energies

Article

On the Classification of Low Voltage Feeders for Network Planning and Hosting Capacity Studies

Benoît Bletterie [1,*], Serdar Kadam [1] and Herwig Renner [2,*]

[1] Electric Energy Systems, Center for Energy, Austrian Institute of Technology, Vienna 1210, Austria; serdar.kadam@ait.ac.at

[2] Institute of Electrical Power Systems, Faculty of Electrical and Information Engineering, Graz University of Technology, Graz 8010, Austria

* Correspondence: benoit.bletterie@gmail.com or benoit.bletterie@apg.at (B.B.); herwig.renner@tugraz.at (H.R.); Tel.: +43-664-88342832 (B.B.)

Received: 14 February 2018; Accepted: 12 March 2018; Published: 14 March 2018

Abstract: The integration of large amounts of generation into distribution networks faces some limitations. By deploying reactive power-based voltage control concepts (e.g., volt/var control with distributed generators), the voltage rise caused by generators can be partly mitigated. As a result, the network hosting capacity can be accordingly increased, and costly network reinforcement might be avoided or postponed. This works however only for voltage-constrained feeders (opposed to current-constrained feeders). Due to the low level of monitoring in low voltage networks, it is important to be able to classify feeders according to the expected constraint in order to avoid the overloading risk. The main purpose of this paper is to investigate to which extent it is possible to predict the hosting capacity constraint (voltage or current) of low voltage feeders on the basis of a large network data set. Two machine-learning techniques have been implemented and compared: clustering (unsupervised) and classification (supervised). The results show that the general performance of the classification or clustering algorithms might be considered as rather poor at a first glance, reflecting the diversity of real low voltage feeders. However, a detailed analysis shows that the benefit of the classification is significant.

Keywords: hosting capacity; reactive power; voltage control; low voltage feeders; classification

1. Introduction

In order to meet long-term objectives in terms of CO_2 emissions reduction and supply security, the share of renewable generation in the European electricity mix has been steadily growing in the last 10–15 years and must increase further. In particular, wind and solar photovoltaic (PV) power have established themselves as ones of the most promising renewable energy resources, providing a non-negligible share of the overall generation in some regions. This renewable generation has been integrated at the transmission level for the largest wind parks or at distribution level, for most of wind and PV generation. This evolution of the power system results for example in reverse power flows and new power infeed present down to the low voltage level.

However, the integration of large amounts of generation into distribution networks faces some limitations and in some networks the hosting capacity is exhausted [1]. The main two constraints that usually limit the amount of generation that can be hosted by distribution feeders are the maximum admissible voltage and currents. In particular, the voltage rise caused by the power infeed from distributed generators is often considered as one of the most limiting constraints [2]. Having recognised this problem, connection standards or guidelines have been published in most European countries (e.g., [3–5] in Germany and Austria). Besides specifying clear and transparent rules to assess the connection of generation to the distribution network, these guidelines have introduced new possibilities or new requirements. These new possibilities (smart grid solutions) allow using new functionalities of modern generators for a more cost-effective network integration of the generation. In particular, several voltage control options based on reactive power (volt/var control), have been proposed in different standards. By implementing voltage control with e.g., distributed generators, the voltage rise caused by the power infeed can be partly mitigated, and the network hosting capacity can be increased accordingly. As a result, costly network reinforcement might be avoided or postponed.

Almost 10 years after the introduction of new requirements for generators in the connection guidelines or standards, the added value of these new features offered by modern generators are only used to a limited extent. One of the reasons for this limited deployment is probably the fact that, despite the large number of research works on e.g., voltage control and reactive power control, clear practical deployment recommendations are missing.

In particular, the actual potential of some smart grid solutions has not been analyzed on a systematic way. Many research efforts have been devoted to develop new control concepts which are more or less complex [6–8], and to investigate their performance, often on the basis of a case-study. In order to be able to conclude regarding the general performance (e.g., achievable hosting capacity increase) and potential (e.g., share of feeders with a substantial benefit) of a specific control, "representative" (wording see Table 1) networks are necessary.

In this context, the search for "representative" networks has been addressed in several works in the last years. Table 1 provides an overview of the main characteristics of the previous works in this field: the scope, the objective, the data set used, the statistical method, the number of parameters and the number of clusters.

A close look at the relevant studies around the topic of feeder classification shows a rather inhomogeneous picture, as visible with the wording which is used (see "target" in Table 1, e.g., "generic", "typical", "reference", "representative", "prototypal", "common" or "benchmark"). Despite this diversity in the wording, most of these studies have the same basic objective: to identify a set of "typical/representative" feeders to perform "generic" network studies.

Moreover, the vast majority of these studies use a clustering analysis to identify these "representative" feeders despite the inaccurate wording (e.g., confusion between clustering and classification—see Section 2.2). In all the studies using clustering, the popular k-means algorithm has been used. While about half of the studies mentioned in Table 1 were dedicated to LV voltage and half to MV voltage, a fundamental difference between these studies is the system boundary. Most of the studies (eight) considered feeders while the others (four) considered networks as observations. Due to the potential large inhomogeneity of feeders belonging to a same network (a primary substation can for example supply purely urban feeders and purely rural feeders at the same time), a classification at feeder level appears to be more pertinent. This approach has been, as previously mentioned, followed by most of the studies analyzed.

Another important difference between the studies on feeder classification is the size of the data set used as input (from less than 200 [9] to about 88,000 [10], see Table 1). As presented in Section 2.1, about 24,000 feeders have been used in this study, which corresponds to the upper range of the previous studies.

One of the most important parameters of clustering is the number of clusters, which needs to be set for k-means clustering at the beginning. In the considered studies, different metrics have been used to quantify the clustering performance and select the "optimal" or "appropriate" number of clusters. More information is presented in Section 2.2.2. The number of clusters specified in the considered studies varies from three to 35, with a median of five clusters. In this study, similar criteria have been used to try to select a suitable number of clusters, and the best trade-off has been obtained for less than 10 clusters. However, a difference between this and the other studies was the main classification objective: to identify with the highest possible confidence the constraint limiting the hosting capacity of LV feeders (voltage or current constraint). For this reason, the clusters (disregarding their numbers) have been mapped into the two categories. The detailed methodology used for the clustering is presented in Sections 2.2.2 and 3.4.

Table 1. Main characteristics of existing studies on distribution feeder/network classification.

Study	Scope	Target	Data Set	Statistical Method [1]	# of Param.	# of Clusters
Willis et al., 1985 [11] (US)	MV feeders	"representative feeders"	1350	k-means	11	10
Schneider et al., 2008 [12] (US)	MV feeders	"prototypal feeders"	575	hierarchical	35	24
Nijhuis et al., 2015 [10] (NL)	LV feeders	"most common types of feeders"	88,000	fuzzy k-medians	945→8 [2]	8
Kerber, 2011 [13]/ Lindner et al., 2016 [14] (DE)	LV networks	"reference networks"	86/358	"qualitative and statistical analysis"	3	7/5
dena, 2012 [15] (DE)	LV and MV networks	"network area classes"	LV: 177 MV [3]: 20	k-means	4	11 [4]
Dickert et al., 2013 [16] (DE)	LV feeders	"benchmark feeders"	n/a	k-means	6	18
Broderick und Williams, 2013 [17] (US)	MV feeders	"representative feeders	3 000	k-means	12 [5]	22
Gust, 2014 [18] (DE)	LV networks	"reference networks"	203	k-medoids	4	20
Cale et al., 2014 [19] (US)	MV feeders	"representative feeders"	1295	k-medoids/random forest	16	12
Li und Wolfs, 2014 [20] (AU)	LV and MV feeders	"representative feeders"	LV: 8858 MV: 204	hierarchical	LV: 7 MV: 6	LV: 8 MV: 9
Walker et al., 2015 [21] (DE)	LV networks	"cluster reference grids"	>20,000	k-means	5 [5]	10
Dehghani et al., 2015 [9] (IR)	MV feeders	"representative feeders"	195	self-organized maps	7 [5]	9

[1] Further methods are additionally used in some cases (e.g., principal component analysis in [17,19] for visualization purpose); [2] A large number of clusters has been selected (94). Feeder properties have been only provided for the 8 largest clusters (representing only about one third of the whole population of feeders); [3] HV networks have also been considered (out of scope here); [4] The clusters are further grouped within five load density areas; [5] after parameter reduction (based on e.g., correlation analysis).

