INTERPLAN

INTEgrated opeRation PLAnning tool towards the Pan-European Network

Work Package 4

Grid equivalenting

Deliverable D4.1

Method for clustering distributions grid

Grant Agreement No:	**773708**
Funding Instrument:	**Research and Innovation Action (RIA)**
Funded under:	**H2020 LCE-05-2017: Tools and technologies for coordination and integration of the European energy system**
Starting date of project:	**01.11.2017**
Project Duration:	**36 months**

Contractual delivery date:	**31/05/2019**
Actual delivery date:	**31/05/2019**
Lead beneficiary:	2 **AIT Austrian Institute of Technology GmbH**
Deliverable Type:	**Report (R)**
Dissemination level:	**Public (PU)**
Revision / Status:	**RELEASED**

Project co-funded by the European Commission within the H2020 Programme (2014-2020)

Document Information

Document Version:	6
Revision / Status:	RELEASED
All Authors/Partners	Alena Palkhouskaya / AIT Adolfo Anta / AIT Mihai Calin / AIT Ata Khavari / DERlab Marialaura Di Somma / ENEA Roberto Ciavarella / ENEA Giorgio Graditi / ENEA Maria Valenti / ENEA Helfried Brunner / AIT Jan Ringelstein / IEE Malte Hof / IEE Melios Hadjikypris / FOSS Sawsan Henein / AIT Sohail Khan /AIT Venizelos Efthymiou / FOSS
Distribution List	INTERPLAN consortium
Keywords:	Clustering, Network models, TSO level, DSO level, Distribution Grids, Transmission Grids, Planning, Modelling.

Document History

Revision	Content / Changes	Resp. Partner	Date
1	Draft table of contents was prepared	AIT	15.12.2018
2	Most of the contents were included	AIT	20.04.2019
3	Document was finalised for internal review	AIT	22.05.2019
4	Editorial changes	AIT	24.05.2019
5	Document was reviewed	AIT	29.05.2019
6	Document was finalised for release	AIT/ENEA	30.05.2019

Document Approval

Final Approval	Name	Resp. Partner	Date
[Review Task Level]	Ata Khavari	DERlab	27.05.2019
[Review WP Level]	Anna Wakszyńska	IEN	27.05.2019
[Review Steering Com. Level]	Helfried Brunner / Giorgio Graditi	AIT/ENEA	30.05.2019

Disclaimer

This document contains material, which is copyrighted by certain INTERPLAN consortium parties and may not be reproduced or copied without permission. The information contained in this document is the proprietary confidential information of certain INTERPLAN consortium parties and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information in this document may require a licence from the proprietor of that information.

Neither the INTERPLAN consortium as a whole, nor any single party within the INTERPLAN consortium warrant that the information contained in this document is capable of use, nor that the use of such information is free from risk. Neither the INTERPLAN consortium as a whole, nor any single party within the INTERPLAN consortium accepts any liability for loss or damage suffered by any person using the information.

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

Copyright Notice

Table of contents

Abbreviations

DER	Distributed Energy Resource
DG	Distributed Generation
EHV	Extra High Voltage
ENTSO-E	European Network of Transmission System Operators for Electricity
GPD	Global Power Plant Database
HV	High Voltage
IQR	Interquartile Range
KPI	Key Performance Indicator
PV	Photo Voltaic
LV	Low Voltage
MV	Medium Voltage
NDA	Non-Disclosure Agreement
PC	Project Coordinator
PCA	Principal Component Analysis
RMS	Root Mean Square
S	Silhouette Coefficient
SSE	Sum of Squared Errors
WP	Work Package

Executive Summary

The energy transition, from a central fossil fuel-based energy supply to more dispersed sustainable sources is changing the loading of distribution grids and low voltage (LV) grids in particular. The electrification of transportation and heating loads increases the electricity demand at household level, while rooftop PV systems generate electricity at residential level and can lead to bidirectional power flows. LV-grids have been built over the past decades and were not designed to meet these load changes and to host electricity generation. Therefore, LV network planning is becoming more and more important to deal with the reinforcing of the LV-grid in the most cost-effective way [1].

The report focuses on describing an approach for clustering electricity grids covering the different use cases for semi-dynamic and integrated grid operation planning (covering all network levels) and reflects the work done related to identification and characterization of a clustering method in the project INTERPLAN – INTEgrated opeRation PLAnning tool towards the pan-European Network.

The characterisation and clustering of electrical networks has previously been studied. For the evaluation of reliability and susceptibility to threats, clustering based on graph theory is already being used especially for the transmission network [2], [3], [4] and [5]. In [5] a small number of networks are defined, based on the length of the feeders, the number of connected customers and the number of branches. Though some analysis can be performed on these representative networks, they are not classified based on enough detail to be usable for network planning. A more extensive approach is required to be able to create generic grids with a strong relation to the existing low voltage grids.

The report introduces an approach for the identification of a clustering method by analysing the existing clustering methods and proposes, in chapter 2, a specific clustering method to be used for the purpose of the INTERPLAN project. Chapter 3 describes transmission and distribution grid models developed or considered for usage in INTERPLAN project and which will be used to evaluate the clustering method. Chapter 4 presents the different clustering parameters which were analysed in order to be used by the clustering algorithm and the ones selected. Initial clustering parameters were selected based on the impact they might have on differentiating feeder types and on distributed generation (DG) hosting capacity. The initial variables varied among the different available networks as needed to account for differences in availability of data from each partner. The presented parameters could provide a very important input to a clustering process according to the use cases of INTERPLAN project. However, not all of them are available within the given databases. Considering low voltage grids, a limited set of variables will be implemented into the clustering algorithm. In Chapter 5 a clustering algorithm based on several grids characteristics and parameters derived from the analysed parameters is presented. The clustering is applied to the grid data available to the project partners and an initial clustering is performed to show a possible application of the clustering method.

Clustering is an effective machine learning technique being used for the data exploration based on grouping of observations (feeders or networks) into clusters according to their specific features. The data partitioning is proceeding without prior information about the number of clusters. In spite of this, the generated clusters should demonstrate a very high diversity from each other, while the data points within a cluster should be quite similar.

According to the general concept of clustering procedure which is presented in the report, the process is starting with the importing of network data and their processing via DIgSILENT PowerFactory environment scripting with Python. This allows to perform the grid simulations and provides a very high computational functionality and flexibility for data exchange. INTERPLAN operated with a database of 2000 low voltage networks, which includes a large number of LV feeders (about 9500).

1 Introduction

With the gradual increase of load level and especially the fast development of various renewable generation resources, the nodal load and generation capacity may grow significantly within a relative short time. Traditional transmission networks, mainly planned for large capacity fossil/hydro power plants, are facing more and more challenge because of the dilemma between their slow response and the fast integration of various renewable generation [6].

Under this circumstance, reference networks [7], [8] can be introduced to guide transmission network planning in timely manner. The concept of reference network is initially proposed from an economic perspective. It can be seen as a benchmark or an ideal model, which can be compared with a real one. For power system planning and operation purpose, a reference network is usually topologically identical to the existing network. The generators and loads in a reference network are the same as those in the existing transmission system. However, the reference network possesses the optimal transmission line capacity. The optimal capacity of each line is determined by optimizing the annual or monthly or even daily operating cost and investment cost of transmission line with respect to the nodal power balance and line flow constraints. If necessary, reliability aspect can be involved by considering both nominal state and contingency scenarios. Reference network is a useful tool which can inspect and guide the transmission network planning.

The concept of reference network can be used not only in transmission network planning, but also in distribution network planning [8]– [9] [10] to evaluate the influence of investment strategies on users and system at different price levels. In [11], a reference network is used for the planning of high-, medium-, and low-voltage networks considering the street map. If the real time inputs are available, reference networks can also be used for quick analysis of possible contingencies [12]. When the reliability is considered, the reference network model is usually difficult to be solved due to the fact that significant numbers of scenarios are considered. Benders' decomposition technique can be applied to address this problem [13].

The energy transition, from a centrally fossil fuel-based energy supply to more dispersed sustainable sources is changing the loading of distribution grids, at low voltage level (LV) in particular. The electrification of transportation and heating loads increases the electricity demand at household level, while rooftop PV systems generate electricity at residential level and can lead to bidirectional power flows. The LV-grids have been built over the past decades and were not designed to meet these load changes and to host electricity generation. Therefore, LV network planning is becoming more and more important to deal with the reinforcing of the LV-grid in the most cost-effective way [9].

One of the main difficulties with the planning and assessment of the LV-grid is the large number of LV- feeders and MV/LV substations. The analysis of each substation and/or each feeder individually becomes a computationally intensive task. The application of mitigation measures, for a LV-feeder which is found inadequate to deal with the future loading, needs to be standardised in order to ensure a more cost-effective solution. The creation of a limited amount of generic LV- feeders can increase the effectiveness of the LV network planning. As the generic LV-feeders must have a close resemblance to the actual feeders in the field, a clustering approach on the data of the whole LV-grid is the most suitable method for creating these generic feeders.

The characterisation and clustering of electrical networks has previously been studied. For the evaluation of reliability and susceptibility to threats, clustering based on graph theory is already being used especially for the transmission network [2], [3], [4] and [5]. In [5] a small number of networks are defined, based on the length of the feeder, the number of connected customers and the number of branches. Though some analysis can be performed on these representative networks, they are not classified based on enough detail to be usable for network planning. A more extensive approach is required to create generic grids with a strong relation to the existing low voltage grids.

1.1 Purpose of the Document

The report focuses on describing an approach for clustering distribution grids covering the different INTERPLAN use cases for semi-dynamic and integrated grid planning (covering all network levels) and reflects the work done in the project task on identification and characterization of a clustering method.

1.2 Structure of the Document

The report introduces the approach for identification of a clustering method for electricity grids by analysing the existing clustering methods and proposes in chapter 2 a specific clustering method to be used for the purpose of INTERPLAN project. Chapter 3 describes transmission and distribution grid models that were developed or considered for usage in INTERPLAN project and which will be used to evaluate the clustering method. Chapter 4 presents the different clustering parameters which were analysed in order to be used by the clustering algorithm and the ones selected. In Chapter 5 a clustering algorithm based on several grids characteristics and parameters derived from these characteristics is presented. The clustering is applied to the grid data available to the project partners and an initial clustering is performed to show a possible application of the clustering method.

2 Approach

This section proposes a general-purpose methodology for grid clustering enabling classification of different distribution feeders for all use cases considered in the INTERPLAN project. As such, the project considered a wide range of grid parameters as possible regressors, while the clustering criteria are given by the different KPIs associated to each sub-case. From a machine learning perspective, non-supervised clustering algorithms will be explored, where the number of clusters is not known in advance.

2.1 General Considerations

First, it needs to be differentiated between classification and clustering techniques (see Figure 1):

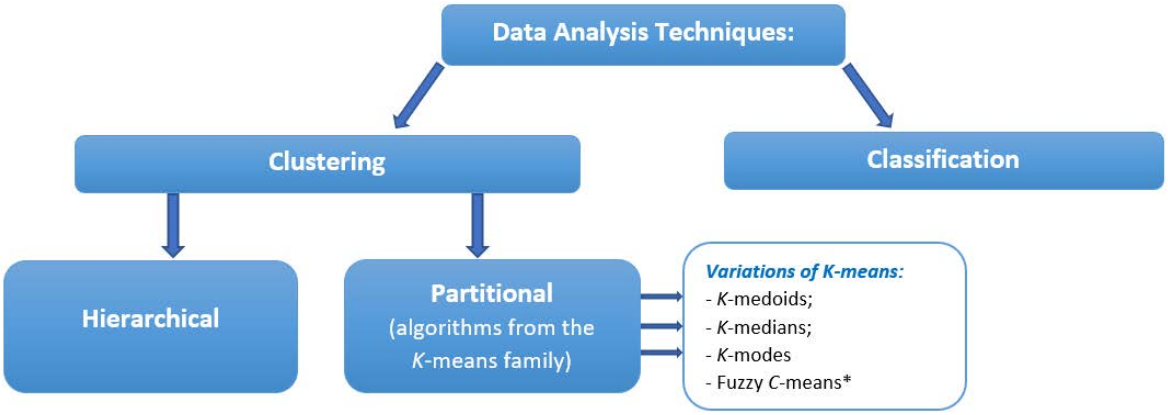


Figure 1 Data Analysis Techniques

Classification strategies take as starting point a set of predefined classes and, given a new object (feeder or network), it assigns it to one of the existing class. From a machine learning perspective, this is considered supervised learning, as there is already prior information and training data to establish these classes.

