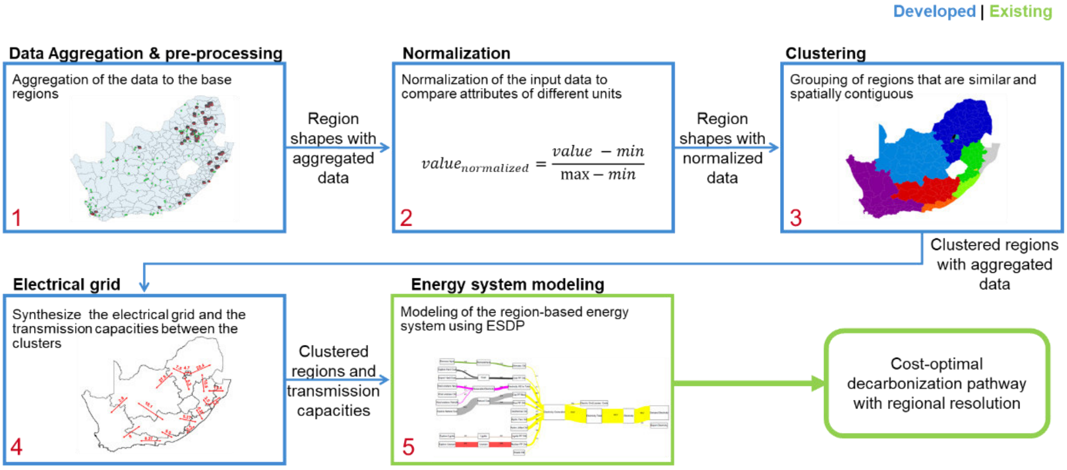
Developed | Existing
Data Aggregation & pre-processing
Aggregation of the data to the base regions
1
Region shapes with aggregated data
Normalization
Normalization of the input data to compare attributes of different units
$value_{normalized} = \frac{value - min}{\max - min}$
2
Region shapes with normalized data
Clustering
Grouping of regions that are similar and spatially contiguous
3
Clustered regions with aggregated data
Electrical grid
Synthesize the electrical grid and the transmission capacities between the clusters
4
Clustered regions and transmission capacities
Energy system modeling
Modeling of the region-based energy system using ESDP
5
Cost-optimal decarbonization pathway with regional resolution

Figura 1: Proceso completo del método desarrollado

6) Resultados

Aplicando la metodología desarrollada al caso de Sudáfrica, se identifican doce clústeres presentados en la figura 2.

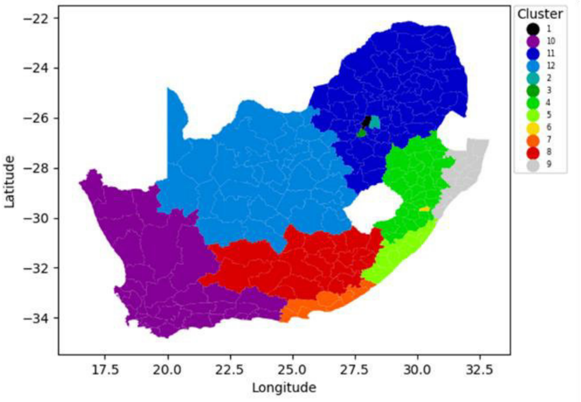

Figura 2: Resultados del clustering territorial de Sudáfrica

Como puede observarse, algunos clústeres consisten en una sola región en torno a las grandes ciudades de Sudáfrica. Se caracterizan por una alta densidad de demanda (clústeres 1, 2, 3 y 6). Además, se han identificado algunas áreas costeras con potenciales altos para la generación eólica pero diferentes densidades de demanda (clústeres 5, 7, 9 y 10). El clúster 12 tiene una densidad de demanda muy baja y una capacidad y un potencial de generación fotovoltaica altos.

La figura 3 presenta los resultados de la red eléctrica con la metodología desarrollada. A la izquierda se representa la red en función de los diferentes niveles de tensión y a la derecha las correspondientes capacidades de transmisión.

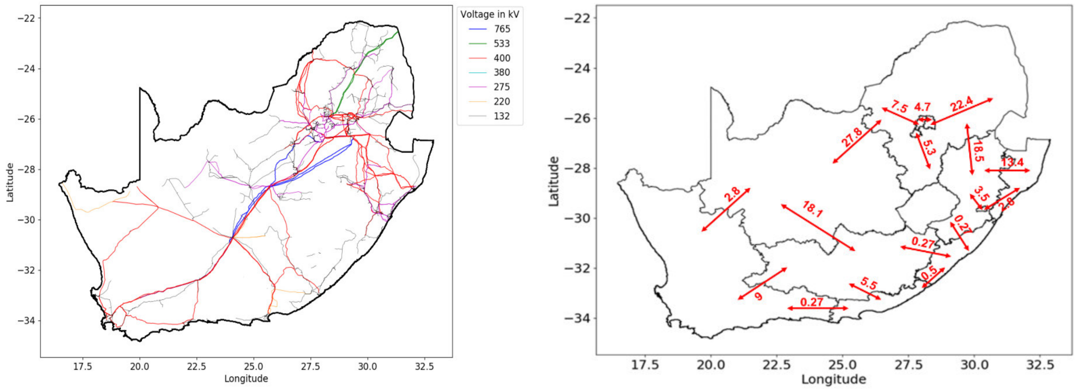

Figura 3: Resultados de la sintetización de la red eléctrica usando datos de OpenStreetMap. A la izquierda se representan las líneas de transmisión en función de los niveles de tensión. A la derecha se presentan las capacidades de transmisión.

Como se puede ver las líneas de transmisión van principalmente del noreste hacia el sur y suroeste del país. Esto parece razonable, dado que las centrales de carbón se encuentran al noreste del país y las demandas al suroeste (principalmente en Ciudad del Cabo).

Al modelar el sistema energético, a partir de las regiones identificadas y las capacidades entre estas, se calcula la vía de transmisión óptima desde un punto de vista económico hacia el futuro. La figura 4 presenta la mezcla de generación en 2015 y 2050 asumiendo una descarbonización de 80%.

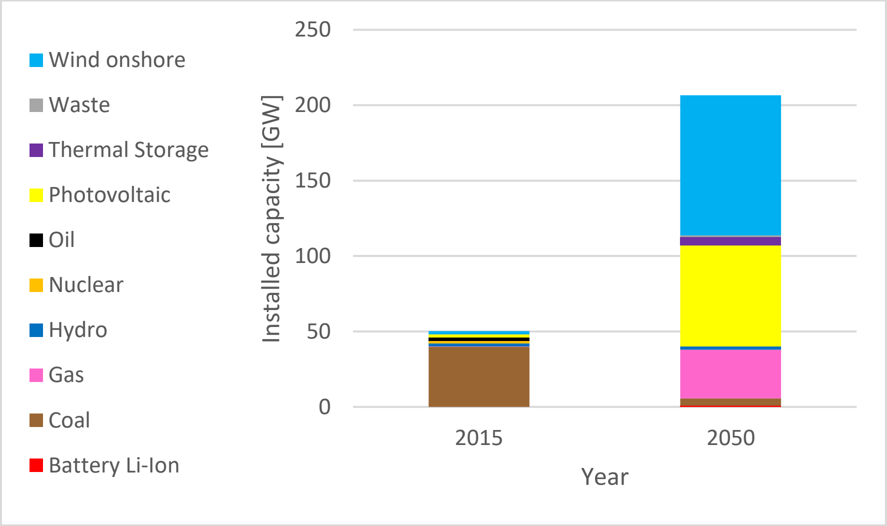

Figure 4: Capacidades de generación de Sudáfrica en 2015 y 2050

Como se muestra, la mezcla de generación en 2015 está principalmente basada en carbón. En 2050 se cambia a una mezcla de generación eólica, fotovoltaica y gas.

La figura 5 presenta la comparación de los resultados del modelo basado en regiones y los resultados de un modelo de un punto. Esta comparación indica que el modelo de un punto no considera los potenciales de generación renovable correctamente. Al comparar el coste de la transición del sistema, se ve que el modelo de un punto no considera costes de 16 billones € relacionados a la red eléctrica. Los porcentajes que figuran a la derecha de las columnas indican la relación entre los costos de la tecnología respectiva y los costos totales.

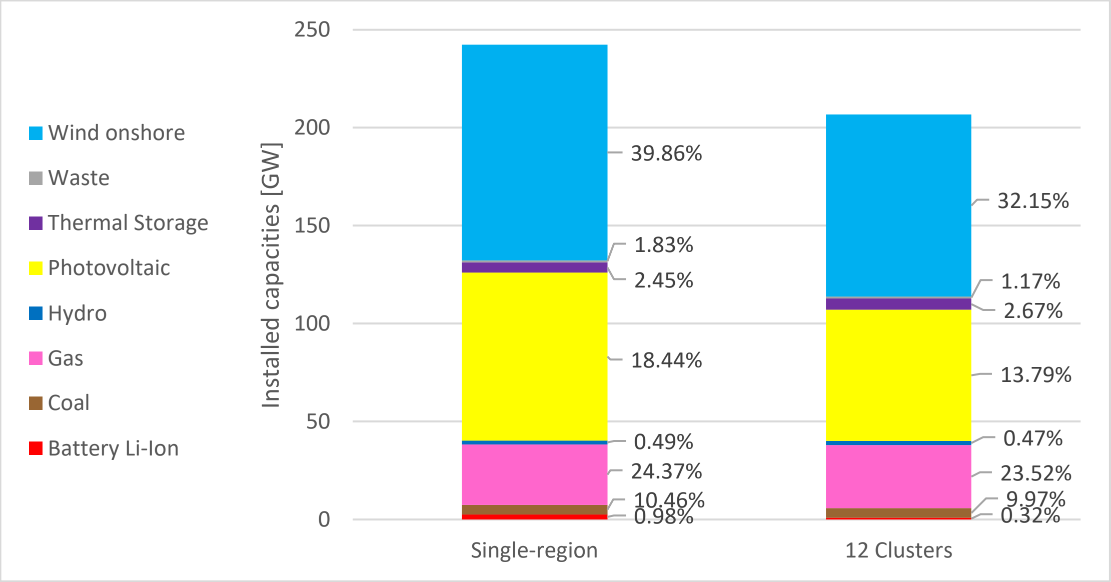

Figura 5: Comparación de la mezcla de generación en 2050 en el modelo de un punto y el modelo basado en regiones. Los porcentajes indican la ratio a los costes totales.

7) Conclusiones

En este proyecto se ha desarrollado una metodología que identifica un número adecuado de regiones dentro de un sistema energético. La red eléctrica correspondiente está sintetizada usando datos de OpenStreetMap y las correspondientes capacidades de transmisión de energía se calculan utilizando valores estándar en función del nivel de tensión. La optimización de la transición hacia la descarbonización del sistema ha mejorado usando las regiones identificadas. Esta metodología también regiones que son de especial importancia para el sistema energético del futuro y dónde deberían instalarse qué capacidades de las diferentes tecnologías. No obstante, el número de regiones puede ser arbitrario debido a un límite que se ha implementado para reducir el tiempo de cálculo. La metodología desarrollada tampoco se aplica a grandes números de regiones (varios miles) debido al elevado tiempo de cálculo del postprocesamiento. Posibles desarrollos adicionales en el futuro incluyen:

1) La especificación de cables de alto voltaje y corriente directa que todavía están simuladas como líneas de corriente alterna
2) Consideración del área disponible para generación renovable
3) Conexiones a otros países con la posibilidad de transferir energía

4) La generación eólica offshore, incluyéndola en el proceso de clustering o después

MII EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

DEVELOPMENT OF A GLOBALLY APPLICABLE METHOD TO ANALYZE REGIONAL DIFFERENCES IN COUNTRY ENERGY SYSTEMS

Autor: Christian Perau

Director: Martin Küppers

Co-Director: Marco Franken

Madrid

Marzo de 2020

Table of Contents

Page

List of figures

page

List of Tables

page

1 Introduction

Meeting the Paris agreement [1], limiting the increase of average temperature to less than 2°C, preferably 1.5°C, requires drastic changes in different sectors. One of the main drivers for climate change is carbon dioxide (CO_2) emissions. Therefore, decreasing emissions is one of the main objectives which the participating countries have. This requires a change of infrastructure in different sectors, mainly electricity, mobility, heat, industry, buildings, and agriculture [2]. Each of those must reduce its emissions. For example, in the electricity sector investments in renewable energy technologies, like wind, solar, hydro, and biomass are necessary. On the other hand, the generation by conventional power plant technologies using coal or lignite which are the technologies with the highest emissions must be reduced. Therefore, one characteristic of future energy systems is their high share of renewable energy and the volatile generation that they entail. Because of this, the transformation of current energy systems towards the future is under investigation.

Besides the investigation of the stability and sustainability of the system, costs are another important factor. Energy system modeling, especially optimization models, are used to calculate the cost-optimal pathway towards a decarbonized future energy system. In such decarbonization pathways the model calculates which capacities of given technologies are installed or shut down under the constraint of being able to meet the specified demand of the energy system and e.g. limits of CO_2 emissions.

1.1 Motivation

In energy system modeling one important factor that has to be considered is the spatial resolution of the model [3]. Optimally each consumer and generator would be represented by one point in space, with the electrical grid connecting them. This approach might be applied to very small cities or countries. However, for most energy systems this results in a number of variables which is too high for an optimality solver to handle. Another approach is to model the energy system of a country as one point, without considering the electrical grid and interactions inside the country. However, this might lead to the neglect of important factors regarding the spatial differences of load, generation and especially transmission and distribution of energy. Because of that an in-between-approach of the two first mentioned approaches is to model the energy system based on regions by modeling their respective energy systems and the interactions between them [4]. Considering this approach, the question emerges on how to define those regions. Intuitively, one might consider the different states of a country as regions. However, this may not be reasonable, since state borders are mostly defined by historical and political reasons. Their regional distribution can be very inhomogeneous regarding renewable potentials, load, or other factors that are important for energy system models. Therefore, the need for a so-called regionalization method rises. In this context, regionalization can be defined as the identification of a small number of groups of internally homogeneous and spatially contiguous regions out of a large set of spatial objects [5]. Using such homogenous regions, an optimality solver can model energy systems taking into consideration the electrical grid between them. This way the validity of energy system modeling might be further improved in comparison to single-region models. Furthermore, the significance of the results improves in comparison to the modeling of administrative levels due to the consideration of regional occurrences.

1.2 Goals and Structure of the Thesis

The goal of this work is to develop a methodology to automatically identify a suitable number of regions inside a country´s energy system which can be used for region-based energy system modeling. On the one hand, the loss of data inside those regions should be as low as possible. On the other hand, the number of regions should be in a reasonable range and reflect the complexity of the country. Another important factor is the electrical grid. Since the regions themselves are modeled as single-region models, they should have a considerable grid to justify this assumption. The methodology shall improve the significance of the results of an already existing energy system model, capable of identifying the cost-optimal decarbonization path by considering the regional differences. On the other hand, computational time and complexity should be kept low.

This thesis is structured as follows. Chapter 2 describes the theoretical basics used in the developed methodology, starting with the energy system model approach used. For the representation of the interconnection between the regions, the basic theory of Net Transport Capacity models is described. Afterwards, different clustering approaches that can be suitable for the regionalization problem and different metrics to evaluate their results are explained. Chapter 3 presents the developed methodology. Starting with an overview of the whole framework, it describes the individual steps in detail. In chapter 4 two use cases show the results which the developed methodology achieves, including a sensitivity analysis of the input data. After the presentation of the cases, a discussion addresses whether the obtained results suit the purpose of this work. Chapter 5 concludes and provides an outlook on possible future developments.

2 Theoretical Basics

This chapter presents and explains the theoretical basics needed for the developed methodology. Firstly, the general aspects of energy system modeling are described. The second section describes Net Transfer Capacity (NTC) models later used for the representation of the electrical grid. Thirdly, different clustering approaches that might be suited for the regionalization problem of this work are explained. The last subsection describes different metrics, called Clustering Validity Indicators (CVI), used to evaluate the clustering results.

2.1 Energy system modeling

This section covers the theoretical basics that are used in energy system modeling. When analyzing supply and demand of energy, energy system models are considerable options. They can provide deep insight and currently receive increased relevance in climate policy, economic development concerns and energy security [4]. Furthermore, they can forecast future energy systems and support a government in its decision-making process [3]. Generally, energy system modeling is divided into two approaches “bottom-up” and “top-down” [11]. The bottom-up approach focuses on the technologies whereas top-down approaches represent a macroeconomic approach. Besides other existing models, this work focuses on a sub-group of the bottom-up approach called optimization models, since they are often used for country-level modeling.

2.1.1 Optimization models

Optimization models are a subclass of the bottom-up approach used in energy system modeling. The focus on technological detail is the main characteristic of bottom-up energy system modeling. However, this technological detail can make them incompatible with very long periods of observation. If the corresponding technologies have short reinvestments cycles - less than 20 years due to developments, cost reductions or other improvements of the considered technologies - it might lead to wrong conclusions [11]. Important inputs for those models include:

- Available commodities (physical and electrical)
- Import, export and conversion processes of those commodities
- Different sectors (energy, heat, mobility)
- Currently installed capacities of different generation technologies
- Technology conversion and storage efficiencies
- Resulting emissions
- Load patterns
- Renewable profiles and potentials
- Operational expenditure, capital expenditure and lifetime of the equipment

Optimization models calculate the cost-optimal mix of generation technologies in the future, minimizing the total costs of the energy system´s operational and investment costs towards a specified goal, e.g. decarbonization of 100%. These approaches can be based on Linear Programs (LP), Mixed-Integer Linear Programs (MILP) or even Non-Linear Programs (NLP) [3]. Further details, such as the maximum installed capacities or the maximum yearly emissions, can be implemented as constraints to the optimization problem. Fixed input parameters of the optimization methods usually include the prices, efficiencies, etc. of the technologies and the total energy demand of the modeled energy system [11].

Optimization models are therefore suitable to calculate the total costs of an energy system. They can also identify the cheapest mix of technologies considering investment decisions, e.g. for the transformation of the energy system from a primarily conventional technology-based to a sustainable system. Furthermore, bottom-up models can reflect technologies that work between sectors like combined heat and power [3].

2.1.2 Regional modeling

Important factors of modeling an energy system are the spatial and temporal resolution. Increasing either of those resolutions leads to an increase of complexity and, therefore, computation time [3]. Nevertheless, this is also a crucial factor depending on the research question, i.e. questions regarding the stability should have a higher temporal resolution than questions regarding future investments. This work, however, focuses primarily on spatial resolution.

Modeling the energy system using a spatial resolution is important for considering the spatial distribution and temporal fluctuations inside those regions regarding renewables, load, and generation. Considering the energy system of a country, neglecting different regional properties, might lead to an overestimation of the spatial availability of renewable power generation within one or more regions [4]. This is especially crucial regarding decarbonized future energy systems with a high share of renewables. Spatial models, on the other hand, require infrastructure to model interactions between the regions. This infrastructure, in the case of energy system modeling, can be the electrical grid.

The important trade-off that has to be considered is spatial and temporal resolution against computation time and complexity. While the highest spatial resolution would be to consider each consumer and each producer in space as one point, doing so would lead to an unfeasible problem or unreasonably long computation times [3]. Therefore, the need for a methodology that aggregates the spatial points into bigger regions emerges. This increases the spatial resolution compared to country-level modeling while ensuring to solve the problem in a reasonable amount of time. Using the electrical grid as interconnection the transport of electricity can be modeled. Considering this, it becomes clear which way the power flows have to go and, therefore, between which regions additional transmission capacities need to be installed considering the future energy system. The next section presents how the electrical grid can be included. Different approaches of how to identify regions within a country are explained in 2.3.

2.2 Net transfer capacity models

An important question that has to be answered regarding the regional modeling of an energy system is how those regions can interact with each other. This can be done by considering the electrical grid as an interconnection between the regions [12]. Often-used examples for this are NTC models between countries [13]. In those cases, different countries are modeled, and their interconnecting lines are seen as the corresponding connections between the different countries. Regarding NTC models, following distinctions between capacities have been agreed on by the European Transmission System Operators (TSO) [13]:

- Net Transfer Capacity
- Available Transfer Capacity (ATC)
- Transmission Reliability Margin (TRM)
- Notified Transmission Flow (NTF)
- Total Transfer Capacity (TTC)

The TTC is the maximum transferable capacity between two systems, which does not violate any system security constraint. Those constraints can either be thermal limits of the transmission lines, voltage limits or stability limits [13]. The maximum transferable capacity however depends on the state of the two systems, i.e. the generation and load situation in the specific moment.

(2.1) and (2.2) describe the relations between NTC, TTC and TRM and ATC, NTC, and NTF, respectively

$$NTC = TTC - TRM \tag{2.1}$$

$$ATC = NTC - NTF \tag{2.2}$$

Because the TTC depends on the load and generation situation the TSOs use the TRM to cover uncertainties of power flows that result from imperfect information and unexpected real-time events. This value can be evaluated by past experiences or statistical methods [13]. One way is to set the NTC to a suitable fraction of the theoretically possible capacity [14]. Therefore, if in an N-0 case, i.e. all equipment is available, one line utilizes 70% of its maximum capacity, it is expected to be overloaded in the N-1 case. This means the line will fail if another equipment fails and is not available anymore. Therefore, the NTC can be seen as the expected maximum power flow that can go from one system into another, which does not violate any system security constraint. However, it has to be kept in mind that the NTC from region A to region B is not necessarily equal to the NTC from B to A, since it depends on the current system load and generation situation [13].

The NTF, therefore, describes the part of the NTC that, in a particular moment, is already assigned to one or more market participants by transfer contracts. The ATC, on the other hand, is the part of the NTC that is still available for market participants.

2.3 Spatial clustering

Spatial Clustering is investigated as an unsupervised method of machine learning used for classification [15] and in the fields of spatial data mining [16]. Many different approaches exist in the literature. Nevertheless, none has been found optimal for all kinds of clustering problems [17]. Therefore, there is no generally optimal clustering algorithm available and choosing the clustering algorithm highly depends on the specific problem [15]. In general, clustering a data set of *n* data points should result in a number of *k* Clusters, *k<<n*, grouping this data set. The objectives are to minimize the sum of the intra-cluster distances, i.e. the dissimilarity of the data points inside a cluster, and to maximize the sum of the inter-cluster distance, i.e. the similarity between two clusters. The intra-cluster distance can be formulated as

$$\sum_{x \in c} \sqrt{\sum_{i \in I} (x_i - \bar{x}_{c,i})^2} \tag{2.3}$$

Where *c* is one of the *K* clusters, *x* is a data point and member of cluster *c*, *I* are the attributes of member *x* and $\overline{x_c}$ is the centroid of cluster *c*, which attributes are the mean values of the corresponding attributes of all members of *c*. Therefore, one objective is to minimize the sum of the intra-cluster distance over all clusters. The inter-cluster distance, on the other hand, is the distance between two cluster centroids. A clustering algorithm should maximize this distance since similar clusters are not desirable. As shown in (2.3), the Euclidean distance is used. Nevertheless, any other form of distance measure can be applied as well.

Choosing an appropriate clustering algorithm strongly depends on the problem that it is applied to and on the results that are desired. Hence, only because one approach is suited for a specific problem it does not necessarily perform well on another kind of problem. Apart from traditional clustering [18], different subgroups can be categorized. Examples are the clustering of time series data [19], spatial clustering [15] [17], and spatio-time clustering [20]. Since the regionalization problem requires spatial contiguity of the identified clusters, spatial clustering has to be applied. The basic principles, however, are the same.

Spatial clustering can be categorized into different approaches. Some of the most widely used are partitioning methods, hierarchical clustering, density-based methods, grid-based methods, and the max-p regions algorithm. Different spatial clustering approaches have been used in different thematic fields, e.g. image processing, genetics, social science, earth science, public policy and natural resources management [21], [22], [23]. Even though the previously mentioned approaches assign each data point to one and only one cluster other approaches have been made [24]. Since in energy system modeling, however, each data point should belong to only one cluster, those approaches are not further considered. In the following, the different approaches are explained and their suitability for the regionalization problem is discussed.

2.3.1 Partitioning methods

Partitioning methods are widely used in the literature. This is mainly due to its simplicity and fast performance. Examples of partitioning methods are K-means, Expectation-Maximization, and K-medoids. K-means is one of the most used clustering algorithms in different applications since it is very fast. Its time complexity is *O(nkl)* [25], *n* being the number of data points, *k* the number of clusters and *l* the number of iterations. In the following, K-means is explained to show the basic principles of partitioning methods.

The K-means algorithm tries to minimize a linear objective function which usually is the sum of the intra-cluster distance (2.3) over all clusters, but can also differ. As input, the algorithm expects the number of clusters, *K*, the user wants to obtain by clustering. Hence, the algorithm does not identify the optimal number of clusters. This number has to be obtained by other means, for example by considering clustering validity indicators, which are discussed in section 2.4. With the information on the number of clusters that is to be obtained the algorithm sets *K* so-called centroids in the data space, which each represent one cluster. Then it performs iterations of the following steps. First, each data point is assigned to the closest centroid. After that, the cluster centroids are recalculated as the mean of all data points belonging to the cluster. The last step is calculating the value of the objective function.

The algorithm terminates when either the change in the objective function is below a given threshold or a maximum number of iterations is reached. The algorithm is presented in Figure 2.1.

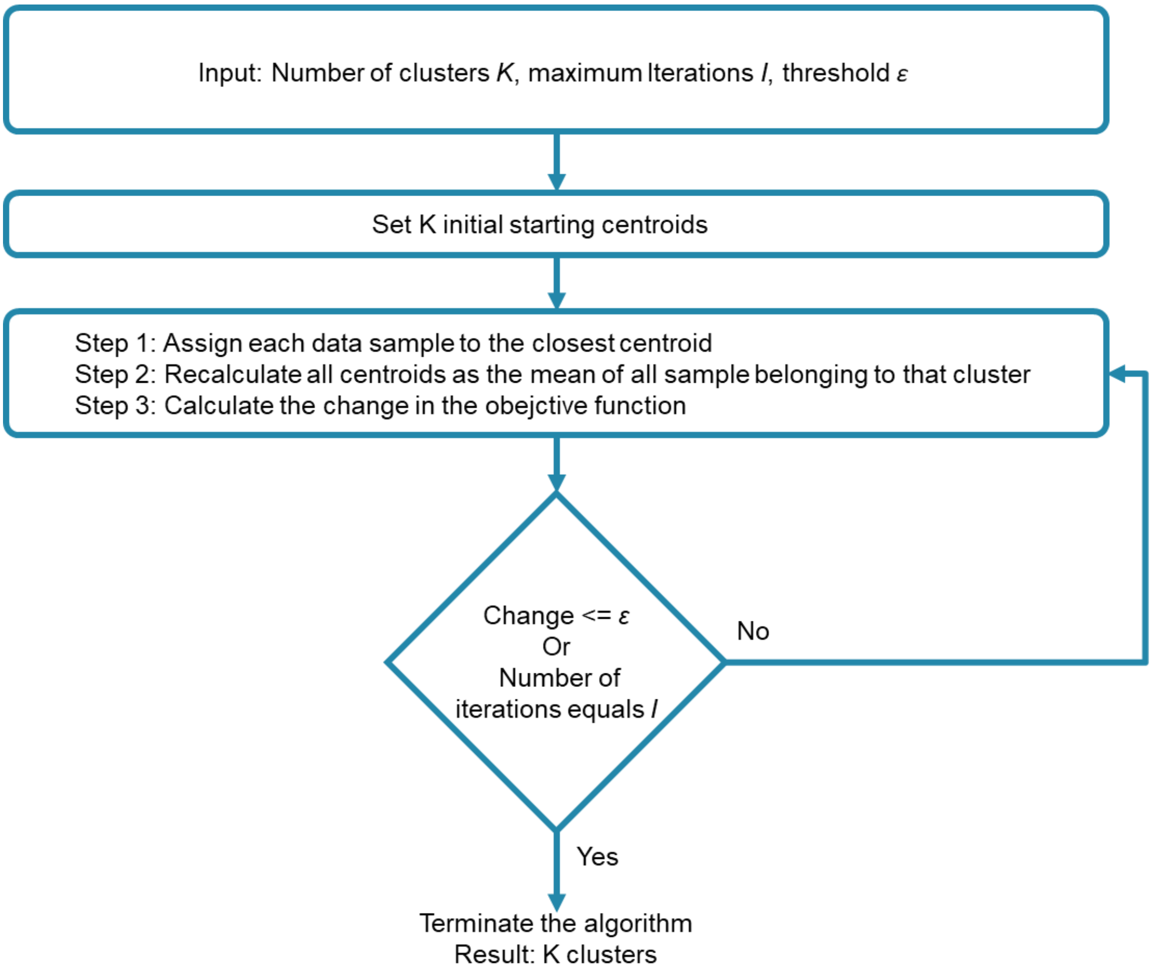
Input: Number of clusters *K*, maximum Iterations *I*, threshold *ε*
Set K initial starting centroids
Step 1: Assign each data sample to the closest centroid
Step 2: Recalculate all centroids as the mean of all sample belonging to that cluster
Step 3: Calculate the change in the obejctive function
Change <= *ε*
Or
Number of
iterations equals *I*
No
Yes
Terminate the algorithm
Result: K clusters

Figure 2.1: K-means algorithm

A general advantage of the K-means algorithm is its simplicity. Furthermore, it is very fast and likely to find very dense clusters. Nevertheless, its results are highly dependent on the initialized cluster centroids. There are methods to improve the setting of initial cluster centroid, such as K-means++ [26], but the results stay dependent on random inputs. This mostly leads to the solution that the K-means algorithm is executed various times and the result with the lowest sum of intra-cluster-distance is chosen as the optimum. This indicates that the algorithm finds local optima, but it is not assured to find the global optimum. Another problem with the K-means algorithm is its tendency to identify spherical clusters of the same size [19]. In the case of the regionalization problem, this is not necessarily intended since there might be regions with similar attributes but resulting in different sizes and forms, e.g. coast or deserts. Furthermore, K-means is not compatible with a spatial contiguity constraint. Therefore, often the spatial coordinates are used as attributes with a higher weight than the other attributes so that only spatial contingent regions are grouped. This might lead to a distortion in the cluster results since the other attributes are weighted less.

Tobler´s first rule of geospatial analysis says that points close to each other tend to be alike [27]. Although this is certainly true in cases of wind and solar potential, considering load distribution in urban areas, one region may be very distinct from its spatial neighbors.

Some examples where K-means performed spatial clustering are to perform denoising in image processing [22], and as a first step process in the regionalization for Europe considering different input data [28].

2.3.2 Hierarchical clustering

Hierarchical clustering is another approach that comes from standard clustering. It can be divided into two approaches agglomerative and divisive. In the agglomerative approach, the algorithm starts by assigning each of the *n* data points to an own cluster i.e. it starts with *n* clusters. Afterwards, the two most similar clusters are merged. This procedure is done until only one cluster is left, i.e. after step *m*, there are *n-m* clusters left. This way the output of the hierarchical clustering approach is a hierarchy indicating which regions should be grouped to obtain a certain number of clusters. This hierarchical structure can be depicted in so-called dendrograms. An example is presented in Figure 2.2.

In this figure, the vertical axis represents the distance between two clusters, whereas on the horizontal axis the clusters are represented. In this example the clusters 6 and 7 are very similar, hence the distance between them is small. Due to this similarity, those two are merged into one new cluster. Afterwards, the clusters are merged until all regions are in one cluster. Like K-means, the algorithm does not calculate an optimal number of clusters. The optimal number of clusters can be found by using CVIs, which are described in 2.4.

The divisive approach works the opposite way. Starting with one cluster containing all data points the cluster is divided into sub-clusters until there is one cluster for each data point or a given number of clusters has been obtained. This can be done for example by optimum cuts of a minimum spanning tree [15]. Even though divisive clustering can obtain better results than the agglomerative approach, in the following it is not considered, because it usually only takes into account the distances between two adjacent regions but not the other regions within the same cluster.

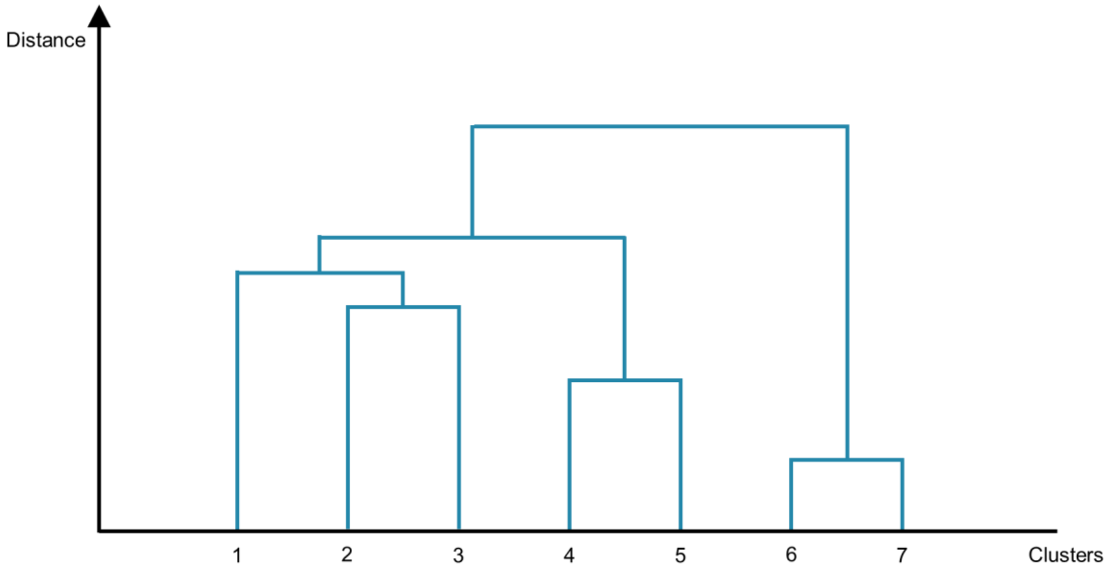

Figure 2.2: Example of agglomerative hierarchical clustering dendrogram

Two important factors have to be considered when using the agglomerative hierarchical clustering: the linkage criterion and the affinity. The latter describes which distance metric is used to calculate the distance between two data points. *Euclidean* and *Manhattan* or C*ityblock* distance are examples [18]. The linkage criterion, on the other hand, describes the distance between two clusters A and B that contain certain numbers of data points n_A and n_B, respectively. The most common approaches for the linkage criterion are [10]:

- Single-linkage: In the single-linkage, the distance between two clusters is the minimum distance of two data points belonging to the two different clusters. Mathematically it can be expressed as:

$$distance(A,B) = \min\big(distance(x,y)\big)\ \forall\ x \in A, \forall\ y \in B \tag{2.4}$$

- Complete-linkage: In the complete-linkage, the distance between two clusters is the maximum distance of two data points belonging to the two different clusters. The equation can be expressed as:

$$distance(A,B) = \max\big(distance(x,y)\big)\, \forall\, x \in A, \forall\, y \in B \tag{2.5}$$

- Average-linkage: The average-linkage defines the distance between two clusters as the average distance between all data points belonging to those two clusters. It is calculated as:

$$distance(A,B) = \frac{\sum_{x \in A} \sum_{y \in B} distance(x,y)}{n_A * n_B} \tag{2.6}$$

- Centroid-linkage: The centroid linkage does not compare the data points as they are but the centroids of the clusters, which are calculated as the mean of all the data points in all attributes in a cluster:

$$distance(A,B) = distance(\bar{x}, \bar{y}) \tag{2.7}$$

- Ward´s method: Instead of defining a distance between two clusters, in each merge step, it is considered which merge results in the lowest increase in intra-cluster distance. When finding this minimal merge combination, it is applied to the clusters. This linkage criterion, however, is only defined for the Euclidean distance.