Regarding the input data, the previous studies differed significantly in terms of feeder parameters (variables) used for the clustering. However, a common group of parameters used in several studies has been identified:

- average distance between nodes
- average impedance at the point of connection
- total cable length
- feeder length
- cable or line rating

These parameters, together with many others have been used in this study. The whole set of parameters as well as a correlation analysis are presented in Sections 2.1 and 3.1. Finally, the vast majority of the previous studies did not perform a full validation of the clustering results. Even if the clustering results themselves are good, it is of prime importance to ensure that the clustering results are relevant for the considered question, in our case the distinction between voltage and current-constrained feeders. In this study, the feeder category (in our case voltage and current-constrained feeders) is known since the hosting capacity has been determined (see Section 2.1). This information has been used as an input for the classification (see Sections 2.2.3 and 3.3) or as external validation for the clustering (see Sections 2.2.2 and 3.4).

The main purpose of this paper is to investigate to which extent it is possible to predict the behavior of LV feeders in terms of hosting capacity constraint (voltage or current constraint) on the basis of a large set of real LV feeders. The main motivation behind this work is to allow DSOs to easily discriminate between LV feeders in which reactive power-based voltage control can help in increasing the hosting capacity, and LV feeders for which reactive power control would be counterproductive.

Two machine-learning techniques are implemented and compared: clustering (used in most of the previous studies) and classification. The results show that the general performance of the classification or clustering algorithms is limited and might even be considered as rather poor at a first glance. All the data exploration attempts showed that the feeder population exhibit continuous distributions and that no natural cluster structure has been observed.

However, even with a rather poor general performance, the added value of the feeder classification can be considered to be significant. By reliably identifying a sub-group of the feeders (voltage-constrained feeders) with a high confidence level (sensitivity close to 100%), voltage control concepts can be deployed by DSOs in the target networks, without the need for additional studies.

2. Method and Data Set

The method followed in this study is mainly used on data analysis and statistical methods. The basic research design is based on explanatory research following the objective to formulate a posteriori hypotheses by analyzing the data with different data exploration techniques.

The data used in this study consists in a large dataset of real low voltage networks as explained in Section 2.1. Before performing the main part of the analysis (classification and clustering), several data processing-steps have been necessary. An overview of all these steps is provided in Section 2.2. This subsection presents in particular the core of the analysis based on the two machine learning techniques (classification and clustering) with more details (Sections 2.2.2 and 2.2.3).

2.1. Available Data Set

The analyses presented in this paper have been performed on a set of about 7300 low voltage networks from a DSO supplying a geographical area in Austria which is dominantly rural. The data has been collected and pre-validated in the frame of the IGREENGrid project [22,23].

The network data have been exported from the geographical information systems (GIS) of the DSO and imported into the simulation software DIgSILENT PowerFactory [24] with the exchange format DGS. The automation of the data import and pre-validation has been implemented in the DIgSILENT Programming Language (DPL) and in Python [25]. In addition, load-related data (e.g., annual elecctricity consumption) has been obtained from other data sources such as metering databases.

As a first step after the data import, feeders have been defined for each LV network and a set of automated scripts has been executed to validate the data. Feeders with unexpected properties (non-radial feeders, special purpose feeders) have been eliminated from the analysis, leading to a set of more than 24,000 LV feeders.

In a second step, two types of parameters have been computed for every single feeder:

- descriptive indicators or explanatory variables
- hosting capacity related indicators.

The purpose of the first type of indicators (descriptive indicators) is to provide variables to be used for the statistical analyses that are in the focus of this paper: clustering and classification. All these parameters can be computed with tools generally available for DSOs (e.g., GIS). The purpose of the second type of indicators is to provide some valuable quantitative information on the capacity of the investigated supply area to host distributed generation while fulfilling the planning criteria (see Section 1). A detailed analysis of the networks considering different scenarios in terms of e.g., generation distribution has been presented in [23] and is not in the scope of this paper.

In this paper, the focus is laid on classifying LV feeders. For this purpose, hosting capacity-related indicators have been used. The calculation of these hosting capacity-related parameters is more complex than the first type of parameters. For this, an adapted Newton-Raphson algorithm has been implemented in the simulation environment [26]. This tool increases the generation connected to the considered feeder until one of the planning limits (voltage or current constraint) is reached. The amount of generation obtained to reach this limit is considered as the feeder hosting capacity. For this hosting capacity computation, different scenarios have been considered. In this paper, only three scenarios that are related to the distribution of the generation along the feeders are considered: "uniform", "weighted" and "end of feeder—eof"). The scenario "uniform" assumes a uniform distribution of the generation long the feeder, the scenario "weighted" assumes that the size of the generator is proportional to the annual electricity consumption of the LV-customers (motivated by e.g., self-consumption) and the scenario "end of feeder" assumes that the generation is located at the end of the feeder, which represents a worst-case. For all these scenarios, load an generation are assumed to be balanced (equally distributed over the three phases) Further scenarios have been defined and analyzed in [23]—a complete overview of these scenarios is provided in [26].

The initial set of more than 80 parameters has been reduced to a subset of 12 parameters, which are considered to be relevant to characterize feeders in terms of hosting capacity, and in particular in terms of being prone to voltage or current constraints. This parameter set has been selected by reviewing the relevant literature (Table 1, [9–21]). These parameters are given in Table 2 and in the explanatory text below (Equations (1) to (3)).

Table 2. Feeder parameters (variables) to be used for the clustering analysis and the classification.

Feeder Parameter (Variable)	Description
ADTN	Average Distance To Neighbors (m)
ANON	Average Number of Neighbors (-)
LastBusDist.	Last Bus Distance: path length between secondary substation and the bus with the lowest voltage (last bus [1]) under the considered scenario [2] (m)
FeederLength	Feeder length: largest distance between the secondary substation and any of the busses (m)
TotLineLength	Algebraic sum of the cable or overhead line length in the whole feeder (m)
km/load	Quotient between TotLineLength and the number of loads in the feeder (km)
Rsc	short-circuit resistance at the last bus [2] (Ω)
Rsum	Equivalent sum resistance (real part of the impedeance): see explanation below and Equation (1) (Ω)
kWm	see Equation (2) (kWm)
kWΩ	see Equation (3) (kWΩ)
In_avg	Average rated current for all the cable or lines of the feeder (A)
In_max	Maximum rated current for all the cable or lines of the feeder (A)

[1] the "last bus" is the bus with the lowest voltage in the feeder under the considered scenario; [2] the three considered scenarios are: "uniform", "weighted", and "eof (end of feeder)".

The concept of equivalent sum-impedance *Rsum* which can be computed on the basis of the feeder topology and the impedance of each line segment, captures the combined information about the feeder impedance and the load distribution along the feeder [27]. For a simple case consisting of one radial feeder with a uniform distribution of N loads without laterals, the equivalent sum impedance can be computed by (1):

$$Rsum = \frac{1}{N} \sum_{k=1}^{k=N} \left(R'_k \cdot l_k \cdot (N - k + 1) \right), \tag{1}$$

where R'_k and l_k are the specific resistance and length of segment k.

The parameters kWm and kWΩ can be computed by Equations (2) and (3) respectively [26]:

$$\mathrm{kWm} = \sum_{k=1}^{k=N} (P_k \cdot d_k), \tag{2}$$

$$\mathrm{kW\Omega} = \sum_{k=1}^{k=N} (P_k \cdot R_k), \tag{3}$$

where P_k is the power of the load connected to node k, d_k the distance between node k and the secondary substation and R_k the short-circuit resistance at node k.

2.2. Presentation of the Concept Used for the Feeder Clustering and Classification

As stated in the introduction, one of the main purpose of the analyses is to implement a feeder classification and to investigate to which extent the feeder behavior can be predicted on the basis of statistical parameters which are usually available for DSOs. In this study, the feeder behavior means whether feeders tend to experience over-voltage or over-current when increasing the generation along the feeders to reach the hosting capacity.

In addition to the feeder classification (supervised learning), clustering (unsupervised learning) has also been implemented since it has been used in almost all the previous work analyzed in Section 1. The results of both analyses (classification and clustering) are compared and discussed (Sections 3.3 and 3.4).

Before explaining the general concept, the basic principles of clustering and classification are recalled here:

- Clustering consists in grouping a set of observations into clusters, on the unique basis of some observed variables, and without knowing a priori the number of clusters. Observations within a cluster should have at the same time a high similarity between each other and a high dissimilarity with observations in other clusters.
- Classification consists in finding a way to identify to which sub-set of observations (category or class) a new observation belongs. This is done on the basis of an algorithm trained on a set of data containing observations whose category or class is known.

2.2.1. General Concept

The analysis presented in this paper has been performed on the basis of the concept shown on Figure 1 and explained below:

I. Data import

In a first step, the data (network data, load data) is imported from different databases (see [23,28]. After a successful import of the network data into the simulation environment (PowerFactory), feeders are automatically defined in order to perform the analysis at feeder level. These feeders are considered as the observations to be clustered or classified.