On the other hand, clustering implies grouping of observations into clusters, based on the similarity of their specific parameters, to define an interrelation between these objects. No prior information, number of clusters or number of classes is known in advance. Thus, it can be seen as unsupervised learning.

According to the goals of INTERPLAN project the clustering analysis will be further performed to generate grid equivalents for different types of networks and different objectives. Existing approaches mainly focus on a limited amount of objective functions, for instance, voltage violations or hosting capacities.

The most applicable algorithms for clustering of electricity grids at the feeder or network level are hierarchical and partitional clustering. In spite of its flexibility and the fact that it doesn't require setting the number of clusters a priory, the hierarchical clustering is not recommended due to its high computational complexity for a large data set. Therefore, the partitional approach will be implemented.

Partitional methods (belonging to the k-means family) are the most efficient and at the same time relatively simple clustering algorithms, which require however to pre-define the number of clusters with their initial centers and to adjust them iteratively. Along with the basic k-means algorithm its different variations (such as k-medoids, k-medians, k-modes, fuzzy c-means clustering etc.) should be considered. For instance, k-medians clustering is less sensitive to outliers compared to k-means. Likewise, to avoid overlapping effects on the cluster boundary, the membership value of each observation to cluster centers can be assigned using the fuzzy c-means clustering algorithm. Other

properties of each clustering methodology will be explored as well; it is then expected that not a single clustering algorithm provides an adequate level of accuracy for all use cases in INTERPLAN.

2.2 Proposed Method

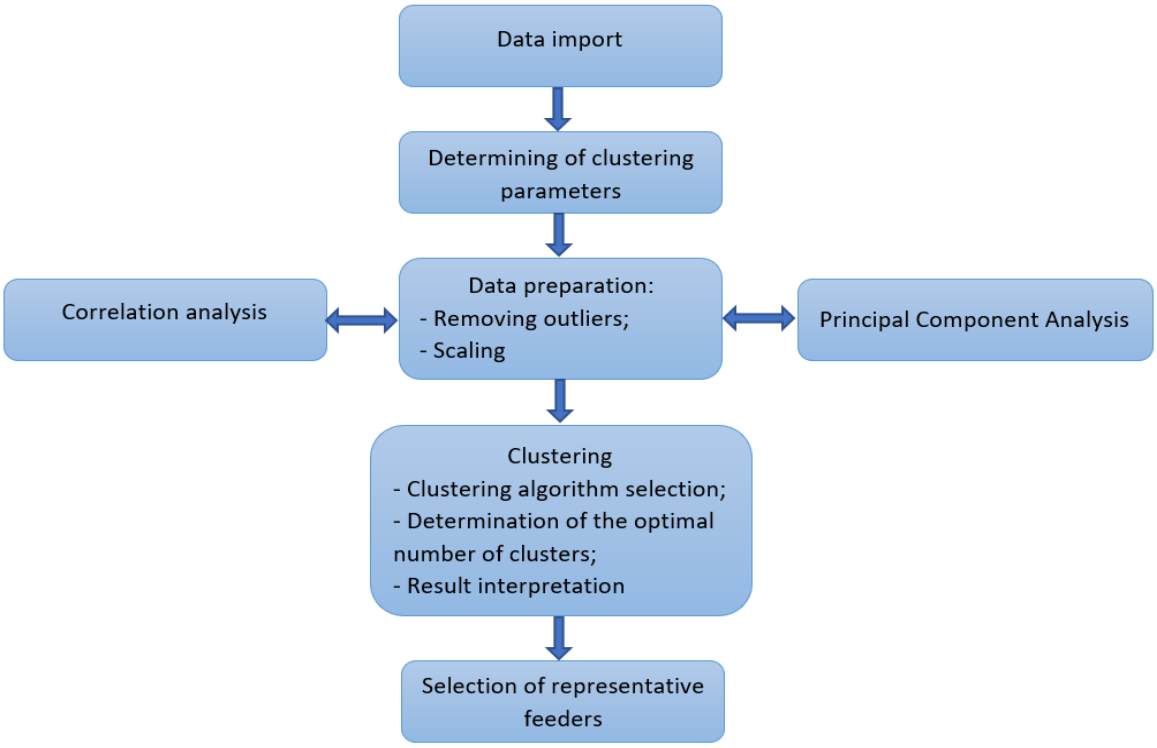


Figure 2 General concept of the clustering approach

The chosen parameters have to be independent from each other to provide more correct cluster partitioning. These parameter dependencies can be explored by using the following data analysis methods:

- Correlation analysis;
- Principal Component Analysis (PCA).

The suggested techniques require the data scaling and normalizing, what can be performed by subtracting the average value from each variable and dividing by the standard deviation. Then the Principal Component Analysis can be implemented. The main purposes of PCA are an extraction of unique feeder information from the available database and a minimization of data redundancy. These goals are achieved by selection of principal components, which are orthogonal to each other, in the direction of the highest data variance. Principal components indicate the implicit variables explained the variation in a data set. In order to investigate the data complexity, a correlation analysis will be also used based on the determination of specific correlation coefficients (Pearson correlation Coefficient, Spearman's rank correlation coefficient, Kendall correlation coefficient, among others). After that, the clustering analysis will be performed, using the algorithms mentioned in Section 2.1.

As a last step of the clustering approach, a set of representative feeders will be identified. Given a new feeder, analyzing and identifying the characteristics of such feeder is tantamount to a classification step that is, assigning it to one of the already identified classes of feeders. A descriptive value for the KPIs associated to these particular feeders can then be directly obtained from the representative feeders for further analysis.

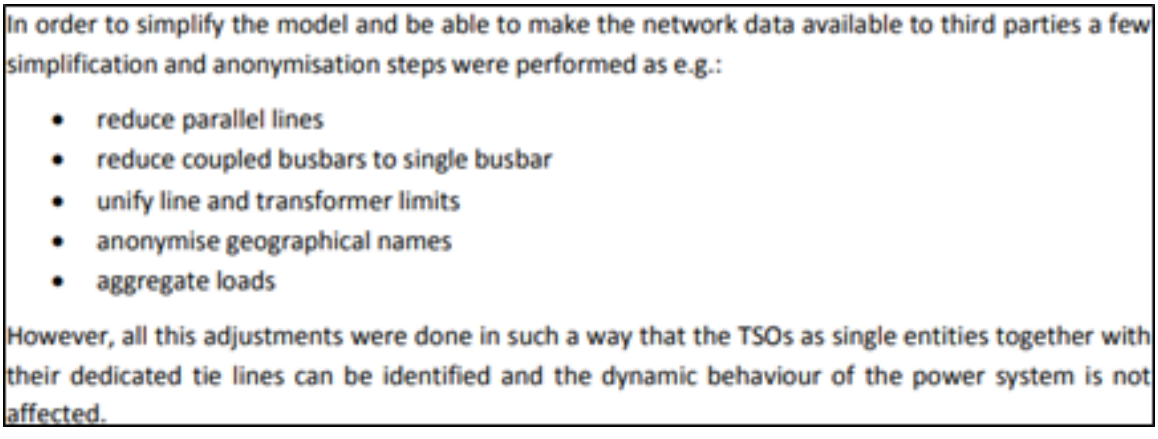
3 Grid Models

The following section describes transmission and distribution grid models developed or considered for usage in the INTERPLAN project. The chapter does not describe all grid models that were used for clustering and development of grid equivalents carried out in the project task of-identification and characterization of a clustering method and the task dedicated to developing a detailed approach for generating grid equivalents for different use cases. The requirements for the grid models to be used in the project were derived from the use case definition done in the beginning of the project.

3.1 Transmission Grid models

3.1.1 ENTSO-E Initial Dynamic Model of Continental Europe

Traditionally academic power engineering research uses commonly documented test grids which are often based on existing grids such as the IEE 39-bus system (New England, USA), or more recently, the Nordic 32 test system (Northern Europe power system) [14]. Grid models used by electric utilities for operation and planning are often considered a company secret for various reasons (e.g. avoidance of asset value estimation, electricity market disadvantages or critical infrastructure protection). Within companies that operate transmission systems, usually dedicated departments provide and update grid models for internal use (which may further differ for operation and planning purposes). Successful approaches have been done to obtain some grid data (e.g. overhead line data, substation position) from published data and sources like Open Street Map (OSM) or Google Earth [15], [16]. Simulation grid models are always oriented on their intended simulation purpose and “even approximate mathematical models of one aspect of the power grid are heterogeneous, messy and difficult to analysis” [17].
The ENTSO-E grid model (ENTSO-E Initial Dynamic Model of Continental Europe) contains load flow data originating from electric utilities. It is expanded by ENTSO-E with standard models for generators and appropriate controller models in order to represent the dynamics of the interconnected continental Europe power system. Concept, scope, data adjustment, dynamic data expansions, and validation (grid frequency characteristic) are given in the ENTSO-E Instruction Manual [18]. A distinguishing property of the model is that it is available in PowerFactory, Netomac, Eurostag and PSS/E format. According to Figure 3, the ENTSO-E grid model does not contain any graphical data, and names of objects - usually relating to geographic names – which are replaced by numbers. The only localization information that is left for the grid model assets is the European country which they’re situated is. In PowerFactory, the assets are grouped into one folder per country. The lack of a fully geographically localized graphic model is a huge drawback for working with the grid model in PowerFactory since the program’s philosophy is built around a visual grid in combination with grid data (e.g. adding a graphical line also changes the impedance matrix). In consequence a (partial) rebuild of the ENTSO-E grid model is necessary for efficient working with PowerFactory.


In order to simplify the model and be able to make the network data available to third parties a few simplification and anonymisation steps were performed as e.g.:

- reduce parallel lines
- reduce coupled busbars to single busbar
- unify line and transformer limits
- anonymise geographical names
- aggregate loads

However, all this adjustments were done in such a way that the TSOs as single entities together with their dedicated tie lines can be identified and the dynamic behaviour of the power system is not affected.

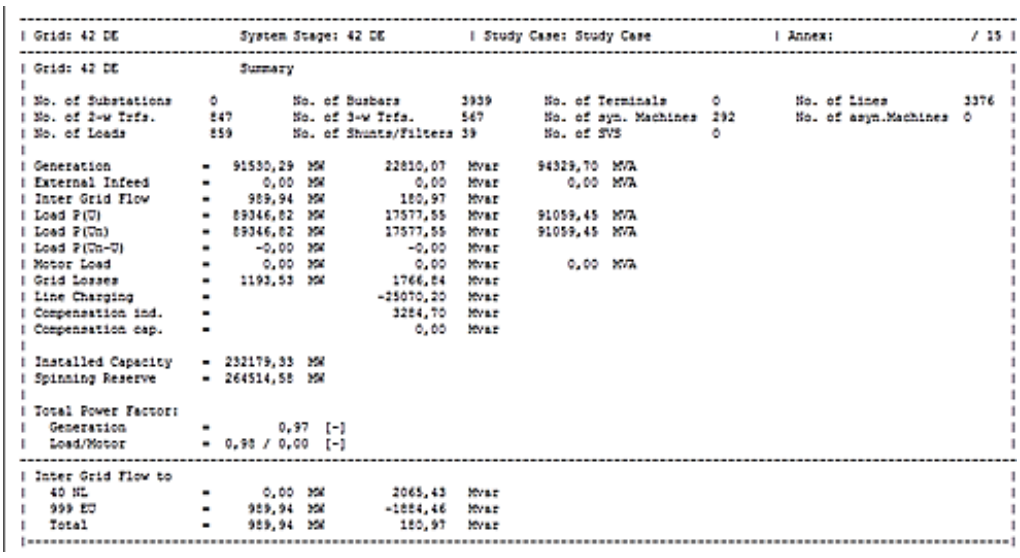
Figure 3 Data adjustments of the ENTSO-E grid model [18].