If spatial contiguity is considered within the hierarchical clustering it can be achieved via a contiguity constraint, i.e. only two clusters that are spatial contiguous are considered for the next merge. This reduces the computational complexity and computation time of the hierarchical clustering without the contiguity constraint of *$O(n^2)$* since fewer combinations of merges exist. There are typically two ways to handle the spatial contiguity, *first-order* or *full-order* constraints [29]. In the first approach, the distance between two clusters is defined using only the regions that are spatially contiguous between those two clusters. On the other hand, the *full-order* constraint considers all regions belonging to the clusters when calculating the distance. Not including the Ward´s method, the *full-order* constraint most of the time achieves better results [29].

One disadvantage of this approach is that after two clusters are merged, there is no reassignment of the different data points to other clusters like e.g. in the K-means algorithm. This might lead to non-optimal results and therefore shows the hierarchical clustering is a heuristic approach. Nevertheless, a greedy optimization algorithm, called *fine-*

tuning, can reassign regions into other clusters [7]. In the algorithm it is considered to shift a region belonging to one cluster in another cluster that is spatially contiguous to that region. Afterwards, it is observed whether the result improves the starting solution of the hierarchical clustering. This way possible sets of changes are identified which are applied, if improving the result. For a detailed description of the metrics indicating whether a clustering result improved see 2.4. Figure 2.3 presents the algorithm.

As an input, the fine-tuning algorithm receives an already clustered set of regions, and a criterion to evaluate the clustering results. It also creates a new attribute of the regions called *"moved"*, which indicates whether the region has been moved already trying to improve the given criterion. This attribute is set to "*Fals*e" for each region at the beginning of the algorithm.

Firstly, it is checked which regions can be moved without violating the spatial contiguity of the clusters. This means a region cannot be moved if by its removal the former cluster has to be split into two clusters, because it would not be spatially contiguous anymore without the moving region. With those candidates, it is checked which change of all possible changes to another spatially contiguous cluster improves the criterion the most or – in case of no improvement – worsens it the least. This best change is applied, and the *moved* attribute of that region is set *"True"*. If all potentially movable regions have been changed already the algorithm is terminated.

Afterwards, the best sequence of changes is applied to the clustering result. If the value of the criterion stays the same or deteriorates, the algorithm is terminated. Otherwise, the *moved* attribute of the regions is reset and the next cycle of fine-tuning with the recently identified clusters is started. The problem with the proposed method is its complexity of $O(n^2rd)$ with n being the number of regions r being the number of iterations of the fine-tuning and d the number of attributes. Therefore, this method might not be suited for big data sets.

Another drawback of the hierarchical clustering is its sensitivity to outliers. In the regionalization problem, this can lead to the formation of small regions that are kept as single clusters while other clusters expand very fast. The second behavior is also called "rich getting richer" and is of especial importance when considering a contiguity constraint and most common when using single-linkage [30].

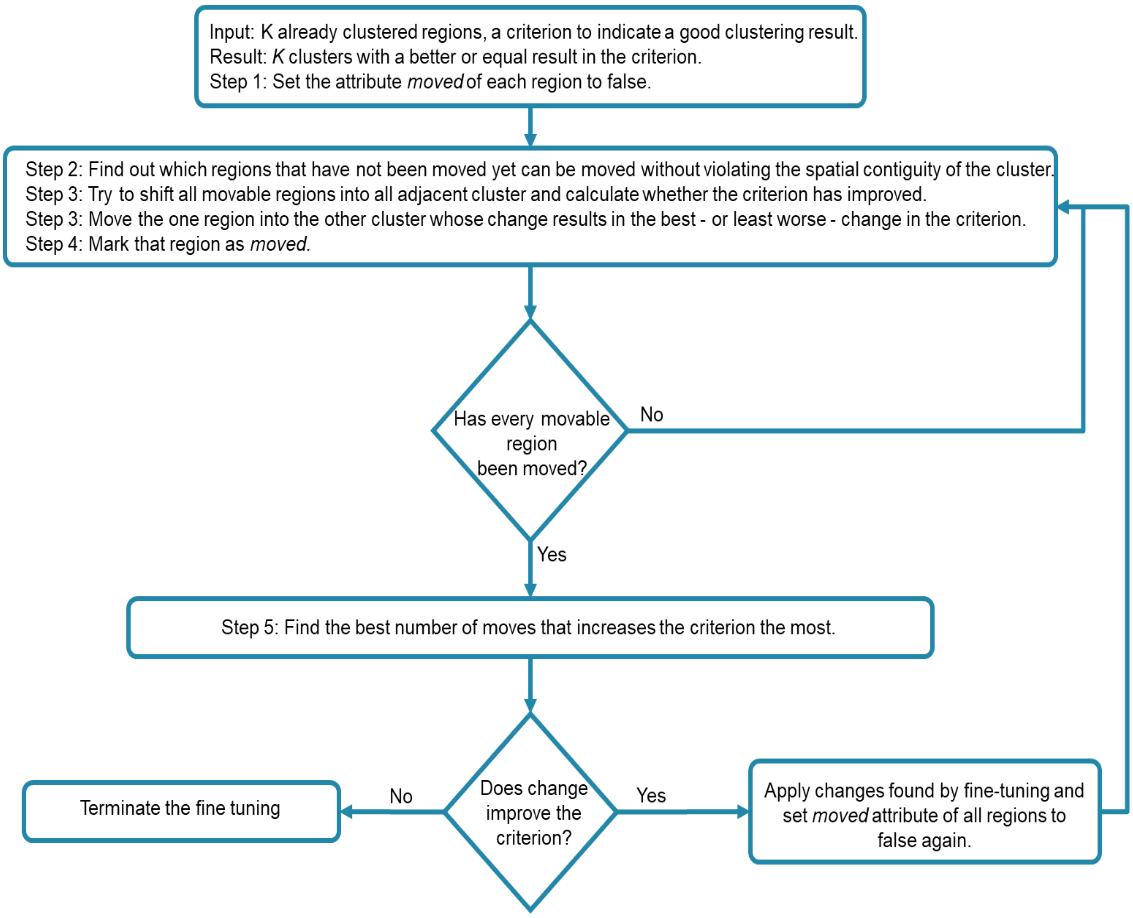
Input: K already clustered regions, a criterion to indicate a good clustering result.
Result: *K* clusters with a better or equal result in the criterion.
Step 1: Set the attribute *moved* of each region to false.
Step 2: Find out which regions that have not been moved yet can be moved without violating the spatial contiguity of the cluster.
Step 3: Try to shift all movable regions into all adjacent cluster and calculate whether the criterion has improved.
Step 3: Move the one region into the other cluster whose change results in the best - or least worse - change in the criterion.
Step 4: Mark that region as *moved.*
Has every movable region been moved?
No
Yes
Step 5: Find the best number of moves that increases the criterion the most.
Does change improve the criterion?
No
Yes
Terminate the fine tuning
Apply changes found by fine-tuning and set *moved* attribute of all regions to false again.

Figure 2.3: Fine-tuning for hierarchical clustering

Regarding spatial clustering, hierarchical clustering has been used e.g. as a second step clustering to identify ecological regions [31] and to define seven climate zones in turkey using the Ward´s method, comparing it to other linkage criteria [32]. It has been used for the regionalization of forest patterns in the United States, also comparing it to the k-means algorithm without spatial contiguity constraint [33]. However, no strict contiguity constraint was implemented, which lead to scattered clusters.

2.3.3 Density-based methods

As described before, partitioning methods use the distance between data points to identify clusters. The density-based methods instead use the density of data points to identify clusters of high-density occurrence and outliers as noise. One common density-based method used is DBSCAN [34]. As input arguments, it needs ε, a distance threshold by which it is determined whether two data points can be in the same cluster, and a threshold of a minimum number of points *MinPts* which must be within the ε-distance to be not seen as noise. The algorithm identifies the clusters by identifying so-called *core objects*. Those are data points in which, within their ε-distance, at least *MinPts* other data points can be found. The algorithm follows four basic rules [15]:

1. A data point can only belong to one cluster and only one which has a *core object* within the ε range of the data point.
2. Two *core objects* which distance is at most ε form a conjoint cluster.
3. If a *non-core object* is within the ε distance of two *core objects* which do not belong to the same cluster the data point must belong to only one cluster.
4. A data point that is further away than ε to any *core object* is considered noise.

An example of how the DBSCAN algorithm performs is shown in Figure 2.4 with *MinPts* equals three and a given ε distance threshold.

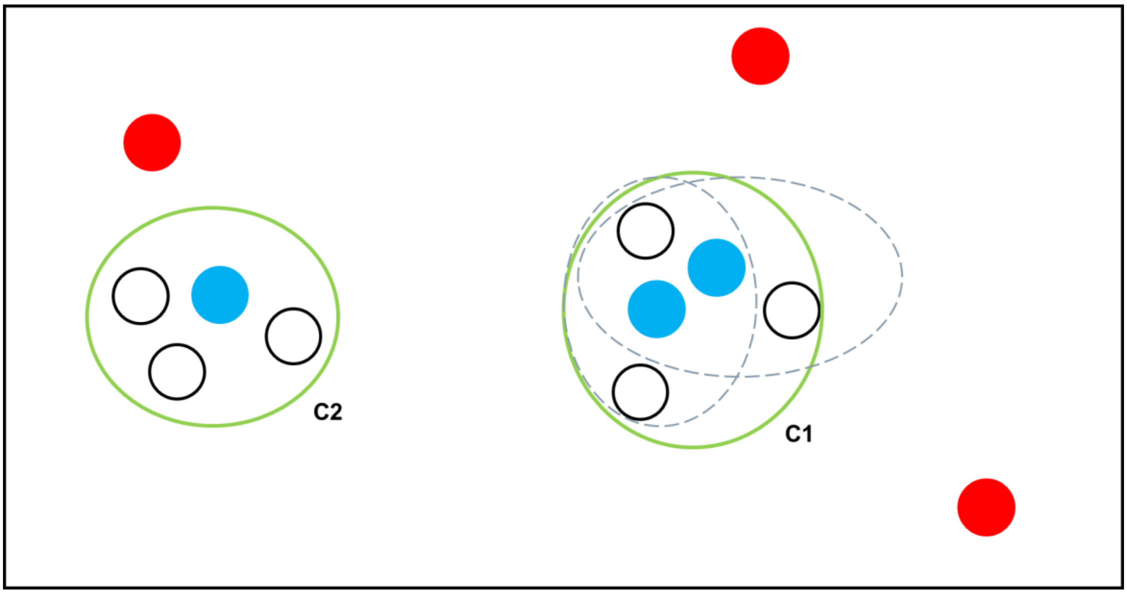

Figure 2.4: Example of the DBSCAN algorithms performance

The blue objects represent *core objects* while the red represents noise. Cluster C1 consists of two core objects that have 3 neighbors within their ε distance.

A general drawback of the DBSCAN algorithm is its input parameters. Depending on which values for ε and *MinPts* are chosen the results can change drastically. Therefore, the user´s choice of those input data is crucial for the algorithm. On top of that, the actual data points influence the choice of the two threshold values as well, which does not suit a generalized methodology. However, some of those inconveniences have been investigated and improved in different works [35].

Another disadvantage is the problem that DBSCAN does not perform well with high dimensional data [15], which could be of importance if time series data are considered as attributes.

2.3.4 Grid-based methods

As mentioned before, density-based methods are not suited for high dimensional data sets. Grid-based methods offer a better approach regarding those kinds of data sets [15]. Commonly used approaches of the grid-based methods are STING (STatistical INformation Grid) [36] and CLIQUE [37]. Grid-based methods divide the space in a finite number of cells forming a grid structure used for the clustering and applying the following steps [25]:

1. Definition of a set of grid-cells to divide the data space.
2. The assignment of the data points to the corresponding grid-cells and calculation of the grid-cells´ density.
3. Neglection of those grid-cells that have a density lower than a user-specified threshold value *t* (CLIQUE) or a hypothesis test with a user-specified confidence interval is not met (STING).
4. Clustering of contiguous grid-cells.

A general advantage of this method is that it is easy to parallelize and therefore efficient [25]. But as in the density-based methods, the grid-based methods rely on user-defined threshold values which makes the algorithm susceptible to inaccurate user inputs [15]. Another input is the definition of grid-cells. In STING it is done by a hierarchical structure in which the resolution is increasing in each layer [15]. But this raises the question of when the correct resolution is reached.

2.3.5 Max-p regions

The max-p regions algorithm is a clustering approach using a MILP [38]. The objective function consists of two terms, one is the intra-cluster distance which is minimized and the other defining the optimal number of clusters. A threshold constraint specifies a condition that has to be met for each cluster of regions, e.g. the solar peak capacity installed has to be greater than 1 MW or the conventional generation has to be lower than 3 GW. This constraint can be anything alike and is used to determine whether a cluster can exist as it is or if regions have to be added to or removed from the cluster to fulfill the constraint. This way the algorithm returns an optimal number of *K* clusters of the former regions.

A drawback of this approach is its high computational complexity of $O(n^3)$. Therefore, a high number of regions lead to high computing time or even termination of the algorithm because of non-feasibility. Nevertheless, different approaches exist trying to decrease its complexity [39]. One example of how to do achieve this is to first solve a relaxed version of the problem and then add constraints iteratively to ensure the spatial contiguity.

Another problem is the constraint that has to be set. Setting such a constraint can be rather subjective. One possibility is to use the variance in the clusters as such a constraint [28], e.g. Setting an upper limit for the variance inside one cluster. Nevertheless, setting this threshold also is subjective and may result in problems considering the global applicability of the methodology.

2.4 Clustering validity indicators

CVIs are used to evaluate the performance of the clustering algorithm [40]. They can be categorized into two types. The first evaluates the results knowing the true clusters. The second one does not need any true labels for the data points, which is the case for the regionalization process since it is not known what the optimal clusters are. For those algorithms which do not result in an optimal number of clusters, those CVI are often used to identify this number. Nevertheless, the optimal number of clusters can also be driven by a specific problem. For example, the number of clusters might need to be below a given threshold. The general concept of finding the optimal number of clusters usually is to plot the CVI to the number of clusters. Either by visual inspection or other methods, the optimal number of clusters can be identified using the resulting curve.

However, finding the optimal point depends on the chosen CVI, since some indicate good results by high scores, whereas others may use a low score to express a well-performed clustering. Other methods of how to determine the optimal number of clusters are e.g. a comparison of the slope of the CVI of the real problem to other randomly computed clustering approaches. This way the statistical significance of the result can be considered [41]. In the next sections, some frequently used CVIs are explained, namely the silhouette coefficient, the Calinski-Harabasz index, the Davies-Bouldin index, and the elbow criterion. Nevertheless a variety of other CVIs exist [18] [42] [40].

2.4.1 Silhouette coefficient

The silhouette coefficient [43] is a frequently used CVI [44] [45]. The formula to calculate the silhouette coefficient *s* for a single data point is the following:

$$s = \frac{b - a}{\max(a, b)} \tag{2.8}$$

Where *a* is the mean distance between the data point and all other data points in the same cluster and *b* is the mean distance between this data point and all other data points in the nearest cluster. The overall silhouette coefficient is the mean of all silhouette coefficients of all data points in the set. The distance can be defined as any kind of distance between two data points for example Euclidean distance.

The advantages of the silhouette coefficient are its boundaries of minus one and one for incorrect clustering and optimal clustering, respectively. Scores around zero, on the other hand, indicate the existence of similar clusters in the results [30]. This means the desired outcome of highly dense and well-separated clusters is indicated by a high score. A drawback of this CVI is that it depends on the clustering approach used, i.e. density-based methods generally result in lower scores than other approaches [30]. In this approach the optimal number of clusters is the one achieving the highest score.

2.4.2 Calinski-Harabasz Index

The Calinski-Harabasz Index [46], also called Variance Ratio Criterion, is another CVI that indicates well-defined clusters with a high score. It is calculated by the ratio between the sums of the squared inter- and intra-cluster distances over all clusters [30]:

$$s = \frac{tr(B_k)}{tr(W_{k)}} \times \frac{n_E - k}{k - 1} \tag{2.9}$$

Where *tr(B_k)* is the trace of the inter-cluster distance

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q) \times (x - c_q)^T \tag{2.10}$$

and *tr(W_k)* is the trace of the intra-cluster distance

$$B_k = \sum_{q=1}^{k} n_q \times (c_q - c_E) \times (c_q - c_E)^T \tag{2.11}$$

With *E* being the entire data set of size n_E divided into *k* Clusters, C_q the set of points in cluster *q*, c_q the center of cluster *q*, c_E the center of *E*, and n_q the number of points in cluster *q*. Its advantages are its fast computing time and presenting well separated and dense clusters with a high value, which is intuitive. The silhouette coefficient depends on the clustering approach used, i.e. density-based methods generally result in lower scores than other approaches [30]. The optimal number of clusters, therefore, is the one that achieves the highest Calinski-Harabasz Index.

2.4.3 Davies-Bouldin Index

The Davies-Bouldin Index [47] compares the similarity between clusters to evaluate whether a clustering performed well. The mathematical formulation of this CVI is:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{2.12}$$

where

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2.13}$$

and s_i and s_j are the inter-cluster distances of cluster *i* and *j*, respectively, and d_{ij} is the distance of the cluster centroids of the clusters *i* and *j*. This can be seen as a trade-off between inter- and intra-cluster distance. But instead of considering all results only the maximum values, i.e. close cluster with low inter-cluster distance are used for the final score. Furthermore, this implies that, in contrast to the first-mentioned CVIs, the Davies-Bouldin Index does not relate a high but a low value to a well-performed clustering, with zero being the best score possible. This has to be kept in mind when evaluating the optimal number of clusters. Although having the same drawback as the first two mentioned CVIs, the Davies-Bouldin index is simple to compute [30]. The optimal number of clusters is the one achieving the lowest score in the Davies-Bouldin Index.

2.4.4 Elbow method

The Elbow method [6] is mostly used in K-means but is also applicable to other clustering approaches, e.g. hierarchical clustering [7]. The score is the sum of the intra-cluster distances (2.3) of all clusters. If those scores are plotted against the number of clusters, it results in a graphic similar to Figure 2.5.

Since a lower value represents a better clustering result the optimal point would be the number of clusters equal to the number of data points since its intra-cluster distance is zero. Therefore, it is not reasonable to choose the lowest value of the score. The curve in Figure 2.5 looks similar to a human elbow. In the point where the elbow is, the optimal number of clusters is expected. This means the optimal number of clusters corresponds to the point where an increase of the number of clusters does not result in a decrease of intra-cluster distance that is considered sufficient, i.e. very low compared to the reduction of the previous steps. For a smooth curve this corresponds to the point where the second derivative is zero.

However, looking at curves of real clustering results they are not as smooth and might contain no point or even multiple points meeting the condition. Identifying this elbow visually can, therefore, be ambiguous. To automatically identify this point there are different methods [48]

1. Select the optimal point as the point where the largest decrease occurs.
2. The point is selected where the largest ratio difference between two points occurs.
3. The second derivative is calculated and the first point where its value is higher than a given threshold is selected as optimal.
4. The optimal number of clusters is the point with the largest second derivative.
5. A line between the smallest and largest number of clusters is drawn and the point on the curve that is furthest away from this line is considered optimal.

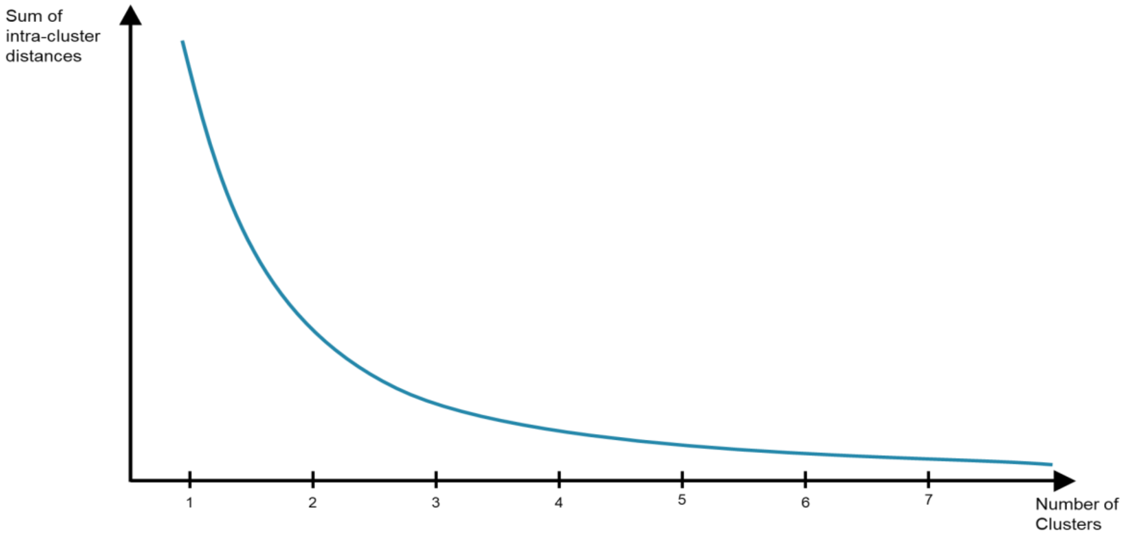

Figure 2.5: Elbow criterion

In some cases, the sums of the intra-cluster distance are also divided by the corresponding number of clusters. This leads to a visually “flatter” curve that looks very similar to the example in Figure 2.5. However, this often leads to an optimal number of clusters in a very restricted interval of small numbers of clusters, since the drop of the sum is decreasing much faster. This is not suited for countries that have a high number of distinct regions and should be represented with a higher number of clusters.

Unlike the first three CVI described, the elbow criterion does not consider the inter-cluster distance. By applying a clustering approach, one might receive a number of clusters as optimal which contains dense clusters but also one or more which are very similar to another cluster.

3 Methodology and Implementation

This section describes the developed methodology. Firstly, the framework in which the methodology is implemented is described. Secondly, the data that are taken into consideration for clustering are presented. Afterwards, the methodology and the different steps are described in detail. The last chapter compares the developed methodology to other existing approaches.

3.1 Framework

The methodology is implemented as a pre-process for energy system modeling. In this work the latter is done using the Energy System Development Plan (ESDP), an existing energy system optimization model [10]. The whole framework is depicted in Figure 3.1, where the blue frame indicates the existing framework and the green the new methodology.

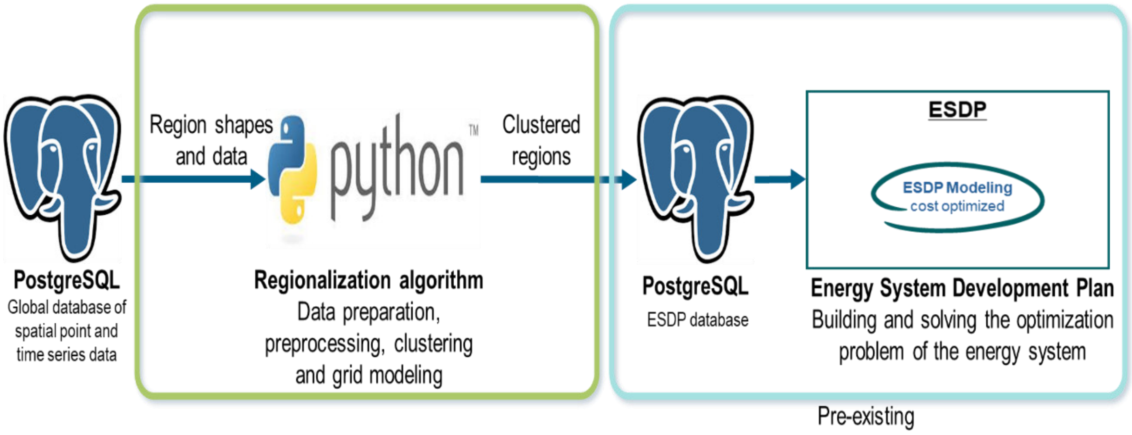
Region shapes and data
python™
Clustered regions
PostgreSQL
Global database of spatial point and time series data
Regionalization algorithm
Data preparation, preprocessing, clustering and grid modeling
PostgreSQL
ESDP database
ESDP
ESDP Modeling
cost optimized
Energy System Development Plan
Building and solving the optimization problem of the energy system
Pre-existing

Figure 3.1: Complete Framework

The workflow is as follows: the PostgreSQL database stores the data of load, generation, renewable profiles, etc. with a spatial resolution on a global basis. The data corresponding to one or multiple countries or other defined regions, including their shapes, are imported into a Python environment. This is where the regionalization algorithm, described below, is applied to the data. After the regionalization, the identified regions, their corresponding data, and the grid connections between them are saved into the PostgreSQL database of the ESDP standard framework. ESDP itself imports those data and uses them for the regional energy system modeling and the calculation of the cost-optimal decarbonization pathway.

3.2 Data used for the clustering process

Since clustering is a data-driven process, the input data are of major concern. This work tries to improve the results of energy system modeling. Therefore, the data that are of special importance for the modeling should be considered in the regionalization. Furthermore, clustering can only perform as well as the input data are. Giving the algorithm incomplete data can result in distorted clusters of regions that are not suitable for energy system modeling. Therefore, an important question is, which data to consider for the regionalization problem of this work. This work uses the following data

- Annual consumption values of the private household (PHH) sector
- Annual consumption values of the commerce, trade, service, and industry (CTSI) sector
- Installed capacities of different available technologies
- Wind and Photovoltaic (PV) profile data for a complete year in an hourly resolution

Other data that are important but due to the lack of data could not be included are

- Spatially dissolved load profile data
- Potential space for the installation of renewables, i.e. available rooftop area for PV and land for wind turbines, respectively

The applied data are stored with Geographic Information System (GIS) data. GIS data are used for the spatial resolution of the global database. They can be either points, lines, polygons or grouped versions of those, i.e. multi-points, -lines, and -polygons, respectively. The geographical information regarding their position, i.e. longitude and latitude, is assigned to the data. Since those might not take into account spatial bends, projections are used. Hence, those data are projected into different spatial reference systems with so-called Spatial Reference IDs (SRID) [49]. Due to the global applicability of the methodology, the SRID used is the ESPG 4326, which can be used worldwide and considers longitude and latitude. Connecting those spatial data to other types of data, e.g. load, generation, wind or solar profiles, GIS data are a powerful tool for spatial analysis in different fields.

Region shape data

One type of data that are stored in the global database are the shape data of different countries. Considering the regionalization of a country´s energy system there are two approaches. Either small predefined regions or point data can be used as input for the spatial clustering. The approach used in this work is to start at a basic level with the so-called *smallest regions* [50], which are the smallest administrative regions of a country available. Those correspond mostly to cities or other forms of districts. This approach is chosen since, from a political point of view, considering one part of a city to be in cluster A and the other part belonging to cluster B is not desirable. Therefore, the data described below are aggregated to these smallest regions. The details are described in 3.3.1.

Load and generation data

The load data used in this work consist of polygons of 50 m x 50 m assigned with the corresponding values of electrical energy consumption of one year in MWh. As described those data are grouped into PHH [51] [52] and GHDI [52] [53]. Since this work considers starting with predefined regions, those data have to be aggregated to those regions. However, the smallest regions can be arbitrarily shaped. Therefore, the polygons of the load values can be located in more than one shape. Because of that, only the centroids of those polygons are considered when assigning the data to the region its center is located in. The generation data, on the other hand, are stored as point data which is why those data can be assigned to the one region, they are located in. Furthermore, their respective generation capacity in MW and the technology it is based on are known [54].

Renewable profile data

Stored as grid cells of 2500 km^2 over the whole world, wind and PV profiles derived from climate observations [55] can be assigned to the different smallest regions that are to be considered within the regionalization process. Figure 3.2 depicts the grid cells for the case of Belgium.

Those data are stored as time series ranging in each time step between zero and one, indicating the relative output of the capacity which a wind turbine or PV system, respectively, can generate at the corresponding moment in time. Since those values also depend on the model of wind turbine and the orientation of the solar panels, respectively, the following is considered within this work. For the wind profiles, this work considers the relative output that a wind turbine of the model Siemens SWT-3.6-107 with a hub height of 100 m can achieve according to the given data. This model is chosen, because of its average height and, therefore, not overestimating the potential power generation of wind turbines. The PV profiles, on the other hand, are available in commonly used steps of the orientation of 22.5° and roof slopes between 0° and 40° in 10° steps. The corresponding value for the considered time series is calculated as the average value of the different orientations and slopes. This is done to better take into account different installation conditions that are likely to occur considering rooftop PV installations.

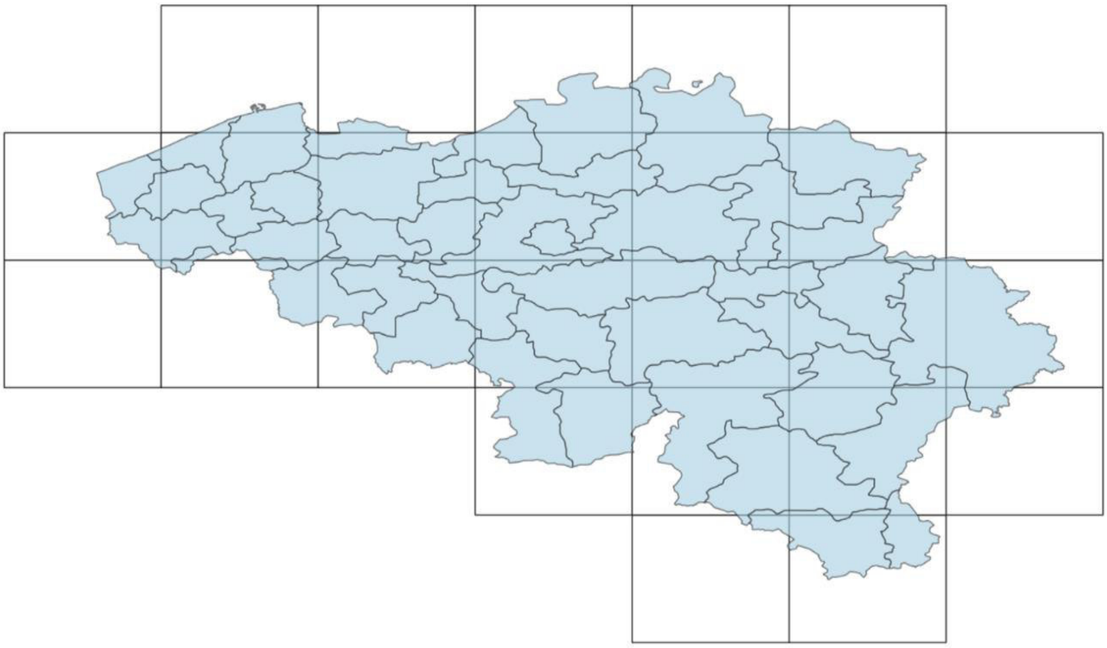

Figure 3.2: Grid cells for wind and PV time series data for the case of Belgium

3.3 Methodology

This section describes the developed methodology in detail. The outcome of any methodology is driven by the requirements that it should be able to fulfill. Therefore, defining the requirements highly influences how the methodology can be developed. The requirements for the regionalization methodology that this work tries to develop are the following:

- The loss of data should be minimized
- The number of regions should be selected regarding the complexity of the investigated county
- The electrical grid should be considered
- Globally applicable
- Adequate run time

The next subsections present the different steps of the developed methodology and considering whether the requirements can be met. Figure 3.3 presents an overview of the complete method.

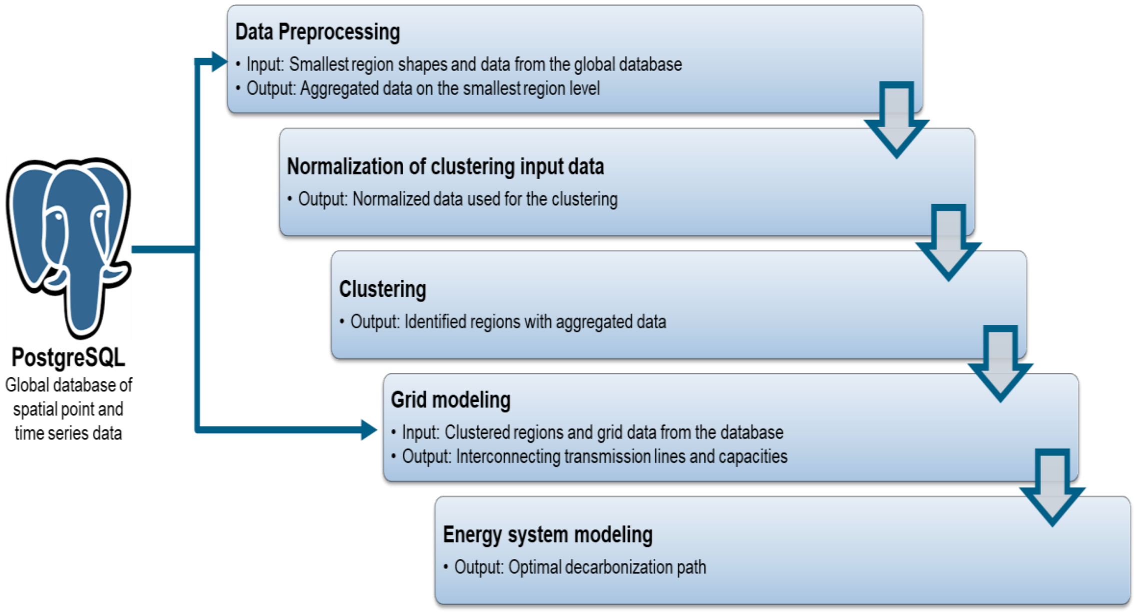
Data Preprocessing
• Input: Smallest region shapes and data from the global database
• Output: Aggregated data on the smallest region level
Normalization of clustering input data
• Output: Normalized data used for the clustering
Clustering
• Output: Identified regions with aggregated data
Grid modeling
• Input: Clustered regions and grid data from the database
• Output: Interconnecting transmission lines and capacities
Energy system modeling
• Output: Optimal decarbonization path
PostgreSQL
Global database of spatial point and time series data

Figure 3.3: Steps of the developed methodology

As shown, first the data and shapes of one or more specified countries are imported into the python environment. Those data are aggregated to the corresponding shape they are located in. Afterwards, the algorithm normalizes the data so that different kinds of input data can be compared, e.g. energy and generation capacities. The third step is the clustering of the normalized input data. This way different regions inside the country are identified. In the fourth step, Open Street Map (OSM) data of transmission lines are imported, and the algorithm identifies those lines that represent interconnections between the previously discovered regions. The last step is to export those regions with their corresponding inter-regional transmission capacities into the ESDP database. This database is used for the setup of the energy system model and calculating the cost-optimal decarbonization pathway.