II. Computation of feeder parameters (variables)

In this second step, all the feeder parameters which have been previously defined (see Section 2.1) are computed through automated scripts [26]. These variables can be divided into two families of variables:

- Descriptive variables (predictors)
- Hosting capacity-related variables (categories or classes)

III. Data validation (plausibilisation)

In this step, the data feeder parameters are carefully analyzed and feeders (observations) having erroneous or unrealistic values are eliminated. In this work, extensive efforts have been devoted to the data validation in order to ensure that only erroneous feeders (bad data) are eliminated. Classical outlier removal methods such as those based on boxplots (using percentiles or other statistical indicators to measure the dispersion from the median) cannot simply be used on multi-dimensional data sets. In some previous works, special outlier removal methods have been used (e.g., error ellipse method [19]). However, fixing a confidence interval (or removing the 1% “more extreme” feeders) results in removing observations which are not usual but are not erroneous. By doing so, the population of feeders is distorted, which can lead to a better classification or clustering performance, and finally to wrong interpretations. For these reasons, only erroneous data has been eliminated from the original data set.

IV. Data preparation

For some of the methods used in this paper (e.g., correlation analysis, see Section 3.1), a scaling of the data is necessary. When necessary, the data have been normalized (subtracting the average value and dividing by the standard deviation).

A further way to prepare data is transforming them to enhance differences between observations (using e.g., logarithm or square root transformation). As in several previous works [17,19] the principal component analysis (PCA) has been used for different purposes: for visualization purpose or to limit redundancy in the data.

V. Data exploration

After the data preparation, the data has been analyzed with classical data exploration techniques such as variable correlation, predictor importance, and variable (feature) selection. One of the objectives of this analysis is to analyze the dependencies between variables and select a limited subset of variables, which best reflects or explains the data structure. Data reduction was not strictly required in this work and has only been applied when a benefit could be obtained.

The detailed statistical analysis of the feeders has been presented in [23] and is therefore not in the scope of this paper. A few results are however provided in Section 3.1.

The next step consists of the analysis itself, which is based either on clustering or on classification. They are described in dedicated sections (Sections 2.2.2 and 2.2.3).

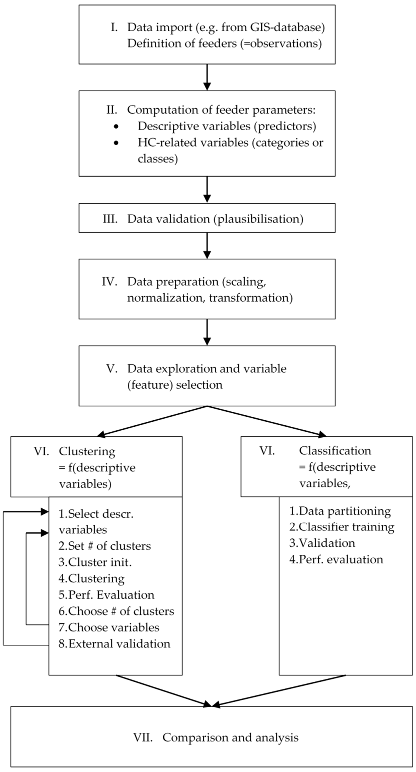
I. Data import (e.g. from GIS-database)
Definition of feeders (=observations)
II. Computation of feeder parameters:
• Descriptive variables (predictors)
• HC-related variables (categories or classes)
III. Data validation (plausibilisation)
IV. Data preparation (scaling, normalization, transformation)
V. Data exploration and variable (feature) selection
VI. Clustering = f(descriptive variables)
1.Select descr. variables
2.Set # of clusters
3.Cluster init.
4.Clustering
5.Perf. Evaluation
6.Choose # of clusters
7.Choose variables
8.External validation
VI. Classification = f(descriptive variables,
1.Data partitioning
2.Classifier training
3.Validation
4.Perf. evaluation
VII. Comparison and analysis

Figure 1. General concept used to perform the clustering and classification.

2.2.2. Feeder Clustering (Non-Supervised Learning)

Clustering is a non-supervised classification (labels or classes are not known) and can be considered as an exploratory data analysis tool. The most popular clustering algorithms are hierarchical clustering and algorithms from the k-means family. From all the relevant papers presented in Table 1, 11 from 12 use clustering: only two are based on hierarchical clustering and nine are based on the family of k-means algorithms (with some variations such as k-medoids or fuzzy k-median).

Hierarchical clustering consists in grouping observations into clusters with an agglomerative or divisive algorithm, based on their proximity. Besides the distance used to evaluate the proximity between observations, the linkage criterion, which defines how to compute the distance between clusters, plays an even more important role. One advantage of hierarchical clustering is that the number of clusters does not need to be a priori set. The partitioning can be directly analyzed by varying the level of resolution (on a so-called dendrogram).

K-means clustering is a heuristic that converges to a local optimum (e.g., minimization of the total distance between observations and their respective cluster centers) for a given number of clusters. One advantage of k-means clustering over hierarchical clustering is that it can be used for large data sets. The main disadvantages are that the number of clusters must be a priory specified, and that the result might be sensitive to the initialization (initial cluster centers which are usually chosen randomly).

There is no consensus whether hierarchical or k-means clustering is better- The decision is usually taken on the basis of the size of the data set. In fact, evaluating whether the clusters identified by a given clustering algorithm are good or not for the considered purpose requires analyzing the data and the results in detail, with domain experts.

Although hierarchical clustering would be feasible for the data set size (24,000 observations) from a computation point of view, k-means clustering has been used in this study to allow a comparison of the results with previous research works. In fact, an extension of k-means clustering called fuzzy c-means (FCM) clustering has been implemented. Fuzzy c-means clustering is a clustering concept introduced in 1981, which uses membership grades instead of hard assignments to cluster observations. With this concept, data points close to the center of a cluster will have a high membership value to that cluster, and low membership values to other clusters.

A variation of k-means clustering is the k-medoids clustering algorithm, which has been used in [18,19]. One of the main differences between these two partitioning clustering algorithms is that the center of the clusters are real observations (instead of a calculated center (centroid) for k-means).

In [28], the authors propose a clustering procedure based on four major steps (see Section 2.2.1):

- Feature selection or extraction
- Clustering algorithm design or selection
- Cluster validation
- Result interpretation

Most of the studies on feeder/network classification presented in the introduction follow to some extend this procedure. However, the validation usually only consists in a comparison between the results of different clustering variables or different number of clusters. None of the considered studies has evaluated to which extent the result of the clustering fulfills the expectations, due to the missing ”prior information”. In this study, this information is available and has been used for validation.

The results of the clustering analysis are summarized in Section 3.4. In particular, an external validation of the clustering is presented.

2.2.3. Feeder Classification (Supervised Learning)

Classification is a supervised machine-learning technique, which significantly differs from clustering (which is, as previously mentioned, unsupervised). It consists in building an algorithm, which is able to identify to which category or class, observations belong. This algorithm is tuned during a learning process with a training dataset, and then used to predict the category of another set of observations.

Among the different techniques usually used (e.g., discriminant analysis, decision tree learning, support vector machine or neural networks), decision tree has been selected based on several criteria for this study. One of the advantages of decision trees (or classification trees) is that the results (trees) are easy to interpret. An important step of classification is to evaluate the classifier performance. The most straightforward method is to compare the predicted class to the true class by building a so-called confusion matrix, from which several statistical indicators such as accuracy, specificity, etc. can be derived. However, this method has the great disadvantage that overfitting cannot be detected, since the validation and the training are done on the same dataset. Alternatively, cross-validation provides a way to test the classifier on a test dataset, whose observations were not used for the training. In this study, the k-fold validation has been used. With the k-fold cross-validation, the original data set is randomly partitioned into k subsets of equal size, from which one is kept for the validation, and k-1 are used for the training. The cross-validation process is then repeated k times and then, an average performance is computed. In this study, k has been set to 10 (10-fold cross-validation). The detailed concept of the classification used in this study is explained in Section 3.4.

3. Results

The results of the analyses are summarized in this section. The first subsection provides a few examples of statistical characterization (descriptive statistics) of the feeder population. The second sub-section provides some insights into the approach used to select the parameters used for classifying and clustering the feeders. The results of the classification and of the clustering are summarized in the last two subsections.

3.1. Statistical Analysis of the Feeders

Before starting the analyses on classification and clustering, the population of feeders has been explored and analyzed with different statistical methods. In this section, only a few examples are provided. A more detailed analysis has been presented in [23].