Figure 4 and Figure 5 show other properties of the ENTSO-E grid model. Further characteristics are:

- Initial load flow converges but gives warnings.
- Several bus bars are isolated from the rest of the system.
- RMS simulation related parameters of power plants seem to be equal for a country.
- Line lengths seem to be altered for countries.

Code	Country	Lines	Buses	Loads	Generators
AL	Albania	193	339	110	77
AT	Austria	123	104	40	31
BA	Bosnia & Herz.	312	294	164	37
BE	Belgium	178	140	36	52
BG	Bulgaria	787	798	419	77
CH	Switzerland	244	193	82	88
CZ	Check Rep	106	288	76	113
DE	Germany	3378	3939	859	898
DK	Denmark	250	397	66	71
ES	Spain	1338	1385	646	496
FR	France	2599	2665	991	1564
GR	Greece	1133	1312	367	128
HR	Croatia	334	329	171	73
HU	Hungary	94	120	37	28
IT	Italy	817	1264	341	458
LU	Luxembourg	41	38	11	12
ME	Montenegro	80	94	35	18
MK	Macedonia	146	163	85	25
NL	Netherlands	905	1031	244	178
PL	Poland	988	648	199	140
PT	Portugal	365	506	87	158
RO	Romania	1194	1171	654	185
RS	Serbia	619	553	303	62
SI	Slovenia	111	230	65	73
SK	Slovakia	52	48	18	82
TR	Turkey	1943	4888	1245	1022
EU*	Europe*	9	316	26	1
	Total	18339	23253	7377	6147

Figure 4 Number of elements in the dynamic model [19].



```
| Grid: 42 DE              System Stage: 42 DE          | Study Case: Study Case              | Annex:                  / 15 |

| Grid: 42 DE              Summary                                                                                                      |
|                                                                                                                                       |
| No. of Substations    0          No. of Busbars        3939      No. of Terminals     0          No. of Lines          3376  |
| No. of 2-w Trfs.      847        No. of 3-w Trfs.      567       No. of syn. Machines 292        No. of asyn.Machines  0     |
| No. of Loads          859        No. of Shunts/Filters 39        No. of SVS           0                                        |
|                                                                                                                                       |
| Generation          =   91530,29  MW        22810,07  Mvar     94329,70  MVA                                                       |
| External Infeed     =       0,00  MW            0,00  Mvar         0,00  MVA                                                       |
| Inter Grid Flow     =     989,94  MW          180,97  Mvar                                                                          |
| Load P(U)           =   89346,82  MW        17577,55  Mvar     91059,45  MVA                                                       |
| Load P(Un)          =   89346,82  MW        17577,55  Mvar     91059,45  MVA                                                       |
| Load P(Un-U)        =      -0,00  MW           -0,00  Mvar                                                                          |
| Motor Load          =       0,00  MW            0,00  Mvar         0,00  MVA                                                       |
| Grid Losses         =    1193,53  MW         1766,84  Mvar                                                                          |
| Line Charging       =                      -25070,20  Mvar                                                                          |
| Compensation ind.   =                        3284,70  Mvar                                                                          |
| Compensation cap.   =                           0,00  Mvar                                                                          |
|                                                                                                                                       |
| Installed Capacity  =  232179,33  MW                                                                                                  |
| Spinning Reserve    =  264514,58  MW                                                                                                  |
|                                                                                                                                       |
| Total Power Factor:                                                                                                                   |
|   Generation        =        0,97  [-]                                                                                                |
|   Load/Motor        =  0,98 / 0,00  [-]                                                                                              |

| Inter Grid Flow to                                                                                                                    |
|   40 NL             =       0,00  MW         2065,43  Mvar                                                                          |
|   999 EU            =     989,94  MW        -1884,46  Mvar                                                                          |
|   Total             =     989,94  MW          180,97  Mvar                                                                          |
```

Figure 5 Grid summary for the Germany Grid (DE) of the ENTSO-E grid model [19].

In order to make the ENTSO-E model better usable for the project, a “rebuilding process” was started which generally aimed at localizing the grid model nodes by assigning PowerFactory terminals to substation names without changing the reference model itself. To this end, the model is quite detailed: each substation is typically modelled with several in-substation busbars. This property is a

unique feature of the ENTSO-E model when compared to e.g. the PyPSA model (cp. section 3.1.3).

Initially the rebuilding process was intended to start at generators with large nominal power. Usually nuclear power plants use very large synchronous generators (> 1GVA) and, especially for Germany, there are not many of these kinds of plants. The idea was to identify the generators of the German nuclear plants in the ENTSO-E grid and sub sequentially start the further rebuilding process by identifying substations around these generators. However due to a large number of identified and unrealistic large generators (>2 GVA) in the ENTSO-E grid this approach failed.

Hence, an alternative approach was taken by identification the cross-border lines first and then starting the further rebuilding process from these lines on towards the centre of a country, in our case Germany. This approach proved to be more successful. It was recognized cross border connections are organized in the ENTSO-E grid model via the grid folder “EU”. This grid is kind of a linkage element. As an example, a line from Germany to Czech Republic first goes from the “DE” grid into the “EU Grid” and then continues into the “CZ” grid. With this knowledge all cross-border lines for Germany where identified and the rebuilding process was started for North-East Germany.

The rebuilding process itself is semi automatized via object-oriented Python scripts and resulting data is stored in an Excel workbook. A key component is the storing of data within in the Substation PowerFactory object (ElmSubstat). This reduces the amount of visible data (substation interior such as transformers or bus bars can be viewed by double-clicking them) and the resulting grid is much more clearly arranged. In an analogue way the data for plants and interior elements (e.g. generators) are stored in the Plant PowerFactory object (ElmPlant).

Starting from known elements and comparing to the ENTSO-E grid map [20], further elements can be identified, selected and automatically arranged as substations or plants. During this process an air distance calculator [21] and some information about German power plants [22] proved to be helpful. Once a reasonable number of substations or plants are identified, a Python script is run that assembles this data into an Excel workbook. Then by using this data, the grid is built automatically by the PowerFactory Network Diagram Tool. Finally, the process of identifying new substations and plants starts from the beginning on.

As an example of obtained results, Figure 6 shows the identified grid area of the German transmission system operator 50Hertz Transmission GmbH. Voltage levels are 400 kV, 220 kV and 110 kV levels, major cities are Hamburg and Berlin and the cross-border connections to Poland and Czech Republic are displayed.

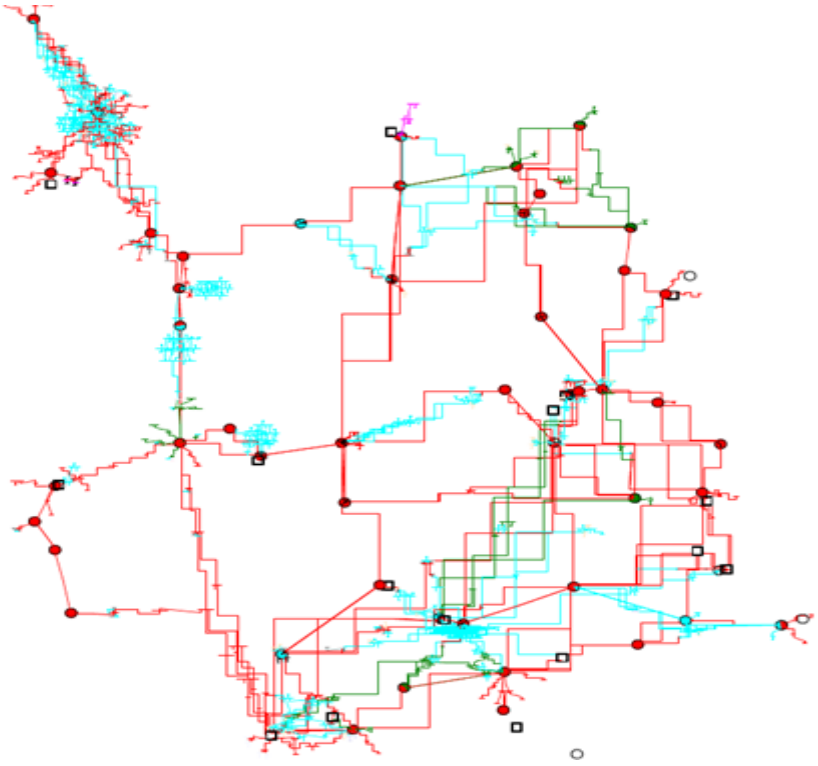


Figure 6 Resulting grid area of the German transmission system operator 50Hertz GmbH

As a component example of this grid area, Figure 7 shows the German plant “Jänschwalde”, featuring 6 generators with a nominal power 588 MVA each and 20 kV nominal generator terminal voltage.

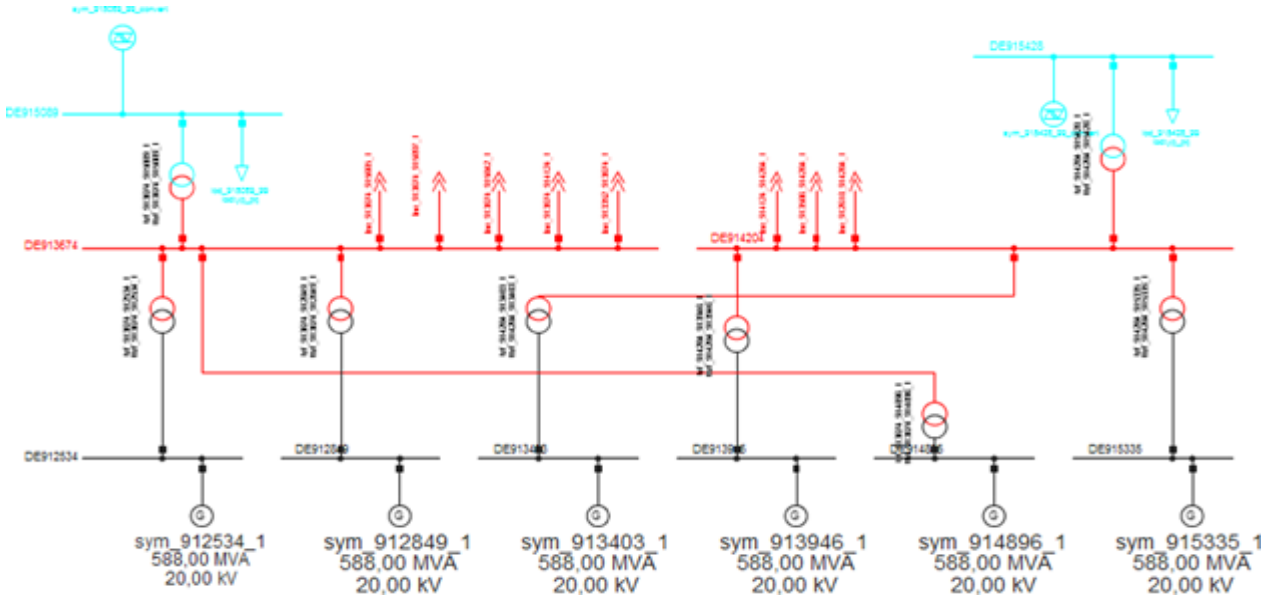


Figure 7 German plant “Jänschwalde” within in the ENTSO-E grid model.

Summarizing, the approach and methods was explained and practically used. Assuming enough time it allows rebuilding the ENTSO-E Grid on the example of map data. The outcome does not match perfectly the map data; however, the existing grid is well recognizable. Further work might also use data available from transmission grid operators such as [23] or [24].