Since the clustering is the core of the developed methodology and different clustering approaches might require different preprocessing of the data, the clustering approach must be chosen first. In 2.3 different approaches of how to apply clustering for regionalization have been discussed. As shown the disadvantage of k-means, being the implementation of the contiguity constraint [28] and the inability to find arbitrarily shaped clusters [15], is considered unsuited for the regionalization problem, despite its fast performance.

Density- and grid-based methods both entail the same drawback: subjective input parameters [15], which might result in phases of trial and error. Those input parameters might also depend on the considered country and therefore do not imply global applicability. The sorting out of outliers is another disadvantage that is not desirable for the regionalization problem this work is facing due to the high numbers of clusters it can result in. Therefore, those two approaches are not considered further either.

The only two approaches that seem suitable for the problem are the hierarchical clustering and the max-p algorithm. Both come with contiguity forcing constrains. On the one hand, hierarchical clustering is fast due to the contiguity constrain but does not do any reassignment of regions on its own. However, fine-tuning can solve this issue but, on the other hand, does not apply to high numbers of regions. Furthermore, hierarchical clustering is sensible to outliers and has the *rich getting richer* characteristic. On top of that, the algorithm does not give an optimal number of clusters as an output which leads to the necessity of CVI to find the optimal number of clusters in a second process.

On the other hand, the max-p algorithm does give as output the optimal number of clusters but needs as input parameter a user-defined threshold value of a condition that has to be met in each of the output clusters. Furthermore, it does not apply to big sizes of data sets. The first inconvenience can be seen for example as a given threshold of variance inside the clusters that have to be less than a user set value [28]. However, defining this value is a rather subjective question. Its time and computational complexity and therefore long running time are another crucial concern.

Since the considered data in the smallest region approach can contain thousands of regions and variables, the max-p region on its own is not suited for the regionalization problem. Therefore, the hierarchical clustering in combination with the fine-tuning algorithm is chosen to perform a regionalization in this work. The Ward´s method is chosen as a linkage criterion due to its superior performance using contiguity constrain [30]. Since the Ward´s method is only defined using the Euclidean distance, this distance metric is applied.

Regarding the disadvantage of not automatically finding the optimal number of clusters, a CVI is used to find this optimal number. Like choosing a clustering approach, selecting the correct CVI is also a question of what the problem is. One requirement for the identified regions is that there should be a low loss of data, i.e. the regions clustered together should be similar. Nevertheless, considering spatial contiguity, there may be two regions that are very similar to each other, but far apart, for example divided by a desert. If the inter-cluster distance is considered, e.g. by using 2.4.1, 2.4.2 or 2.4.3, the CVI can indicate a very small number of clusters where the two very similar clusters are combined. However, to do so, the algorithm would have to first add another cluster, which might be very distinct from the two similar clusters, to obtain the spatial contiguity for the two similar regions. This implies, the unsuitability for the regionalization problem of the first three CVI which is why in this work the elbow criterion is used to identify the optimal number of clusters even though it can be ambiguous in identifying the optimal number of clusters.

To identify the optimal number of clusters in an automated way, the different possibilities described in 2.4.4 are considered. However, the first two approaches – maximum decrease and maximum decrease ratio – are likely to identify very low numbers of clusters since the curve decreases especially fast for low number of clusters. The third method – surpassing a threshold of the second derivative – rises the problem of how to determine the threshold value, which might be subjective. The fourth and fifth method – largest second derivative and furthest from a connecting line – seem to be better applicable to the elbow method, even though the fourth might also suffer the same problem as the first two methods. Therefore, the fifth method is chosen, i.e. a line is drawn between the considered values of the elbow curve and the point located furthest from this line is considered optimal.

3.3.1 Data preprocessing

Since different occurrences in the data can be existent which might lead to distorted clusters, some preprocesses have to be applied. Since one requirement of this work is to be globally applicable those cases have to be considered at the beginning of the data preparation. The following rules explain the handling of exceptions implemented in the methodology.

Split up of multi-polygon shapes

One problem that has to be faced is the occurrence of "bridge-regions". This work defines those regions as multi-polygons stored as single-region that are not spatially contiguous and connecting different parts of the entirety of regions. Therefore, those regions can build bridges connecting regions that are not spatially contiguous. Figure 3.4 presents two of those bridge regions for the example of Belgium.

The light and dark blue regions are multi-polygons which are not spatially contiguous. When the regions are checked for spatial contiguity it would result in distorted spatial connectivity, since those bridge-regions connect regions that are not spatially contiguous themselves. When using those regions in the clustering it might result in something similar to Figure 3.5.

Since the clusters should be spatially contiguous, it shows the necessity of handling those bridge-regions. Therefore, when importing the region shapes the algorithm checks whether a region is a multi-polygon and splits those into the corresponding number of single polygons. This has two effects. Firstly, cities that are not spatially contiguous might happen to be in different clusters. Secondly, islands that belong to a district or city are separated and, therefore, have to be handled in the later run since they naturally do not have any spatially contiguous region. Nevertheless, bridge-regions can lead to non-spatially contiguous clusters and therefore have to be avoided.

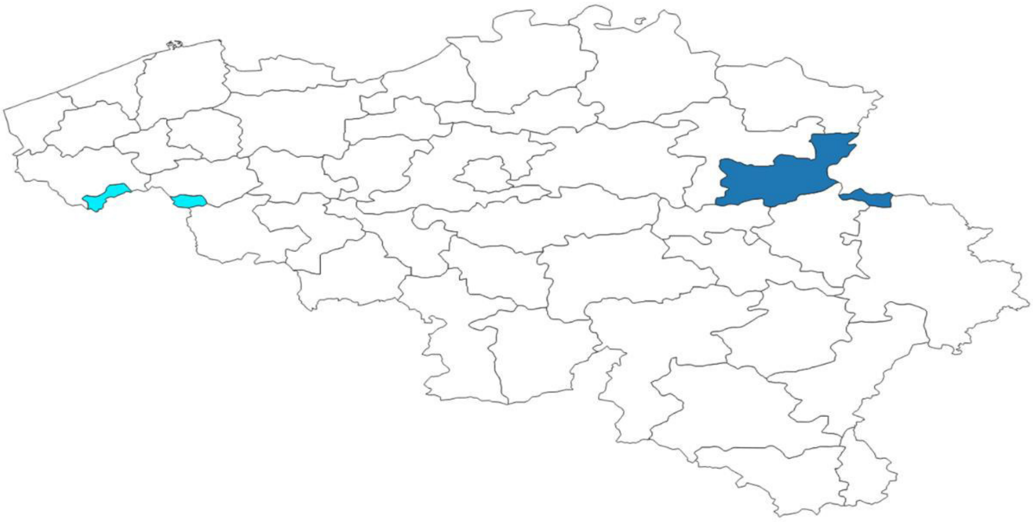

Figure 3.4: Bridge-regions for the example of Belgium

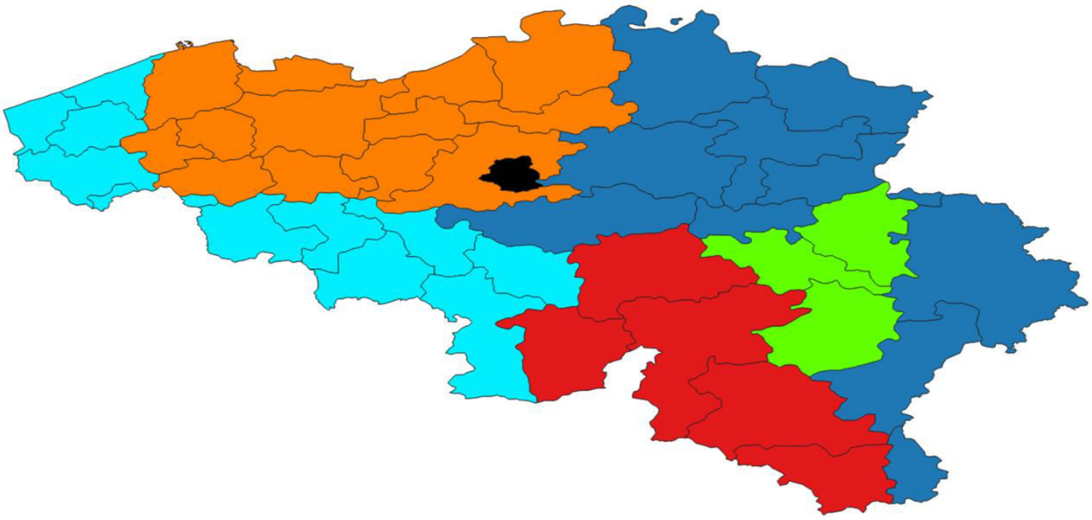

Figure 3.5: Illustrative cluster example of Belgium if bridge-regions exist

Data aggregation

Since the different data considered for the clustering have to be aggregated to the smallest regions, depending on the different attributes, the following steps are executed:

- Load data: Since load data, PHH and CTSI, are likely to be high in big sized regions, those data are aggregated to the smallest region, and the corresponding sums of load are normalized by the area of the regions, i.e. resulting in the load-density of MWh/km².
- Generation data: Generation can be seen as a negative load. To be consistent also the sum of the power generation data for each technology is normalized by the region´s area resulting in MW/km².
- Wind and PV profiles: Since wind and PV profiles are stored as grid cells one region might intersect multiple grid cells. Therefore, the corresponding profile is calculated as

$$p_{r,t} = \sum_{g\ in\ G} \frac{A_{r,g}}{A_r} * p_{g,t} \tag{3.1}$$

  where $p_{r,t}$ is the value of the profile of region r in time step t, g is one of the intersecting grid cells G, $A_{r,g}$ is the area that r intersects with grid cell g, A_r the total area of r, and $p_{g,t}$ is the profile value of g in t. Therefore, the profile of a region is the weighted profile of the intersecting grid cells depending on the intersecting area.

Detection of mainland, islands, and outliers

One problem that different clustering algorithms are susceptible to is the existence of noise [15]. Since hierarchical clustering is sensitive to outliers and noise, it has to be filtered in a pre-process. Small regions without any load or generation can be considered noise. Those regions might be identified as single clusters. However, due to their low importance regarding the energy system model this is not desirable. Those regions can exist, because they are split from bridge-regions. Also uninhabited islands can be such regions. Therefore, for each multi-polygon that is splits it is checked, which is the biggest of the newly constructed regions and islands. After the aggregation of the data to the corresponding regions, it is checked whether a smaller split region has any load or generation data. If this is not the case this region is not considered for the clustering. If it is also not connected to any other region it leads to the removal of this region from the data set. Since islands that do not have any generation nor load are of neglectable importance for the energy system model, this step is implemented to denoise the region data. In a second step, the algorithm identifies all connected groups of landmasses and removes all those from the region set which do not have any load or generation data. This is done, due to the existence of islands consisting of more than one region but not containing any load nor generation capacities.

Another important factor regarding islands is the problem of identifying the mainland or main islands. This is of special importance for countries that consist of various islands, e.g. the Philippines, Japan, or New Zealand. Considering New Zealand as an illustrative example, the methodology has to cluster the different regions on its two main islands, while each cluster can only contain regions from one of the islands. Figure 3.6 depicts the country of New Zealand.

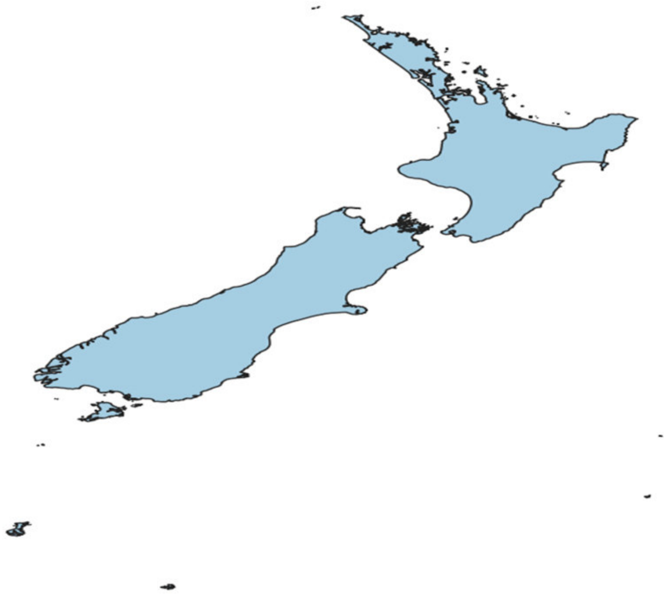

Figure 3.6: Country of New Zealand

What is intended to do in this work, is to identify the two main islands and then cluster them in the same process. Since clustering the two islands separately does not take into account a comparison of either aggregating more regions in one or the other, the clustering must be done in the same process. Because of that, all connected landmasses in the region set are identified. The approach considered in this work is to set a threshold value of what percentage of area one islands must have to be considered as a main island and, therefore, be considered as a set of regions in the clustering process. The investigated country highly influences the threshold value that can be considered optimal. Therefore, finding a threshold value for any given country is not trivial. Considering the examples of New Zealand, Japan and the Philippines a threshold of four percent seems to be suited. However, this threshold is left as an input parameter for the user to be able to change which islands are considered as main islands.

Still, there might be islands in the data containing load or generation data. However, islands cannot be used within the clustering process since they would always be a cluster of their own since they naturally do not have any connection to the mainland. Therefore, the algorithm calculates the distance between each of the remaining islands to every region within the mainland or main islands. For each island, the closest region on the mainland is identified and it is checked whether the corresponding distance is less than 50 km. If the distance is higher the island is removed from the region set, since an island that is more than 50km away is assumed to be unconnected to the electrical grid of the mainland. If the distance is lower the load and generation data of the region are added to those of the closest region in the mainland. Since the true connection point is unknown, this is assumed to facilitate the procedure of the algorithm. This leaves the data set with the number of mainlands and main islands as groups of internally connected regions.

3.3.2 Normalization of clustering input data

Since the clustering compares different attributes like wind profiles and load data, the data must be normalized [56]. However, only normalizing all the data to zero and one can lead to an undesired outcome: the values of the profile for each time step would be as important as e.g. the PHH value for a whole year. Since this is not intended, the normalization has to be done differently.

Firstly, different attributes are defined which can have multiple sub-data this work refers to as entries. The attributes that are considered are load, generation separated in conventional and renewable, and profiles. The load data consist of one value for PHH and one for CTSI consumption, and the generation data consist of the installed capacities of the different generation technologies. The profiles include the wind and solar time series.

The approach used in this work is referred to as *unweighted normalization* in the latter and is calculated as

$$value_{norm,unweighted,a,r} = \frac{\dfrac{value_{a,r} - \min(value_{a,R})}{\max(value_{a,R}) - \min\,(value_{a,R})}}{n_A} \tag{3.2}$$

For each region *r* of the region set *R*, each of the corresponding entries *a* of the attribute *A* is normalized to zero and one and divided by the number of entries within the attribute. For the generation capacities this value is different for renewables and conventional generation technologies. In this case the value is two multiplied by the corresponding number of technologies. This differentiation is done to reflect the special importance of renewable energy generation in future energy systems.

The unweighted normalization considers the similarity of the regions without giving any emphasis on e.g. a specific generation technology. The tables below present a simple example of one country for the normalization approach. Table 3.1 presents the sums of the different entries of the attributes load and generation. Table 3.2 indicates the resulting range of entries within the corresponding column for the unweighted normalization.

Yearly consumption [MWh]		Installed capacities [MW]				
		Conventional			Renewable	
PHH	GHDI	Coal	Nuclear	Gas	Wind	Solar
50,000	100,000	10	15	5	20	5

Table 3.1: Illustrative example of the sum of yearly energy consumption and installed generation capacities for all regions in the region set

Yearly consumption [MWh]		Installed capacities [MW]				
		Conventional			Renewable	
PHH	GHDI	Coal	Nuclear	Gas	Wind	Solar
[0;0.5]	[0;0.5]	[0;$\frac{1}{2*3}$]	[0; $\frac{1}{2*3}$]	[0; $\frac{1}{2*3}$]	[0; $\frac{1}{2*2}$]	[0; $\frac{1}{2*2}$]

Table 3.2: Illustrative example of the range of values for unweighted normalization

3.3.3 Clustering

When receiving the normalized data, clustering is performed as described in 2.3.2 using the open-source code of scikit-learn [30] providing the hierarchy of cluster merges. In this hierarchical clustering, the Ward´s method is chosen due to its superior performance on contiguity constrained clustering.

Selecting the optimal number of clusters

Since the hierarchical clustering does not provide an optimal number of clusters this number is identified by using the elbow criterion. Since the number of clusters should not be too high, keeping the computation time of ESDP low, only the values of the elbow criterion within a certain interval are computed using the computed the hierarchy. The upper limit of clusters that is considered suitable is 30 clusters. Using this number of regions, the ESDP framework can still compute the decarbonization pathway. Using a higher number of regions often results in problems regarding computational complexity. Nevertheless, setting the upper limit of the interval of the elbow criterion to 30 would lead the algorithm to never identify this number of clusters since those points would be on the line of connecting line and therefore the distance would be zero. Therefore, the upper limit of this interval must be higher. Because of those reasons, this work considers the following interval for the elbow criterion:

$$Interval = [\max(2, n_{mainlands})\,; 50] \tag{3.3}$$

Where $n_{mainland}$ is the number of mainlands and main islands of the country. The lower threshold has to be set because the algorithm cannot cluster to a lower number of clusters than the $n_{mainland}$ since the islands cannot be connected due to missing connectivity.

On the other hand, receiving only one cluster would represent a single-region model that does not need any kind of preprocessing like cluster analysis. The maximum value is set to 50 because the automatized method of finding the optimal number of clusters depends on the interval that is considered, i.e. considering an interval of [2;50] might lead to another "optimal" number of clusters than a considered interval of [2;100].

However, in some cases, the optimal number of clusters still results to be very high, e.g. 49 clusters in the case of Algeria. As described before, such high numbers of clusters are not desirable. Therefore, the steps presented in Figure 3.7 are performed.

As the figure shows, firstly the optimal number of clusters is identified by considering the interval between x and 50, x being the maximum of $n_{mainland}$ and two. If this optimal number of clusters is lower than or equal to 30 and is not equal to x the algorithm is terminated providing the optimal number. If this is not the case the upper limit is iteratively reduced by one in each cycle and the optimal number of clusters is recalculated. When the optimal number of clusters meets the specified conditions, this number is selected in the following.

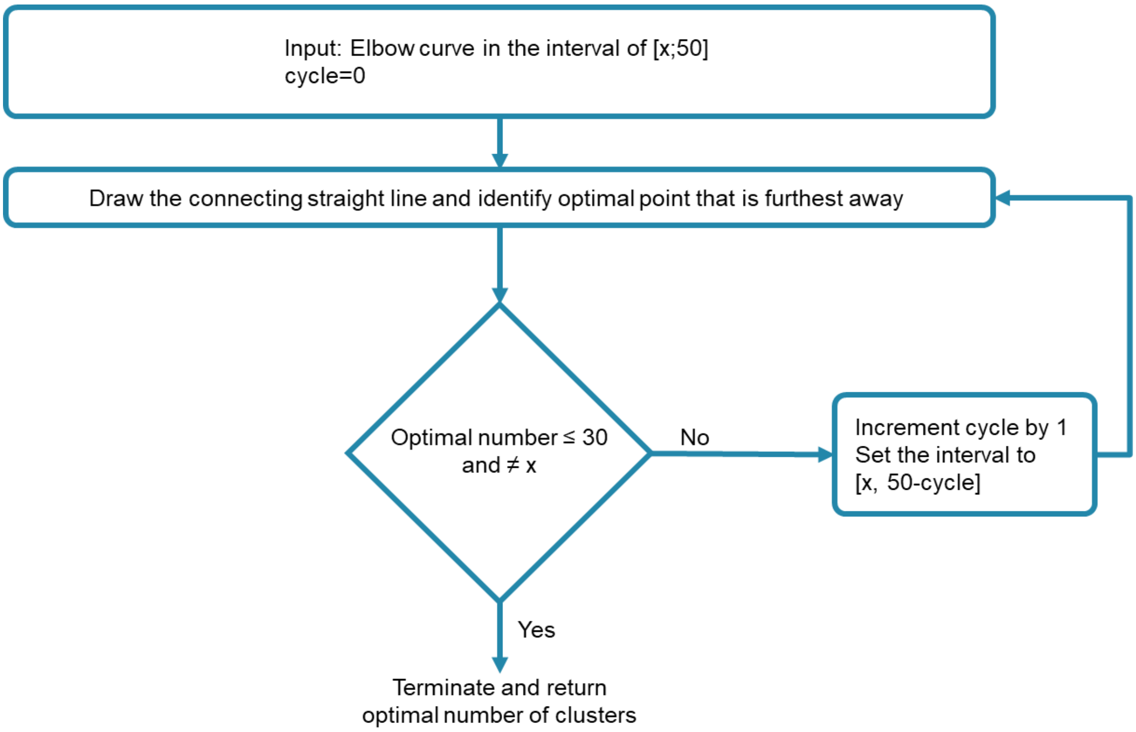
Input: Elbow curve in the interval of [x;50]
cycle=0
Draw the connecting straight line and identify optimal point that is furthest away
Optimal number ≤ 30 and ≠ x
No
Increment cycle by 1
Set the interval to
[x, 50-cycle]
Yes
Terminate and return
optimal number of clusters

Figure 3.7: Algorithm to identify the optimal number of clusters

Fine-tuning

To improve the result of the clustering, the optimal number of clusters is used as a starting solution for the fine-tuning post-process. Since the fine-tuning scales with the power of two to the number of regions, it might take very long to terminate. Therefore, the fine-tuning is not applied to every number of clusters.

Since the Ward´s method is only defined using Euclidean distance, the range of the values for the time series, however, is small compared to the range of the other attributes. Hence, the magnitude of the difference in the profile attributes is small compared to the other attributes, especially when considering the squared difference, the Euclidean distance calculates in the first place. Therefore, using the Euclidean distance metric the time series influence the result of the clustering process very little. Because of this unintended behavior another distance metric is needed and applied in the fine-tuning process to better consider the time-series. In this work the considered distance metric is the following. For each attribute, i.e. load, generation and profiles, the Manhattan distance is calculated. With these three corresponding sums of distances, the Euclidean distance is calculated. The following formulas represent the metric:

$$sum_{distances} = \sqrt{\sum_{A} {distance_A}^2} \tag{3.4}$$

$$distance_A = \sum_{c \in K} \sum_{x \in c} \sum_{a \in A} ||x_a - \bar{x}_{c,a}|| \tag{3.5}$$

In these equations *a* is one of the entries of attributes *A*, and *c* is one of the *K* clusters with *x* being a member of this cluster. Using this method, the drawback of the Euclidean distance in combination with the normalization methods is expected to be overcome, since the Manhattan distance is preferable on high dimensional data [57]. On the other hand, the Euclidean distance in (3.4) is used to better reflect possible outliers since it is more prone to those than the Manhattan distance [58].

In contrast to the standard fine-tuning process, this work implements two additions. Firstly, a time limit is implemented to be able to terminate the algorithm after a specified amount of time and use the identified improvements. Secondly, the standard fine-tuning often results in clusters that are formed like lines. Those lines often have one almost isolated region at the end which was only connected to one other region. In the standard fine-tuning algorithm, this can lead to the following problem. If both regions turn out to be initially in the same cluster including other regions, the region with only one neighbor cannot be assigned to any other cluster than the one its only neighbor belongs to. On the other hand, its neighbor cannot be assigned to any other cluster, because it would break the spatial contiguity of the cluster. Therefore, those two regions cannot be reassigned to any other cluster. Figure 3.8 presents this problem for the case of South Africa, where the red region is the almost isolated one and the green region is its only neighbor.

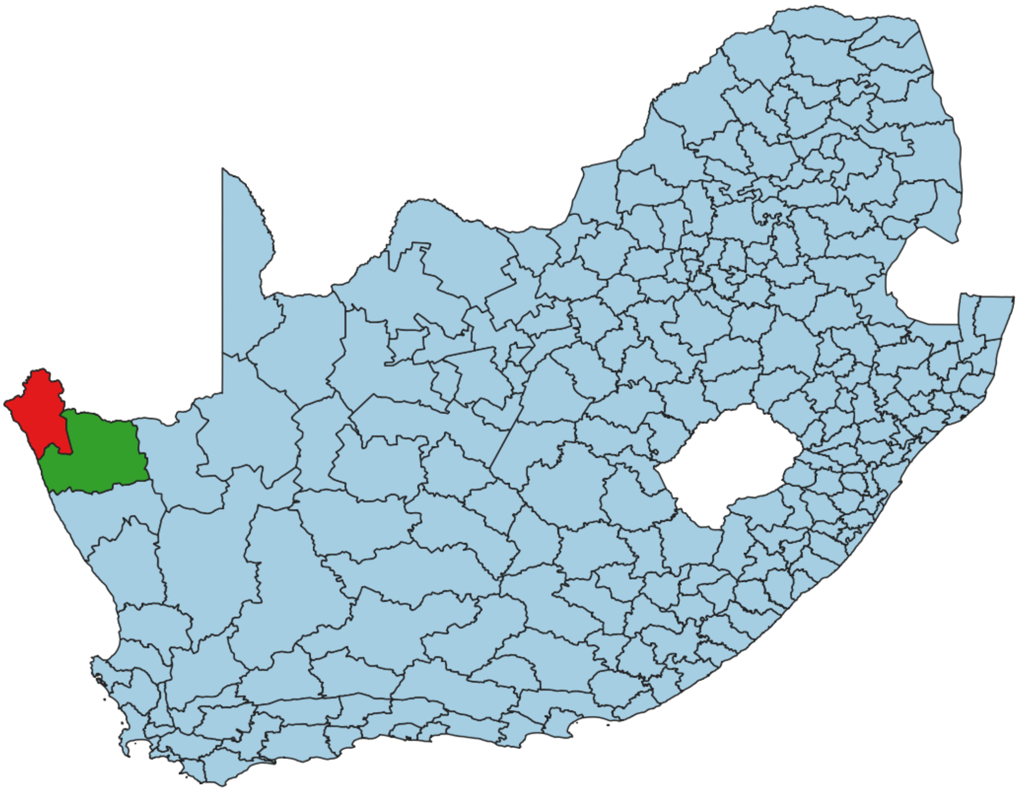

Figure 3.8: Almost isolated regions for the case of South Africa

Figure 3.9 presents an example of a possible cluster outcome during the fine-tuning. The green cluster consists of a line with some coastal regions at each end. The fine-tuning algorithm cannot handle those occurrences since it can only move one region at once. This is where another algorithm is applied. The idea is to shift multiple regions at once into another cluster. However, the combinatorial complexity should be kept low since the fine-tuning can result in high computation times already. Therefore, the following procedure is applied. After each of the cycles in the original fine-tuning the optimal shift of regions is applied, and the new algorithm presented in Figure 3.10 is executed.

Firstly, all regions that are connected to only one other region in its respective clusters are identified. With those regions as starting points, the algorithm creates a list of regions which later identifies a path inside the clusters. For each of the starting regions, the algorithm identifies the region it is connected to and adds it to the list. If this region is only connected to two other regions, including the starting region, the algorithm continues following this path until it reaches one region that has either one or more than two neighbors.

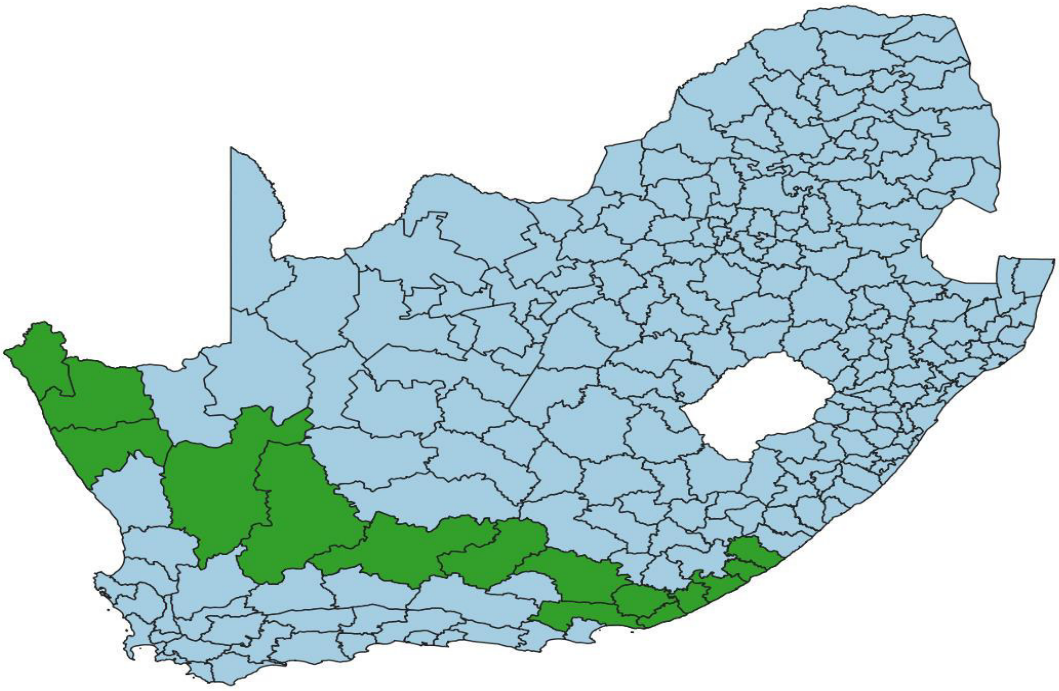

Figure 3.9: Illustrative example of a line-like cluster

If the last region is only connected to one other region it indicates the cluster only consists of one line of clusters, i.e. the last found region is another starting region. To prevent the algorithm from shifting a whole cluster into another one the last identified region is dropped from the list. Because of this behavior, the algorithm cannot lower the number of clusters. If, however, the region has more than two neighbors it must be checked if, removing this region, the spatial contiguity of the cluster is still valid. If not, this region is also not considered for changing the cluster.

In the next step, all other clusters the regions in the list could be shifted to are identified. Afterwards, the algorithm calculates the result of the CVI for each possible subpart of the list always starting from the starting node. For example, if the list contains three regions, region 1 being the starting region, the algorithm calculates the changes of the CVI for the combinations of shifting the regions {1}, {1, 2}, and {1, 2, 3} in any other connected cluster. The change that achieves the best result of the CVI is identified. If this change improves the result it is applied to the former clusters. When no change, which leads to an improvement, can be identified the algorithm terminates. The results of this methodology in contrast to the original fine-tuning are presented in 4.1.2.

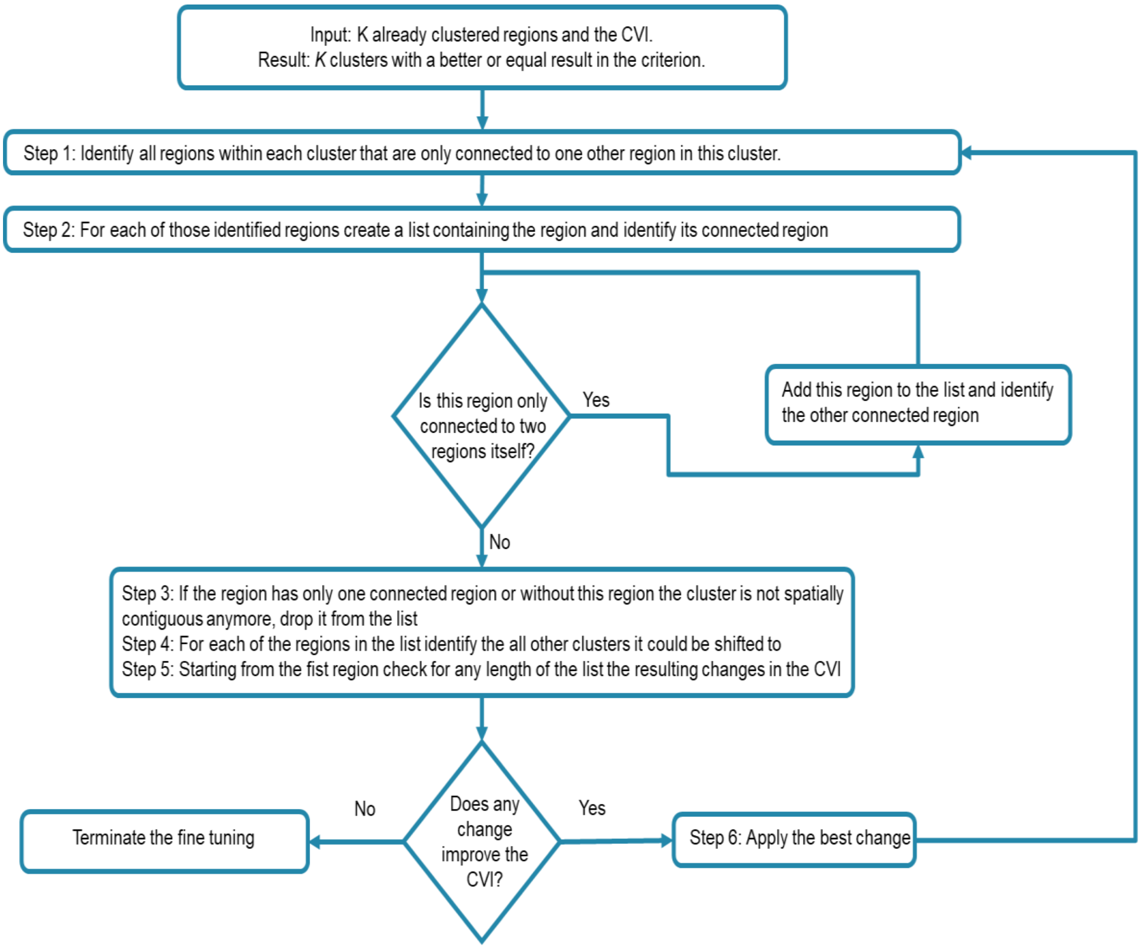
Input: K already clustered regions and the CVI.
Result: K clusters with a better or equal result in the criterion.
Step 1: Identify all regions within each cluster that are only connected to one other region in this cluster.
Step 2: For each of those identified regions create a list containing the region and identify its connected region
Is this region only connected to two regions itself?
Yes
Add this region to the list and identify the other connected region
No
Step 3: If the region has only one connected region or without this region the cluster is not spatially contiguous anymore, drop it from the list
Step 4: For each of the regions in the list identify the all other clusters it could be shifted to
Step 5: Starting from the fist region check for any length of the list the resulting changes in the CVI
Does any change improve the CVI?
No
Terminate the fine tuning
Yes
Step 6: Apply the best change

Figure 3.10: Sub algorithm for the fine-tuning

After the sub-algorithm terminates the original fine-tuning continues its procedure. Therefore, after each of the cycles of the fine-tuning, the sub-algorithm is applied, which is expected to achieve better results than just applying it once after the fine-tuning. This way the methodology developed in this work might overcome the incapability of the original fine-tuning to shift multiple regions, while keeping the additional computational cost low.