Since the R/X ratio of distribution feeders significantly affects the potential of reactive power-based voltage control, a closer look at the R/X distribution has been taken. Figure 2 shows the distribution of the feeder R/X ratio at the end node for the scenario "uniform" (see Section 2.1). For each bar of the histogram (length class), the share of voltage-constrained feeders is show in blue, the share of voltage and current-constrained feeders in magenta and the share of current-constrained feeders in red (for the scenario "uniform"—see Section 2.1).

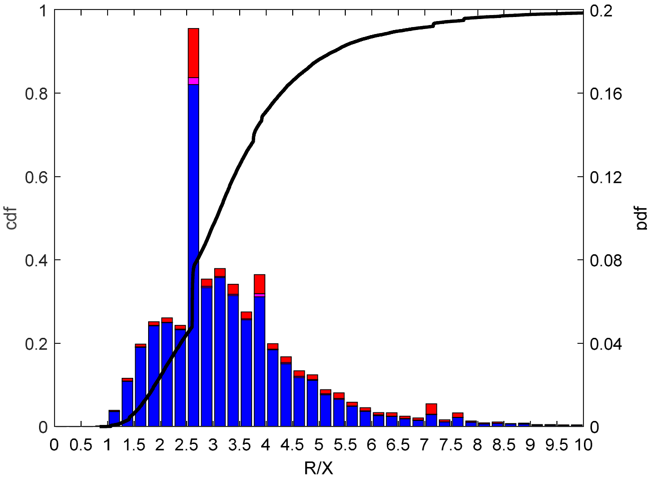

Figure 2. Distribution of the R/X ratio at the end node for the uniform generation distribution.

This R/X ratio plays an important role in the effectiveness of reactive power control for limiting the voltage rise as shown in Equation (4) [29] (for one generator connected at the end of the feeder):

$$\Delta U \approx \frac{R \cdot P}{U_N^2} \cdot \left(1 - \frac{\tan\varphi}{R/X}\right), \tag{4}$$

where ΔU is the voltage rise caused by the power infeed, U_N the nominal voltage, R and X the feeder resistance and reactance, P the injected active power, and φ the displacement angle. By consuming reactive power (increasing tan φ), the voltage rise can be partly compensated: the lower the R/X ratio, the more effective the control.

This figure shows that the R/X ratio is almost always above 1, which is in accordance with the usual assumption that LV feeders have a "large" R/X ratio. Only about 21% of the feeder have a R/X ratio below 2.4, which allows a compensation of the voltage rise by 20% with cos ϕ = 0.90. The large peak for a R/X ratio of 2.6 corresponds to the most common cable type Al 150 mm^2. Another aspect to consider in this context, is that the reactive power consumption leads to an increase of the current, which might bring originally voltage-constrained (without reactive power-based voltage control) to be current-constrained (with control).

The increase of current due to the reactive power flows caused by the voltage control, can be estimated with Equations (5) and (6) ((5) is obtained by neglecting the voltage drop which is in quadrature with the voltage [30], and (6) is derived from a first order approximation of the current, considering small voltage variations):

$$\Delta U_{max} \approx \frac{R \cdot P_{HC}^{noQ}}{U_N^2} \approx \frac{R \cdot P_{HC}^{Q}}{U_N^2} \cdot \left(1 - \frac{\tan\varphi}{R/X}\right), \tag{5}$$

$$\frac{I_{max}^{Q}}{I_{max}^{noQ}} \approx \frac{1}{\left(1 - \frac{\tan\varphi}{R/X}\right) \cdot \cos\varphi}, \tag{6}$$

where ΔU_{max} is the maximum allowed voltage rise (3% for LV feeders according to [5]), defining the hosting capacity for voltage-constrained feeders, R and X are the feeder resistance and reactance, P_{HC}^{noQ} and P_{HC}^{Q} the hosting capacity without and with reactive power control and U_N the nominal voltage. Putting this simple equation in relation with the R/X statistics shown on Figure 2 lead to the following conclusion: about 50% of the feeders have a R/X ratio below 3, which results in an increase of the current by a factor 1.33 (+33%). This means that these feeders could only fully benefit from a reactive power-based voltage control, if the maximum loading (without control) is below 1/1.33 = 0.75 p.u.

The remaining of this section is devoted to the main objective of this study: perform and analyze feeder classification and clustering methods.

3.2. Parameter Selection and Data Reduction

Before performing the clustering analysis and classification (see Section 2.1), several data exploration techniques have been used to get a better understanding about the relation between the parameters intended to be used for clustering and classification. Although the number of variables (feeder parameters) available for the analysis does not require the selection of a subset (i.e., the use of data reduction/feature selection techniques), the proximity between these variables has been analyzed in a first step. For this, three methods have been used:

- Correlation analysis
- Variable clustering (details not shown here)
- Principal Component Analysis (PCA)

In order to investigate the correlation between parameters, the Spearman correlation has been computed for the whole data set (for all the feeders) and a threshold of 0.70 has been used to identify significantly correlated variables.

Figure 3 shows the correlation between parameters (only the lower triangle is shown). Red crosses indicate a high correlation (>0.70), and blue crosses a low correlation (≤0.70). The set of parameters with a correlation lower than the considered threshold of 0.70 is indicated with blue squares: three "poorly correlated parameters" and two more which can be taken out of the group of correlated parameters.

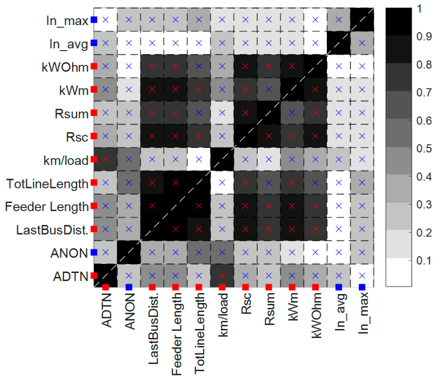

Figure 3. Correlation map (Spearman correlation). Red and blue crosses show a correlation higher or lower than the threshold of 0.70.

Finally, a principal component analysis has been performed in order to validate the variable selection and allow a graphical representation of the clustering result.

Principal component analysis consists in an orthonormal transformation into components which are a linear combination of variables. The components are selected in a way to maximize the variance explained by the first components. In this case, the first two principal components lead to a ratio of explained variance of about 66%. These two first components have the following main parameters (see Figure 4): component 1 is dominated by the parameter Rsum (and its correlated parameters such as the three distance metrics) while component 2 is dominated by the parameter *km*/*load*, *ANON* (and *ADTN*). The first component therefore mainly relates to the feeder length and impedance, and the second component to the structural properties (different metrics related to load density).

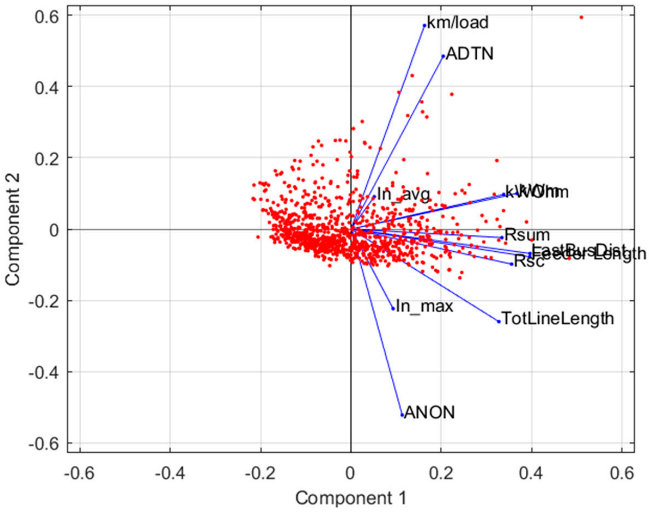

Figure 4. Projection of the parameters on the first two components obtained from a principal component analysis, including a random subset (1000) of the feeder data set (red points).

The results from all these analyses are coherent and confirmed the parameter selection. The final set of "poorly correlated variables" has been determined by selecting the parameters with the highest variance within a cluster (within a group of correlated parameters). These parameters are:

- *In_max*
- *In_avg*
- *km/load*
- *ANON*
- *Rsum*

3.3. Classification of LV Feeders

As explained in Section 2.2, the data set of about 24,000 LV feeders has been classified, using as explanatory variables the feeder parameters mentioned un Section 2.1 (Table 2) and as category the following characteristics:

- voltage-constrained feeders
- current-constrained feeders

Among all the different supervised machine-learning techniques such as neural networks or discriminant analysis, classification trees have been used for their simplicity and interpretability. The software package Matlab has been used for this purpose.