3.1.2 PyPSA-EUR

PyPSA-EUR is an open model of the ENTSO-E network area of the European power system at the transmission system level [25]. It contains 6001 lines at and above 220 kV voltage levels, 3657 substations, a database of conventional power plants and time series for electrical demand and variable infeed. The model developers provide an open toolset [26] for generating the model in the Python for power system analysis (PyPSA) framework. The intention of the developers is to provide a freely available model for “both operational studies and generation and transmission expansion planning studies” [26], and to overcome the restrictive licensing of the ENTSO-E initial dynamic model of continental Europe. The grid topology data for the model is obtained automatically from the geographical vector data of the online ENTSO-E interactive grid map. The cable electrical parameters were obtained by assuming one standard cable type per each of the three voltage levels (220 kV, 300 kV, and 380 kV). DC lines are also modelled. Since the ENTSO-E map does not contain transformer information, a single transformer capacity of 2 GW with a reactance of 0.1 per unit was assumed in order to avoid introducing artificial constraints. The power plant database is derived from different sources, including OPSD, ENTSO-E PPL and the Global Power Plant Database by the World Resource Institute (GPD). Since the sources contain different plants and different plant information, a dedicated powerplant matching tool was used to harmonize them. For obtaining time series for intermittent generation, the CORINE Land Cover database was used for wind power, and the Surface Solar Radiation Data Set - Heliosat for PV.

Because of the assumptions and the usage of the ENTSO-E interactive map, the model is not totally accurate. On the positive side it features an openly available and complete model for continental Europe. The PyPSA model could be built by the provided sources, but a load flow calculation failed thus far for reasons unknown.

3.1.3 German transmission grid model

This model is based upon the German transmission system comprising the 380 kV and 220 kV voltage levels. It is confidential to Fraunhofer IEE but can be used for validation of controllers in INTERPLAN by serving as real physical grid model running on an OpSim component hosted at IEE. Public available data sets and grid maps of the four German transmission system operators provide the basis for this model. It contains nearly 400 substations and more than 600 overhead lines and cables. One third of the substations and lines or cables are situated in the 220-kV voltage level and the remaining substations and cables or lines are in the 380 kV voltage level. In addition, over 50 transformers between the two voltage levels are included in the model. The information about line and transformer parameters has been extracted from the data sets of the TSOs. Future grid development plans are not yet taken into account. The model is available in DIgSILENT PowerFactory including graphical representation.

The used information about the different types of power plants have been obtained from the public available power plant list from the German “Bundesnetzagentur”. This list includes information like e.g. location of the power plant or the voltage level the power plant is connected to. Using a geo-referencing, the power plants can be assigned to the substations extracted from the data sets of the transmission system operators.

The information about the loads at the different substations are obtained from the energy system model “SCOPE” developed at Fraunhofer Institute for Energy Economics and Energy System Technology in Kassel.

The Figure 8 shows a schematic plot of the currently available model.

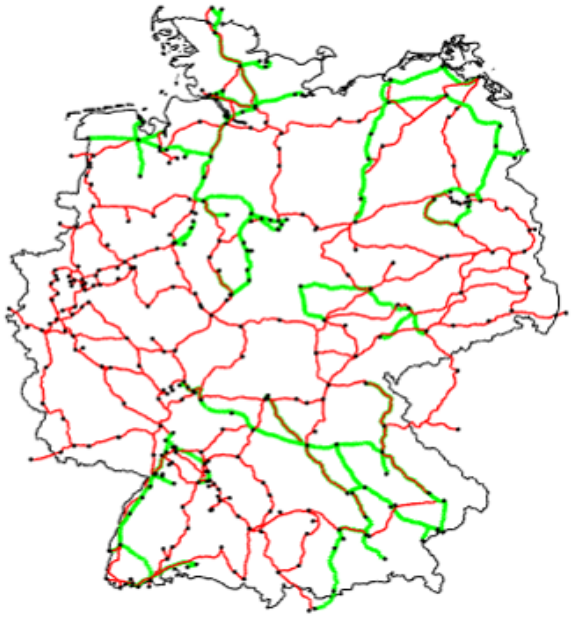


Figure 8 A schematic plot of the currently available model.

3.2 Distribution Grid models

3.2.1 SimBench models

SimBench is a project aimed at developing a public available benchmark dataset for grid models at all voltage levels. The project is running until April 2019. 13 basic grid models with different voltage levels can be created by this dataset until now. The following list shows the voltage levels that are included:

- 1 extra high voltage (EHV) grid
- 2 high voltage (HV) grids
- 4 medium voltage (MV) grids
- 6 low voltage (LV) grids

After finishing the project, each grid model is planned to have 3 different scenarios. At the time of writing the summary, only one such scenario is available. For the medium and low voltage levels, load and generator profiles are available at this time.

Until now these grid models can only be generated from the benchmark dataset by using the pandapower software. It is also possible to merge the grid models with each other if more than one voltage level is needed. Overall the SimBench project provides 246 possible grid model combinations.

In the next two subchapters a detailed description is given for a medium and a low voltage grid model. The combination of these models is an option for a distribution level network for WP6.

3.2.1.1 SimBench semi urban medium voltage grid

The basic data for this grid model is originating from the SimBench Project. It is available at IEE and is confidential. The model represents a synthetic benchmark grid. It comprises the 20 kV voltage level with additional 2 transfer points into the 110 kV voltage level. The model contains a total of 124 lines, 118 busbars, 2 transformers, 120 loads, and 126 static generators. In addition, it provides 10

switches to divide the network into subnets. It is available in the pandapower software format and is expected be also available in the PowerFactory software soon. A graphical representation is plottable using the pandapower software. The model provides detailed data for cables and transformers. The lines include parameters for length, resistance, reactance, capacity, and the maximum current. The transformers are modelled with parameters regarding tap-characteristics, transformer losses and relative short-circuit voltage. The grid model provides static model data for generators. The generators are assigned to energy sources: photovoltaic, wind power, biogas and hydro power. The model assets are not attributed to geographic coordinates or substation names. The model does include time series data for loads and generators for the scenario year 2016 in 15-minute resolution. This time series data was generated by a combination of real measurement data and synthetic data. For developing the synthetic data, a profile generator was used with multiple input data sources (Weather data, location assumptions, etc.).

The given file format may be converted to pickle, excel, Json, SQL, PYPOWER, MATPOWER and PowerFactory by the pandapower software; however, the PowerFactory conversion currently does not fully support all elements and does not take over the node coordinates, thus resulting in a generic PowerFactory network graph. Hence, the conversion result needs to be manually corrected.

At the time of writing this document, the grid model was used for load flow calculation with predefined data.

3.2.1.2 SimBench semi urban low voltage grid

The basic data for this grid model is originating from the SimBench Project. It is available at IEE and is confidential. The model represents a synthetic benchmark grid. It comprises the 0.4 kV voltage level with an additional transfer point into the 20 kV voltage level. The model contains a total of 42 lines, 44 busbars, 1 transformer, 41 loads, and 1 static generator. It is available in the pandapower software format and is expected to be also available in the PowerFactory software soon. An interactive graphical representation is not available at this time. The model provides detailed data for cables and transformers. The lines include parameters for length, resistance, reactance, capacity, and the maximum current. The transformer model includes parameters regarding tap-characteristics, transformer losses and relative short-circuit voltage. The grid model provides static model data for a photovoltaic generator. The model assets are not attributed to geographic coordinates / substation names. The model does include time series data for loads and the generator for the scenario year 2016 in 15-minute resolution. This time series data was generated by a combination of real measurement data and synthetic data. For developing the synthetic data, a profile generator was used with multiple input data (weather data, location assumptions, etc.).
The given file format may be converted to pickle, excel, Json, SQL, PYPOWER, MATPOWER and PowerFactory by the pandapower software.
At the time of writing this document, the grid model was used for load flow calculation with predefined data.

3.2.2 Cyprus distribution grid model

The basic data for the distribution grid model considered for INTERLAN is based on a part of the distribution grid of Cyprus provided by EAC. It is available at FOSS and the partners of INTERPLAN through a NDA and it is confidential for the public. The model represents the real physical grid area of Alambra transmission substation in Cyprus in the year of 2019. The model represents a synthetic benchmark grid. It comprises the voltage levels of 11 kV and 400 V line - line.

The model contains a total of 964 lines, 1867 busbars (2423 terminals), 552 transformers, 551 loads, and 559 generators (558 PV systems, 34 of which active, and 1 biomass unit) and consequently 35

active generators. In addition, it provides 552 protection devices (552 fuses) and 372 breakers/switches. It is available in “pfd” format / DIgSILENT PowerFactory software.

An interactive graphical representation is available with feeder noted by different colour (7 representative feeders) and distribution substations either represented by a circle (overhead transformers) or rectangle (ground mounted transformers) at the top level of hierarchical graphical representation.

Each distribution substation node includes a distribution transformer of a specific type depending on its nominal power, general/aggregated load, low/medium voltage busbars and/or aggregated PV system. The cable and transformer models are set according to information provided by the manufacturer for each specific type. It is readily available for performing both steady state and dynamic simulations. Steady state information of the network is already set properly therefore steady state simulations are feasible and can provide results of high credibility/accuracy.

The distribution grid model can also be used for producing general results for dynamic studies. However, fine tuning of models based on technology type, control etc. is required to acquire results closer to reality. Storage is not included (as no storage is installed on the specific distribution network) neither provisioned in the model, but it can be incorporated at a later stage based on project requirements. The model assets are not attributed to geographic coordinates, but geographical information/ coordinates can be incorporated into the model in a future stage if this is required. The substation name, unique ID and nominal power of each distribution substation is noted at the as well as the rating and type of cables at the top level of graphical hierarchy. The model does include time-series (per hour) for load consumption and PV power production for a single day which can be extended to a whole year or shorter time period of second resolution. The model can be used in both steady state and RMS simulations which can provide results from per hour to per cycle resolution and which is suitable for steady state/dynamic studies and for appropriate control development.

3.2.3 DECAS Synthetic Network

This synthetic test network covers all elements between the transmission (220 kV) level, distribution (110 kV and 20 kV) level and the LV connection points. A 110 kV looped grid is supplied from two sides, with two three-winding 220/110 kV transformers. At the MV level, two feeders (rural and urban) are modelled in detailed, representing the generic model of two typical feeders. The consumption of other MV feeders was determined by appropriately scaled equivalent loads. At the LV level, several LV (radial) feeders were modelled in detail to represents the situation in rural and urban LV networks. The consumption of other LV networks that were not modelled in detail is determined by scaled equivalent LV loads.

Modelled HV network is supplied with a slack bus and includes one 110 kV loop. The underlying MV networks are supplied through this HV loop. On the MV networks, the transformer stations have been modelled with two types of generic radial feeders

- Rural (overhead) feeder
- Urban (cable) feeder

The length of the rural feeder is 8 km, which represents a typical length of a rural feeder, where larger (400 kVA) transformers (5 in total) are uniformly distributed along the beginning of the feeder, followed by 13 transformers of 250 kVA and 4 transformers with a nominal power of 160 kVA. Transformers towards the end of the feeder are supplying remote small settlements, with low consumption. The length of the urban feeder is 2 km, with 400 kVA transformers (10 in total) uniformly distributed along the feeder. Rural feeder mainly supplies household consumers, while a portion of industrial or business consumers is attached to the urban feeder (approximately 30% of total consumption).

The LV networks consist of radial feeders. Network is supplied through one MV/LV transformer with nominal powers ranging from 50 kVA up to 800 kVA, depending on the size of the supplied LV network. The loading rate varies from 15% to 70%, with majority of the transformers being loaded between 50% and 60%.

When compared to rural LV networks, urban ones are different in the following terms:

- increased amount of the supplied customers,
- shorter lengths of the feeders,
- consequently, increased transformers ratings.

Radially operated urban LV networks are supplied by one transformer of nominal power from 160 kVA to 1250 kVA, dependent on the size of the individual network. Loading of the transformer is similar as it is in case of the rural network, with majority of the transformers being loaded between 50% and 60%.

The LV feeders are connected to the MV urban feeder at different points along the feeder. Consumption is more concentrated than in the rural case, and the feeders are shorter and amount to about 400 m. The network supplies both, household and business customers. In both cases, with the rural and urban LV networks, the transformers tap positions are set so that the nominal voltage at the LV busbar in the transformer station is always achieved, thus allowing a 10 % voltage drop in the LV network.