3.3.4 Grid modeling

Since the identified regions inside the energy system, shall be able to interact with each other, the next step is to define this capability. For this purpose, an NTC model as described in 2.2 is built. Since detailed grid data from the TSOs are not available for each country this work investigates the potential use of OSM data.

OSM data

OSM data are publicly available data of different kinds of spatial attributes [59]. Those data include categorizations of land use, building types, line classification like highway, railway and more. The fact that OSM data are not perfect, is an important factor that has to be taken into account. Since there is no globally valid rule on how the contributors must upload the data, they have to be denoised. This indicates that the results of the grid model have to be validated. The data of interest in the regionalization problem in the context of grid modeling are transmission lines. The data corresponding to the electricity sector are specified by an entry in the so-called "*power*" tag. Apart from other entries, the data this work takes into consideration include those with the *power* tag of "*line*" and "*cable*". Those data are said to correspond to the high voltage levels. In contrast, the tag "*minor_line*" and "*minor_cable*" mostly refer to the distribution grid. Since including the distribution grid would lead to a high additional computational effort and do not perform the task of transportation of energy, those tags are not considered. Furthermore, those line data can include other tags. The ones important to this work are the tags *voltage*, *frequency*, *cables*, and *name*. These tags indicate the voltage level of one or more transmission lines, its frequencies, the number of cables the transmission line carries and the name of the installation, respectively. However, due to the incompleteness of the data, lines exist that, for example, do not have a *voltage* tag, i.e. the voltage level is unknown.

The corresponding lines of the investigated country are imported into the Python environment and the following steps are executed.

Creation of the NTC model

Firstly, all lines that interconnect different regions and provide necessary information must be identified. Therefore, all lines that do not have any entry in all of *voltage*, *frequency*, and *cables* are dropped from the data because those entries do not contain any information apart from the spatial position of the transmission line. Afterwards, the algorithm identifies the corresponding regions of the start and endpoint of the lines. The algorithm keeps those lines connecting two different regions, i.e. the region of the start point is not the same as the region of the endpoint of the line. The rest of the lines are removed as they do not provide transfer capacities between regions.

Secondly, for each line in the data set it is checked if the tag of frequency is 50 Hz or 60 Hz, which are the most used frequency levels in the electricity sector in the world [60]. However, if the frequency is zero the line could potentially be a High Voltage Direct Current (HVDC) line. Since there is no tag for this kind of line, the name is the only tag possibly providing this information. If it contains the word ‘HVDC’ this line is considered as such. Unfortunately, those names can be the corresponding translation of HVDC, e.g. “HGÜ” in German. Since it would be unreasonable to look up every possible translation of this technology HVDC lines without an English indication in its name are considered Alternate Current (AC) lines. In this case, their frequency is set to the most frequent value found in the data set. However, other frequencies can occur. The Deutsche Bahn AG, for example, has its own electrical grid with a frequency of 16.67 Hz which can also be in the OSM data. Since those lines are not subject to the general electricity sector, any line with a non-zero frequency value of neither 50 Hz nor 60 Hz is removed from the line data set.

Occurrences of transmission lines carrying different voltage levels or possibly different frequency levels is another aspect that has to be considered building the NTC model. For example, one line-object of the OSM data can contain two values of frequencies, three values of voltages and four values of cables. This example leads to the question of which of the voltage levels has how many cables and is of which frequency. Since without a detailed consideration no answer can be given, the following rules are applied to generalize the data filtering:

1. If there is only one voltage level, the number of elements in cables is compared to the number of elements in frequency. If it is the same number a new line-object is created for each element and the corresponding value of frequency and cables is assigned. If, however, only one value is given in cables, while the number of values in frequency is greater than one, for each frequency a new line-object is created. Afterwards, the number of cables is calculated as:

$$n_{cables} = \frac{n_{cables,total} - n_{cables,total}\ mod\ (3 * n_{frequencies})}{n_{frequencies}}$$

   Where *n* n_{cables} is the number of cables for each frequency, $n_{cables,\ total}$ is the total number of cables, and $n_{frequencies}$ is the number of frequencies of the original line-object. The division by three times $n_{frequencies}$ ensures the number of cables to be a multiple of three which is needed because three-phase systems are considered.

2. If the voltage level is different from one the numbers of elements in each of the three tags are compared. If all numbers of elements correspond, each is set as own line-object. However, if they do not correspond it is checked whether the tag cables only has one element and either the number of frequencies is equal to the number of voltages or the number of frequencies is one. In this case, the number of cables for each new line object is calculated as:

$$n_{cables} = \frac{n_{cables,total} - n_{cables,total}\ mod\ (3 * n_{voltages})}{n_{voltages}}$$

3. If any combination occurs that does not lead to a classification as described above the line-object is removed from the line data set
4. If an element of cables is one, set the number of cables to three
5. If an element of cables is two, set the number of cables to six
6. If an element of cables is greater than three, set the number of cables to

$$n_{cables} = n_{cables,total} - n_{cables,total}\ mod\ 3$$

7. If the voltage is zero, set it to the most frequent value of the voltages
8. Remove all lines with a voltage equal or lower than 100kV

Rules 1, 2, and 3 ensure that lines that cannot be interpreted easily are removed from the considered lines. Those occur mainly because of incomplete data upload to the OSM data. On the other hand, the rest of the rules ensure consistency within the line-objects.

Rules 4, 5, and 6 are applied because those values probably resulted from poor knowledge of the person uploading the data in the database considering one line is probably meant to be one three-phase-system which corresponds to three cables in the AC system. Respectively two cables probably resemble two three-phase systems, i.e. six cables. Furthermore, a number of four cables, for example, could imply a three-phase system and an earthing cable, therefore three cables are considered.

Rule 7 is a strong simplification. Nevertheless, the most frequent value corresponds to the voltage level that most transmission lines have in the respective country. Therefore, the probability of a line having this voltage level is high. The user uploading the data to the OSM database often does not know the voltage level and therefore leaves this tag empty.

Rule 8 on the other hand, assures a lower computational complexity. Furthermore, lines with a voltage level below 100 kV do not transport high capacities between regions. Therefore, those lines are not considered in this work.

Having identified the corresponding transmission lines between countries and their voltage levels, the transmission capacity between the regions has to be calculated. Therefore, standard transmission capacities are assigned to the lines according to Table 3.3 [8] [9].

Voltage level [kV]	110	220	380	750	1150
Transmission capacity [MW]	343	520	1790	5404	13750

Table 3.3: Transmission capacities depending on the voltage level [8], [9]

However, depending on the country other voltage levels might exist. Due to the lack of such standard data, voltage levels in between those defined above are interpolated. This way, for each pair of regions, the sum of the transmission capacity of all the lines connecting them is calculated. The transmission capacities are also multiplied by 0.7 as described in 2.2 to consider security constraints. Data for high voltage transmission cables, on the other hand, are more difficult to find in a standardized format. Since the number of cables compared to overhead lines is low [61], assuming the same transmission capacities should result in a minor distortion of the actual transmission capacities.

3.3.5 Energy system modeling

The region-based energy system modeling this work tries to achieve should use the identified regions their aggregated data and the grid interconnection between them. Therefore, those data are exported into the ESDP database. Which the existing framework can use to build the model and calculate the decarbonization pathway. Another input for ESDP is the length of the transmission lines which are needed to calculate transmission losses and costs of operation and maintenance of the lines. For each combination of two regions, the length of the connecting transmission lines is automatically estimated by the distance between the two regions centroids. This helps not to overestimate the transport capacity of power generation of those power plants that are located at the opposite side of the region, regarding the connecting transmission line.

Using those data ESDP tries to minimize the costs of transforming the current energy system, based on the given regions, to a future energy system characterized by high renewable shares and low emission. Input parameters for ESDP are the time horizon to be considered, i.e. until 2050 and the goal of decarbonization, e.g. 100%. On the other hand, the different technologies that are considered for the current and future energy system including their investment, and operation costs, technical lifetime, etc. have to be declared.

3.4 Previous work – state of the art

Region-based energy system modeling has been investigated in different works. Regarding regionalization by clustering, however, the only other work that is known to the author of this work is [28]. A multi-step clustering approach is used. Instead of defining the smallest regions on the city or district level, the point data are used directly as input for the clustering. The developed algorithm first cuts the considered map into grid cells. Afterwards, a clustering of the point data of each grid cell is performed using K-means in a first step. K-means is chosen due to its fast performance and the large data set which has to be handled. The received clusters are clustered in a second step using the max-p algorithm due to its superior performance on smaller data set. Furthermore, the max-p regions algorithm identifies the optimal number of clusters automatically using a threshold value for a given condition. The variance inside the clusters is used as such. In the last two steps, the algorithm recombines the clustered grid cells and executes the max-p algorithm a second time to be able to combine clusters that were located in different grid cells before. However, using the K-means algorithm the primarily resulting clusters are almost all Voronoi-shaped, which implies the weakness of K-means not being able to consider a contiguity condition. Therefore, it receives the location as an attribute with a high weight. Even though spatially close points tend to have similar attributes, this is not necessarily the case for all kinds of data, e.g. load. Furthermore, the use of the specified threshold condition for max-p regions might be arbitrary and it was not described why the corresponding formula was chosen. On top of that, only the results of the clustering receiving one attribute at a time were shown, even though it can receive multiple attributes. This could make sense from a business perspective view, identifying, for example, good locations for wind energy. However, modeling the system as a total, the suitability of this approach has not been made clear.

4 Analysis and Assessment Procedures

This section presents the results of the developed methodology, applying it to two use cases. The first use case is South Africa. This country is chosen due to its high conventional power generation and simple overall structure. The second use case is Germany, which is a complex country to model because of its diverse generation mix and its high number of distributed load centers. For the case of South Africa, also a sensitivity analysis is presented to show the methodology´s behavior towards different input data. In the end a discussion addresses the suitability of the developed algorithm to the initial problem of the region-based modeling using regionalization.

In the use cases the following procedure is applied. Firstly, the countries are described, and points of special importance are highlighted. Afterwards, the input data are presented to show the extent of the different attributes Next, the results of the developed methodology are inspected. For the case of South Africa also the achievements of the new fine-tuning method in comparison to the one in the literature are presented. In the next step, the results of the NTC model are compared to available presentations of the corresponding electrical grid. The results of ESDP, i.e. the decarbonization path, are evaluated for the first use case of South Africa.

4.1 Use Case 1 – South Africa

4.1.1 Data

Since the input data for the clustering show a major impact on the results, those data have to be described. The data presented below are those that are identified after the pre-processing of the data and shapes of South Africa. The data consists of 236 smallest regions. Figure 4.1 presents the load-density distribution of South Africa for PHH and CTSI data, respectively. As shown in the figure, the centers of high load-density are the areas around Johannesburg (J), Cape Town (C), Pietermaritzburg (P) and Durban (D).

Figure 4.2 presents the local distribution of the power generation of South Africa. In this picture the size of the circle indicates the installed capacity. As shown, the main part of the power generation relies on coal. Furthermore, those coal power plants are mostly located in the north-east of the country. On the other hand, in the west and center of the country a high share of renewables can be observed. For a comparison of total installed capacity see Appendix A.

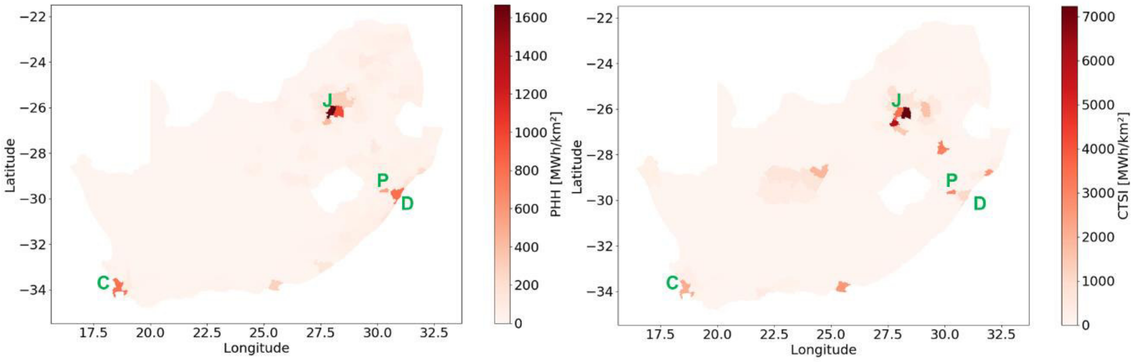

Figure 4.1: PHH (left) and CTSI (right) load-density distribution in South Africa

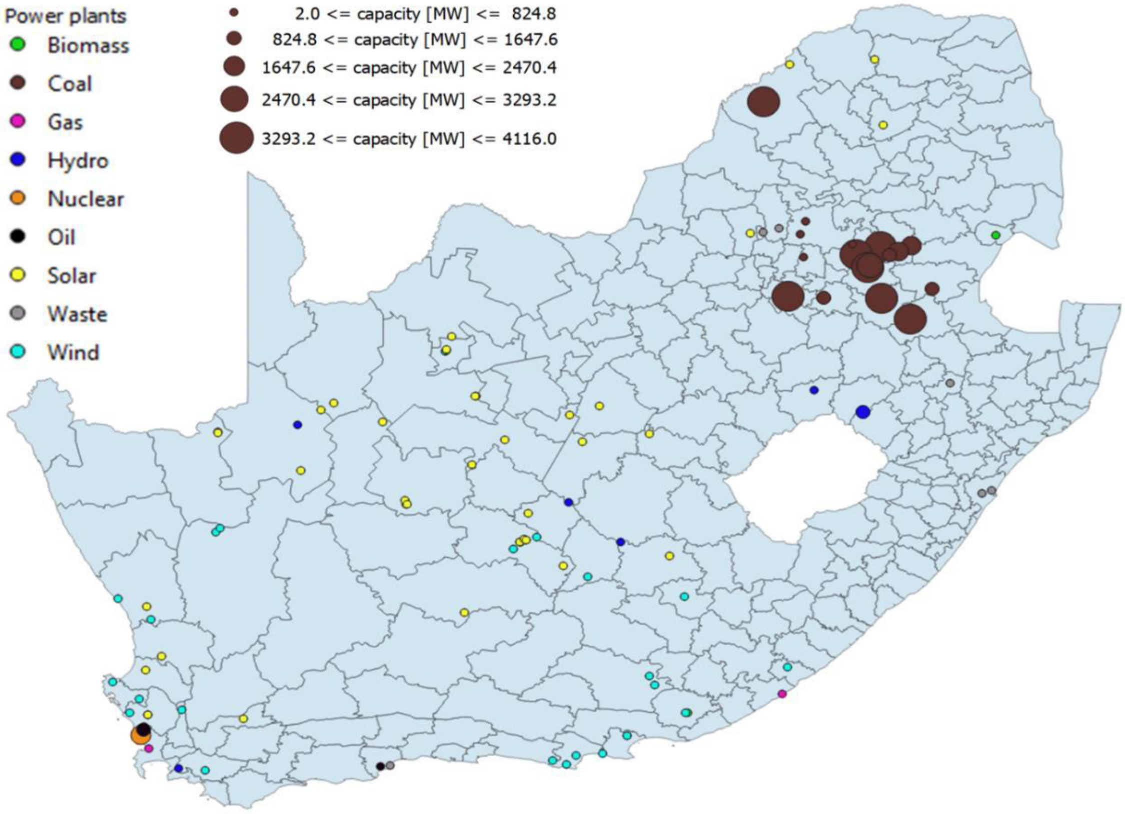
Power plants
Biomass
Coal
Gas
Hydro
Nuclear
Oil
Solar
Waste
Wind
2.0 <= capacity [MW] <= 824.8
824.8 <= capacity [MW] <= 1647.6
1647.6 <= capacity [MW] <= 2470.4
2470.4 <= capacity [MW] <= 3293.2
3293.2 <= capacity [MW] <= 4116.0

Figure 4.2: Power generation mix in South Africa

Since time series are difficult to present regarding their spatial resolution, Figure 4.3 presents the Full Load Hours (FLH) for the smallest regions to indicate their distribution in South Africa. As shown, the areas with the highest wind potential are along the coast of the country, whereas the solar generation potential increases towards the north of the country.

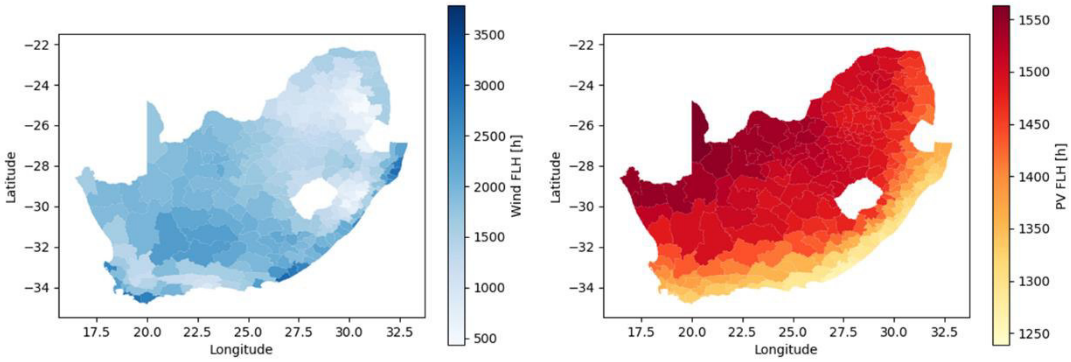

Figure 4.3: Wind (left) and PV (right) FLH in South Africa

4.1.2 Sensitivity analysis of cluster results

This section presents the results of the developed methodology described in 3.3. Firstly, the output of the standard methodology as described in 3 is shown. Afterwards, the behavior of the algorithm towards different input data is presented.

Standard data input

Using the input data as described in 3.2, the optimal number of clusters is 28 clusters. The curve of the elbow criterion is depicted in Figure 4.4. However, since the complexity of the energy system modeling is meant to be kept low and this optimal number of clusters is very high regarding the complexity of the energy system of South Africa. This optimal number has to suit the purpose of its application. Comparing this number of clusters to another regional modeling of South Africa [62], considering nine regions, or the modeling of Europe [28], considering 28 regions, it does not seem suitable to model the energy system of South Africa including this many regions. Therefore, using the elbow curve, the number of clusters is determined by a considerable bend in the curve. Such a bend can be observed for at twelve clusters. This number is chosen for sensitivity analysis.

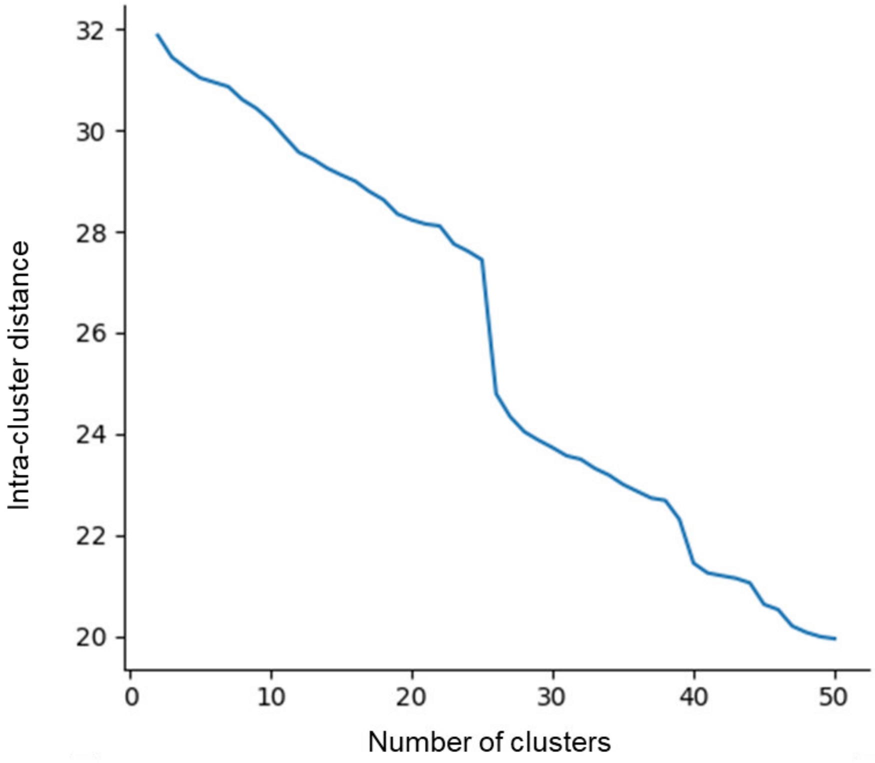

Figure 4.4: Elbow curve in the use case of South Africa

Figure 4.5 presents the outcome of the methodology. As shown, four of the identified clusters are single-regions around Johannesburg (clusters 1, 2, and 3) and Pietermaritzburg (cluster 6). The number of regions belonging to each cluster is shown in Table 4.1. As shown, the number of regions in the other clusters is in the range of 21 to 64.

In Appendix A the different mean values and standard deviations of the different attributes of the identified clusters are presented. As shown, the single-region clusters 1, 2, 3, and 6 are mainly characterized by their high load-density values of PHH and CTSI, respectively. Furthermore, clusters 1, 3 and 6 do not contain any generation capacities. On the other hand, cluster 2 is very similar regarding the generation structure to its surrounding cluster 11 but they have very distinct loads, which is considered by the algorithm. Another special occurrence is the low wind FLH for cluster 6 compared to the surrounding cluster 4. The clustering also seems to give special importance to the time series, especially to that of wind, which can be seen by the formation of the different coastal clusters 5, 7, 9, and 10. Especially clusters 5 and 9 have very similar FLH. Nevertheless, cluster 9 occurs to have a higher mean load. Furthermore, the influence of the temporal component has to be considered, which can lead to considerable differences between the clusters. This factor is discussed during the sensitivity analysis.

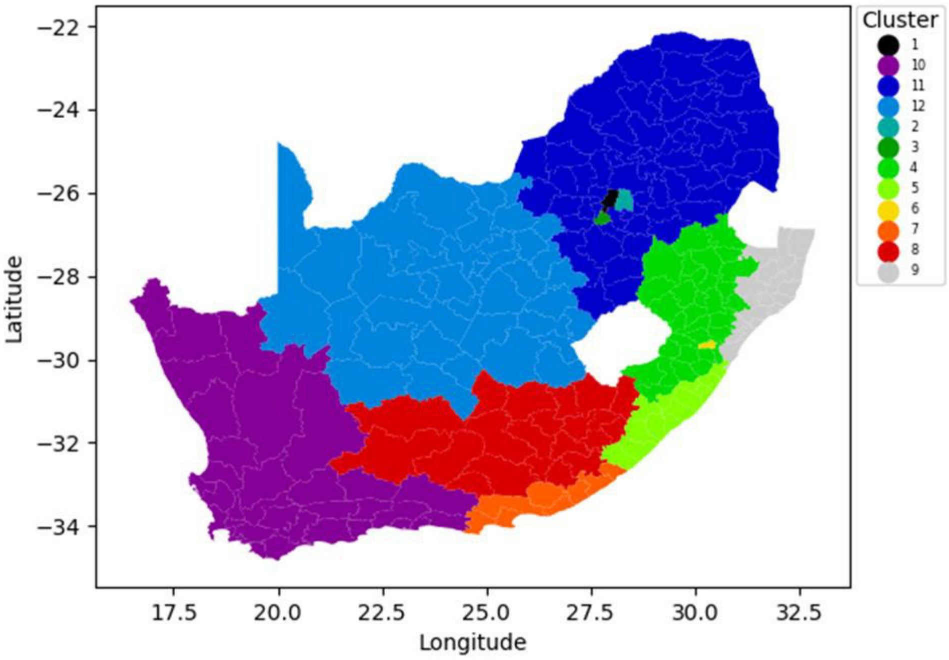

Figure 4.5: Results of the developed Methodology for South Africa using regional model of 12 Clusters and unweighted normalization

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Number of regions	1	1	1	32	24	1	38	21	25	34	64	40

Table 4.1: Number of regions in each cluster

Regarding the standard deviation within the clusters it can be observed that a high standard deviation in the case of yearly electricity consumption is most present for low mean values of the cluster. This is mainly because some clusters contain regions with a very low load but also some that are higher. This effect is also influenced by the usage of load-density instead of the actual load values. For the generation data the opposite occurs. Clusters that contain a considerable amount of installed capacities are likely to have a high standard deviation. This mainly results from some regions that contain high generation capacities belonging to the same cluster as other regions containing low or no capacities. The highest standard deviation, in this case, occurs for coal power plants due to their higher share in the generation technology mix of South Africa. The FLH, on the other hand, show that most variation results from the wind profiles, which normally have higher spatial variation than PV time series since the later usually follow a very similar daily and nightly profile.

Comparison to the standard fine-tuning

Comparing the adapted fine-tuning algorithm to the standard from the literature, Figure 4.6 presents the achieved results of the intra-cluster distance. The resulting clusters are shown in Figure 4.7. As shown the newly developed methodology improves the intra-cluster distance by another 3% compared to the standard fine-tuning. On the other hand, the computational time increases from 3.5 hours to 6.2 hours using the adapted fine-tuning. Comparing the resulting clusters, the adapted version changes the two almost isolated regions of cluster 12 to other clusters, which the standard fine-tuning is not able to do. Furthermore cluster 7 in the case of standard fine-tuning stays a single cluster, almost not visible, in the south of South Africa.

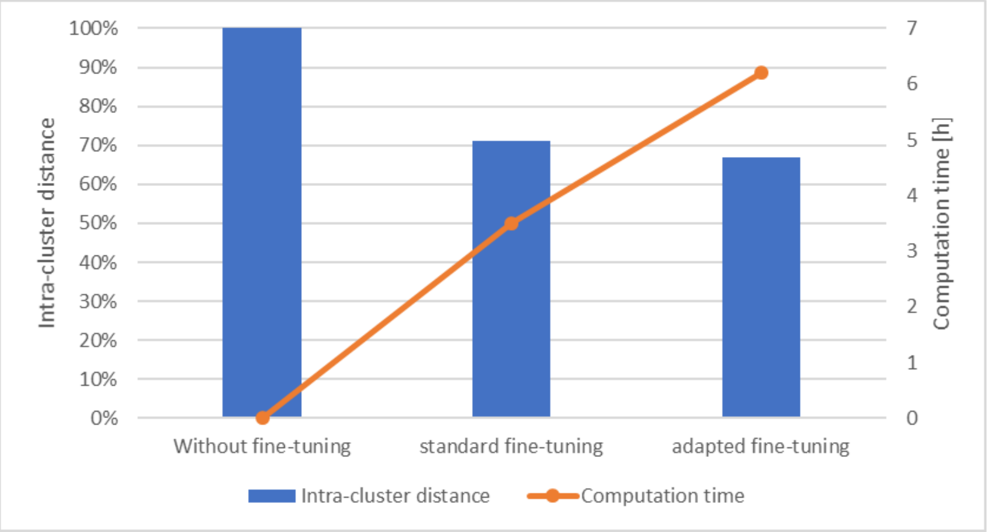

Figure 4.6: Comparison of intra-cluster distance depending on the fine-tuning used

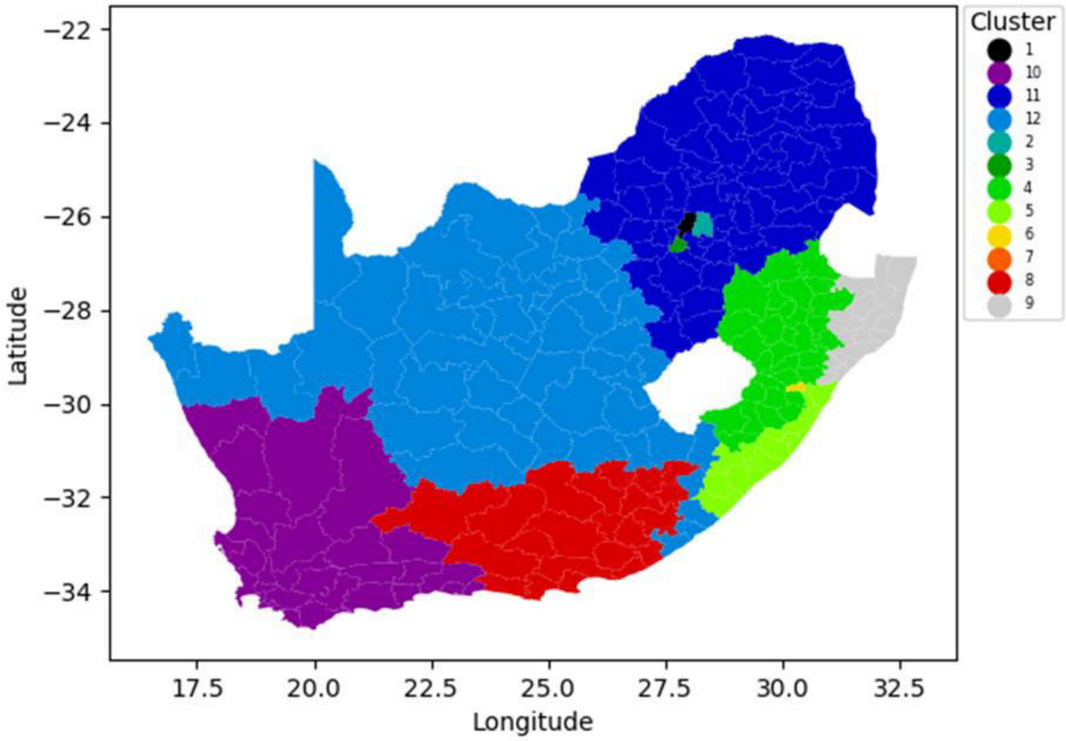

Figure 4.7: Regional model of 12 clusters of South Africa using standard fine-tuning

Considering the FLH instead of time series

One important part of the clustering is the temporal resolution that is considered using the time series as input data. Therefore, another aspect that should be considered is the output of the methodology using only the FLH instead of the time series as an input for the clustering process. The result is depicted in Figure 4.8. As shown, the result changes, grouping some of the single clusters together and also identifying Cape Town as a single cluster. Furthermore, clusters 3, 9, and 11 seem to be very similar to some of the PV potential zones and cluster 2 resembles a very low wind potential area shown in Figure 4.3. On top of that, clusters 4, and 5 are very similar to high wind potential zones. This indicates the high influence of renewable potentials in the clustering process when using either profiles or FLH.

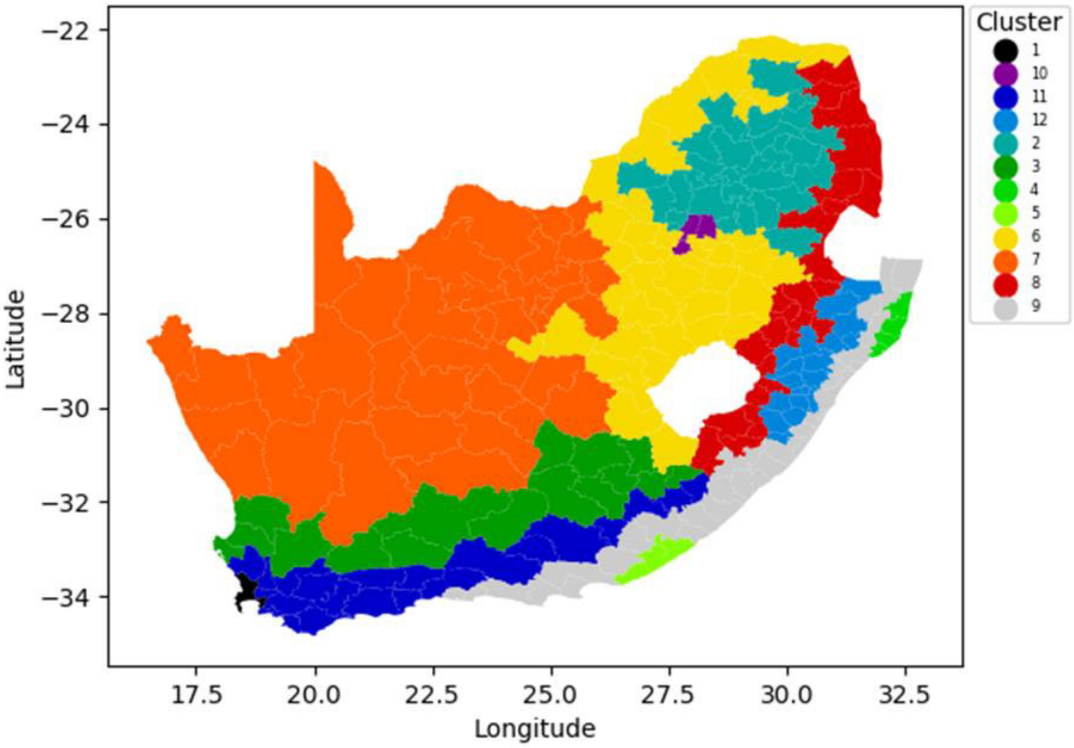

Figure 4.8: Regional model of 12 clusters of South Africa using FLH instead of time series

Only considering load- and generation-density

Another sensitivity that must be considered is are the load- and generation-densities of the respective regions. When only including those in the clustering process, the result is shown in Figure 4.9. As presented, this influences the methodology to identify the big cluster 12 including most of the regions. On the other hand, the load and generation structure of South Africa is implied since most of the other clusters are single-regions with a high value of load-density, i.e. load centers. Cluster 11 is the only cluster showing an influence of the coal generation that is located inside those regions. However, other connected regions of cluster 12 with high coal generation capacities are not assigned to cluster 11. This shows the low importance of generation capacities compared to the load-density.