In a first step, a fully-grown classification tree (without restriction on the tree depth, i.e., a deep tree) has been built. It uses all the 12 parameters (or predictors) available. Before looking at the tree properties, the predictor importance can be evaluated. It quantifies the contribution of each variable (predictor) to split the tree. Figure 5 shows the predictor importance for the fully-grown tree: the most important variable is the *kWm*, followed by the parameters *kWOhm*, *FeederLength*, *In_avg* and *LastBusDist* (about 10 times less important).

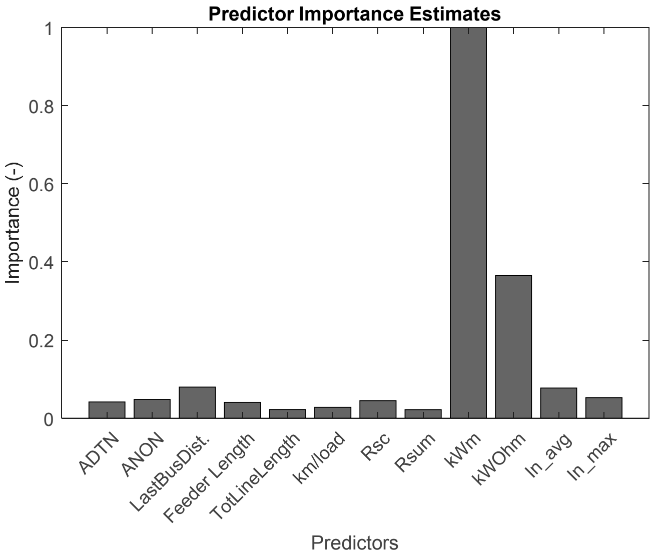

Figure 5. Estimation of the importance of each predictor (variable) for the fully-grown tree.

Deep classification trees (such as the one obtained from this first attempt (fully-grown tree)) are known to be prone to overfitting, which means that the good fitting obtained on a training set cannot be reproduced with a different set (testing set). In such cases, the tree has memorized a learning set instead of learning the general data structure.

Several alternatives are possible in order to avoid overfitting. One possibility is to constrain the tree building process by specifying a maximum number of splits or a minimum leaf size. The drawback of this method is that the constraints must be set from the beginning, i.e., before having some good understanding of the data. Another widely used possibility is to prune the tree (merge leaves) in order to reduce its complexity [31]: this is known as cost-complexity pruning. In order to evaluate the performance of a classifier, two generic indicators are widely used: the resubstitution error and the cross-validation error. The first one is simply evaluated by counting the misclassified observations on the whole data set while the second one requires to separate the data set into a training set and a testing set (usually with the proportion 90%/10%). The advantage of using the cross-validation error is that over-fitting can be detected.

Figure 6 shows the misclassification errors (resubstitution and cross-validation) as a function of the pruning-level: a low number of splits (only two) is enough to obtain the best achievable classification performance (any increase of the number of splitting does not reduce the cross-validation error and might lead to over-fitting).

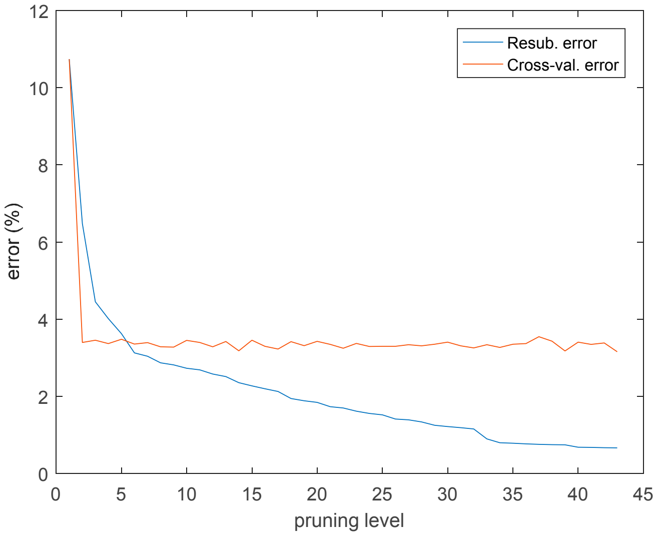

Figure 6. Evaluation of the optimal pruning on the basis of the cross-validation error.

This figure also shows that a low classification error (cross-validation error) can be achieved (about 3.4%). This result should however be carefully interpreted. Indeed, the data set is rather heavily unbalanced (skewed) with about 90% of the observations falling in one category (voltage-constrained feeders), and about 10% falling in the other category (current-constrained feeders)—see Section 3.1 [23]. This means that a random guess would already lead to a rather low misclassification error (10%). In order to consider this, several options are possible. The first one is to partition the data set (training and testing sets) in order to have a balanced proportion of both classes. The second one is to adjust misclassification costs and force the classification to be "equally good" for both categories. In this work, the second option has been selected and some cost weights have been used with the ratio between voltage-constrained and current-constrained feeders (about 90/10). When doing so, the corrected cross-validation error increases to 15.3%.

The final classification tree which is obtained from this analysis (specifying misclassification costs to "balanced" the data set, and pruning to obtain the best compromise between complexity and accuracy) is shown on Figure 7.

In the considered application (classification of LV feeders into voltage-constrained and current-constrained feeders for network planning purpose), misclassification does not have the same impact for both categories (voltage- and current-constrained feeders).

Indeed, one of the options to extend the hosting capacity is to implement a reactive power-based voltage control. As explained in the introduction, this type of control allows reducing the voltage rise caused by the infeed from distributed energy resources, at the cost of an increase of the current resulting from the additional reactive power flow. Since the current is not observed with the considered voltage control concepts, the deployment of such solutions should be limited to feeders which are actually voltage-constrained and not current-constrained. For this reason, a heavily unsymmetrical cost function has been introduced to force the classifier to avoid misclassification of current-constrained feeders. With this cost function, none of the current-constrained feeder is classified as voltage-constrained feeder: there is no misclassified feeder I→U (true class = I, and predicted class = U). This is reached at the cost of a significantly higher misclassification, as seen in Table 3. In order to look into this, the confusion matrix can be used (Table 3): the left part of this table shows the confusion matrix of the pruned tree with "balanced" misclassification costs (i.e., reflecting the ratio between classes). The ratio of problematic misclassified feeders (I→U) is 3.3%. In order to bring this ratio down to 0, high (I→U) "selective" misclassification costs are specified. As a side effect, the ratio of (U→I) misclassified feeders increases strongly from 11.4% to 53.8% (right part of the table).

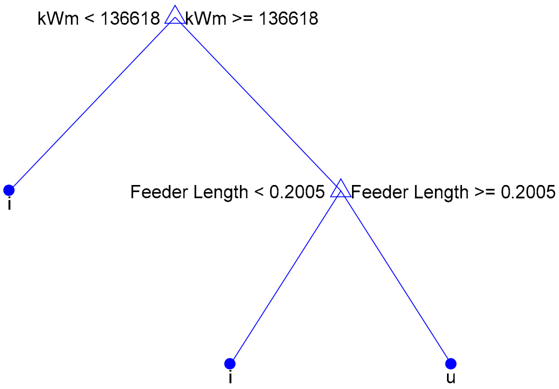

Figure 7. Graphical representation of the pruned tree (FeederLengh in km).

Table 3. Confusion matrix for a pruned classification tree with "balanced" (reflecting the data structure) and "unbalanced" (to avoid misclassification I→U misclassification costs).

Class	"Legend"		"Balanced" Misclassification Costs		"Selective" Misclassification Costs	
True→ Predicted ↓	**U**	**I**	**U**	**I**	**U**	**I**
U	TU [1]	FI [2]	88.6	11.4	46.2	53.8
I	FU [3]	TI [4]	3.3	96.7	0	100

[1] TU: true U-constrained feeders (normalized here to the actual number of U-constrained feeders); [2] FI: false I-constrained feeders (normalized here to the actual number of U-constrained feeders); [3] TI: true I-constrained feeders (normalized here to the actual number of I-constrained feeders); [4] FI: false U-constrained feeders (normalized here to the actual number of I-constrained feeders).

To further analyze the classifier performance, different indicators have been computed:

- Accuracy: probability of a correct classification among the data set (Equation (7))
- Sensitivity: the ability to classify correctly I-constrained feeders among the I-constrained feeders (Equation (8))
- Specificity: the ability to classify correctly U-constrained feeders among the U-constrained feeders (Equation (9))

- False positive rate (false alarm rate): the rate of U-constrained feeders which have been classified as I-constrained feeders (Equation (10)):

$$accuracy = 1 - Error = \frac{\#TU + \#TI}{\#U + \#I}, \tag{7}$$

$$sensitivity = \frac{\#TI}{\#I}, \tag{8}$$

$$specificity = \frac{\#TU}{\#U}, \tag{9}$$

$$false\ alarm\ rate = \frac{\#FI}{\#U}, \tag{10}$$

where # stands for number of elements in the corresponding subset.