The network has been modelled in DIgSILENT PowerFactory v.2018 software. At the time of writing this document, the grid model was used for load flow calculations and to evaluate different control strategies for distributed generation.

4 Clustering parameters

Initial variables were selected based on the impact they might have on differentiating feeder types and on DG hosting capacity. The initial variables varied among the different available networks as needed to account for differences in availability of data from each partner.

Table 1: Clustering parameters

ID	Parameter symbol	Unit of measure	Parameter Description
1	*V*	[kV]	Nominal voltage of the grid equivalent
2	*dv/dP*	[p.u./MW]	Sensitivity of critical-bus voltage magnitude due to active-power change at any bus of the feeder
3	*dv/dQ*	[p.u./Mvar]	Sensitivity of critical-bus voltage magnitude due to reactive-power change at any bus of the feeder
4	*ADTN*	[m]	Average distance to the neighbour nodes
5	*N*	–	Number of nodes in feeder
6	*L*	[m]	Feeder total line length
7	*L_critical node*	[m]	Critical node* length
8	*L_max*	[m]	Feeder maximal line length
9	*Z_∑*	[Ohm]	Equivalent sum impedance per feeder
10	*R_k*	[Ohm]	Short circuit resistance
11	*I_nom*	[A]	Nominal current of the most loaded line in the feeder
12	*I_max*	[A]	Maximal current of the most loaded line in the feeder
13	*S_RES*	[kVA]	Average installed RES Power per point of connection
14	*S_DG*	[kVA]	Average installed DG Power per point of connection
15	*S_load*	[kVA]	Load rated power per point of connection
16	*H*	[s]	Inertia Constant
17	dP_n/dP_i	[MW]	It is the effect of the injection of active power variation ΔP at busbar i for the adjacent branch n
18	dP/df	[MW/Hz]	Total grid freq. droop contribution
19	*Kp,q,v,f*	[p.u.]	Controller parameters of RES, DG, Storage
20	V_n/V_l	[%]	Feeder maximal voltage drop
21	dP_n/dP_l	[%]	Maximum loading
22	L_{min_V}	[m]	Length to the feeder node with minimal voltage
23	*Vmin*	[p.u.]	Minimal voltage of feeder node

** The feeder critical node is the node with the highest voltage deviation from the nominal value (in terms of voltage drop (or voltage rise in the presence of DG)).*

The presented parameters could provide a very important input to a clustering process according to the use cases of INTERPLAN project. However, not all of them are available within the given

database. Considering low voltage grids and the available parameters, the following variables will be implemented into clustering algorithm:

Table 2: Parameters chosen for feeder clustering

ID number in Python Data frame	ID	Parameter symbol	Unit of measure	Parameter Description
1	7	*L_critical node*	[m]	Critical node length
2	2	*dv/dP*	[p.u./MW]	Sensitivity of critical-bus voltage magnitude due to active-power change at any bus of the feeder
3	3	*dv/dQ*	[p.u./Mvar]	Sensitivity of critical-bus voltage magnitude due to reactive-power change at any bus of the feeder
4	9	$Z\_\sum$	[Ohm]	Equivalent sum impedance per feeder
5	8	*L_max*	[m]	Feeder maximal line length
6	20	V_n/V_l	[%]	Feeder maximal voltage drop
7	23	*Umin*	[p.u.]	Minimal voltage of feeder node
8	12	*I_max*	[A]	Maximal current of the most loaded line in the feeder
9	21	dP_n/dP_l	[%]	Maximum loading
10	22	L_{min_V}	[m]	Length to the feeder node with minimal voltage
11	11	*I_nom*	[A]	Nominal current of the most loaded line in the feeder
12	5	*N*	−	Number of nodes in feeder
13	10	*R_k*	[Ohm]	Short circuit resistance

5 Clustering Algorithm

A general concept of the clustering procedure is presented in figure 9. The process is starting with importing network data and it's processing via DIgSILENT PowerFactory environment scripting with Python. This allows to perform the grid simulations and provides a very high computational functionality and flexibility for data exchange. INTPERLAN operates with a database of 2000 low voltage networks, which includes a large number of feeders (about 9500).

The next step of the clustering process implies a calculation of feeder parameters described in Section 4.

5.1 Statistical analysis

To represent the obtained data points of each feature graphically statistical methods such as box plot and violin plot are applied. For this purpose, a Python statistical data visualization library `Seaborn` (based on a Python 2D plotting library `Matplotlib`) has to be imported: `sns.violinplot`, `sns.boxplot`. The data points are necessarily going through a normalization process to make it possible to compare the properties of selected variables. In Python it can be realized by using an estimator `MinMaxScaler`, implemented in a `Scikit-learn` library:

$$x' = \frac{x - min(x)}{max(x) - min(x)}, \tag{5.1}$$

where x is an original value, x' is the normalized value.

As shown in Figure 9, the majority of features related to cable characteristics present a normal (or close to normal) distribution, excepting the parameter "nominal current of maximal loaded line", which demonstrates an asymmetrical distribution curve with the isolated area on the extreme level. The features related to feeder topology present a positively skewed distribution with the mean lying to the right of the peak. Only the variable "feeder terminal minimal voltage" indicates the highest normalized values and slightly tends to a left-skewed distribution.

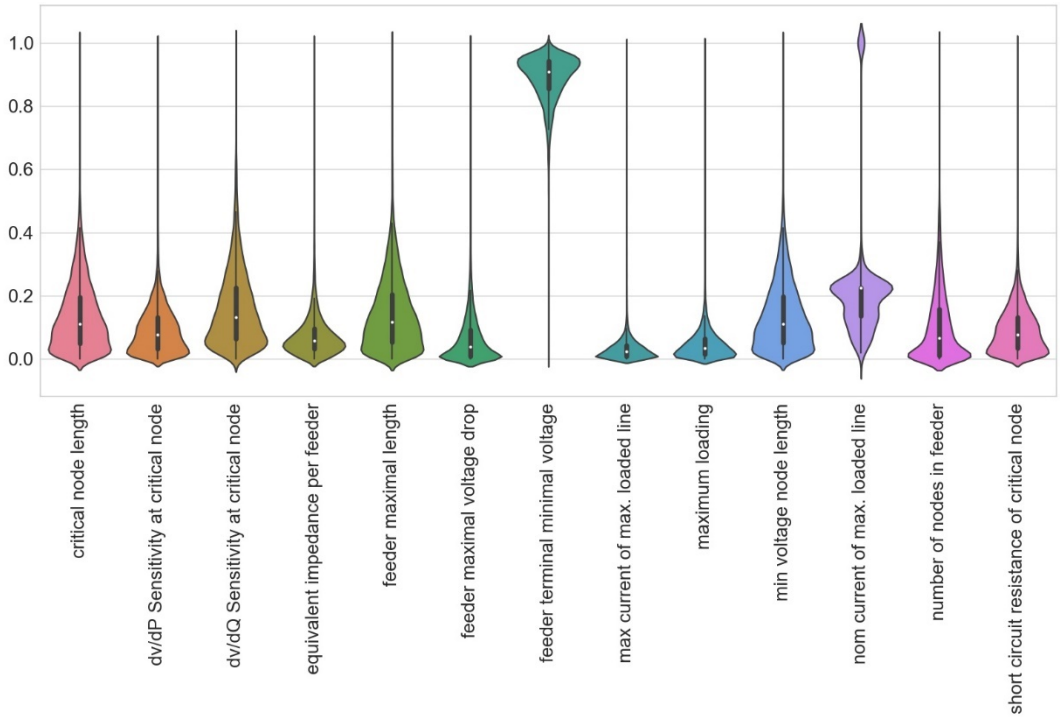


Figure 9 Violin plots of parameters chosen for clustering process

Figure 10 illustrates the box plots for each variable, which contain the information about the median of the given data set, its interquartile range (IQR), namely the difference between 75th and 25th percentiles, and the extremes of 1.5 IQR as well. The grey, isolated points correspond to the outliers. It is obvious, that the medians for the features presenting a normal distribution are located closer to the centre of the IQR and approximate to the means (red rhombuses), while the variables with skewed distribution indicate a displacement of the median values to the lower or upper quartile.

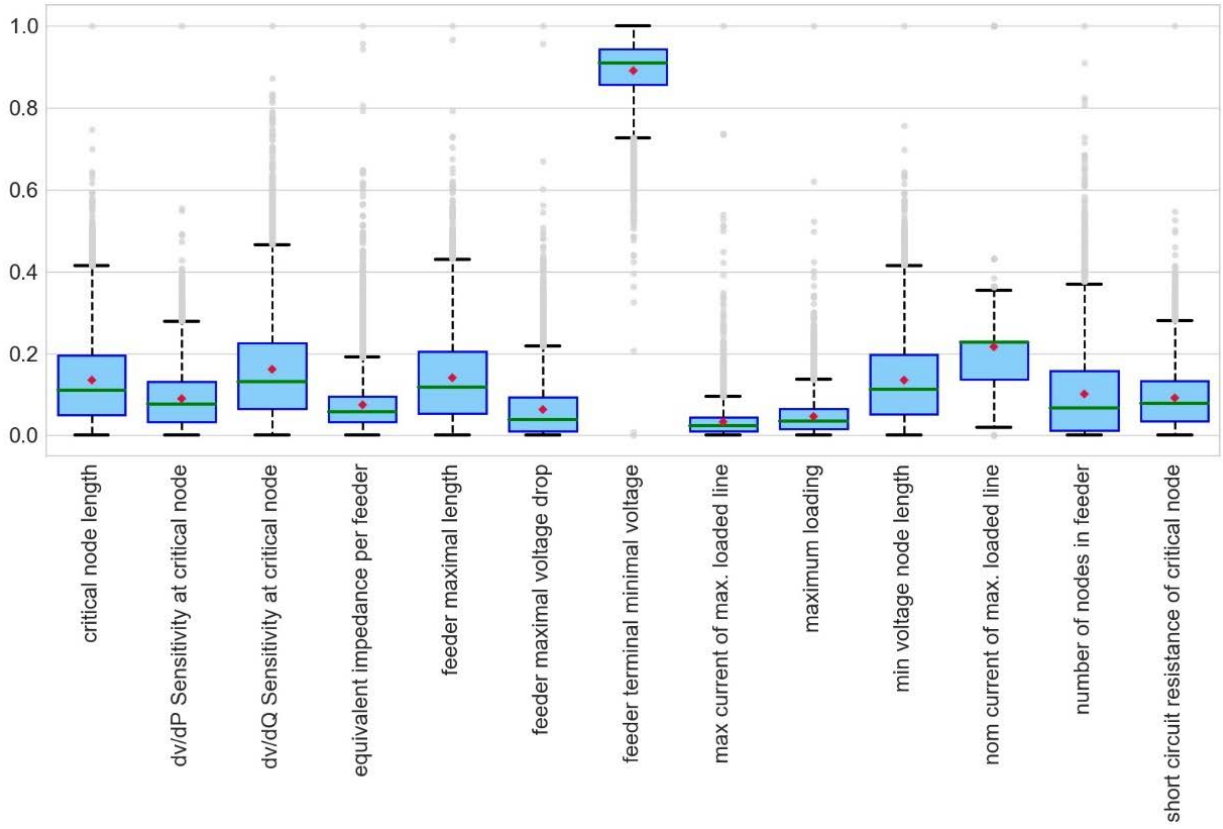


Figure 10 Box plots of parameters chosen for clustering process

For a better understanding of interdependence between feeder variables it is also recommended to perform a correlation analysis. In Python `corr()` function is used to measure pairwise correlations between columns of a given data frame. As a standard method the Pearson correlation is implemented: `DataFrame.corr(method='pearson')`.

Figure 11 represents a correlation matrix with correlation coefficients between variables varying within the interval from -1 to +1. The set limits indicate a total negative linear correlation (deep blue colorization) and total positive linear correlation (red colorization), respectively.

Although the variables “critical node length” and “minimum voltage node length” are characterized by a perfect positive correlation, they will be both taken into consideration. Indeed, in presence of distributed generations (DGs) voltage rises can occur in the distribution system and the node with the highest voltage difference won’t be associated with the node representing the highest voltage drop. Likewise, the farthest node can not only be considered in terms of the lower voltage limits. Therefore, the variable “feeder maximal length” will also be presented in the analysis.