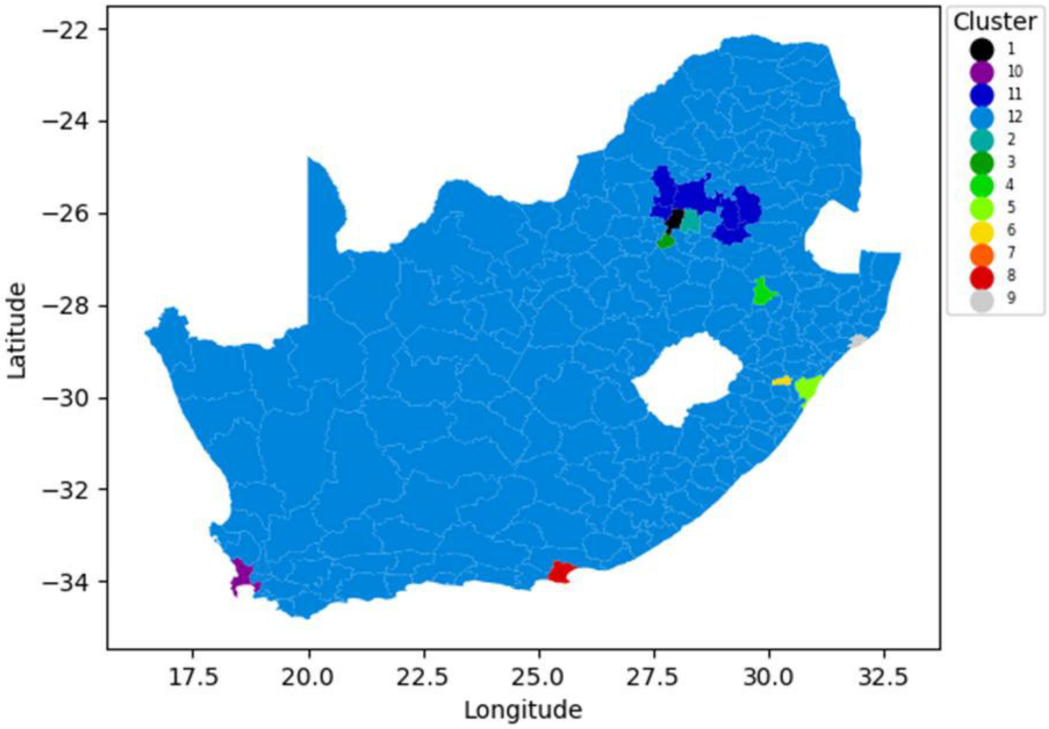

Figure 4.9: Regional model of 12 clusters of South Africa only using load- and generation-densities

Only time series or FLH as input data

To better understand the influence of the time series and FLH this subsection focuses on those using them as the single input for the clustering. The results are presented in Figure 4.10. As can be seen, there are some similarities between the clusters, e.g. clusters 9 and 2 of in this graphic seem very similar to clusters 4 and 5 of the standard clustering result shown in Figure 4.5, respectively. However, using only the time series, the figure does not resemble much the two maps of the FLH shown in Figure 4.3, which indicates the impact of the temporal component of the time series for the clustering input. Using only the FLH for wind and PV generation the methodology, the identified clusters resemble very much the important FLH centers shown in Figure 4.3. For example, cluster 1 identifies the high wind FLH, and clusters 7, 8, and 11 are different PV potential zones. This shows again the high influence of the renewable potentials in the clustering.

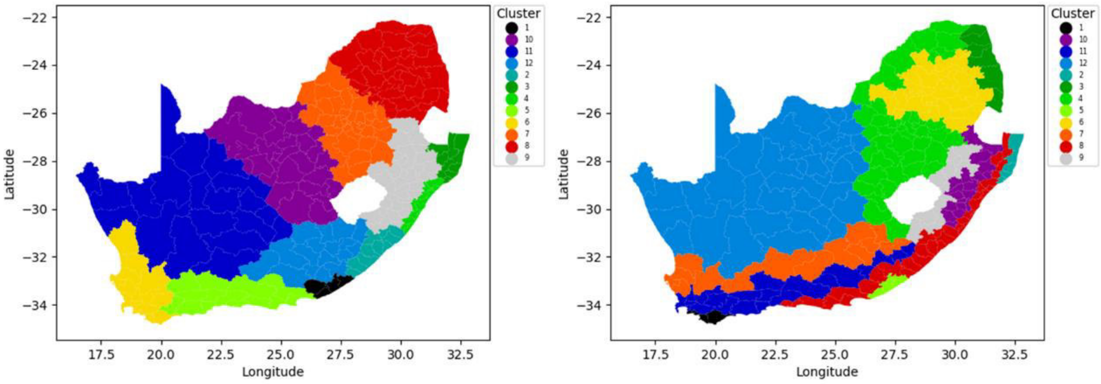

Figure 4.10: Regional model of 12 clusters of South Africa only using time series (left) and only FLH (right)

4.1.3 Results of the electrical grid

Using the OSM data belonging to the country of South Africa and applying the steps as described in 3.3.4, the algorithm results in the grid structure shown in Figure 4.11.

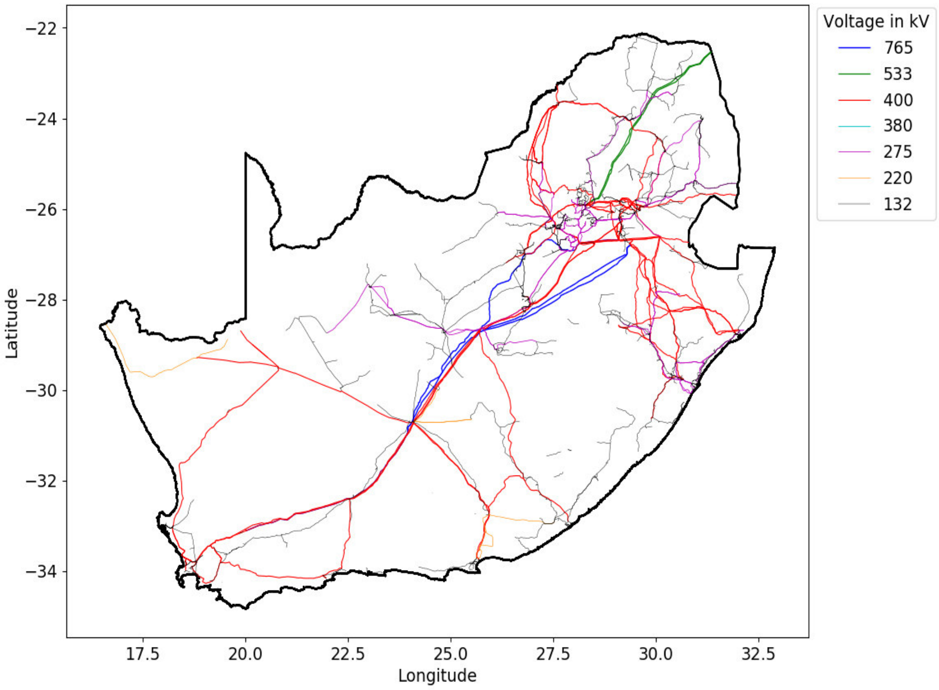

Figure 4.11: Electrical grid build from OSM data

Since no official data set is available, the validation of the developed methodology is done visually comparing it to another presentation of the electrical grid in South Africa shown in Appendix A. Even though some of the voltage levels of the lines are underestimated, e.g. the 765 kV from Bloemfontein to Cape Town or the transmission line in the southeast, the overall picture seems to fit the compared grid structure well.

Figure 4.12 presents the total transmission capacities between regions that result from the constructed grid shown in Figure 4.11. Taking as an illustrative example the connection between the clusters 11 and 12, there are following transmission lines connecting them: three of 765 kV, two of 400 kV, one of 380 kV one of 275 kV, and five of 220 kV. Assuming the standard transmission capacities as presented in Table 3.3, the resulting transmission capacity is at least 23.6 GW, i.e. of each line there is only one three-phase system installed. This means, only one of the 765 kV lines has to carry two three-phase systems to meet the calculated transmission capacity. Therefore, the resulting transmission capacities seem.

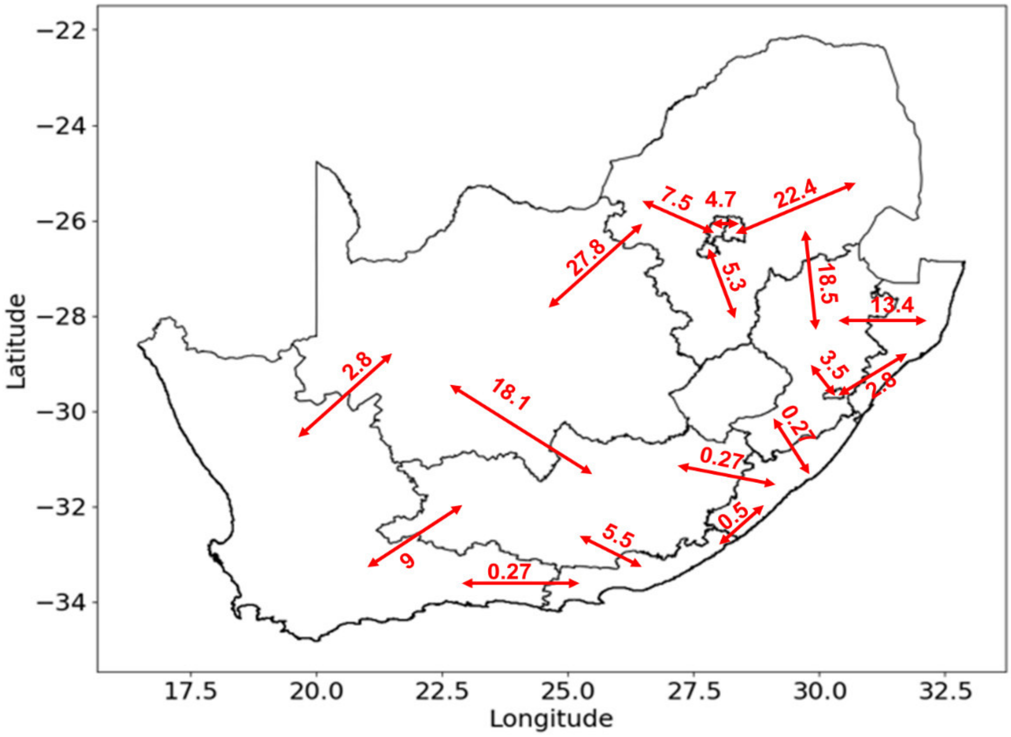

Figure 4.12: Transmission capacities between identified regions in GW

4.1.4 Results of the decarbonization path

When using the clustered regions, installed capacities, renewable profiles, and load distributions, assuming a standard load pattern, as input for the energy system model, its decarbonization pathway to the year 2050 is calculated. In this simulation a decarbonization of 80% is assumed. This can be adapted and is used in this work as an illustrative example of the regional modeling. Using different given technologies and commodities the solver calculates the cost-optimal mix of generation technologies [63]. Figure 4.13 presents the total installed capacities for the years 2015 to 2050 in the complete country.

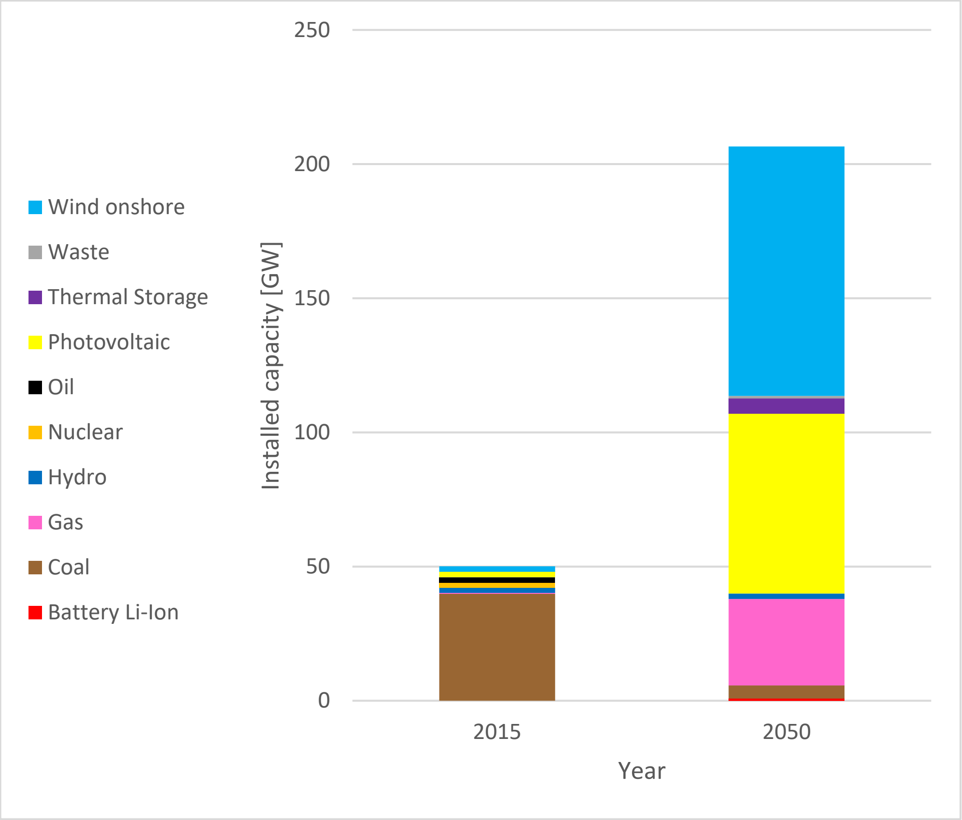

Figure 4.13: Capacities installed in South Africa from 2015 to 2050

As shown, the technology mix in 2050 differs significantly from the current state. The biggest differences are the reduction of coal as well as oil power plants and their substitution with mainly wind, PV, as well as gas power plants. The presentations of the installed capacities depending on the regions in 2015 and 2050 are shown in Appendix A.

It is shown that the main region for new PV generation capacities is cluster 12 due to its high transmission capacities to other regions and its high PV potential. New wind capacities are primarily installed in clusters 7, 8, 9, 10 and 12. Those clusters have been shown to have high wind potentials. However, also cluster 5 has high potentials of wind generation but low yearly demands and rather low transmission capacities to other regions which is probably why no further capacities were installed in that region. As shown in the figure, gas power plants are installed in almost every region. Especially in clusters 1, 2, and 11 which are characterized by high demands but rather low renewable potentials.

Figure 4.14 and Figure 4.15 present the net sum of the annual exchanged electrical energy between the identified clusters of South Africa in 2015 and 2050, respectively. As shown, the transported energies in 2015 are mainly distributed from cluster 11, where the coal power plants are located, to the other regions, especially to the west. On the other hand, the transferred energies in 2050 in most cases change their direction indicating the change in energy generation distribution. As shown, in 2050 high amounts of energy are produced on the south coast and in the center of the country.

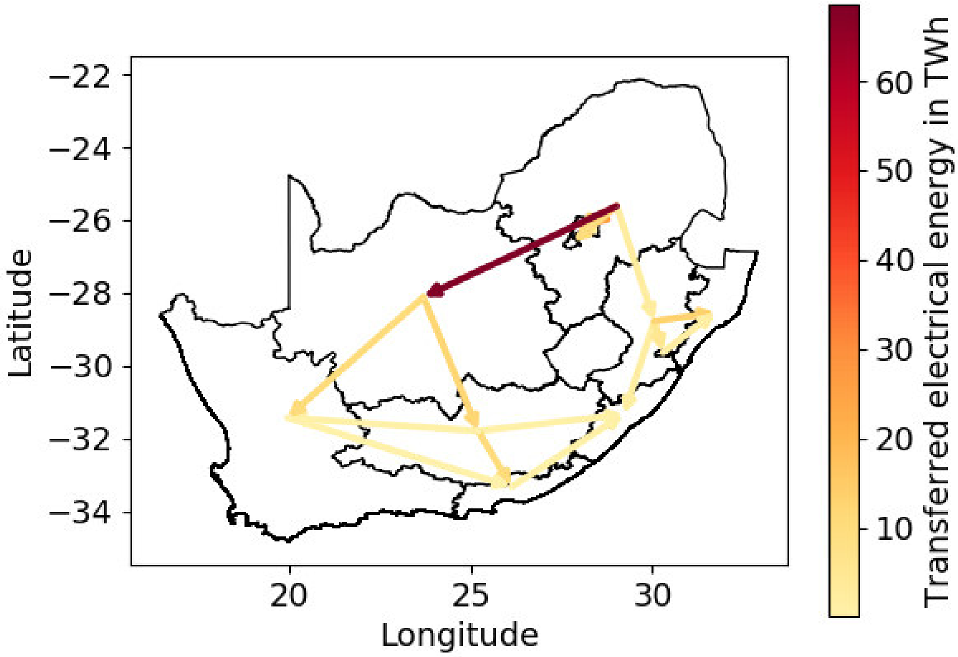

Figure 4.14: Transported energies between the regions of South Africa in 2015

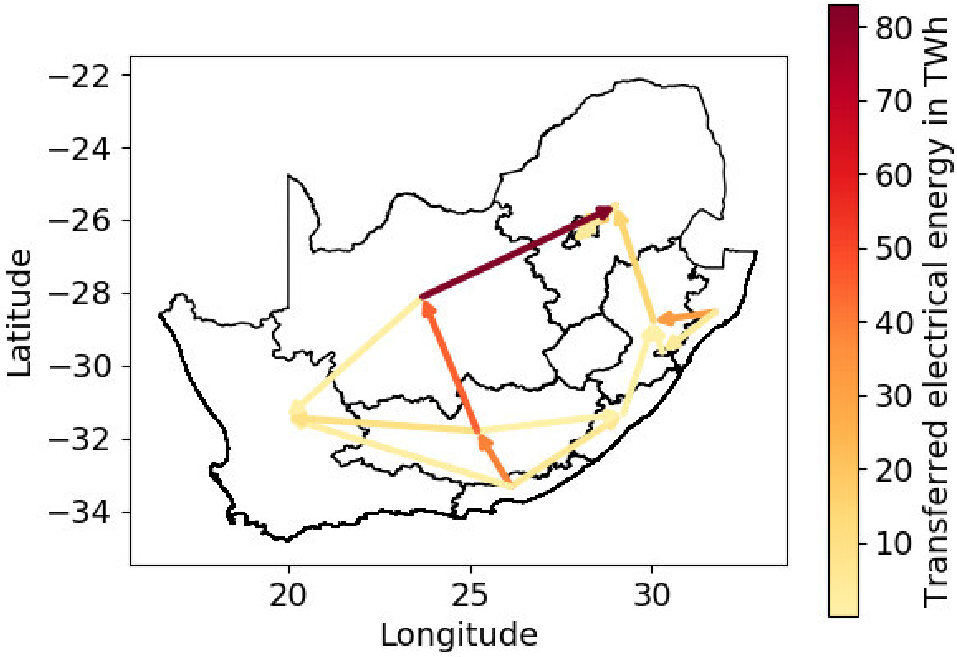

Figure 4.15: Transported energies between the regions of South Africa in 2050

New transmission capacities are only installed in the south of the country indicating the growing transmission capacities needed to transfer the wind energy produced in those regions. Regarding the line utilization, most transmission lines do not get close to their respective limits. Only one transmission line in the north and one in the south show high utilizations. The energy transmitted via the line in the south is small compared to the rest of the transferred energies. The transmission line in the north is crucial and, therefore, should be further considered regarding stability. The corresponding figures are presented in Appendix A.

As explained in 3.3.5 the costs of the different technologies and transmission capacities are the main target function for the optimization problem. Therefore, the total costs of the energy system transformation are one of the most important key performance indicators of the modeling. The discounted costs of the transformation of the energy system calculated by the optimization problem are 110.9 billion €.

4.1.5 Comparison to the country-level model

Since the regional modeling of the country energy system is supposed to improve the significance of the results of the single-region model, the two results are compared in this section. Two compare the two models the same initially installed capacities and demand as in the regional model is given as an input for the single-region model. The renewable profiles for the country are calculated using the procedure described in 3.3.1 with the country shape being the only regions. The results of this modeling are presented in the following using the same input parameters for the modeling in ESDP. The total costs of the transformation of the single-region model are 105 billion €. This means that the costs of regional modeling are 5 billion € higher. This difference in costs results from the costs of the grid which are neglected in the single-region model. Figure 4.16 presents the difference between the total installed capacities in 2050.

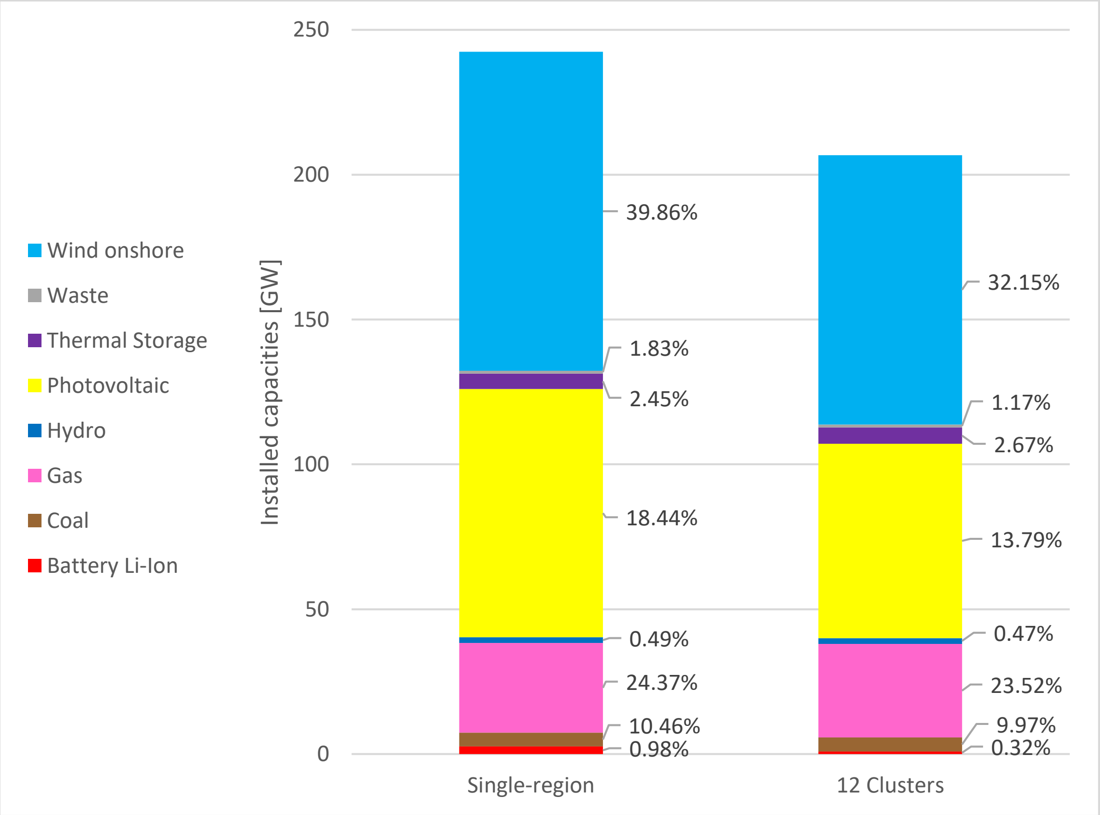

Figure 4.16: Comparison of installed capacities in the single-region model and the regional model of the 12 clusters. Percentages on the right side indicate the technology´s respective share of the total costs.

As shown, the amount of installed gas capacities is slightly higher in the regional model. On the contrary, the installed capacities of PV and wind are higher in the single-region model. Furthermore, in the single-region model three GW of batteries and five GW of thermal storage are installed, whereas in the regional modeling only one GW of Batteries but six GW of thermal storage are installed. This gap in installed capacities results from the underestimation of renewable potentials in the single-region model. Since in the single-region model the corresponding renewable profiles are calculated as the weighted average of the grid cells, in some hours a higher total capacity of renewables is needed. On the contrary, in the regional model the power plants are installed in those regions that have higher relative outputs what leads to the same energy output with less installed capacities.

Figure 4.17 presents the shares of generated electricity of the different technologies in 2050 and compares the results of the single-region model and the regional model. As shown, the generated energies by the different technologies are very similar. However, in the single-region model the amount of energy produced by solar is higher than in the regional model, whereas for the energy generated by wind the opposite applies.

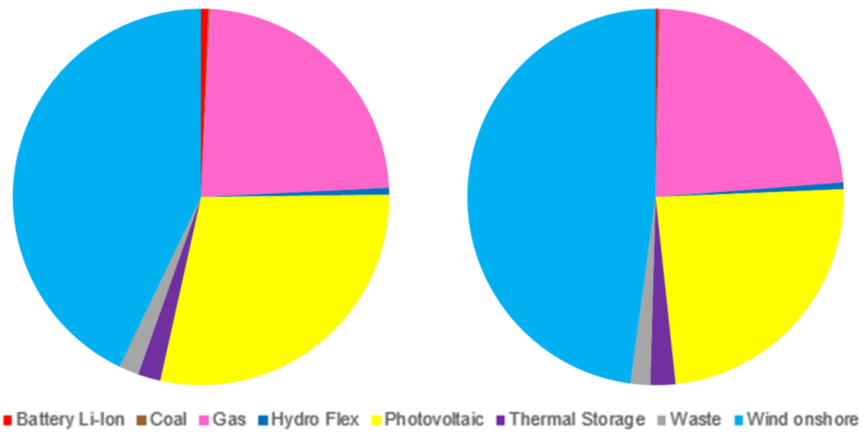

Figure 4.17: Electrical energy generated by the different technologies in South Africa in 2050 in the single-region model (left) and regional model (right)

4.1.6 Conclusion

Evaluating the results that have been achieved applying the developed methodology, it can be stated that the spatial resolution gives a deeper insight into the general energy system of the country. Different regions have been identified depending on different attributes. Those regions have significant differences between each other, and high standard deviations can be observed.

The generation of the grid structure using OSM data fits the real grid structure well. Even though the voltage level of some transmission lines is underestimated and therefore, also their transmission capacities, those capacities are still well dimensioned since the maximum utilization level is in a reasonable range. However, if underestimated the energy system modeling is still able to identify additional needs of transmission capacity. Nevertheless, this has to be validated cautiously since it could also occur that some transmission lines are overestimated. For example, a line of 10 kV that does not have a corresponding voltage label in the OSM data might receive the most frequent voltage level and, therefore, would be overestimated.

Furthermore, it has been shown that the spatial distribution does achieve different results regarding the decarbonization path. The amount of wind and PV generation need is overestimated by more than 30 GW. Furthermore, the electrical energy generated by wind turbines is significantly lower in the single-region model. The curtailed energy of the single-region model is 6.5 TWh of PV and 9.6 TWh of wind, respectively. In comparison, in the regional model 2.4 TWh of PV and 18 TWh of wind generation are curtailed. This implies the underestimation of wind potential in the single-region model since in the regional model less capacity is installed but even more energy is generated and curtailed.

The difference in the systems transformation cost is 5 billion € over the time period of 35 years. This difference mainly results from the additional costs of the grid operation in the regional model which account for about 15% of the total costs, i.e. 16 billion €. One aspect of the results, which has to be further investigated, is the installation of gas power plants within the clusters 1, 2, and 3 which are dense-populated cities. However, the transmission capacities between those clusters and its surrounding cluster still have unused potential. This means that the gas power plants could also be installed in cluster 11 instead and those capacities are used to a higher extent.

Comparing the results to another regional modeling of the decarbonization pathway towards 2050 in South Africa [62], the results of this work see a lower increase of PV and wind generation. This, however, also results from the use of a different decarbonization objective since this work used an 80% goal, whereas in the aforementioned article 100% decarbonization is considered. To compare those results the ration of wind and PV capacities is considered. This work indicates that the wind generation capacities are 40% higher than those of PV whereas in the other study the share of wind generation is only 20% of the installed PV capacities. This difference might result from the fact that this work considered a regionalization approach. On the contrary, the compared results were calculated considering the nine provinces of South Africa. This might have led to an underestimation of the wind potential. However, also other technologies have been considered. Especially automatically aligning PV capacities might make a difference to the average profiles that are considered in this work.

4.2 Use Case 2 – Germany

4.2.1 Data

The input data for the use case of Germany consists of 4869 regions. Figure 4.18 presents the PHH and CTSI load-density distributions. As shown the PHH load-density centers of Germany are mainly metropolitan areas such as Berlin, Hamburg, Munich, the area of Rhine and Ruhr, Stuttgart, and the area of Rhine-Main. Apart from that many different middle-load centers exist within the country of Germany. This implies the complexity of modeling Germany with only a few regions. Considering the CTSI demand, the cities having the highest load-density are very small and hard to identify visually, e.g. Bremerhaven.

Figure 4.19 presents the power generation distribution in Germany. As shown the technology mix in Germany is very distinct in different regions. For example, coal centers can be identified in the west and east of Germany, whereas nuclear power generation is mainly located in the south. Furthermore, solar power generation is located more frequently in the south and east of Germany. Unfortunately, the data considering wind generation are very few. This does not reflect the true German capacities. For a comparison of the total installed capacity see Appendix: B.

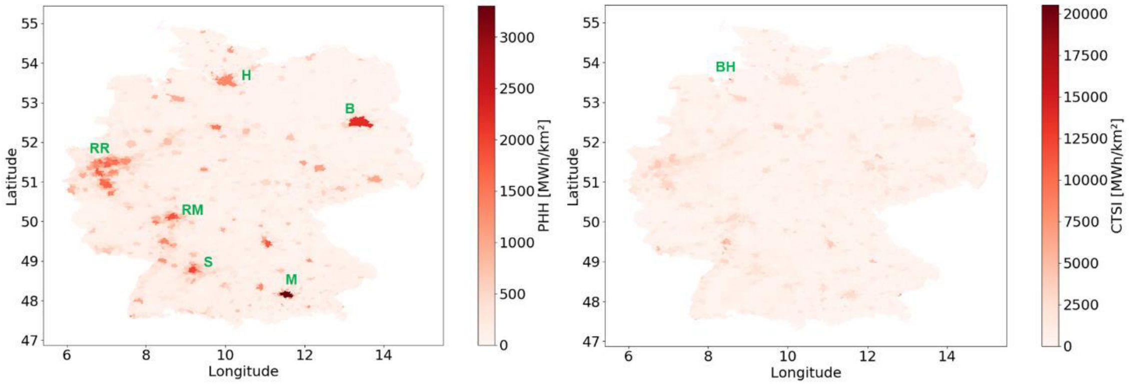

Figure 4.18: PHH (left) and CTSI (right) load-density distribution in Germany. Cities of high importance are Berlin (B), Bremerhaven (BH), Hamburg (H), Munich (M), the areas of Rhine-Ruhr (RR) and Rhine-Main (RM), and Stuttgart

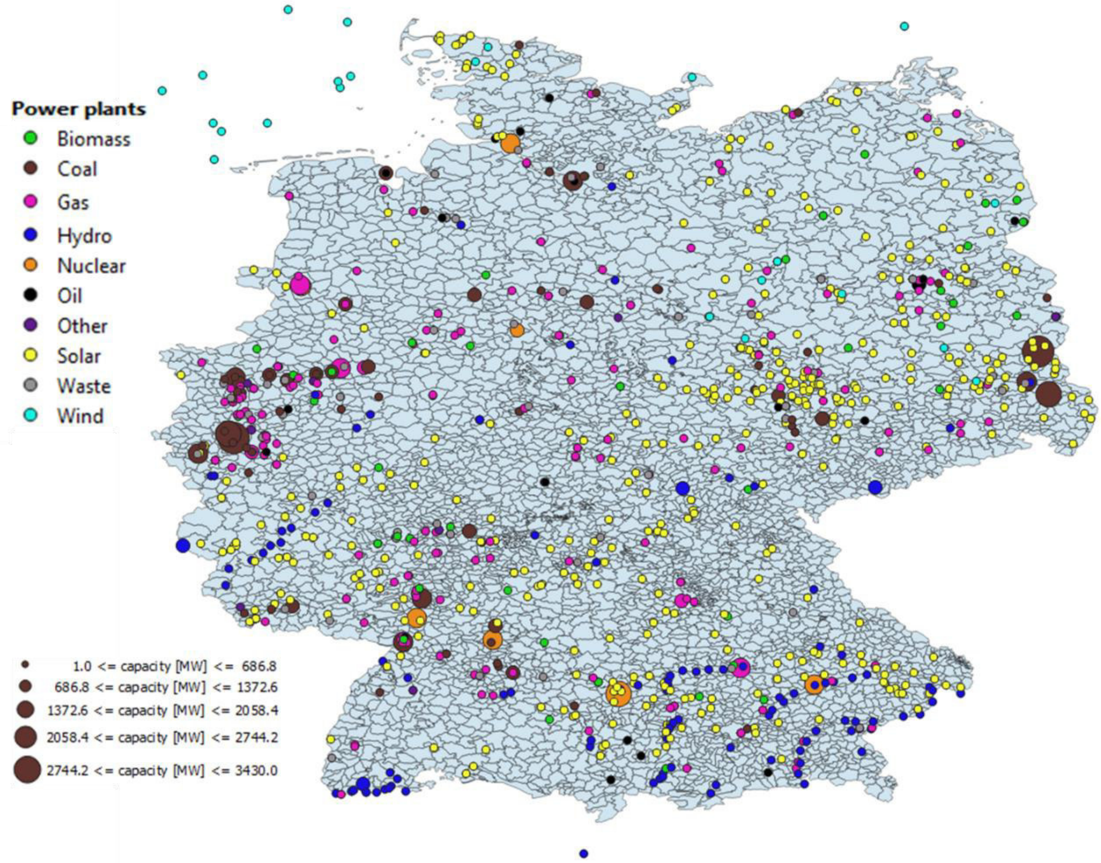
Power plants
Biomass
Coal
Gas
Hydro
Nuclear
Oil
Other
Solar
Waste
Wind
1.0 <= capacity [MW] <= 686.8
686.8 <= capacity [MW] <= 1372.6
1372.6 <= capacity [MW] <= 2058.4
2058.4 <= capacity [MW] <= 2744.2
2744.2 <= capacity [MW] <= 3430.0

Figure 4.19: Power generation distribution in Germany

Lastly, the renewable potentials are presented in Figure 4.20. As shown, the north of Germany is characterized by high wind potentials whereas the PV potentials increase to the south. However, those PV FLH are low compared to other countries since the maximum is around 1000 hours. Furthermore, the original grid cells of the database can be observed due to the comparably small regions of Germany.