Besides the indicators which are commonly derived from the confusion matrix (first three indicators), the fourth one plays an important role in this study as previously explained (reactive power-based voltage control should not be implemented in I-constrained feeders).

Table 3 shows that using high costs for the misclassification of I-constrained feeders allows reaching 100% sensitivity at the expenses of a rather high false alarm rate (53.8%), which represents a strong loss of potential. In fact, a trade-off between selectivity and false alarm rate must be found. In order to visualize this, Receiver Operating Characteristics (ROC) graphs can be used [32]. A ROC curve is a technique to visualize the performance of classifiers, and in particular, the trade-off between sensitivity (y-axis) and false alarm rate (x-axis). A random classifier would have a diagonal (0, 0)-(1, 1) as ROC-curve while a perfect classifier would follow the y and then the x-axes: (0, 0)-(0, 1)-(1, 1).

Figure 8 shows the ROC-curves obtained by varying the ratio of the misclassification costs between FI and FU between 10^{-4} and 10^{4}. For each misclassification cost, a classification tree has been grown and pruned as previously explained, and the ROC curve has been built. This figure shows that a sensitivity rate above 90% can be reached with a rather low false alarm rate (about 5%). However, in order to reach a sensitivity rate of 100%, the false alarm rate increases to about 54%. In fact, these ROC-curves are suitable for a decision-making process implemented in the frame of a probabilistic network planning by specifying a risk level.

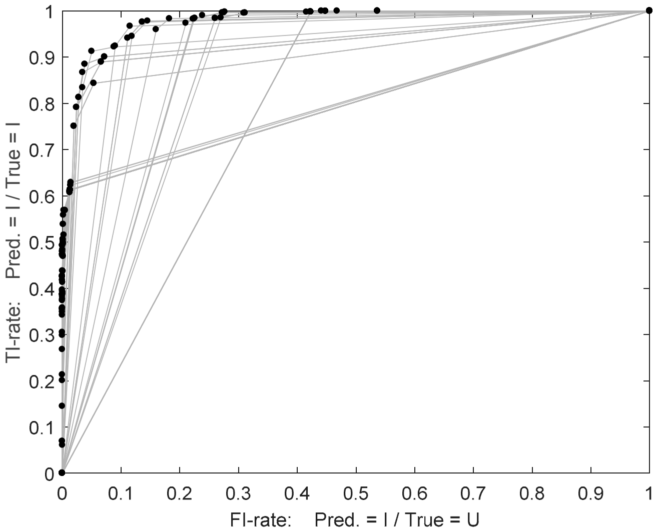

Figure 8. ROC curves (x: false alarm rate (FI)/y: sensitivity (TI)) obtained for different misclassification costs (cost ratio between 1:10^{-4} and 1:10^{4} for the FI and FU costs).

These results might be interpreted as a poor performance of the classifier (high false alarm rate). In fact, using very unsymmetrical misclassification costs (very high costs for FU) forces the classifier to exclude many U-constrained feeders due to only few I-constrained feeders, which are in a region of the variable space dominated by U-constrained feeders. This effect has been indeed observed for the feeder data set. Figure 9 shows the scatter plot for the whole data set, projected on the first two principal components obtained by a principal component analysis (PCA), which are dominated by the equivalent sum impedance *Rsum* (first component) and the parameters *km*/*load* or *ANON* (second component)—see Section 3.2.

In this figure, the observations (feeders) are colored according to the constraint (blue for voltage-constrained feeders and red for current-constrained feeders). This figure shows that current-constrained feeders seem to be located close to the lower left corner (small *Rsum* and small *km*/*load* or *ANON*) while voltage-constrained-feeders are found further from the origin. However, a careful look at the partial distributions (left and lower part of the figure) shows that there is a strong overlap in the region close to the origin, in which most of the feeders are found. This figure confirms the difficulty to discriminate between both classes (here on the sole basis of the two first principal components). In fact, the decision tree allows identifying those I-constrained feeders, which force to exclude numerous U-constrained feeders. Excluding them from the data set is however not a valid approach since these feeders are not outliers.

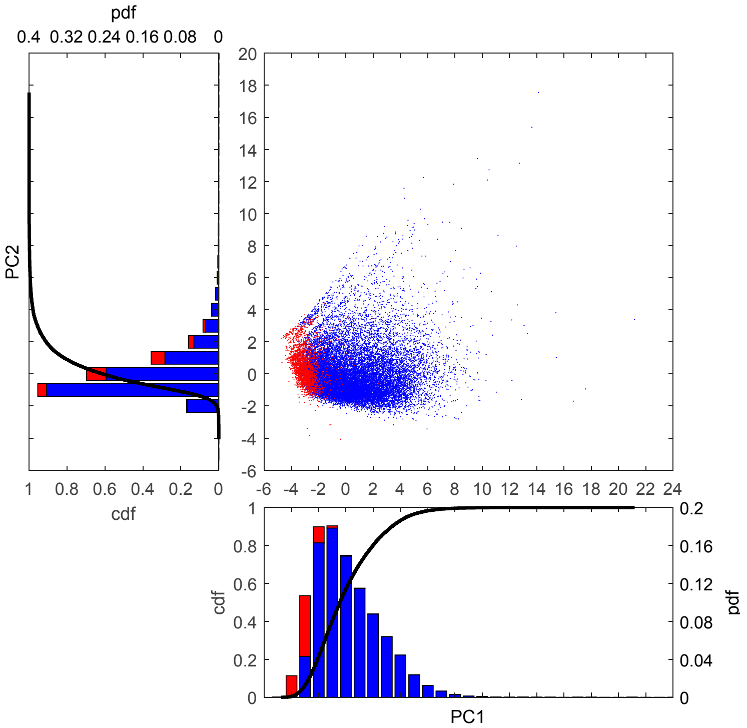

Figure 9. Scatter-plot of the two first principal components—coloring according to the constraint: blue: U-constrained feeders/red: I-constrained feeders.

Finally, a careful look at the misclassified feeders U→I, which necessarily lead to a loss of potential, shows that a great share of these feeders (70%) have a loading greater than 70%, as visible on Figure 10. On this figure, the right axis is for the probability density function (pdf—bars) and the left axis for the cumulative density function (cdf—curve).

These feeders are voltage-constrained but have a rather high loading, which means that they would probably turn to be current-constrained when reactive power-based voltage control is implemented due to the increased active and reactive power flows (see Section 3.1). This is, in fact, confirmed by analyzing the hosting capacity values obtained from the scenario with reactive power-based voltage control (see Section 2.1). As a result, only about one third of the misclassified feeders U→I remain voltage-constrained when implementing a reactive power-based voltage control. This means that the actual reduction of the potential due to the U→I misclassification ("false alarm") is reduced from 53.8% to about 18%.

As a conclusion, a decision tree classifier has been trained with the data set in a way to avoid over-fitting. Besides its generic performance which can be evaluated by the cross-validation for example, unsymmetrical misclassification costs have been introduce to avoid problematic errors (false U-constrained feeders). The side-effect of reaching a high sensitivity (close to 100%), is a rather high false alarm rate, which represents in fact a loss of potential in terms of feeders potentially benefiting from voltage control (the main application here). However, this loss of potential is limited since most of the affected feeders are in fact close to experience over-loading when implementing reactive power control to increase the hosting capacity.

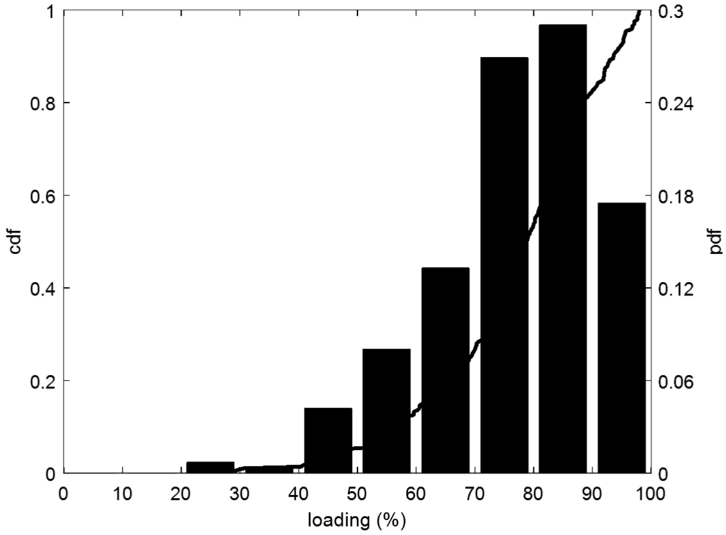

Figure 10. Evaluation of the lost potential with the distribution of the maximum loading for misclassified feeders (true class = U and predicted class = I).