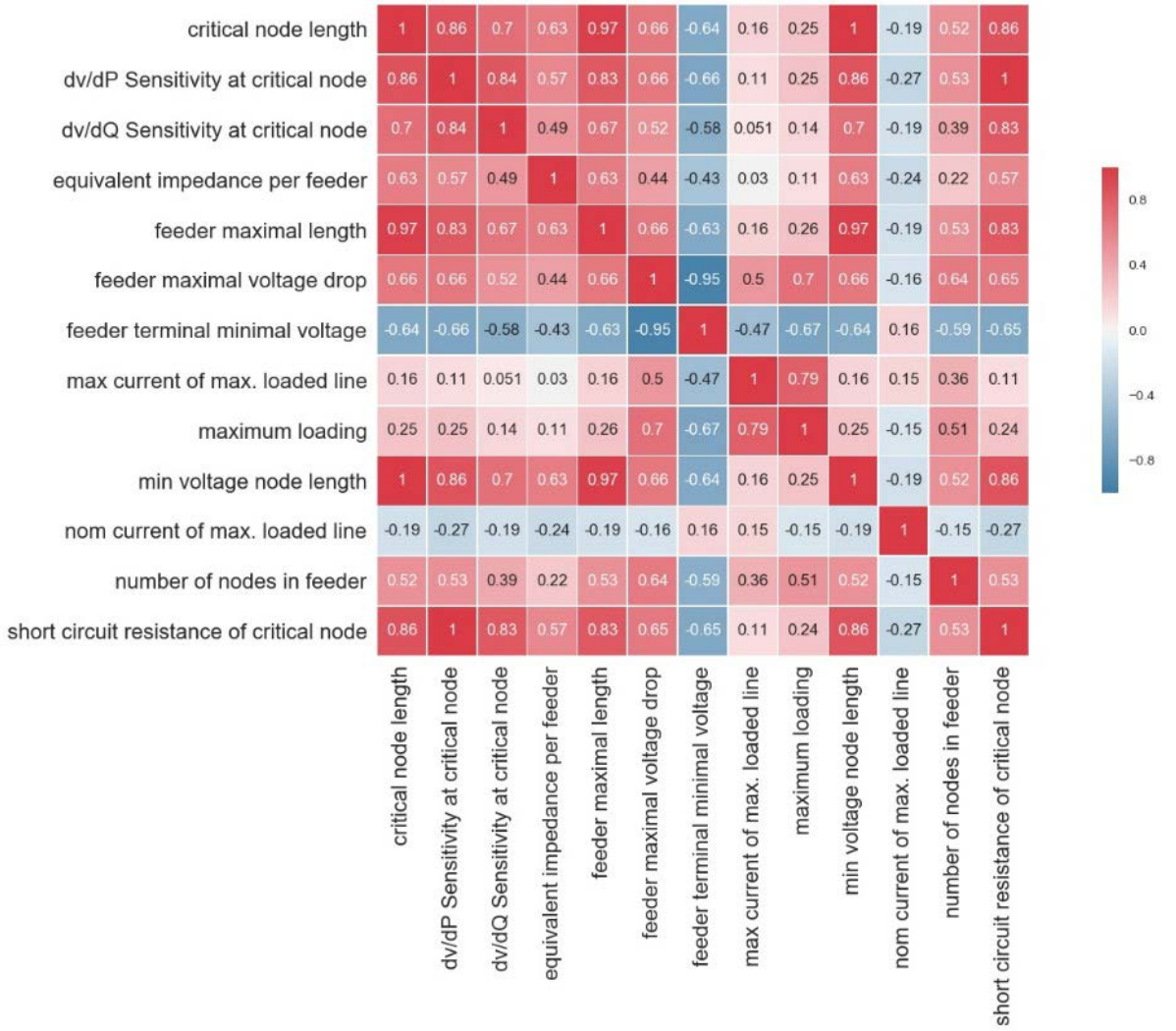


Figure 11 Correlation matrix of feeder variables

Since the variable “short circuit resistance of critical node” has been calculated based on the sum of active power losses of the individual feeder line, it demonstrates a perfect positive correlation with the variable “dv/dP sensitivity at critical node”. The correlation matrix indicates furthermore a very high negative correlation between variables “feeder terminal minimal voltage” and “feeder maximal voltage drop”. It would be possible to proceed with the analysis with a reduced number of parameters, as we have observed clear correlation factors. Taking into account, that computation time and complexity is fairly low in these applications and it is not the highest priority for INTERPLAN, all thirteen parameters should be considered to get more accurate clustering output.

5.2 Principal component analysis

The last step of the data preparation process assumes the implementation of a principal component analysis (PCA). This procedure is used to convert the original data, performed in a multidimensional space, into a subset of lower dimensional variables, which describe maximal data variance and manage the data organization. PCA is thereby a very useful method to reduce the data redundancy and to enable the visualization of clustering results.

Before realization of PCA the data points for each variable have to be scaled using a standard z-score (estimator `StandardScaler` in Python):

$$z = \frac{x - \mu}{\sigma}, \tag{5.2}$$

where μ is the mean of the feeder sample, σ is its standard deviation.

The `Scikit-learn` library support the performing of PCA by applying the `PCA()` function into a given data set. The number of principal component has to be specified in the parameter list: `PCA(n_components=2)`.

Figure 12 represents a PCA-biplot (scatter plot of each feeder in the coordinate system of the first two principal components), combined with the vector plot of the thirteen original variables. The first principal component explains 57,55 % of the global data variability and mainly accounts for the parameter “feeder maximal length”, while the second principal component explains 16,8 % of the data and principally relates to the parameter “maximal current of max. loaded line”.

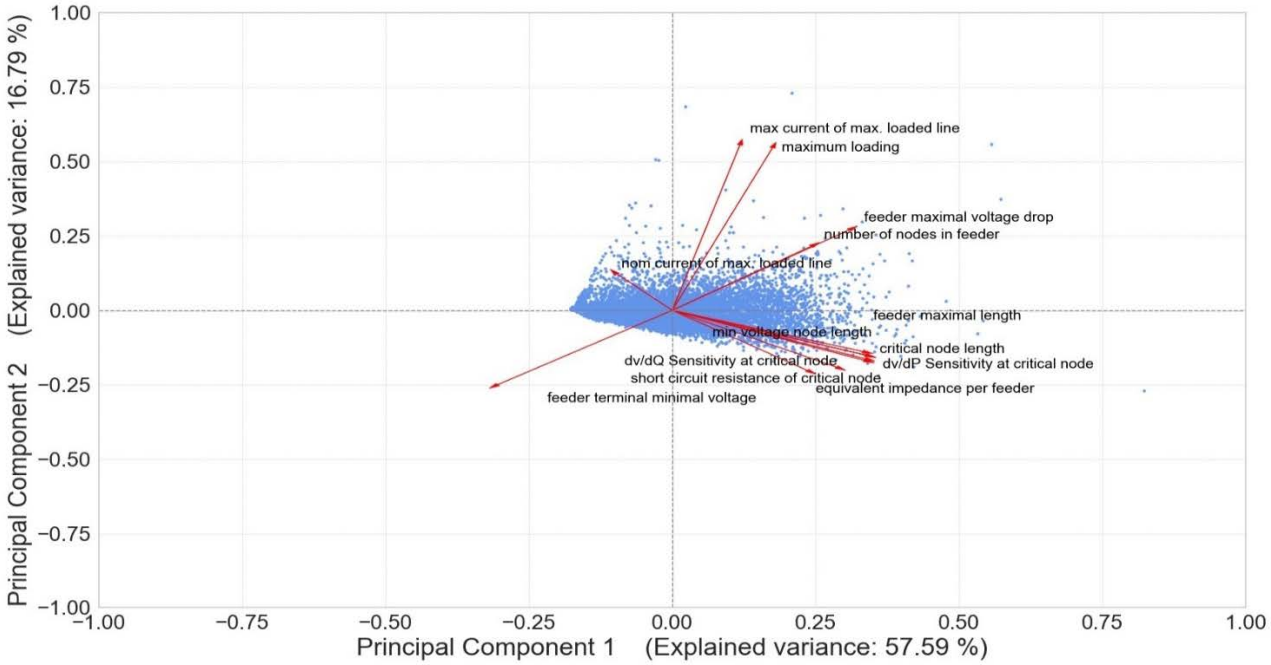


Figure 12 PCA biplot

5.3 Clustering analysis

After the data preparation the clustering analysis can be performed. As explained in Section 2.1 the partitional clustering approach would be the most appropriate technique in terms of the project objectives and data availability. We will apply the K-means-based algorithms, k-medoids and k-medians among them, because of their high efficiency by operation with the large data set.

To realize partitional clustering in Python the `KMeans` function should be imported from the `Scikit-learn` library. The number of clusters has to be predefined by specifying the parameter `n_clusters`: `KMeans(n_clusters=5)`. The cluster centroids can be computed using the command `cluster_centers_`. The closest point to the centroid will give us the special parameters of a representative feeder.

The `Scikit-learn` library doesn’t include k-medoids and k-medians clustering algorithms, because of this, the additional package `PyClustering` has to be installed. The command `get_clusters` organize a data points to a list of allocated clusters; the command `get_medoids` returns the list of cluster medoids (analogously, `get_medians` for the k-medians algorithm).

Table 3 gives the overview of the feeder characteristics for 5 clusters, obtained after applying different partitional algorithms.

Table 3: Characteristics of representative feeders according to different variations of K-means

Parameter / Cluster	critical node length, [km]	dv/dP Sensitivity at critical node, [p.u./MW]	dv/dQ Sensitivity at critical node, [p.u./Mvar]	equivalent impedance per feeder, [Ohm]	feeder maximal length, [km]	feeder maximal voltage drop, [%]	feeder terminal minimal voltage, [p.u.]	max current of max. loaded line, [kA]	maximum loading, [%]	min voltage node length, [km]	nom current of max. loaded line, [kA]	number of nodes in feeder	short circuit resistance of critical node, [Ohm]
Centroids													
1	0.163	0.363	0.258	0.043	0.180	0.396	1.019	0.025	11.202	0.168	0.216	6	0.058
2	0.956	1.930	0.969	0.171	1.011	3.034	0.989	0.041	18.346	0.959	0.229	24	0.309
3	0.637	1.364	0.646	0.099	0.690	4.858	0.971	0.109	49.385	0.640	0.226	48	0.216
4	0.461	0.994	0.537	0.084	0.505	1.549	1.005	0.043	19.345	0.464	0.225	23	0.160
5	0.057	0.064	0.166	0.004	0.065	0.026	1.022	0.058	5.766	0.059	1.000	2	0.009
Medoids													
1	0.163	0.326	0.258	0.041	0.224	0.379	1.018	0.022	11.737	0.163	0.185	4	0.052
2	0.921	1.879	1.058	0.165	0.921	3.393	0.992	0.040	22.759	0.921	0.176	21	0.301
3	0.578	1.330	0.648	0.116	0.658	5.222	0.963	0.099	53.714	0.578	0.185	43	0.210
4	0.497	0.968	0.519	0.108	0.500	1.934	1.005	0.038	17.799	0.500	0.211	27	0.157
5	0.079	0.036	0.140	0.000	0.079	0.000	1.024	0.056	5.557	0.079	1.000	3	0.005
Medians													
1	0.076	0.205	0.177	0.023	0.083	0.082	1.023	0.008	4.310	0.080	0.211	2	0.033
2	0.805	1.632	0.801	0.119	0.859	3.309	0.987	0.053	23.160	0.808	0.270	28	0.261
3	0.425	0.937	0.459	0.066	0.459	1.573	1.005	0.047	20.746	0.428	0.270	23	0.150
4	0.228	0.472	0.276	0.048	0.246	0.496	1.017	0.022	10.068	0.232	0.270	7	0.076
5	0.029	0.038	0.140	0.000	0.032	0.000	1.024	0.009	0.948	0.031	1.000	1	0.005

5.4 Internal validation of clustering results

The most difficult issue of a clustering approach is determining the optimal number of clusters. For internal validation of the clustering results INTERPLAN will utilize two distance-based indicators: Sum of Squared Errors (SSE) and Silhouette Coefficient (S).