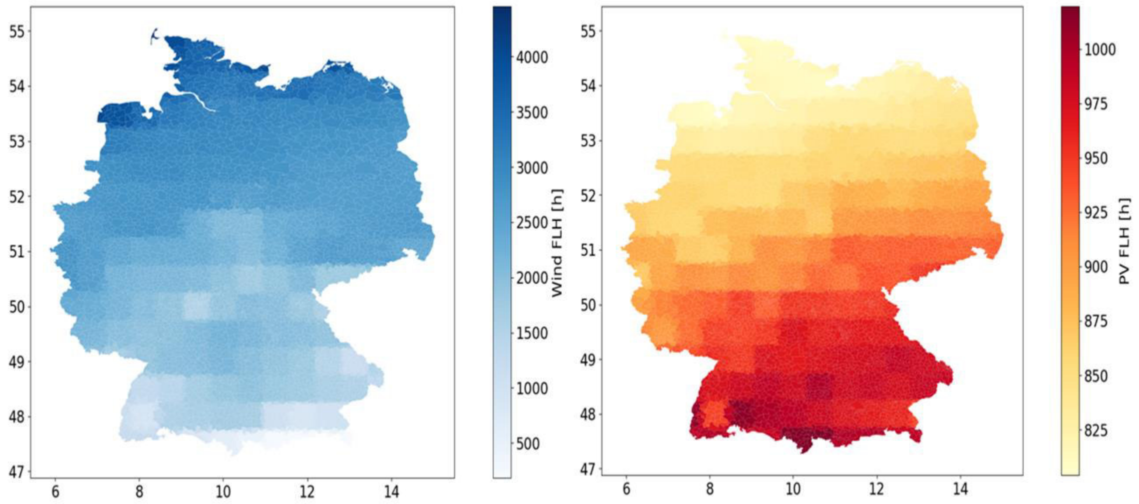

Figure 4.20: Wind (left) and PV (right) FLH in Germany

4.2.2 Cluster Results

When applying the developed methodology to the use case of Germany, the resulting curve of the elbow criterion is depicted in Figure 4.21, indicating an optimal number of clusters of five. Since the German case seems too complex to be reflected by only five regions another bending point on the elbow curve is searched. The optimal number of clusters considered in the further is 17. Another important factor that has to be considered in the case of Germany is the time limit of the fine-tuning since with this high number of regions the algorithm does not even terminate the first cycle in a week. Therefore, a threshold has to be set. In this work, the comparison is done setting this time limit to either three days or one week. To receive better results the fine-tuning can be applied for a longer time. However, due to time restrictions in this work, one week is set as the maximum time. Figure 4.22 presents the results of the methodology for three days and one week of fine-tuning.

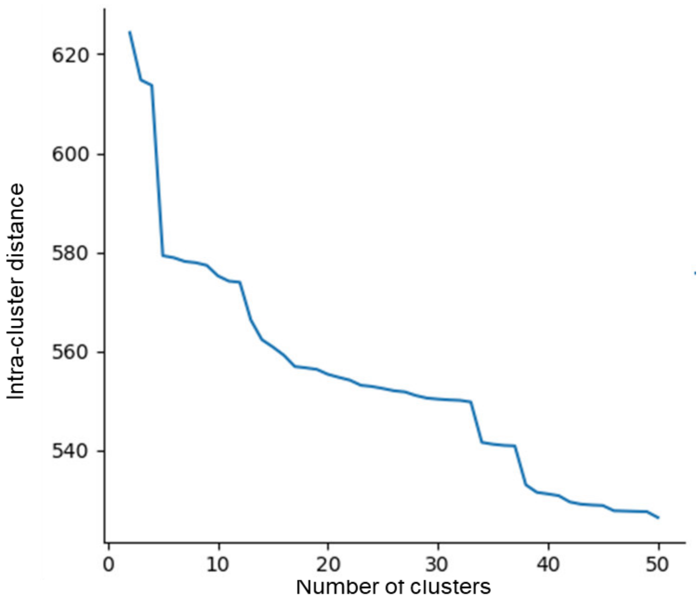

Figure 4.21: Intra-cluster distance for the case of Germany

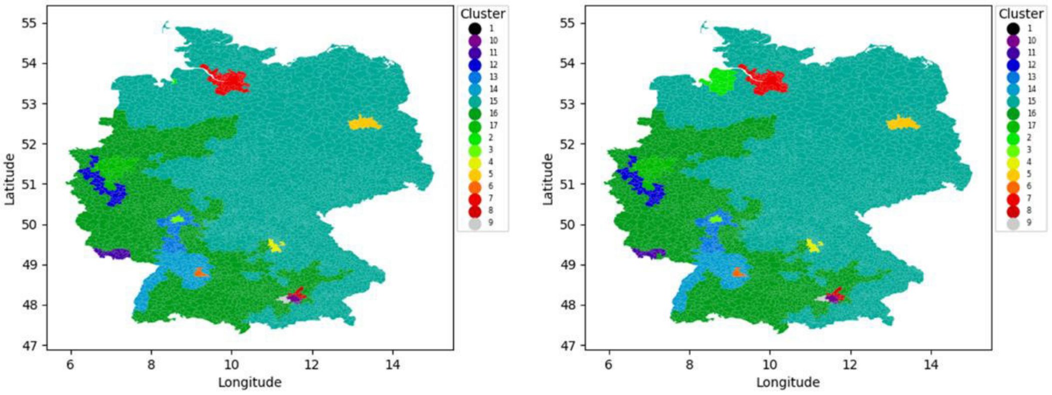

Figure 4.22: 17 clusters identified for the case of Germany applying the fine-tuning for three days (left) and one week (right)

As shown, the main difference that is achieved within the week is the increased size of cluster 2 around Bremerhaven. The rest of the clusters visually do not differ from the results of only three days of fine-tuning. Comparing the intra-cluster distance of the base case and the different time limits for fine-tuning, the decrease of intra-cluster distance within the first three days is around 3.3 %, whereas the last four days of fine-tuning only decrease the value by another 0.5%. This is due to the fine-tunings behavior of first applying the best possible changes to the data set. A corresponding graphic is presented in Appendix B.

In the following, the comparison of the different clusters is done regarding the results of the one week of fine-tuning due to its lower intra-cluster distance. The number of clusters belonging to each cluster is presented in Table 4.2. As shown, cluster 15 and 16 contain the highest number of regions, which are 64.5% and 30.7% of the total number of regions, respectively.

Cluster	Number of regions
1	1
2	19
3	5
4	16
5	12
6	7
7	60
8	19
9	14
10	8
11	28
12	51
13	133
14	223
15	3140
16	1496
17	44

Table 4.2: Number of regions belonging to the clusters when fine-tuning is applied for one week

The presentations of the means and standard deviations of the 17 clusters of Germany for load, generation, and renewable potentials can be observed in Appendix B. As shown, the load centers identified are mostly clusters 3, 4, 5, 6, 7, 10 and 17 which correspond to the areas around Frankfurt, Nuremberg, Berlin, Stuttgart, Hamburg, Munich, and the area of Rhine-Ruhr. The standard deviation again is higher as the mean value is higher. This is probably due to other regions belonging to the same cluster due to their similarity in one of the other attributes. The standard deviation in both PHH and GHDI is especially high in cluster 5 which is the area around Berlin. This is also influenced by using the density-load which is more similar than the total load of the different regions belonging to the cluster.

Regarding the installed capacities it is shown that the main part of the generation mix consists of coal, gas, and nuclear power generation. This, however, is due to the incompleteness of the data since it is known that the German electricity generation mix already consists of around 33% renewable power generation [64]. Again, the standard deviations of the different clusters are very high, which is influenced by the centralized character of conventional power generation technologies.

The FLH show that wind generation generally achieves more energy output. On the other hand, wind generation does vary more than the generation of PV units. The highest FLH of wind can be achieved in clusters 2, 5, 7, 12, and 17 which correspond to Bremerhaven, Berlin, Hamburg, and the area of Rhine-Ruhr, and its surrounding region. The big clusters 15 and 16 are not characterized by high wind FLH because both regions are characterized by comparably low wind generation which has been depicted in Figure 4.20.

4.2.3 Results of the electrical grid

The results of the electrical grid constructed by OSM data are presented in Figure 4.23. Since the transmission lines of 100 kV and higher are included, it complicates the validation of the grid structure of Germany since those data are difficult to obtain and the German grid is highly meshed. Therefore, only the transmission lines of 200 kV and higher are used to compare the constructed electrical grid structure to an official presentation shown in Appendix B.

Comparing those two grid structures, the lines and voltage levels seem to fit well. However, some of the transmission lines are underrated in terms of their voltage level, e.g. the transmission lines in the south of Baden-Württemberg, close to Konstanz, and the transmission lines in the state of Saarland.

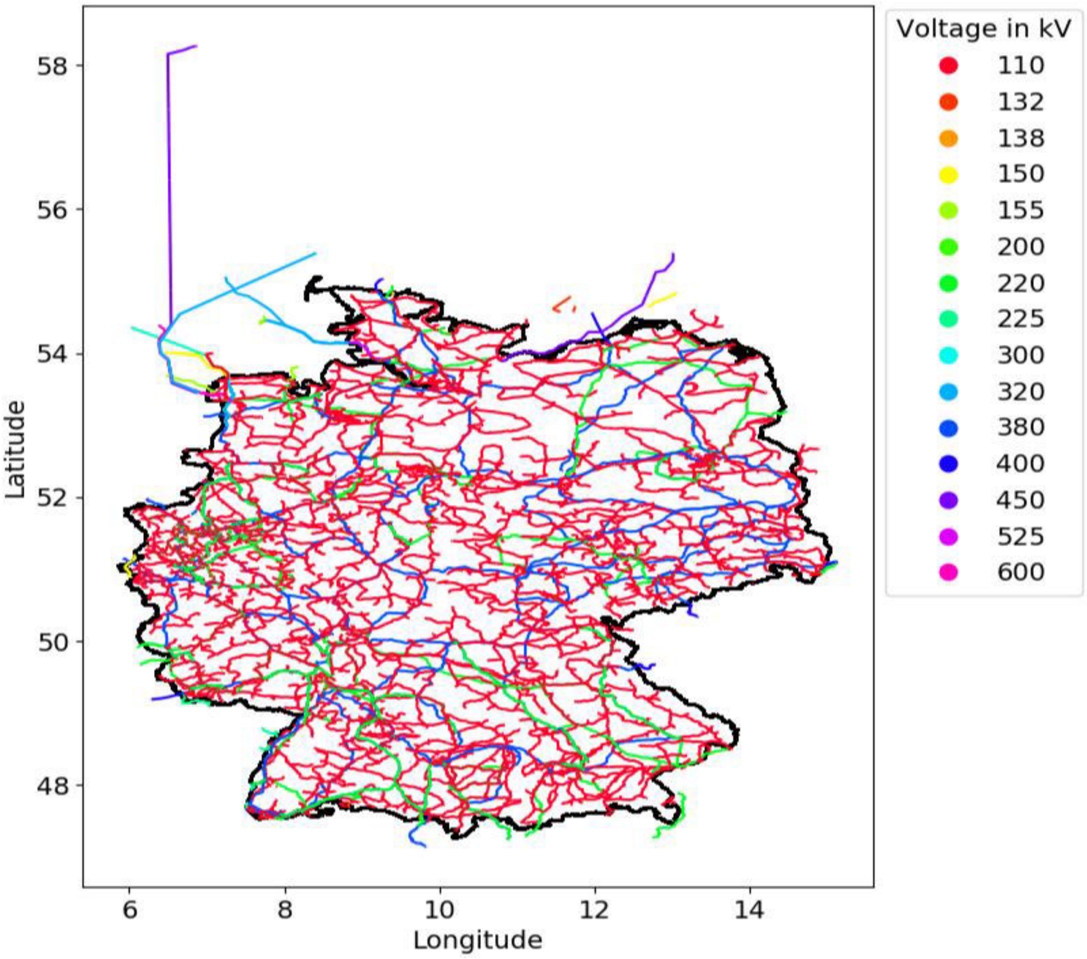

Figure 4.23: Electrical grid of Germany built from OSM data

4.2.4 Conclusion

The case of Germany is very complex due to multiple reasons. The number of regions is very high and there are a lot of load centers and generation structures within the country. The developed methodology identified clusters of very distinct sizes which might be reasonable. Unfortunately, clusters like cluster 15 and 16 are very big and reach from the north of the country till the south. Furthermore, cluster 16 is also only connected by some small corridors which assure the spatial contiguity. This results in the problem that the flows between north and south would be completely ignored when modeling the energy system. Since the north of Germany has high potentials for wind generation those flows will play a crucial role in the future energy system of Germany and, therefore, should not be neglected in the modeling. Applying the fine-tuning for more time might help overcome this issue. However, even one week did not make a high difference and it is unknown how long it would take the algorithm to terminate. If the first cycle has not been finished in one week it might take months to calculate. Whether the user can wait that long is questionable.

The grid data, constructed from the OSM data, seems to fit well to the overall structure of the German transmission grid. Some of the line´s voltage levels are underrated. Nevertheless, the energy system modeling still would be able to install additional transmission capacities if needed. However, since the number of cables is unknown it is not assured that the calculated transmission capacities correspond to the actual values. Furthermore, lines with a voltage level between 100 and 200 kV could not be validated due to the lack of official data. However, this voltage level usually is very meshed and therefore hard to validate visually which would have led to validation problems anyway.

4.3 Discussion

In this section the developed methodology and the results that can be achieved applying it to the two use cases are discussed. As shown in the previous sections, the methodology identifies regions that can be used within a regional energy system model to better consider regional differences within a country´s energy system. The case of South Africa implied that a single-region model of a country might poorly evaluate different aspects, e.g. renewable potential, that exist in different regions of the country. The developed methodology identified well-separated clusters that can be used as input for regional energy system modeling. However, the case of Germany showed that the methodology might lead to unsuited results regarding decentralized energy systems. Even though some important regions, e.g. the area of Rhine-Ruhr, can be observed, two very big regions are identified as well. Those regions are not suitable for the energy system modeling because they neglect important transmission paths, e.g. north to south. Since the fine-tuning algorithm only terminated by the time constraints, either this time limit has to be increased or a higher number of clusters must be chosen to improve the results of the developed methodology.

The importance of fine-tuning and renewable potentials can also be shown by the use case of Germany. Considering cluster 2 in the last four days of fine-tuning the algorithm assigns different other regions to the beforehand single-region cluster. Regarding the attributes of demand and generation the difference to other surrounding regions is seemingly very high as shown in Figure 4.18 and Figure 4.19. The only attribute that it shares similarity with compared to the surrounding regions is the renewable profiles. This shows, that without the fine-tuning the renewable profiles have a low influence on the clustering results. However, this attribute is of high importance to future energy systems and, in the use case of Germany, it must be taken into account to identify the differences in the regions of the north and south.

One of the most crucial parts of the methodology is the normalization since it highly influences the similarity of the clusters. Apart from the one explained in 3.3.2, different approaches have been considered during development. One more possibility that is described in Appendix C is called *weighted normalization*. This method emphasizes those technologies or demands that are more present in the present energy system in the country. However, the objective of this work relates to the transformation path. Therefore, emphasizing on the technologies that are most present today, which are mostly conventional power plants, is not reasonable.

Another variation to the developed methodology that was also considered was to include the electrical grid already in the clustering process. This means defining two regions to be spatially contiguous if a connecting transmission line with substations in each region exists. However, since only high voltage levels are considered many smallest regions were not connected to any other region, i.e. not having any substation or transmission line. Therefore, the merging of all those regions without any substation to that region containing the closest substation was considered. Nevertheless, this approach restricts the clustering in a way that the number of clusters increased drastically due to unconnected islands of grids which can result from incomplete data. Since the objective of the modeling is not to do a detailed power flow analysis but to investigate general insights like locations of future energy generation this approach has not been considered suitable.

Regarding the requirements of this work, the developed methodology identifies well-separated clusters and reduces the intra-cluster variance significantly in the fine-tuning process. However, the optimal number of clusters might be arbitrary which is a common problem using the elbow criterion. This is because that the bending point does not necessarily relate to the complexity of the energy system but to those points where two clusters are merged that are very distinct from each other. Another approach that has been considered is setting a threshold of intra-cluster distance considering the number of smallest regions of a country to find the optimal number of clusters. However, such a characteristic value could not be observed. Since this work does not apply a detailed power flow analysis but tries to identify general insights in future energy systems, the neglect of the electrical grid in the clustering process, and, therefore, the possible lack of transport capacities is within the regions, is considered reasonable. The developed methodology also fulfills the requirement of global applicability. However, regarding the results of countries with a high number of regions, as in the use case of Germany, it has to be further investigated whether or not the fine-tuning algorithm is capable of improving the initial results of the hierarchical clustering to an extent where the clusters are seen suitable for the energy system modeling. The examples of this methodology for New Zealand, Spain, and France can be observed in Appendix D. Those results show the methodology´s achievements regarding countries with considerable islands.

5 Conclusion and outlook

In this work a methodology has been developed that identifies regions with significant differences within a country´s energy system. Applying a spatially constrained clustering, the smallest administrative regions of a country are merged depending on their respective demand, generation mix, and renewable potential. Since there is no globally optimal solution finding the optimal number of clusters and slightly different parameters can change this number drastically, the user has to asses between the level of detail and computation time of the energy system model. The developed methodology might identify a number of clusters as optimal that is unintuitive or unsuited to a country´s energy system. The electrical grid is constructed using OSM data and indicates the possible interaction between the identified clusters. As shown, the grid structure fits the reality well, despite the underestimation of few transmission capacities. When using those identified regions, including their respective attributes, in a regional energy system modeling, different factors can be considered that a single-region model would neglect.

As shown in the use case of South Africa, the methodology is very sensitive to the inclusion of renewable profiles or FLH, as the identified regions are very similar to the clusters when only including the profile time series. There is also a significant difference in the results when considering the temporal resolution of the renewable profiles compared to the FLH. Different load centers have been identified within the country. However, the existing generation capacities seem to have a lower influence on the clustering results. It has been shown that a single-region model might underestimate the potentials of renewable power generation in different regions of a country due to the aggregated profile time series. Therefore, higher more capacities have to be installed which are unnecessary in the regional modeling. Depending on the country and the renewable potentials, this could potentially be the opposite as well, i.e. the overestimation of renewable potentials, which are limited by grid restrictions. Furthermore, the costs of operating the electrical grid and additional installations of transmission lines are neglected in the single-region model.

On the other hand, the use case of Germany has shown some limitations of the developed methodology regarding the number of smallest regions. The identified regions, in this case, are not suited for a regional energy system modeling due to the neglect of important transmission capacities from the north to the south of Germany. The potentials of wind in the northern regions of Germany have been underestimated because these regions have been merged with southern regions that are characterized by lower wind potentials. However, applying the fine-tuning algorithm for a longer time might help to improve the regions.

Important considerations for the developed methodology that might be implemented in the future include the integration of offshore wind generation potentials. Since the FLH of offshore wind turbines are considerably higher than those of onshore turbines, the developed methodology might underestimate the importance of different coastal regions. The algorithm has to be extended so that coastal areas with offshore time series can be compared to regions which do not have any coastline. Furthermore, the methodology must detect which regions are coastal and which are along the border to other countries on the mainland to identify which regions have access to offshore wind generation.

Apart from the recognition of coastal regions, the methodology could also be adapted to identify which regions are connected to the energy system of other countries by interconnecting transmission lines. This way, it could be identified which regions can import electricity from or export it to other regions and the respective capacity limits.

Another factor regarding the construction of the electrical grid is the inclusion of transmission capacity values for cables and HVDC transmission lines, respectively. Since those are currently seen as AC overhead lines with their respective transmission capacities, some of the capacities might be inaccurate.

The available space for power plants, especially renewables, is not considered yet in the methodology. Including this attribute, in the energy system modeling a constraint could be implemented to restrict the available space and, therefore, maximum capacity of renewable potentials which better reflects the reality.

References

[1] U. N. C. Change, "unfccc.int," [Online]. Available: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement. [Accessed 18 February 2020].

[2] K. Jörling y T. Sach, «Klimaschutz in Zahlen (2019),» Druck- und Verlagshaus Zarbock GmbH & Co. KG, Berlin, 2019.

[3] P. Lopion, P. Merkewitz, M. Robinius y D. Stolten, «A review of current challenges and trends in energy systems modeling,» *Renewable and Sustainable Energy Reviews,* vol. 96, pp. 156-166, November 2018.

[4] S. Pfenninger, A. Hawkes y J. Keirstead, «Energy systems modeling for twenty-first century energy challenges,» *Renewable and Sustainable Energy Reviews,* vol. 33, pp. 74-86, May 2014.

[5] R. M. AssunÇão, M. C. Neves, G. Câmara y C. Da Costa Freitas, «Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees,» *International Journal of Geographical Information Science,* vol. 20, nº 7, pp. 797-811, August 2006.

[6] R. L. Thorndike, «Who belongs in the family?,» *Psychometrika,* vol. 18, nº 4, pp. 267-276, 1953.

[7] D. Guo, «Greedy Optimization for Contiguity-Constrained Hierarchical Clustering,» de *2009 IEEE International Conference on Data Mining Workshops*, Miami, FL, USA, 2009.

[8] D. Oeding y B. Oswald, Elektrische Kraftwerke und Netze, Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[9] R. Puffer, *Optimierungspotenziale bei Freileitungen - Was ist machbar,* 2012.

[10] C. Müller, A. Hoffrichter, L. Wyrwoll, C. Schmitt, M. Trageser, T. Kulms, D. Beulertz, M. Metzger, M. Duckheim, M. Huber, M. Küppers, D. Most, S. Paulus, H. Heger y A. Schnettler, «Modeling framework for planning and operation of multi-modal

energy systems in the case of Germany,» *Applied Energy,* vol. 250, pp. 1132-1146, September 2019.

[11] A. Herbst, F. Toro, F. Reitze y E. Jochem, «Introduction to Energy Systems Modelling,» *Swiss Journal of Economics and Statistics,* vol. 148, nº 2, pp. 111-135, April 2012.

[12] W. Schade, E. Jochem, T. Barker, G. Catenazzi, W. Eichhammer, T. Fleiter, A. Held, N. Helfrich, M. Jakob, P. Criqui, S. Mima, L. Quandt, A. Peters, M. Ragwitz, U. Reiter, F. Reitze, M.-J. Schelhaas, S. Scrieciu y H. Turton, «ADAM – 2 degree scenario for Europe – policies and impacts. Deliverable M1.3 of ADAM (Adaptation and Mitigation Strategies: Supporting European Climate Policy),» Karlsruhe, 2009.

[13] ETSO, «Net Transfer Capacities (NTC) and Available Transfer Capacities (ATC) in the Internal Market of Elecricity in Europe (IEM) Information for User,» 2000.

[14] Bundesnetzagentur, «Bedarfsermittlung 2017-2030 Bestätigung Netzentwicklungsplan Strom,» Germany, 2017.

[15] J. Han, M. Kamber y A. Tung, «Spatial clustering methods in data mining: a survey,» *Data Mining and Knowledge Discovery - DATAMINE,* vol. Data Mining and Knowledge Discovery, January 2001.

[16] J. Mennis y D. Guo, «Spatial data mining and geographic knowledge discovery—An introduction,» *Computers, Environment and Urban Systems,* vol. 33, nº 6, pp. 403-408, November 2009.

[17] J. C. Duque, R. Ramos y J. Suriñach, «Supervised Regionalization Methods: A Survey,» *International Regional Science Review,* vol. 30, nº 3, pp. 195-220, June 2007.

[18] D. Xu y Y. Tian, «A Comprehensive Survey of Clustering Algorithms,» *Annals of Data Science,* vol. 2, nº 2, pp. 165-193, June 2015.

[19] T. Warren Liao, «Clustering of time series data—a survey,» *Pattern Recognition,* vol. 38, nº 11, pp. 1857-1874, November 2005.

[20] Z. Shi y L. Pun-Cheng, «Spatiotemporal Data Clustering: A Survey of Methods,» *ISPRS International Journal of Geo-Information,* vol. 8, nº 3, 28 February 2019.

[21] M. Chavent, V. Kuentz-Simonet, A. Labenne y J. Saracco, «an R package for hierarchical clustering with spatial constraints,» *Computational Statistics,* pp. 1799-1822, 20 January 2018.

[22] P. Chatterjee y P. Milanfar, «Clustering-Based Denoising With Locally Learned Dictionaries,» *IEEE Transactions on Image Processing,* vol. 18, nº 7, pp. 1438-1451, July 2009.

[23] S. Yuan, P.-N. Tan, K. S. Cheruvelil, S. M. Collins y P. A. Soranno, «Constrained spectral clustering for regionalization: Exploring the trade-off between spatial contiguity and landscape homogeneity,» de *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Campus des Cordeliers, Paris, France, 2015.

[24] K. Kim, Y. Chun y H. Kim, «p-Functional Clusters Location Problem for Detecting Spatial Clusters with Covering Approach: Functional Clusters Location Problem,» *Geographical Analysis,* vol. 49, nº 1, pp. 101-121, January 2017.

[25] T. Soni Madhulatha, «An overview on clustering methods,» *IOSR Journal of Engineering,* pp. 719-725, April 2012.

[26] D. Arthur y S. Vassilvitskii, «k-means++: The Advantages of Careful Seeding,» *Proceedings of the eighteenth annual ACM,* 2007.

[27] W. Tobler, «A Computer Movie Simulating Urban Growth in the Detroit Region,» *Economic Geography,* vol. 46, p. 234, June 1970.

[28] K. Siala y M. Y. Mahfouz, «Impact of the choice of regions on energy system models,» *Energy Strategy Revies,* pp. 75-85, 05 June 2019.

[29] D. Guo, «Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP),» *International Journal of Geographical Information Science,* vol. 22, nº 7, pp. 801-823, June 2008.

[30] «scikit-learn,» [En línea]. Available: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering. [Último acceso: 28 December 2019].

[31] J. L. Fitterer, T. A. Nelson, N. C. Coops y M. A. Wulder, «Modelling the ecosystem indicators of British Columbia using Earth observation data and terrain indices,» *Ecological Indicators,* pp. 151-162, September 2012.

[32] Y. Unal, T. Kindap y M. Karaca, «Redefining the climate zones of Turkey using cluster analysis,» *International Journal of Climatology,* vol. 23, nº 9, pp. 1045-1055, July 2003.

[33] J. A. Kupfer, P. Gao y D. Guo, «Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning,» *Ecological Informatics,* vol. 9, pp. 11-18, May 2012.

[34] M. Ester, H. P. Krigel, J. Sander y X. Xu, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,» de *Proc.ofthe2ndInternationalConferenceonK*, Portland, WA, 1996.

[35] S. U. Rehman, S. Asghar, S. Fong y S. Srasvady, «DBSCAN: Past, present and future,» de *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, Bangalore, India, 2014.

[36] W. Wang, J. Yang y R. Muntz, «STING : A Statistical Information Grid Approach to Spatial Data Mining,» de *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB´97)*, 1997.

[37] R. Agrawal, J. Gehrke, D. Gunopulos y P. Raghavan, «Automatic Subspace Clustering of High Dimensial Data for Data Mining Applications,» de *Proc. 1998 ACM-SIGMOD Int. COnf. Management of Data (SIGMOD´98)*, June 1998.

[38] J. C. Duque, L. Anselin y S. J. Rey, «THE MAX-P-REGIONS PROBLEM*,» *Journal of Regional Science,* vol. 52, nº 3, pp. 397-419, August 2012.

[39] J. C. Duque, M. C. Vélez-Galego y L. C. Echeverri, «On the Performance of the Subtour Elimination Constraints Approach for the p-Regions Problem: A Computational Study: On the Performance of the Subtour Elimination Constraints Approach for the p -Regions Problem,» *Geographical Analysis,* vol. 50, nº 1, pp. 32-52, June 2017.

[40] G. Chicco, «Overview and performance assessment of the clustering methods for electrical load pattern grouping,» *Energy,* vol. 42, nº 1, pp. 68-80, June 2012.

[41] K. S. Cheruvelil, S. Yuan, K. E. Webster, P.-N. Tan, J.-F. Lapierre, S. M. Collins, C. E. Fergus, C. E. Scott, E. N. Henry, P. A. Soranno, C. T. Filstrup y T. Wagner, «Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method,» *Ecology and Evolution,* vol. 7, nº 9, pp. 3046-3058, May 2017.

[42] G. Chicco, R. Napoli y F. Piglione, «Comparisons Among Clustering Techniques for Electricity Customer Classification,» *IEEE Transactions on Power Systems,* vol. 21, nº 2, pp. 933-940, May 2006.

[43] P. J. Rousseeuw, «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,» *Journal of Computational and Applied Mathematics,* vol. 20, pp. 53-65, November 1987.

[44] A. M. Aryal y S. Wang, «Discovery of patterns in spatio-temporal data using clustering techniques,» de *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, China, 2017.

[45] H. Assem, L. Xu, T. S. Buda y D. O´Sullivan, «Spatio-Temporal Clustering Approach for Detecting Functional Regions in Cities,» de *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA, USA, 2016.

[46] T. Calinski y J. Harabasz, «A dendrite method for cluster analysis,» *Communications in Statistics - Theory and Methods,* vol. 3, nº 1, pp. 1-27, 1974.

[47] D. L. Davies y D. W. Bouldin, «A Cluster Separation Measure,» *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vols. %1 de %2PAM-1, nº 2, pp. 224-227, April 1979.

[48] M. Ramesh Kanthan y S. Naga Nandini Sujatha, «Automatic Grayscale Classification using Histogram Clustering for Active Contour Models,» *International Journal of Current Engineering and Technology,* 03 April 2013.

[49] H. Butler, C. Schmidt, D. Springmeyer y J. Livni, «spatialreference.org,» [En línea]. Available: https://spatialreference.org/. [Último acceso: 07 February 2020].

[50] C. GADM, «GADM,» [En línea]. Available: https://gadm.org/. [Último acceso: 30 March 2020].

[51] D. f. R. a. U. P. (. R. o. t. E. C. G. H. P. I. Joint Research Centre (JRC). [En línea]. Available: https://ghsl.jrc.ec.europa.eu/. [Último acceso: 05 March 2020].

[52] IEA, «IEA webstore,» [En línea]. Available: https://webstore.iea.org/statistics-data. [Último acceso: 05 March 2020].

[53] O. S. Map, «Open Street Map,» Open Street Map contributors, [En línea]. Available: https://www.openstreetmap.org/export#map=7/50.552/10.657. [Último acceso: 05 March 2020].

[54] G. K. R. I. o. T. i. S. E. W. R. I. Global Energy Observatory, «Global Power Plant Database,» 2018. [En línea]. Available: http://datasets.wri.org/dataset/globalpowerplantdatabase. [Último acceso: 05 March 2020].

[55] NASA, «MERRA-2,» [En línea]. Available: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/. [Último acceso: 05 March 2020].

[56] N. X. Vinh, J. Epps y J. Bailey, «Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance,» *Journal of Machine Learning Research,* vol. 11, pp. 2837-2854, 10 October 2010.

[57] C. C. Aggarwal, A. Hinneburg y D. A. Keim, «On the Surprising Behavior of Distance Metrics in High Dimensional Space,» de *Database Theory — ICDT 2001*, vol. 1973, G. Goos, J. Hartmanis, J. van Leeuwen, J. Van den Bussche y V. Vianu, Edits., Berlin, Heidelberg, Springer Berlin Heidelberg, 2001, pp. 420-434.

[58] B. Boehmke y B. Greenwell, Hands-On Machine Learning with R, vol. Chapman & Hall/CRC The R Series, Chapman and Hall/CRC, 2019.

[59] O. Contributors, «openstreetmap.org,» [En línea]. Available: https://www.openstreetmap.org. [Último acceso: 18 February 2020].

[60] «Next-Kraftwerke,» Next-Kraftwerke, [En línea]. Available: https://www.next-kraftwerke.com/knowledge/utility-frequency. [Último acceso: 05 March 2020].

[61] U. Leprich, M. Ritza, M. Igel, H. Guss, U. Macharey, J. Diegler y K. Weiler, «Ausbau elektrischer Netze mit Kabel oder Freileitung unter besonderer Berücksichtigung der Einspeisung Erneuerbarer Energien,» 2011.

[62] A. S. Oyewo, A. Aghahosseini, M. Ram, A. Lohrmann y C. Breyer, «Pathway towards achieving 100% renewable electricity by 2050 for South Africa,» *Solar Energy,* vol. 191, pp. 549-565, October 2019.

[63] S. N. Paredes Pineda, *Development and Evaluation of a Simplified Approach,* Munich, 2019.

[64] BMWI, «Bundesministerium für Wirtschaft und Energie,» [En línea]. Available: https://www.bmwi.de/Redaktion/DE/Dossier/erneuerbare-energien.html. [Último acceso: 19 March 2020].

[65] «GET invest,» [En línea]. Available: https://www.get-invest.eu/market-information/south-africa/energy-sector/. [Último acceso: 30 March 2020].

[66] «VDE FNN,» Forum Netztechnik/Netzbetrieb im VDE, January 2018. [En línea]. Available: https://www.vde.com/de/fnn/dokumente/karte-deutsches-hoechstspannungsnetz. [Último acceso: 22 March 2020].

[67] D. Wang, L. Yang, A. Florita, S. Shafiul Alam, T. Elgindy y B.-M. Hodge, «Automatic Regionalization Algorithm for Distributed State Estimation in Power Systems,» de *Global Conference on Signal and Information Processing*, Washington, D.C., 2016.

[68] D. Müllner, «Modern hierarchical, agglomerative clustering algorithms,» *arXiv:1109.2378 [cs, stat],* 12 September 2011.