3.4. Clustering of LV Feeders

As explained in Sections 1 and 2.2.2, all the studies mentioned in Section 1 are based on clustering analysis (i.e., process of grouping a set of observations into clusters). In these studies, the results of the clustering analysis have been analyzed and validated through an internal validation, whose purpose was in most cases to support the decision on the number of clusters to be used. In this paper, a clustering analysis has been conducted in a similar way as in most of the considered studies (see Table 1). The results of this analysis are analyzed through an external validation since the information of the class membership (voltage or current-constrained feeders) was available.

The feeders have been clustered with the k-means clustering (using the squared Euclidean distance), which is, as mentioned in Section 1 (Table 1), the most widely used clustering method.

The most important parameters of the clustering analysis which can have a significant impact on the results are:

- Variables used
- Number of clusters used

As in the previous studies, the variables used in the clustering analysis have been selected based on several analyses (e.g., correlation analysis, PCA). However, this variable selection is still subject, to some extent, to subjective decisions and is hard to fully justify. For this reason, the number of variables and the variables themselves have been varied. Regarding the number of clusters, there is no universal method to determine the "optimal" or "appropriate" number of clusters. Instead, there is a number of established methods supporting to some extent the decision.

In the studies reviewed in Section 1, different metrics have been used to quantify the clustering performance and select the "optimal" or "appropriate" number of clusters (R^2 (coefficient of determination) [16], sum of squared errors [10,12,33], silhouette value [10], cubic clustering [17] and Calinski-Harabasz [19] criterions).

In this study, the following two metrics have been evaluated: the silhouette value and the normalized sum of squared errors ($nSSE$).

The silhouette coefficient for each observation (here for each feeder i) is a measure of how similar that observation is to observations in its own cluster, compared to observations in other clusters:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{11}$$

where a_i is the average distance to all other objects in the cluster to which feeder i belongs to, and b_i is the minimum over all the clusters not containing feeder i, of the average distance between all the feeders in the considered clusters (not containing feeder i), and feeder i. The silhouette value lies between -1 and +1. A high silhouette value indicates that the observation i is well-matched to its own cluster, and poorly-matched to neighboring clusters.

The normalized sum squared error is the sum of the squared errors (here distance between observations and centroids (cluster centers)), normalized to the error obtained for a single cluster.

Figure 11 shows the clusters obtained by specifying the number of clusters to 3, 4 or 5. The left part shows the clusters on a scatter plot, using the first two principal components to allow an easy visualization (as in [17,19]). The obtained normalized sum squared error ($nSSE$) is given on the top of each scatter plot. The right part of the figure shows the corresponding silhouette plots: the silhouette value is computed for each single observation and shown in a sorted way for each cluster (the average value over all the clusters is given above the silhouette plots).

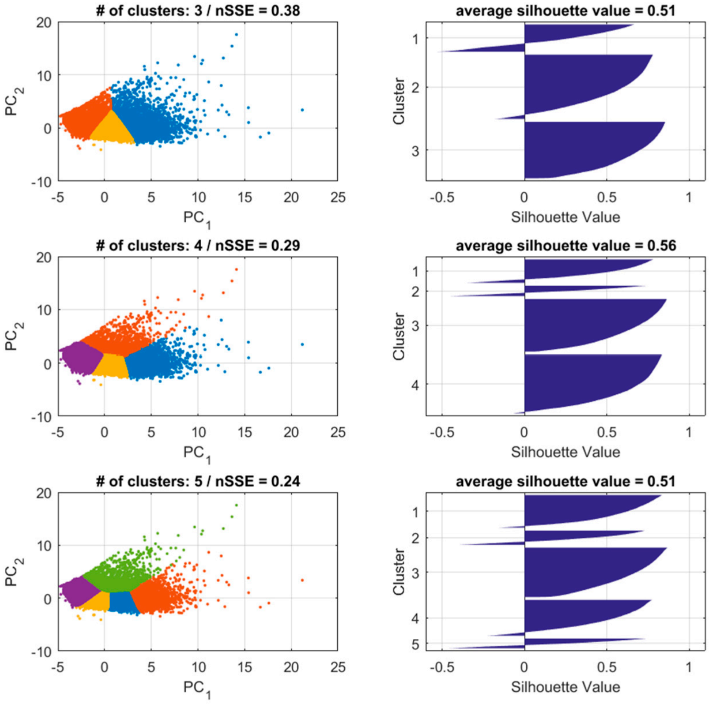

Figure 11. Scatter-plots for the first two principal components (PC_1 and PC_2) and silhouette plots for 3, 4 or 5 clusters.

This figure shows that, as expected, the normalized sum squared error decreases when the number of clusters increases (0.38 for three clusters and 0.24 for five clusters). However, the silhouette value does not show such a monotonic behavior: the best (highest) silhouette value among these three cases is obtained for four clusters (0.56). This means that, in this case and for this metric (silhouette), increasing the number of clusters does not necessarily lead to a better clustering result (in terms of how close observations are within a cluster and how far they are from other cluster).

The cluster plots also show that most of the observations are very close to each other (continuous distribution of feeders) and that there does not seem to have any "natural" cluster structure in the feeder population. The result of the clustering seems to divide the areas with a high density of observations into sectors or roughly equal size.

Figure 12 shows the impact of varying the number of clusters (between 1 and 20) on the two metrics used to quantify the clustering performance (note that the silhouette value is only defined for at least two clusters).

This figure shows that, as expected and previously observed, the normalized sum squared error (*nSSE*) decreases monotonously when increasing the number of clusters. A usual way of selecting the "optimal" number of clusters is to use the elbow criterion. In this case, the *nSSE*-curve does not exhibit a clear elbow shape and a visual selection of the "optimal" number of clusters is questionable (this has also been observed in previous studies with sufficient large data sets). A number of clusters between six and 16 (for the cluster number ranges of most previous studies, see Table 1) could be somehow justified. However, the silhouette curve does not exhibit a monotonic behavior and even shows that the highest clustering performance according to this metric is reached for only two clusters. These analyses confirm the known difficulty to interpret the clustering results and select the "optimal" or "appropriate" number of clusters.

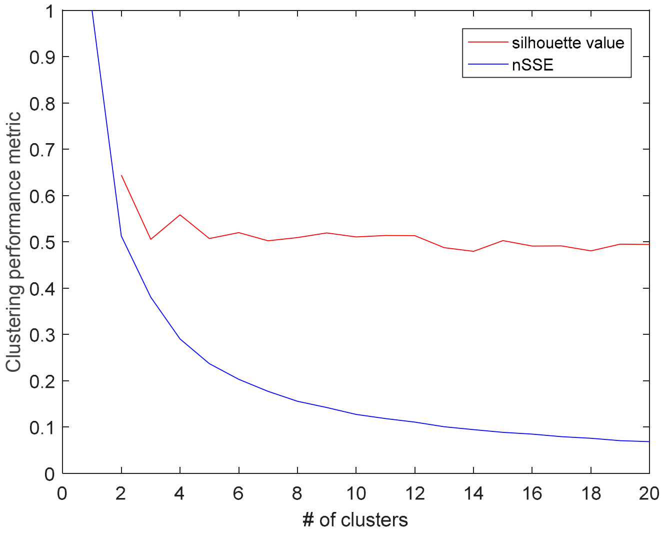

Figure 12. Impact of the number of clusters (between 1 and 20) on the used metrics (average silhouette value and normalized sum squared error).

Finally, the results of the clustering and their suitability for the considered problem (reflect the behavior of feeders in terms of hosting capacity) have been analyzed through an external validation. For this, the classes, which had also been used in the classification, are used to measure how close the clustering is to the predetermined classes (voltage or current-constrained feeder). The results of this external validation for a clustering with four clusters is shown on Figure 13 (projection on the first two principal components).

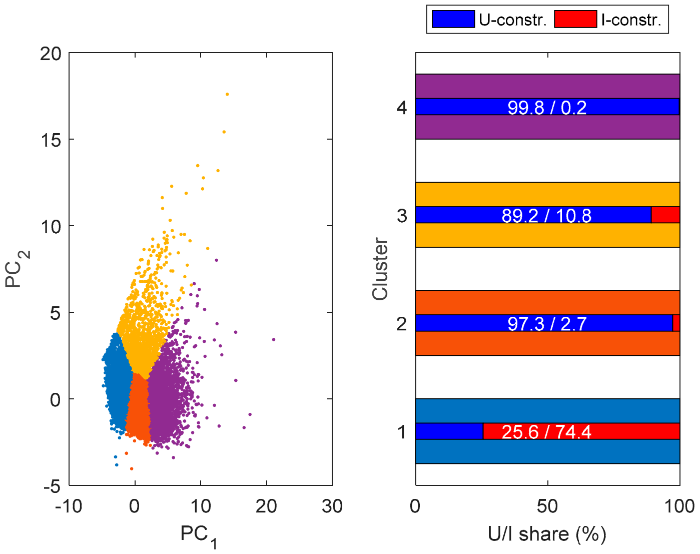

Figure 13. External validation of the clusters—cluster vs. classes.