The Sum of Squared Errors for the K-means clustering with the resulting clusters $C = \{C_1, C_2, \dots C_k, \dots C_K\}$ is defined as following:

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2, \tag{5.3}$$

where C_k is the k^{th} cluster, x_i is a point in C_k and c_k is the centroid of cluster C_k. The main objective consists of minimizing the Sum of Squared Errors.

By calculating the Silhouette Coefficient both internal and external cluster distances are assumed:

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}, \tag{5.4}$$

where a_i is the average of the distances between feeder i and all other points in the same cluster; b_i is the minimal average distance of feeder i to all the data points in any other cluster, not containing $feeder_i$. The Silhouette Coefficient of a given feeder i indicates its reference to a specified cluster and lays within the limits from -1 to +1. In Python the average Silhouette value for all samples can be computed using the function `silhouette_score`, imported from the `Scikit-learn` library. For the visualization of results, it is recommended to use the `Yellowbrick` visual diagnostic tool.

The left part of Figure 13 illustrates the scatter plot for three different sets of clusters in the coordinate system of the first two principal components; the red crosses indicate cluster centroids. The areas with a high concentration of points have been divided into sections, which represent the similar feeder characteristics. The separately located points, having more specific features, have been assigned to the nearest clusters. The right part of the figure 5.5 shows the corresponding silhouette plots: vertical axis is related to the number of observations; horizontal axis is related to the silhouette width. It was assumed, that the average silhouette width greater than 0,5 indicates a reasonable cluster structure. The negative silhouette values mean that the observations share similar

characteristics with other clusters.

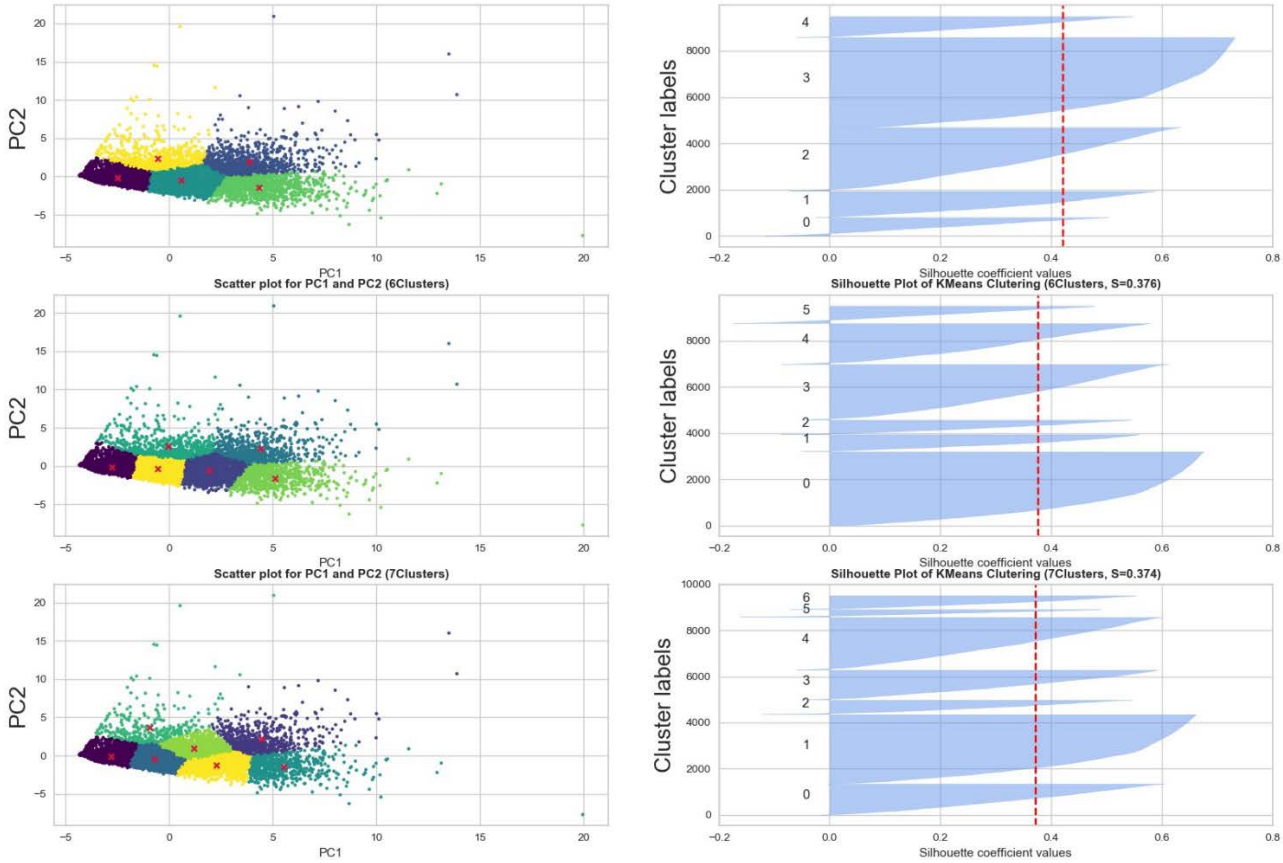


Figure 13: Scatter plot and Silhouette plot of different cluster sets

It can be seen, that practically all feeders have positive silhouette coefficients, so they have been correctly assigned to the clusters. The highest overall average silhouette width (0,421) is observed for the set of 5 clusters. This can be interpreted as an acceptable cluster partitioning.

Figure 14 illustrates the changes of SSE and Silhouette indices by varying the number of clusters.

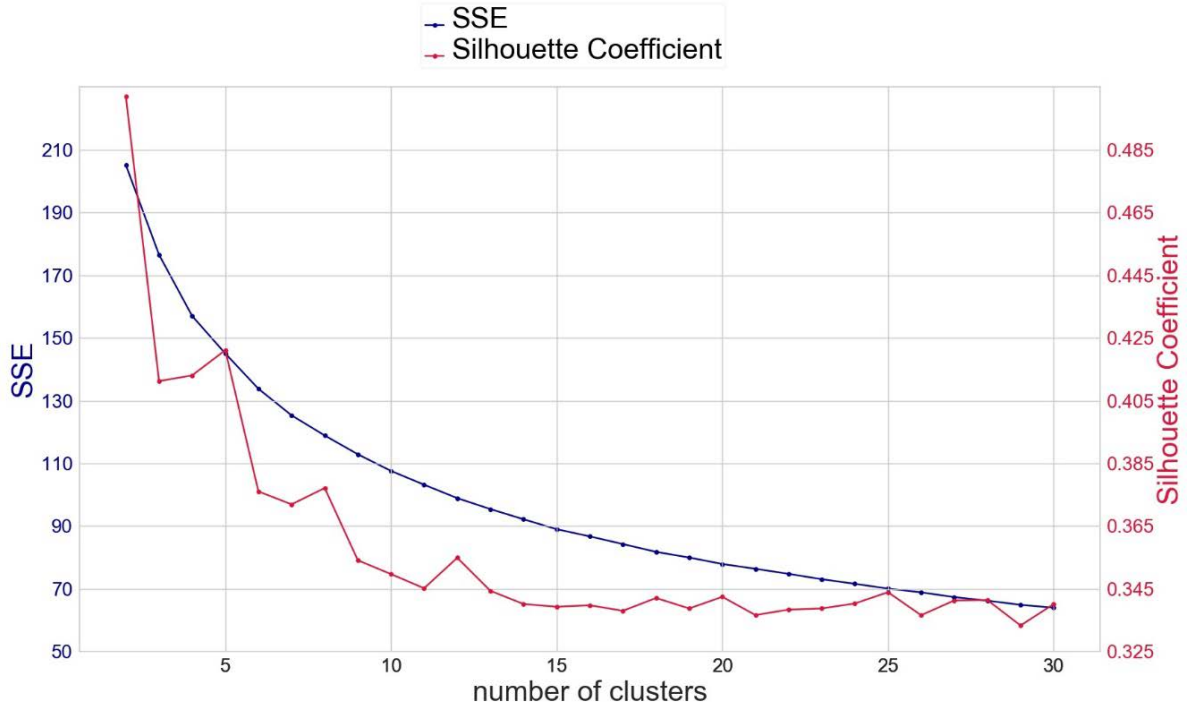


Figure 14 Determining of optimal number of clusters (based on PC-dataset)

Figure 14 reveals that the SSE-curve monotonously decreases. The curve goes dawn intensively on the interval between two and five clusters and then decreases smoothly. The Silhouette curve has two visible peaks among the others with the maximal values occurring at two and five clusters. From this point of view, the set of five clusters would provide an optimal clustering output.

The cluster grouping will also go through the external validation based on the KPIs of the INTERPLAN project. The obtained results of feeder clustering will be compared with the distribution of the other feeder parameters, which are a priory known and have not been considered in the clustering process.

Figure 15 shows the box plots of the KPI “active power losses” for different sets of clusters. The corresponding representative feeders have been marked as the yellow points. It can be seen, that for the set of five clusters two representative feeders (see Figure 15 a) are located outside the interquartile range (cluster 1 and cluster 3). For the set of 6 clusters (see Figure 15 b), cluster 2 and cluster 4 represent similar power losses within the range between approx. 1,5 % and 3,0 %. The parameter distribution of these two clusters has to be compared with the distribution of the other KPIs, whose values are expected to be significantly different. For the set of 7 clusters (see Figure 15 c) representative feeders are situated closer to the cluster medians. For the set of 8 clusters (see Figure 15 d) cluster 1 and cluster 4 show the lowest power losses, while more than half of feeders in the cluster 7 have the power losses higher than the appropriate value of 4 %. The other clusters represent the various distributions of active power losses within the range between 0,5 % and 4 %. This could indicate a quite effective cluster organization, however the representative feeder for the cluster 2 lays outside the interquartile range.

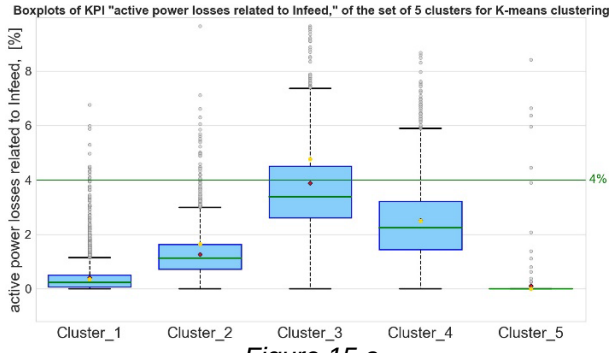


Figure 15 a

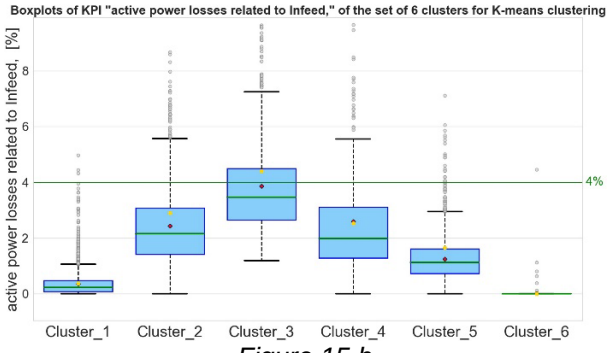


Figure 15 b

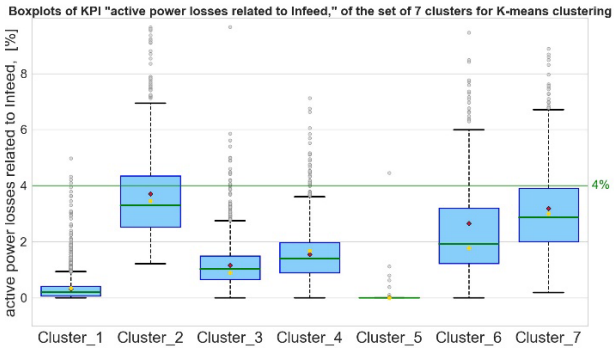


Figure 15 c

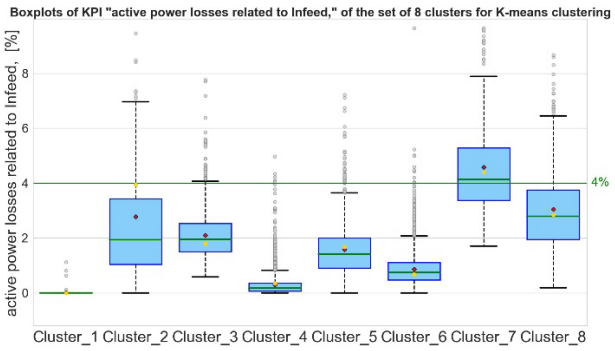


Figure 15 d

Figure 15 Box plots of KPI "active power losses" for different cluster sets

In addition, the correctness of clustering results will be evaluated using Purity, a quantitative criterion for the external validation. Purity describes how homogeneous are the clusters in terms of the predefined classes:

$$P_j = \frac{1}{n_j} Max\left(n_j^i\right), \tag{5.5}$$

where n_j^i is number of feeders of class label i in cluster j and n_j is the number of feeders in cluster j. The global purity:

$$Purity = \sum_{j=1}^{m} \frac{n_j}{n} P_j \tag{5.6}$$

where m is number clusters and n is the total number of feeders.