List of abbreviations

AC	Alternate Current
ATC	Available Transfer Capacity
CTSI	Commerce, Trade, Service, and Industry
CVI	Clustering Validity Indicator
ESDP	Energy System Development Plan
FLH	Full Load Hours
GIS	Geographic Information Systems
HVDC	High Voltage Direct Current
NTC	Net Transfer Capacity
NTF	Notified transmission Flow
OSM	Open Street Map
PHH	Private Household
PV	Photovoltaics
SRID	Spatial Reference ID
STD	STandard Deviation
TRM	Transmission Reliability Margin
TSO	Transmission System Operator
TTC	Total Transfer Capacity

Appendix: A – Use case South Africa

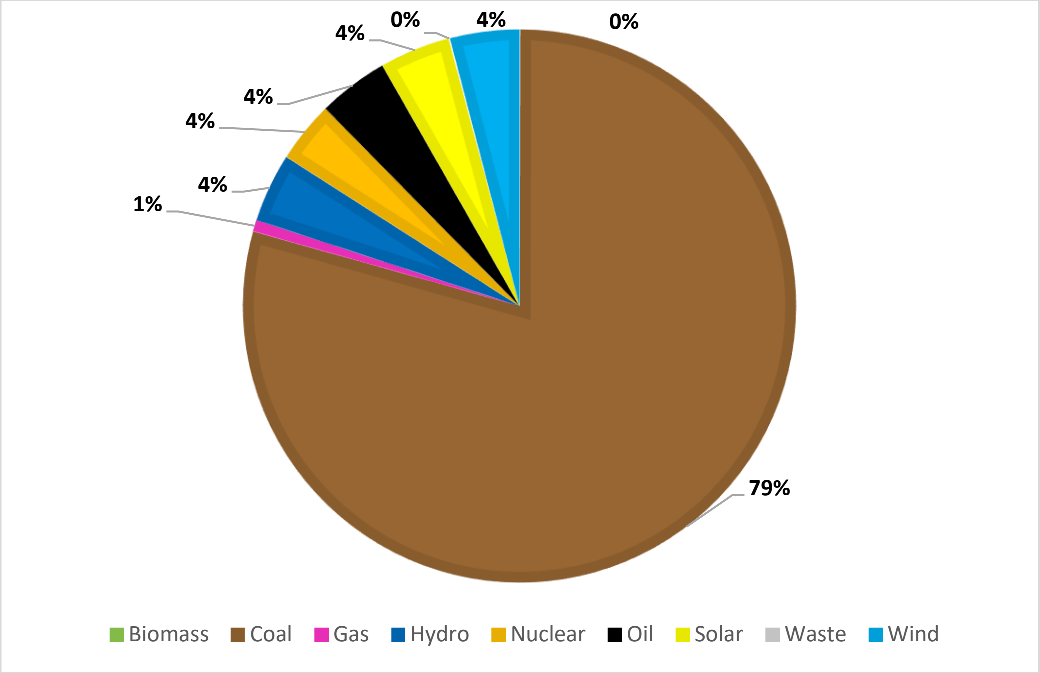

Figure 5.1: Total generation capacities of South Africa in 2015

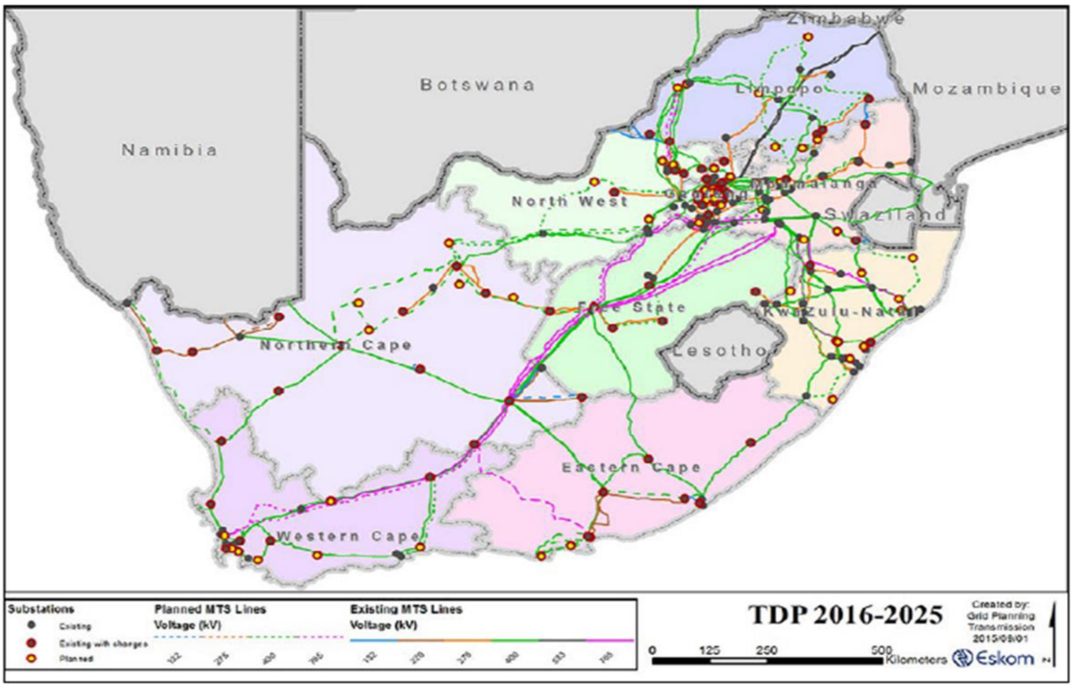

Figure 5.2: Electrical grid structure of South Africa [65]

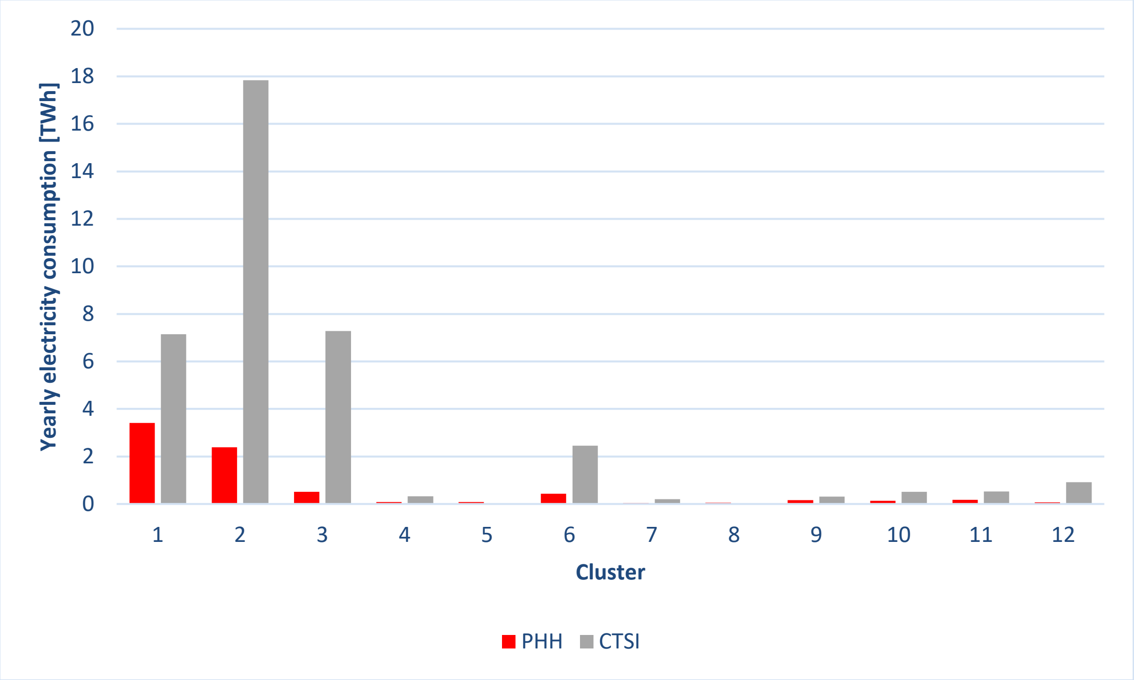

Figure 5.3: Mean demand of the 12 identified clusters in the use case of South Africa

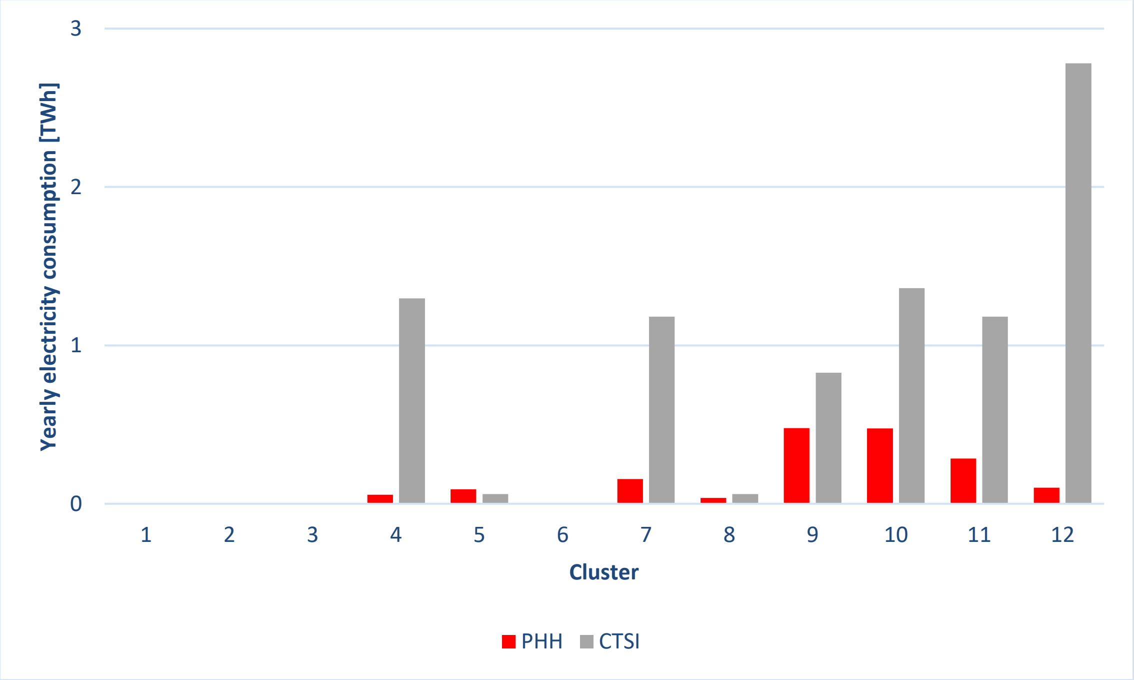

Figure 5.4: Standard deviation of the demand of the 12 identified clusters in the use case of South Africa

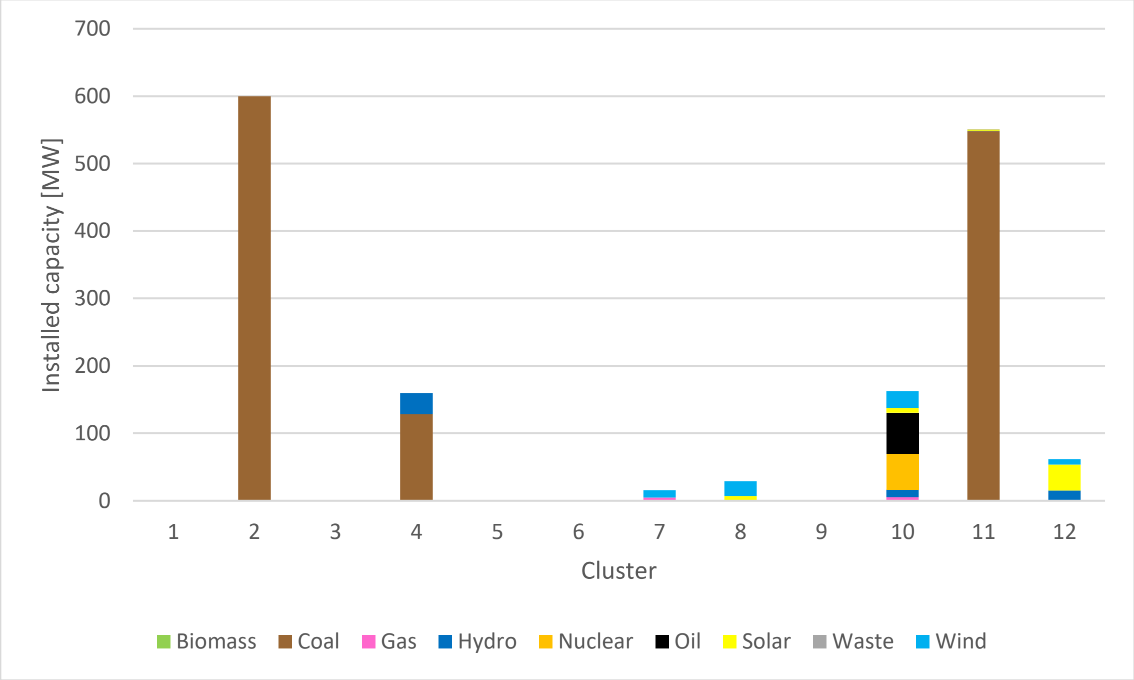

Figure 5.5: Mean installed capacities of the 12 identified clusters in the use case of South Africa

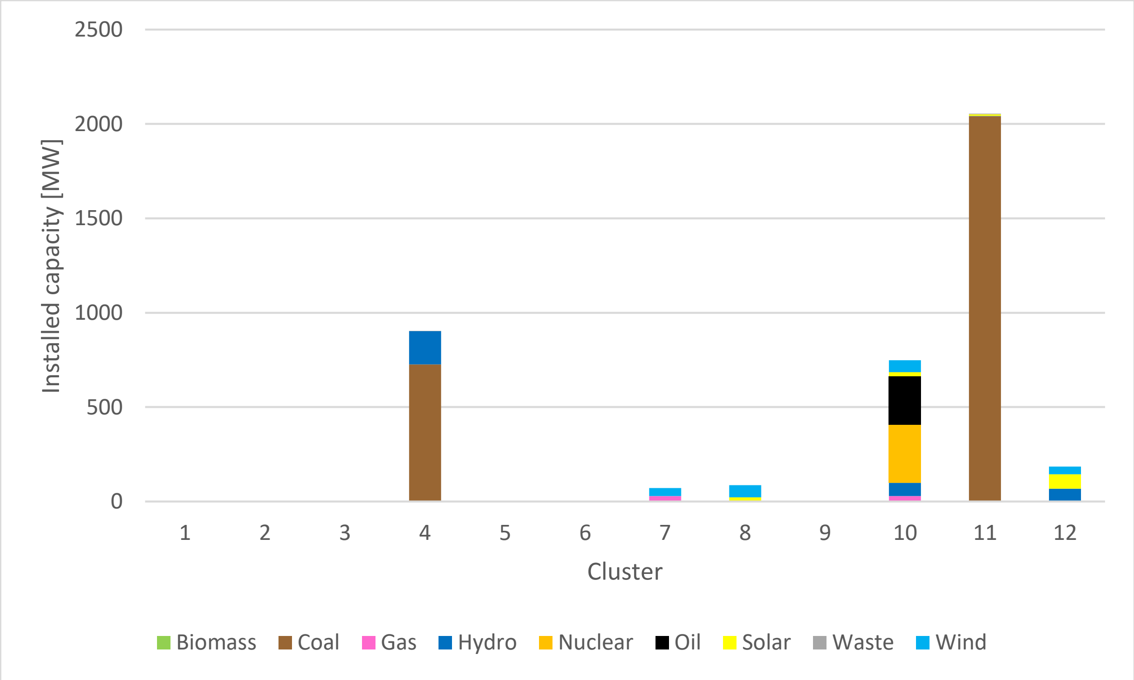

Figure 5.6: Standard deviation of installed capacities of the 12 identified clusters in the use case of South Africa

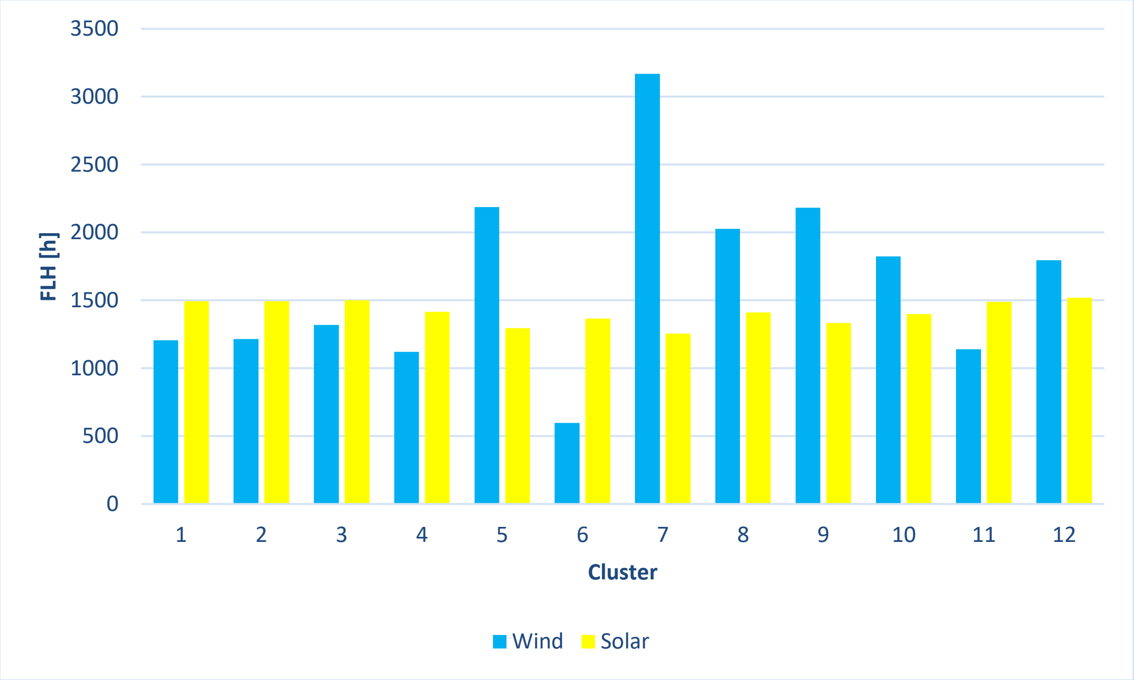

Figure 5.7: Mean FLH of renewables of the 12 identified clusters in the use case of South Africa

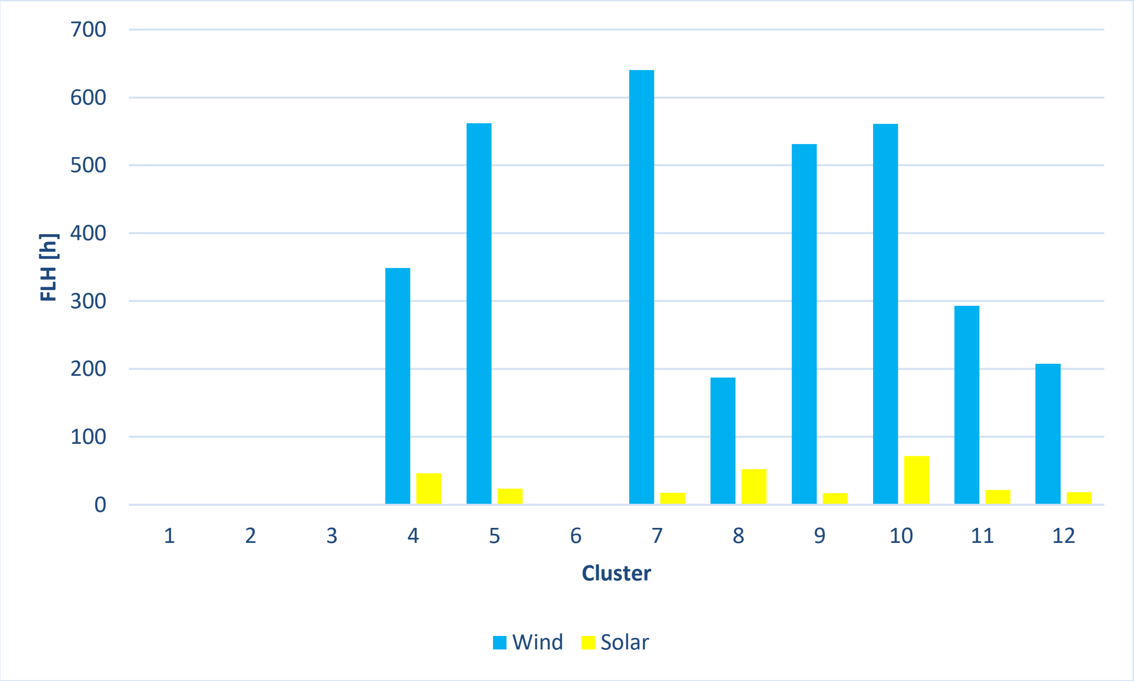

Figure 5.8: Standard deviation of the FLH of renewables of the 12 identified clusters in the use case of South Africa

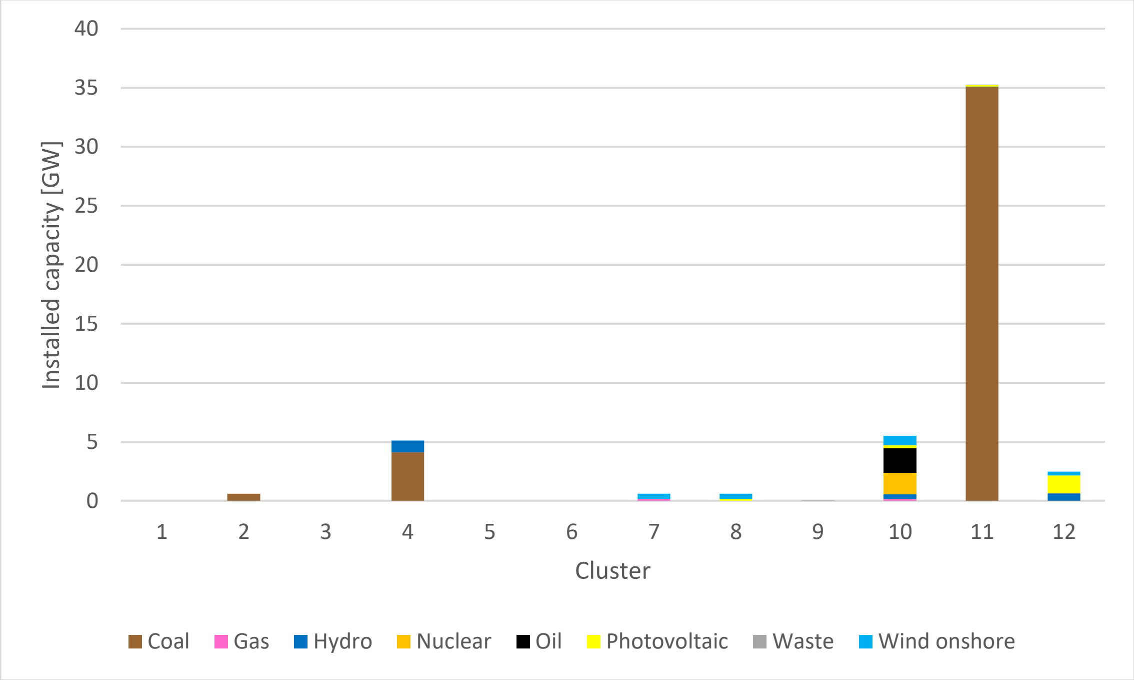

Figure 5.9: Total installed capacities in the 12 clusters of the use case of South Africa in 2015

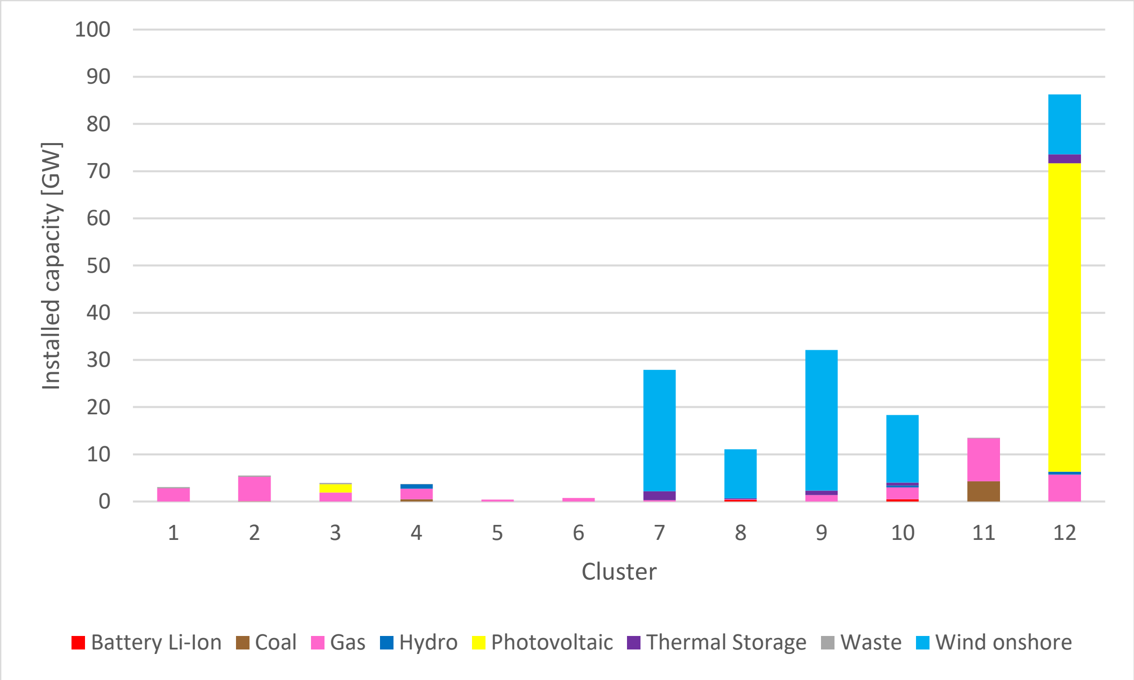

Figure 5.10: Total installed capacities in the 12 clusters of the use case of South Africa in 2050

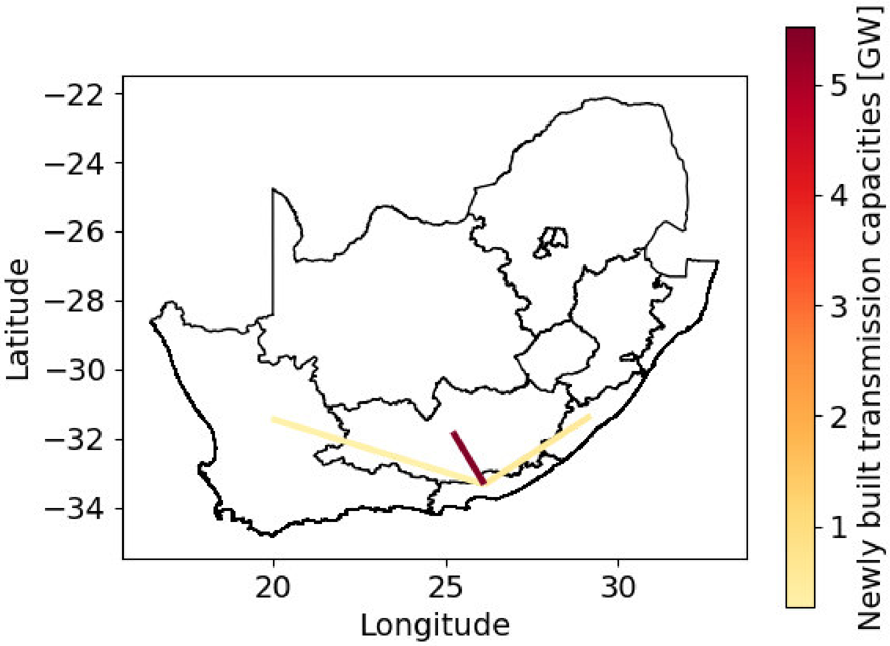

Figure 5.11: Additional installed capacities until 2050 in the use case of South Africa

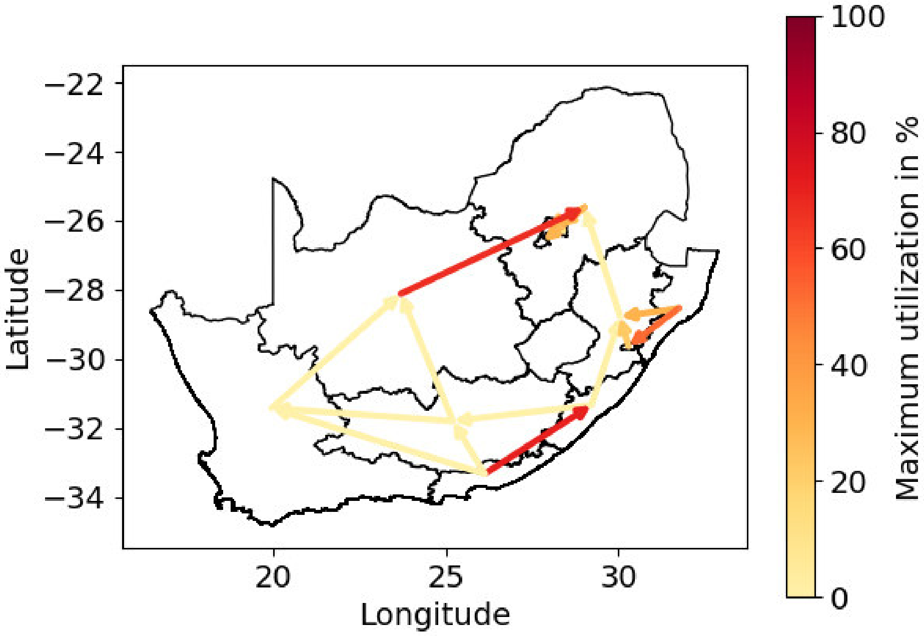

Figure 5.12: Maximum utilization of the transmission capacities between the 12 clusters of the use case of South Africa in 2050

Appendix: B – Use case Germany

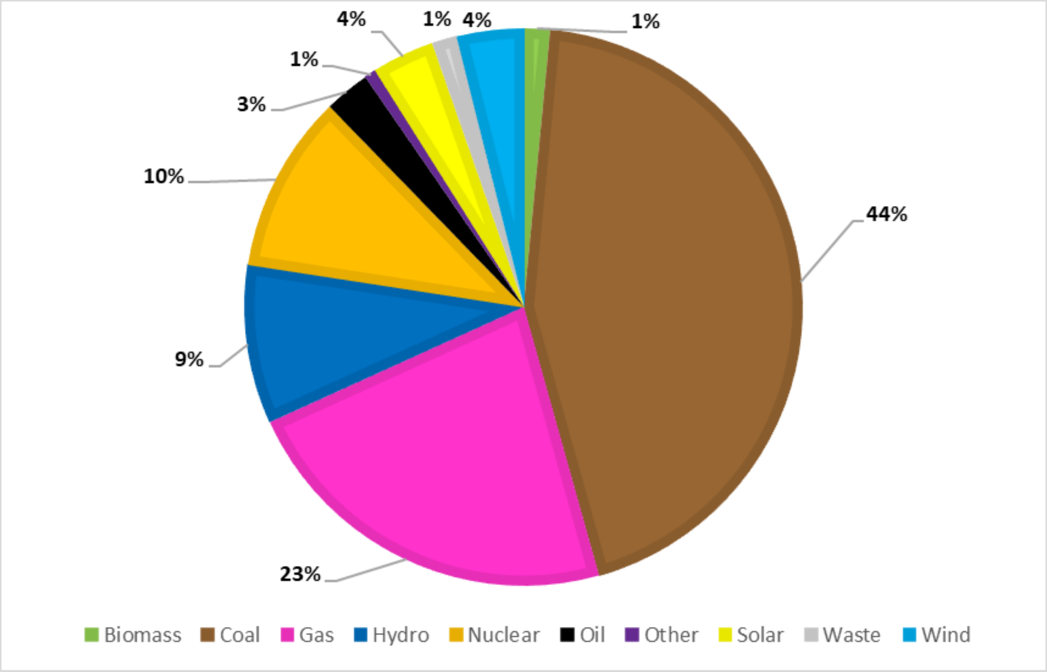

Figure 5.13: Total generation capacities of Germany in 2015

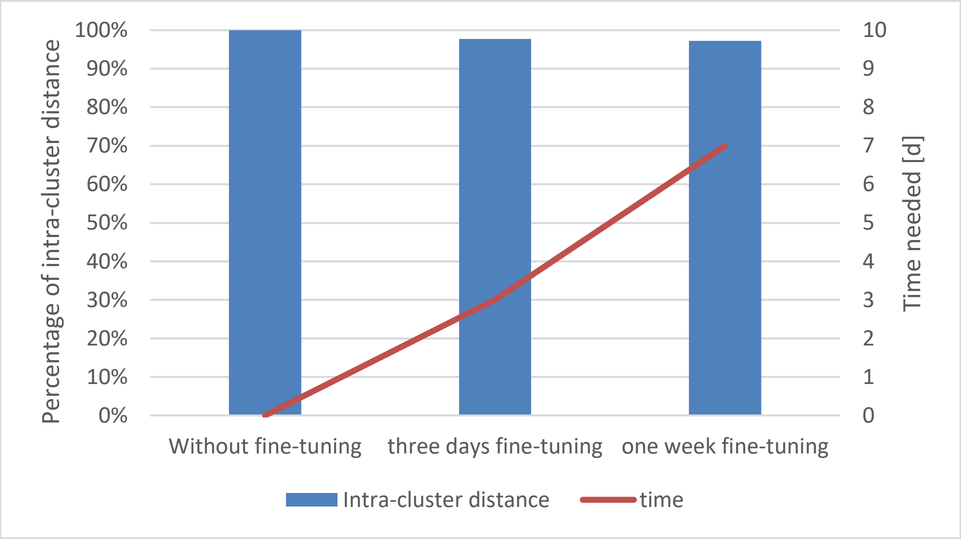

Figure 5.14: Relative intra-cluster distance depending on the time of fine-tuning of the 17 clusters in the use case of Germany

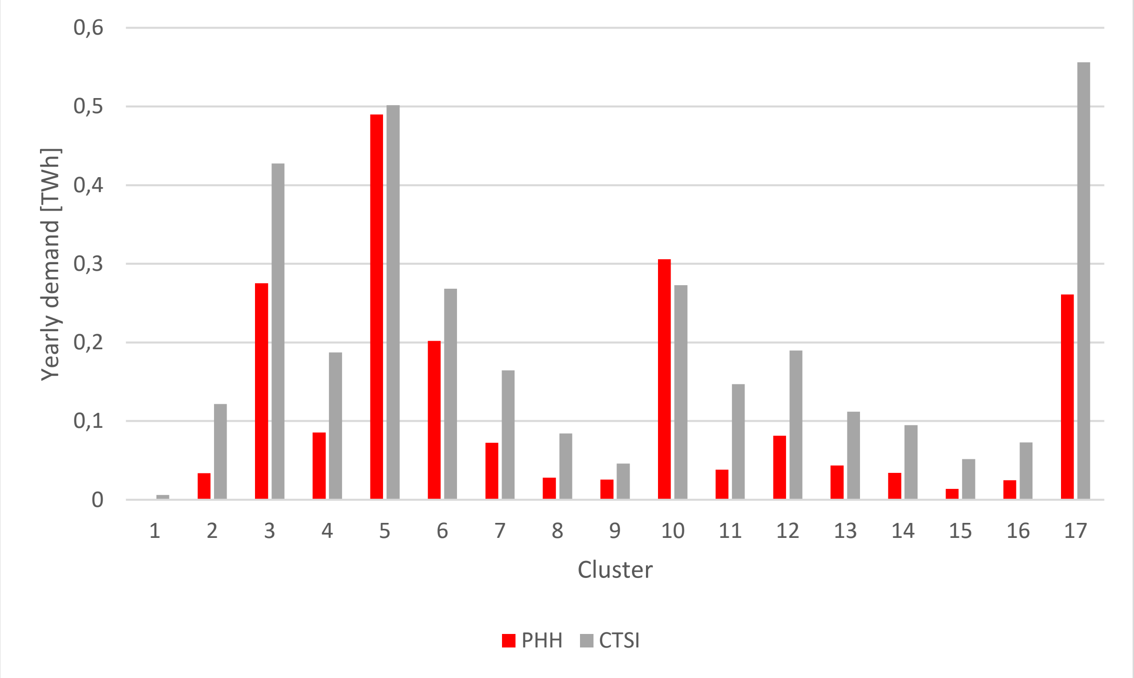

Figure 5.15: Mean yearly demand of the 17 clusters of the use case of Germany

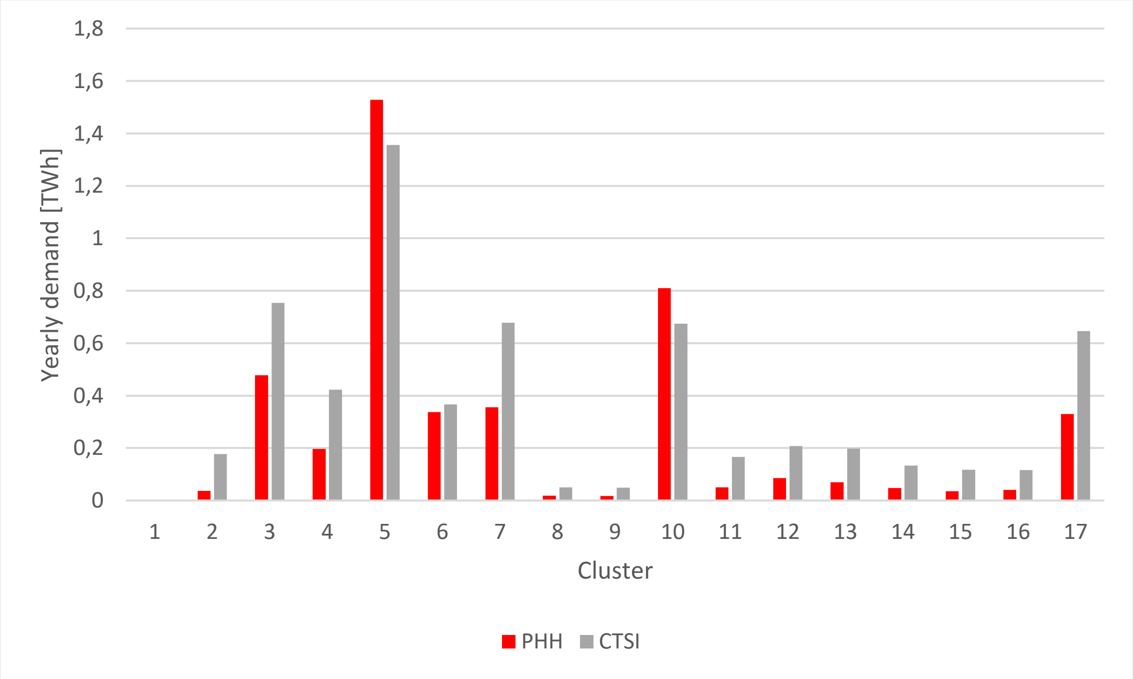

Figure 5.16: Standard deviations of yearly demand of the 17 clusters of the use case of Germany

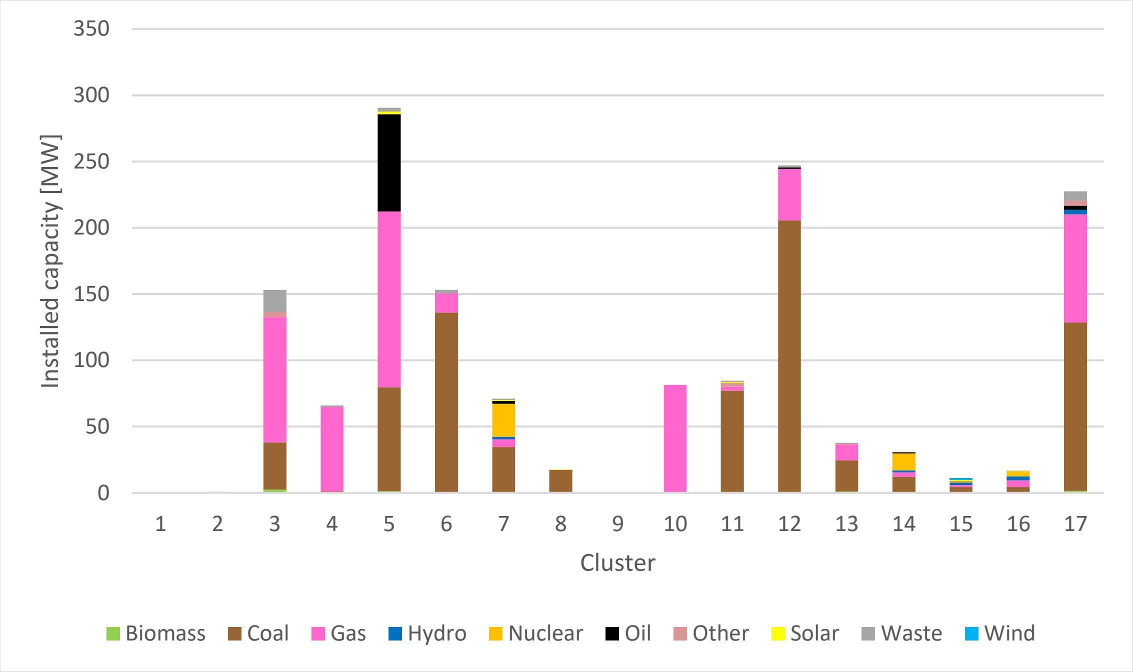

Figure 5.17: Mean installed capacities of the 17 clusters of the use case of Germany

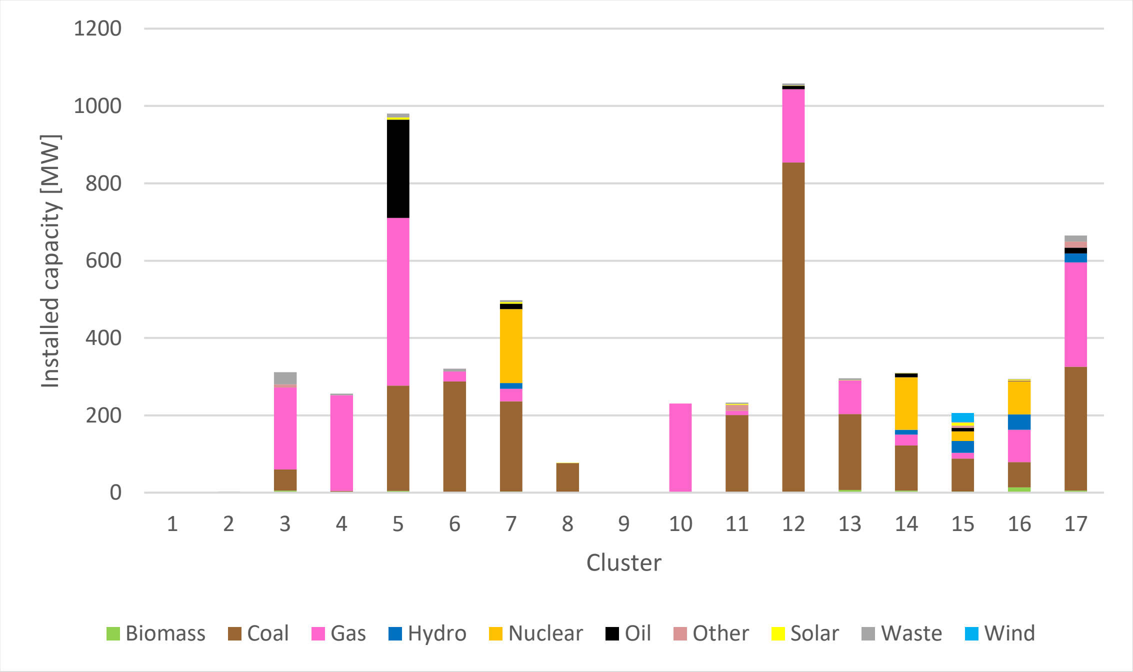

Figure 5.18: Standard deviations of installed capacities of the 17 clusters of the use case of Germany

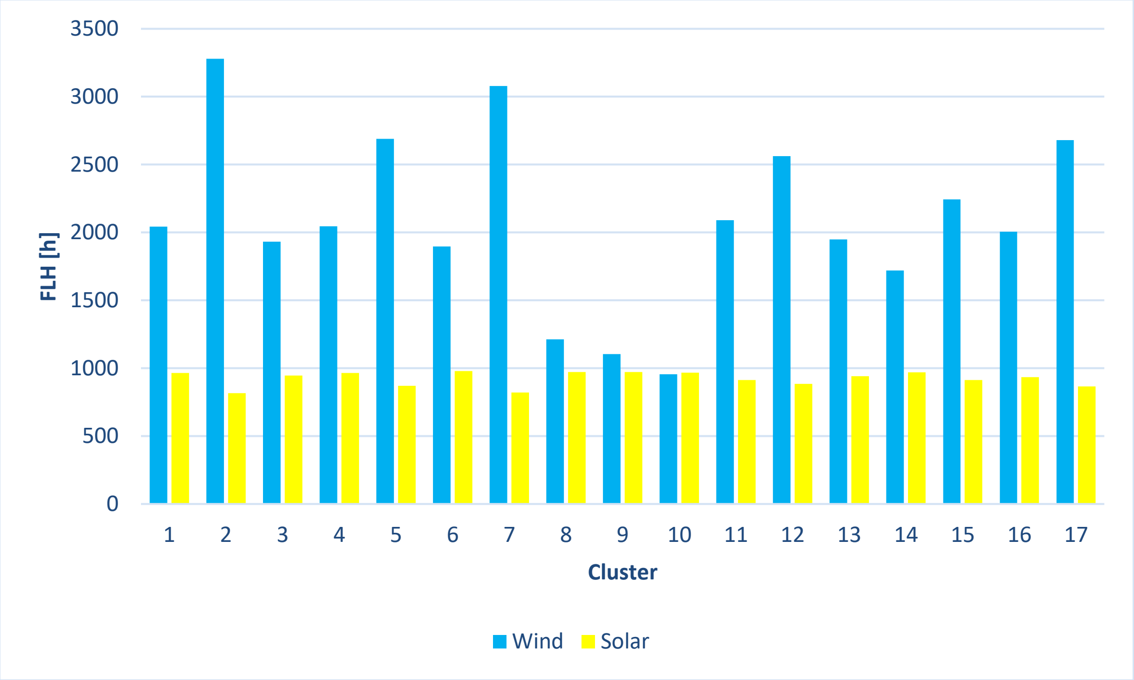

Figure 5.19: Mean FLH of renewables of the 17 clusters of the use case of Germany

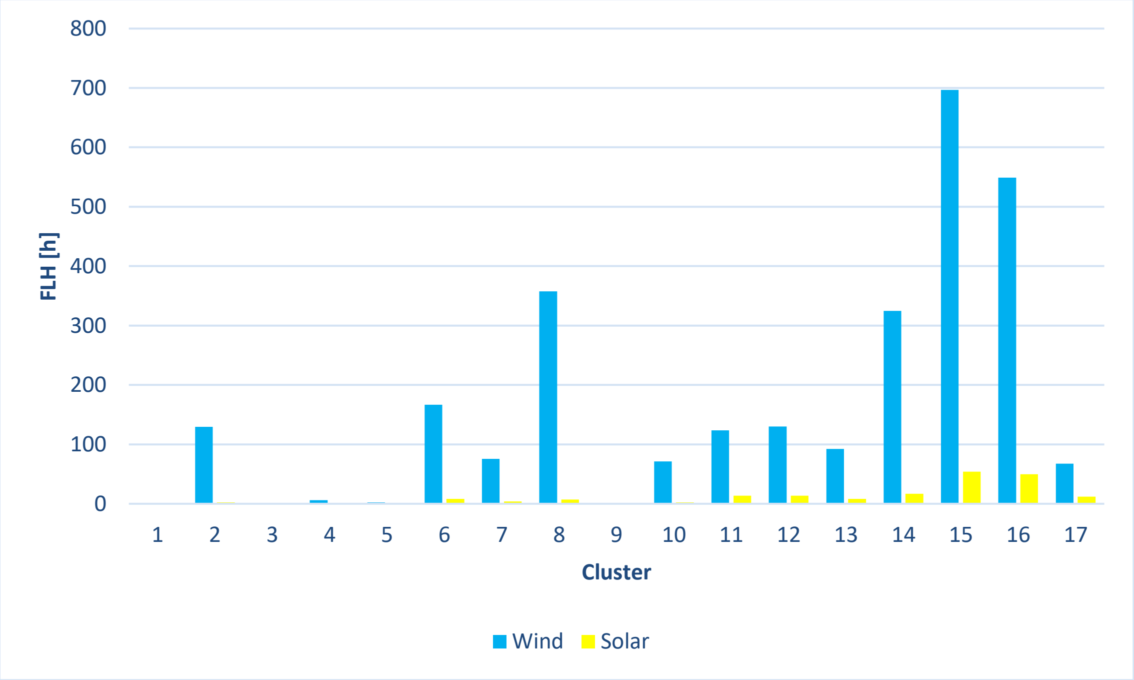

Figure 5.20: Standard deviations of the FLH of renewables of the 17 clusters of the use case of Germany

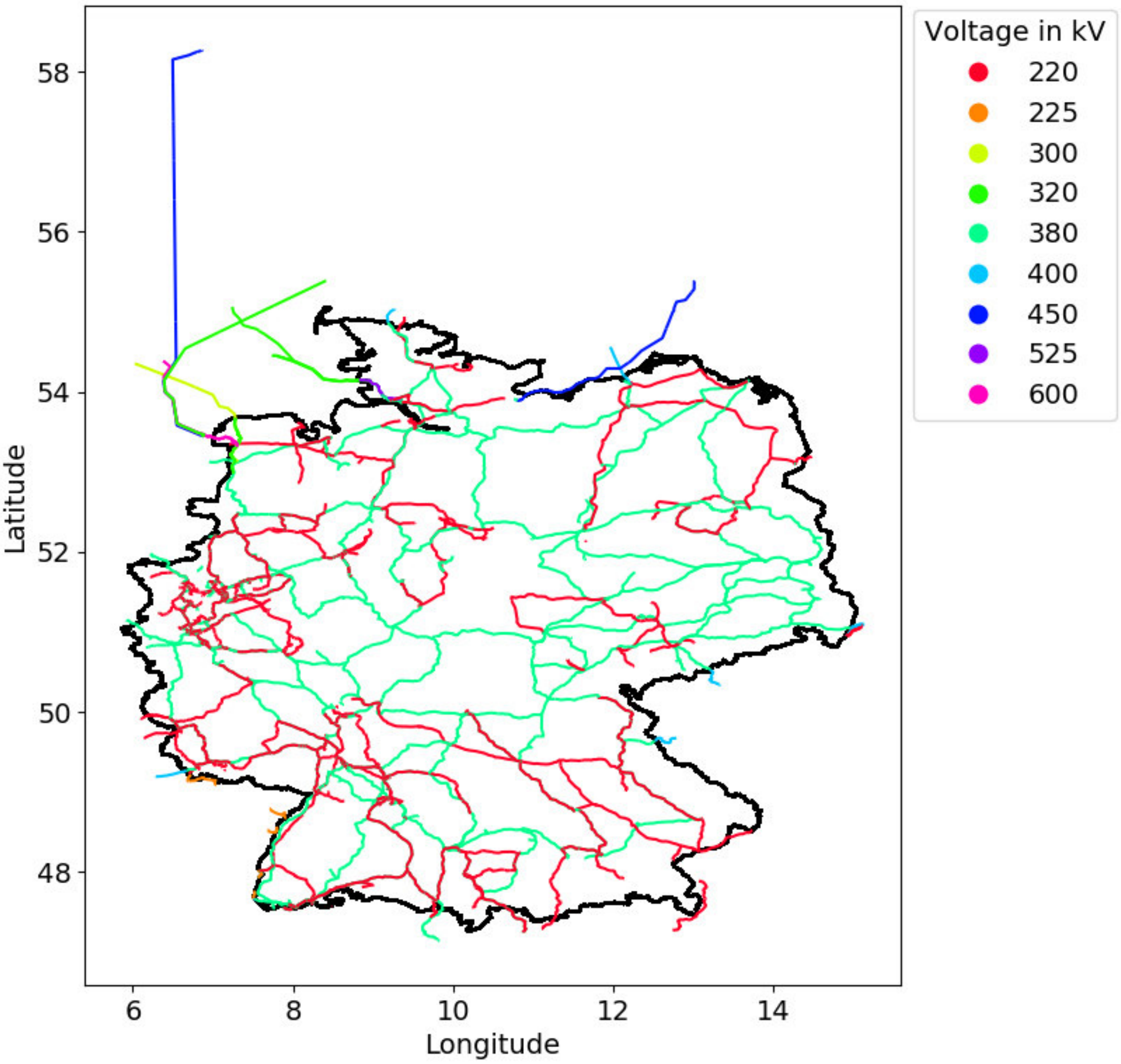

Figure 5.21: Electrical grid of Germany build from OSM data with voltages of at least 200 kV

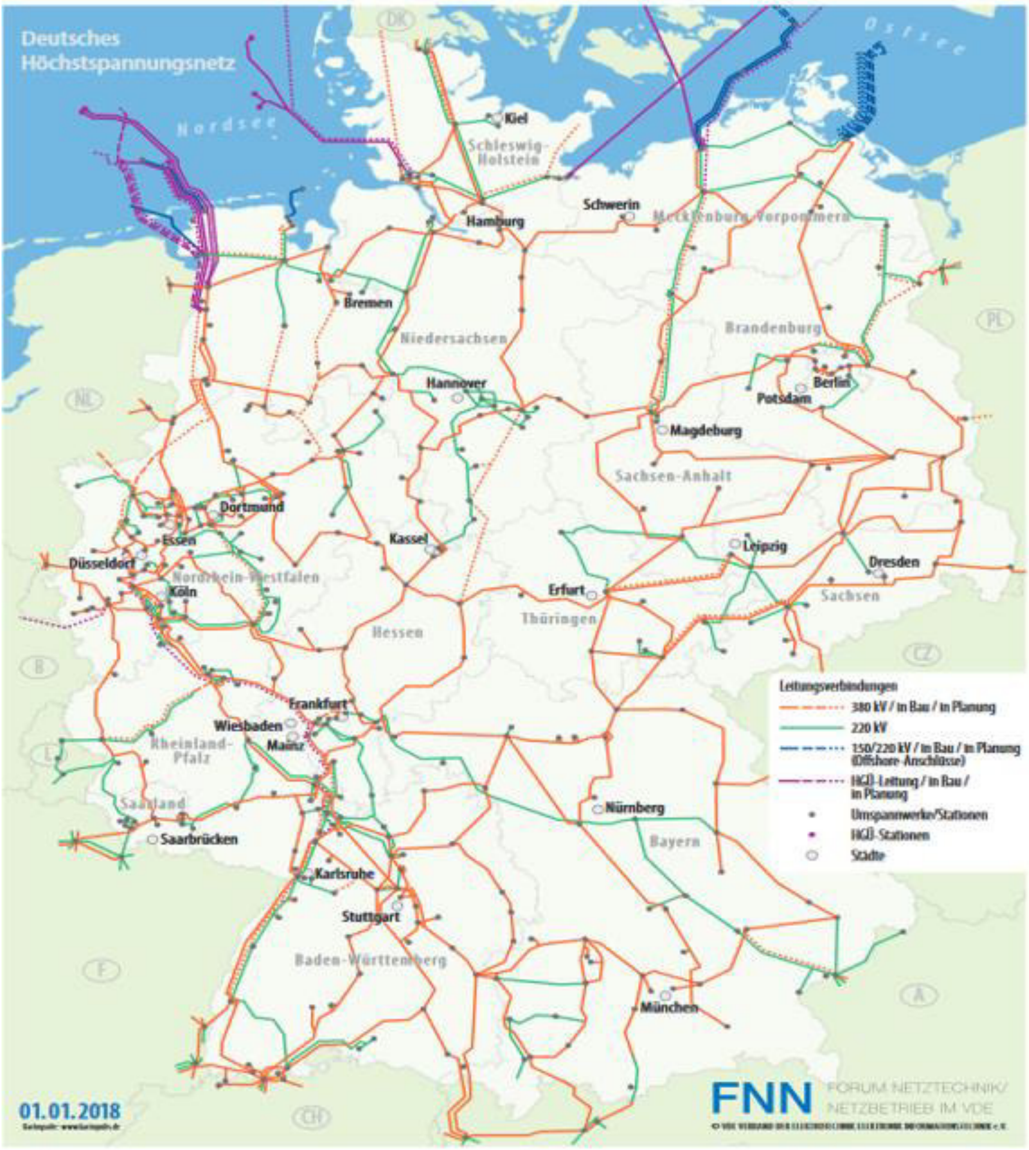

Figure 5.22: Electrical grid in Germany [66]

Appendix: C – weighted normalization

To the other approach, this work refers to as *weighted normalization*, in which each value is calculated as

$$value_{norm,weighted,a,r} = \frac{\frac{value_{a,r} - \min(value_{a,R})}{\max(value_{a,R}) - \min\,(value_{a,R})} \times \frac{\sum_{r\in R} value_{a,r}}{\sum_{a\in A}\sum_{r\in R} value_{a,r}}}{n_A} \tag{5.1}$$

For each region *r* of the region set *R*, each of the corresponding entries *a* of the attribute *A* is normalized to zero and one, multiplied by the fraction of this entry´s sum to the sum of all entries of the attribute, and divided by the number of entries in *A*. For example, the entry of PHH is normalized to zero and one, multiplied by the fraction of its sum and the sum of the entries of PHH and CTSI, and divided by two. The same procedure is applied to the attribute generation. However, the generation is also divided into conventional and renewable generation. For the profiles, however, this procedure would lead to a higher weighting of hours with high generation. Since this is not intended, the profile data are only divided by the number of time steps and the number of different time series considered.

Yearly consumption [MWh]		Installed capacities [MW]				
		Conventional			Renewable	
PHH	GHDI	Coal	Nuclear	Gas	Wind	Solar
50,000	100,000	10	15	5	20	5

Table 5.1: Sum over all regions of load and generation capacities

Yearly consumption [MWh]		Installed capacities [MW]				
		Conventional			Renewable	
PHH	GHDI	Coal	Nuclear	Gas	Wind	Solar
$[0;\frac{50,000}{150,000*2}]$	$[0;\frac{100,000}{150,000*2}]$	$[0;\frac{10}{55*2*3}]$	$[0;\frac{15}{55*2*3}]$	$[0;\frac{5}{55*2*3}]$	$[0;\frac{20}{55*2*2}]$	$[0;\frac{5}{55*2*2}]$

Table 5.2: Illustrative example of the weighted normalization

Appendix: D – Results of the developed methodology for other countries

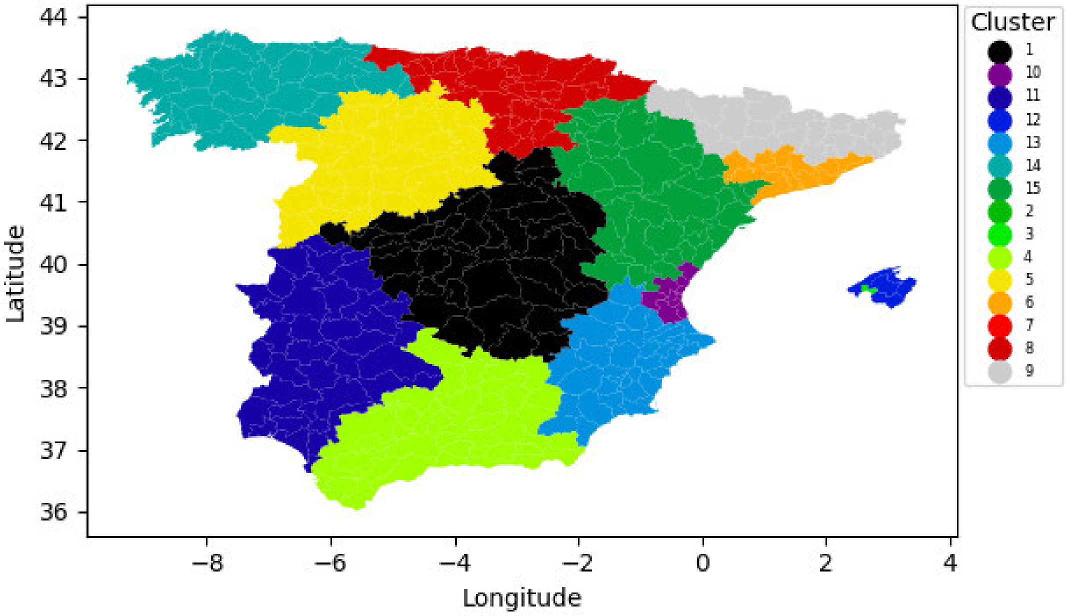

Figure 5.23: 15 clusters identified in Spain

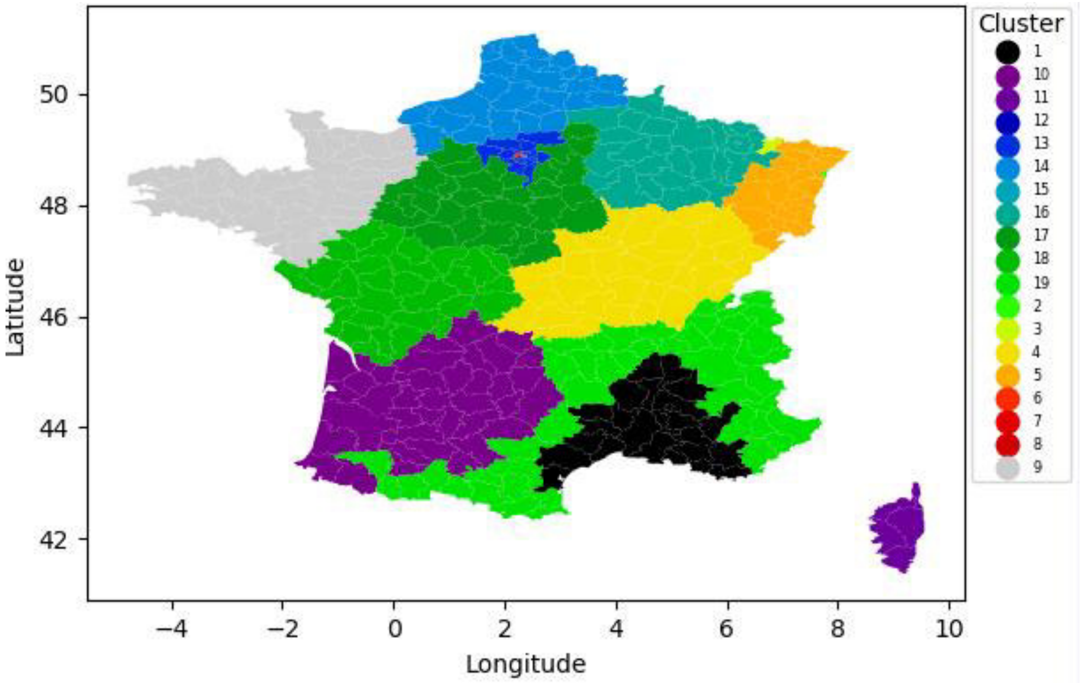

Figure 5.24: 19 clusters identified in France

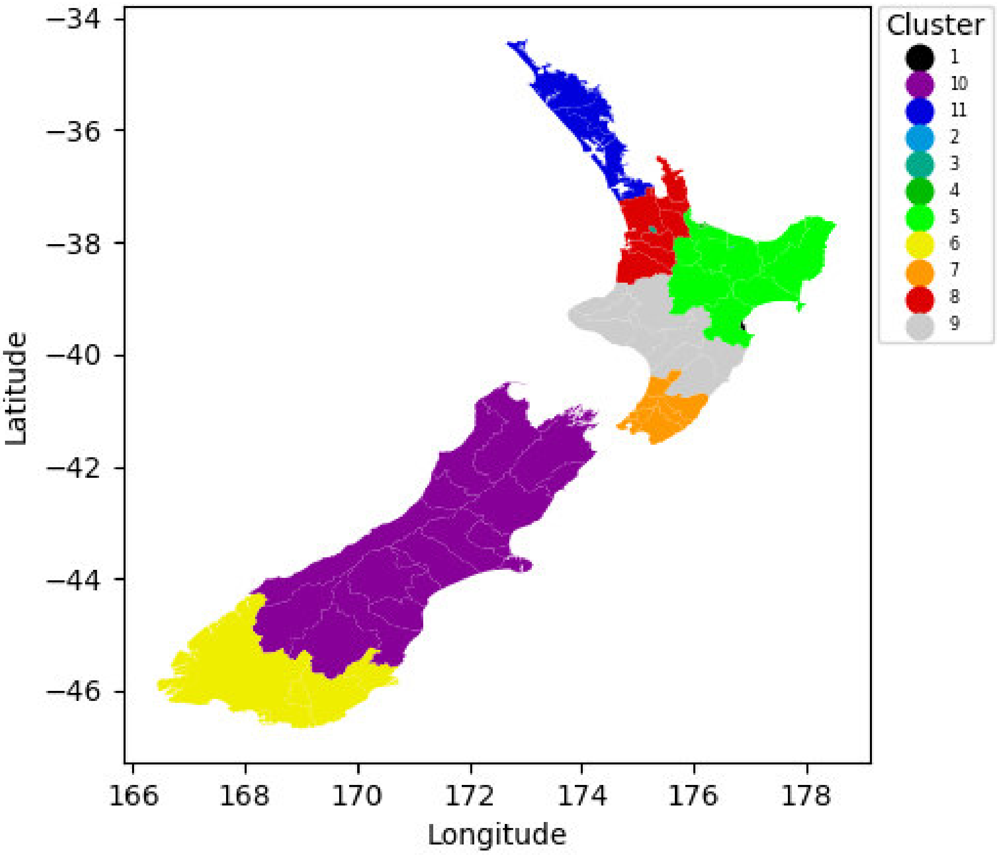

Figure 5.25: 11 clusters identified in New Zealand

Appendix : E – Sustainable development goals

There are two sustainable development goals that are addressed by this thesis. The first and main objective is number 7 – affordable and clean energy – while the second is number 13 – climate action. The sustainable development goal number 7 [https://www.un.org/sustainabledevelopment/energy/] mainly consists of two sub-goal which are, on the one hand, ensuring the global access to affordable electrical energy and, on the other hand, an increase of energy generation by renewable energies like wind, solar, hydro or biomass. While the first sub-goal is not addressed by this thesis, using the developed methodology might save considerable amounts of money and improves the transformation pathway towards decentralized energy systems characterized by high share of renewable power generation. In 2015 the share of renewable energy generation in the final energy consumption has reached 17.5 %. The targets of objective 7 that are addressed are 7.2 – increase substantially the share of renewable energy in the global generation mix – and 7.B – expand infrastructure – which is considered by the developed methodology in the form of the electrical grid, i.e. it is calculated what investments in the electrical grid have to be made when and especially between which regions. Even though the targets of sustainable development goal 7 are defined for 2030, reaching even more ambitious goal in 2050, as calculated in this work, can be seen as equivalent.

As described in this work the costs that could have been accounted for the electrical grid were 16 billion € (in the case of South Africa). Those costs are not considered in the single-region model and, therefore, imply an incomplete view of the total system. Furthermore, the generation (investment and operation) costs of the single-region model and the region-based model were 105 billion € and 94 billion €, respectively. This means that the using the developed methodology costs of approximately 11 billion € could be saved. However, it has to be kept in mind that in the single-region model the wind and PV profiles that have been used where averages of the total are of South Africa. In the real-world new wind turbines, central PV generation units, or rooftop PV units are installed where good conditions, i.e. high full load hours, are. This means that actually the costs of the single-region model might be lower than calculated in this work. However, the neglect of the costs of the grid stays untouched.

The second of the sustainable development goal this work addresses is number 13 [https://www.un.org/sustainabledevelopment/climate-change/]. Global warming is one of the biggest threats humankind is facing in our time. As 2019 was the warmest year ever recorded and the last decade (2010-2019) was the warmest decade, the increasing temperature cannot be denied. This increasing temperature could result into catastrophic events regarding the seas, agriculture, and many more. Therefore, the Paris agreement was signed by a most countries in the world. The goal of the Paris agreement is to agree that the increase of global temperature should be kept considerably lower than 2°C or if possible even 1.5°C. The primarily reason for the global warming is greenhouse gas emissions, especially carbon dioxide (CO_2). To lower the emissions, changes in the infrastructure and generation of energy are required. The main sectors that are involved in this transformation are electricity, heat, mobility, industry, buildings and agriculture. In each of those sectors the emissions have to be reduced by different measures (efficiencies, additional renewable power generation, etc.). The goal of this work was to develop a methodology to increase the significance of an optimization model to calculate the cost-optimal pathway towards a decarbonized energy system. Therefore, this work addresses the sustainable development goal 13 regarding the electrical energy sector. The target goal that is addressed by this work is 13.B – Promote mechanisms for raising capacity for effective climate change-related planning – which relates to a cost-optimal planning of the transformation pathway of a country energy system.

The quantification, however, is difficult since the maximum emission were fixed to be 80% of the emission from today. Therefore, this value will be always given (if there is a feasible solution to the optimization problem). Nevertheless, this value could be adapted in other simulations to compare the effect of other decarbonization goal, e.g. 90% or 100%, to the costs those transformation result in.