It is assumed that the KPI "power losses" can be divided into two classes:

- active power losses > 4%;
- active power losses ≤ 4%, representing the allowable level of power losses for the LV grids [27].

Figure 16 illustrates for the set of 8 clusters the share of feeders belonging to the predefined class. We can see, that cluster 7 contains mainly feeders related to the first class (active power losses > 4%), while the other clusters mainly contain the feeders of the second class (active power losses ≤ 4%). This indicates generally low level of power losses in the considered LV networks. Concerning the clustering aspects, the purity of cluster 7 (0,56) and cluster 8: (0,8) should be higher to provide more adequate feeder partitioning. The cluster purity closer to 1 would represent the optimal clustering results.

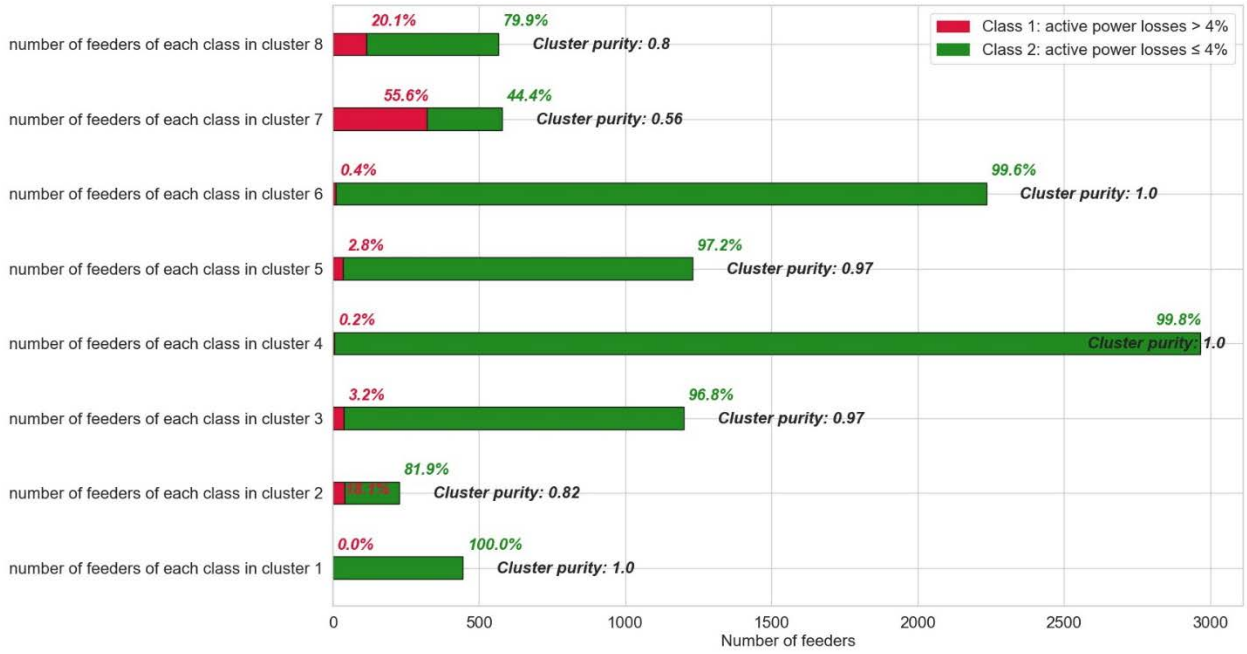


Figure 16 Bar plot of cluster purity

6 Conclusions and Outlook

Clustering is an effective machine learning technique being used for the data exploration based on grouping of observations (feeders or networks) into clusters according to their specific features. The generated clusters should demonstrate a very high diversity from each other, while the data points within a cluster should be quite similar.

Partitional methods (belonging to the k-means family) are the most efficient and at the same time relatively simple clustering algorithms, which require however to pre-define the number of clusters with their initial centers and to adjust them iteratively. Along with the basic k-means algorithm its different variations (such as k-medoids, k-medians, k-modes, fuzzy c-means clustering etc.) should be considered. For instance, k-medians clustering is less sensitive to outliers compared to k-means. Likewise, to avoid overlapping effects on the cluster boundary, the membership value of each observation to cluster centers can be assigned using the fuzzy c-means clustering algorithm. Other properties of each clustering methodology will be explored as well; it is then expected that not a single clustering algorithm provides an adequate level of accuracy for all use cases in INTERPLAN.

Initial clustering parameters were selected based on the impact they might have on differentiating feeder types and on DG hosting capacity. The initial variables varied among the different available networks as needed to account for differences in availability of data from each partner. The presented parameters could provide a very important input to a clustering process according to the use cases of INTERPLAN project. However, not all of them are available within the given databases. Considering low voltage grids, a limited set of variables will be implemented into the clustering algorithm.

The report describes transmission and distribution grid models that were developed or considered for usage in the project. This report does not describe all grid models that were used for clustering and development of grid equivalents carried out.

According to the general concept of clustering procedure which is presented in the report, the process is starting with the importing of network data and their processing via DIgSILENT PowerFactory environment scripting with Python. This allows performing the grid simulations and provides a very high computational functionality and flexibility for data exchange. INTERPLAN operated with a database for 2000 low voltage networks, which includes a large number of LV feeders (about 9500).

7 References

[1] M. G. M. &. C. J. F. G. Nijhuis, “Clustering of low voltage feeders from a network planning perspective,” in *Proceedings of the 23rd International Conference on Electricity Distribution (CIRED), 15-18 June 2015, Lyon, France*, Lyon, 2015.

[2] A. Holmgren, “Using graph models to analyze the vulnerability of electric power networks”,” *Risk Analysis,* vol. 26, no. 4, pp. 955-969.

[3] A. G. P. R. Y. Xu, “Architecture of the Florida power grid as a complex network,” *Physica A: Statistical Mechanics and its Applications,* vol. 401, pp. 130-140.

[4] P. Schulyz, “A random growth model for power grids and other spatially embedded infrastructure networks,” *The European Physical Journal Special Topics,,* pp. 2593-2610, 2014.

[5] ,. M. D. ,. P. S. J. Dickert, “Benchmark low voltage distribution networks based on cluster analysis of actual grid properties,” in *Proceedings PowerTech*, Grenoble, 2013.

[6] M. W. X. N. a. Y. L. Hong Fan, “Transmission network expansion based on reference network concept,” in *IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Xi'an, 2016.

[7] R. A. G. Strbac, “Performance regulation of distribution systems using reference networks,” *Power Engineering Journal,* vol. 15, no. 6, pp. 295-303, 2002.

[8] G. S. V. Levi, “Assessment of performance-driven investment strategies of distribution systems using reference networks,” *IEE Proceedings: Generation transmission and distribution,* vol. 152, no. 1, pp. 1-10, 2005.

[9] N. P. H. G. R. Bhakar, “Reference network development for distribution network pricing,” in *IEEE Transmission and Distribution Conference and Exposition*, 2010.

[10] N. P. H. G. R. Bhakar, “Development of a flexible distribution reference network,” in *IEEE Power and Energy Society General Meeting*, 2010.

[11] T. R. A. S.-M. e. a. C.M. Domingo, “"A Reference Network Model for Large-Scale Distribution Planning With Automatic Street Map Generation,” *IEEE Trans. on Power Systems,* vol. 26, no. 1, pp. 190-197, 2011.

[12] R. B. M. N. N.P. Padhy, “Smart reference networks,” in *IEEE Power and Energy Society General Meeting*, San Diego, 2011.

[13] X. C. X. T. Z. W. J. Zhang, “Benders Decomposition Algorithm for Reference Network,” in *TENCON 2015 IEEE Region 10 Conference*, Macao, China, 2015.

[14] A. F. C. a. G. L. L. D. P. Ospina, “Implementation and validation of the Nordic test system in DIgSILENT PowerFactory,” in *PowerTech*, Mancester, 2017.

[15] “Open energy system databases,” Wikipedia, [Online]. Available: https://en.wikipedia.org/w/index.php?oldid=883471133. [Accessed 22 02 2019].

[16] “Open Power System Data – A platform for open data of the European power system,” Neon Neue Energieökonomik GmbH, [Online]. Available: https://open-power-system-data.org/#. [Accessed 22 02 2019].

[17] I. Dobson, “Synchrony and your morning coffee,” *Nature Phys,* vol. 9, no. 3, pp. 133-134, 2013.

[18] ENTSO-E, “Dynamic Study ModelRange of Applications and Modelling Basis,” 09 01 2015. [Online]. Available: https://docstore.entsoe.eu/Documents/Publications/SOC/Continental_Europe/InitialDynamicModel_Handbook_gen.pdf. [Accessed 22 02 2019].

[19] F. R. S. S. e. al., “Evaluation of the ENTSO-E initial dynamic model of continental Europe subject to parameter variations,” in *2017 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Washington DC, USA, 2017.

[20] ENSTO-E, “Grid Map,” [Online]. Available: https://www.entsoe.eu/data/map/. [Accessed 22

02 2019].

[21] “Entfernungsrechner - Entfernung berechnen und darstellen.,” [Online]. Available: https://www.luftlinie.org/. [Accessed 22 02 2019].

[22] “Bundesnetzagentur - Kraftwerksliste.,” [Online]. Available: https://www.bundesnetzagentur.de/DE/Sachgebiete/ElektrizitaetundGas/Unternehmen_Institutionen/ Versorgungssicherheit/Erzeugungskapazitaeten/Kraftwerksliste/kraftwerksliste-node.html. [Accessed 22 02 2019].

[23] “50hertz.com > Transparenz > Kennzahlen > Statisches Netzmodell.,” 50Hertz GmbH, [Online]. Available: https://www.50hertz.com/de/Transparenz/Kennzahlen/StatischesNetzmodell. [Accessed 22 02 2019].

[24] “Netzbelastung in der Regelzone.,” [Online]. Available: https://www.50hertz.com/de/Transparenz/Kennzahlen/Netzbelastung. [Accessed 22 02 2019].

[25] F. H. D. S. a. T. B. J. Hörsch, “PyPSA-Eur: an open optimisation model of the European transmission system,” *Energy Strategy Reviews,* vol. 22, pp. 207-215, 2018.

[26] “GitHub repository of the PyPSA Model of the European Energy System,” [Online]. Available: https://github.com/PyPSA/pypsa-eur. [Accessed 21 03 2019].

[27] “e-control,” [Online]. Available: https://www.e-control.at/documents/20903/-/-/520753b0-d90a-4c14-9a5f-03f472d42234. [Accessed 15 03 2019].

8 Annex

List of Figures

List of Tables