The left part shows the scatter plot (similar to Figure 11 with different coloring) and the right part shows for each cluster (the wide bars are colored as the clusters) the share of voltage (blue) and current (red)-constrained feeders. This share between both classes is in addition indicated as a numerical value, which allows to evaluate the performance of the clustering for the considered problem (to identify feeders according to their class (voltage or current-constrained)). Indeed, this share can be interpreted as a "partial purity" measurement. A common metric used to quantify the clustering performance with an external validation is the (global) purity, given by Equation (12):

$$purity(\mathrm{CLU}, \mathrm{CLA}) = \frac{1}{\mathrm{N}} \cdot \sum_{clu \in CLU} \max_{cla \in CLA} |clu \cap cla|, \tag{12}$$

where $purity(\mathrm{CLU}, \mathrm{CLA})$ is the (global) purity of the clusters (against the classes), cla are the classes belonging to the set of classes CLA (here voltage or constrained-feeders), clu the clusters belonging to the set of clusters CLU and N the number of observations (feeders). In our case, the "partial purity" values (for each cluster) have been considered.

The more dissymmetric the ratio (the purest), the more the clustering is able to discriminate between both classes. For example, the fourth cluster (purple) almost only contains voltage-constrained feeders (99.8% of the feeders in this cluster are voltage-constrained). On the opposite, the cluster 1 (blue) mostly contains only current-constrained feeders, with however a significantly lower level of purity. The reader should note that these two clusters (with the highest purity levels for both classes) do not have any border.

Following the same reasoning as for the classification (considering that current-constrained feeders should be identified with the highest possible confidence), the most interesting cluster is the fourth cluster which has the highest purity level (only cluster with a purity level greater than 99%). This cluster contains however only about 16% of the whole feeder population (or about 18% of the population of voltage-constrained feeders). This means that this clustering leads to a rather poor result: its ability to discriminate between the two classes (voltage and current-constrained feeders) is low, even when accepting a "risk" (here 0.2% due to the "impurity").

By following the same conservative approach as for the classification (considering only feeders which are (almost) surely voltage constrained), the deployment potential of reactive power-based voltage control would be significantly lower than for the classification. The share of voltage-constrained feeders which would be safely identified as such, is about 46% for the decision tree-based classification (see Section 3.3), and only about 18% for the clustering (loss of potential or 54% and 82% respectively).

4. Discussion and Conclusions

In this study, two main concepts for identifying "typical" or "representative" feeders have been analyzed, on the basis of a large set of real LV feeders in Austria (24,000 LV feeders). The main motivation behind this work is to be able to clearly identity the constraint limiting the feeders hosting capacity (voltage or current). Indeed, reactive power based voltage control concepts have been developed and successfully validated in the past years. By consuming reactive power, the amount of generation that can be integrated into the network (maintaining an acceptable voltage level), can be increased. This increase of the hosting capacity through voltage control can, of course, only be implemented in feeders which are voltage-constrained (and not current-constrained). Due to the low (or almost inexistent) level of monitoring in LV networks, it is of prime importance to be able to classify feeders (i.e., identify the constraint: voltage or current) with a high level of confidence in order to avoid deploying controls which would in fact worsen the situation.

The first concept implemented to identify voltage-constrained feeders, k-means clustering, is an unsupervised machine-learning technique, which has been used in most of the previous studies. The result of this clustering consists in a set of clusters, which intend to reflect the data structure, based on the parameters used to characterize the observations (the feeders).

The second concept implemented to identify voltage-constrained feeders, classification, is a supervised machine-learning technique, which uses prior information (the belonging to the class voltage or current-constrained feeders) on a data set to train a classifier (decision tree here). In opposite to previous works, the hosting capacity and the associated constraint have been determined for a large data set, which allowed implementing a classification and validating the clustering results.

The analyses performed in this study showed that a perfect classification cannot be reached, which means that some feeders are misclassified. Misclassification does not have the same impact for both categories (voltage- and current-constrained feeders). Current-constrained feeders should be identified with a high confidence level (in a conservative way) to avoid implementing a control which would be counterproductive.

The results shown in this paper might be interpreted as rather poor, but they reflect in fact the diversity of the feeders within the data set. All the attempts to identify a clear structure in the data (data exploration and visualization) showed continuous distributions of feeders whatever the variables considered.

The lowest achieved misclassification to ensure that (almost) no current-constrained feeders are wrongly classified as voltage-constrained feeders is rather high: 54% for the decision tree and 82% for the k-means clustering. This can be considered as a significant loss of potential. As expected, the supervised learning based on a decision tree classifier leads to significantly better results than the k-means clustering. With 82% loss of potential, the benefits of using clustering are even questionable.

This performance, which might be considered to be poor, is the costs of reaching a very high sensitivity (close to 100%), which results in a high false alarm rate (loss of potential to implement reactive power control for hosting capacity increase of voltage-constrained feeders). However, a detailed analysis showed that this loss of potential is in fact limited, since most of the affected feeders are close to experience over-loading when implementing reactive power control. When considering this, the loss of potential drops from 54% to about 18%.

As a conclusion, even with a modest performance from a generic point of view, the benefit of the feeder classification can be considered as significant. The concept developed allows predicting the behavior of LV feeders in terms of hosting capacity constraint with a limited number of variables. By identifying a sub-group of the feeders (voltage-constrained feeders) with a very high confidence level (sensitivity close to 100%), the control concepts can be deployed by DSOs in the target networks without the need for additional studies. As a result, more detailed planning studies can be prioritized to those feeders for which the classification results are unsure.

As future research direction, the investigated concepts could be used for further data sets to validate the results. In particular, the classification results might be significantly better for less skewed (unbalanced) data sets. A further research direction could be a formalization of the probabilistic character of the classification and of the clustering to reflect the probabilistic nature of distribution network planning.

Acknowledgments: This work has been partly supported by the project IGREENGrid which has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement No. 308864.

Author Contributions: Benoît Bletterie performed the hosting capacity evaluation and the feeder classification. Serdar Kadam supported this work by writing scripts for simulation automation. Herwig Renner contributed to the analysis and discussion of the results.

Conflicts of Interest: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Bollen, M.; Rönnberg, S. Hosting Capacity of the Power Grid for Renewable Electricity Production and New Large Consumption Equipment. *Energies* **2017**, *10*, 1325.
2. Bletterie, B.; Goršek, A.; Uljanić, B.; Jahn, J. Enhancement of the network hosting capacity—Clearing space for/with PV. In Proceedings of the 25th European Photovoltaic Solar Energy Conference and Exhibition, Valencia, Spain, 6–10 September 2010; pp. 4828–4834.
3. VDE. *VDE-AR-N 4105 Generators Connected to the Low-Voltage Distribution Network—Technical Requirements for the Connection to and Parallel Operation with Low-Voltage Distribution Networks*; VDE: Frankfurt am Main, Germany, 2017.
4. Ministère de L'écologie, de L'énergie, du Développement Durable et de L'aménagement du Territoire. *Arrêté du 23 Avril 2008 Relatif aux Prescriptions Techniques de Conception et de Fonctionnement Pour le Raccordement à un Réseau Public de Distribution D'électricité en Basse Tension ou en Moyenne Tension D'une installation de Production D'énergie Électrique*; Ministère de L'écologie, de L'énergie, du Développement Durable et de L'aménagement du Territoire: Paris, France, 2008.
5. E-Control. *Technische und Organisatorische Regeln für Betreiber und Benutzer von Netzen. Teil D: Besondere Technische Regeln. Hauptabschnitt D4: Parallelbetrieb von Erzeugungsanlagen mit Verteilernetzen*; E-Control: Vienna, Austria, 2016.
6. Bletterie, B.; Gorsek, A.; Fawzy, T.; Premm, D.; Deprez, W.; Truyens, F.; Woyte, A.; Blazic, B.; Uljanic, B. Development of innovative voltage control for distribution networks with high photovoltaic penetration: Voltage control in high PV penetration networks. *Prog. Photovolt. Res. Appl.* **2012**, *20*, 747–759. [CrossRef]
7. Kadam, S.; Bletterie, B.; Lauss, G.; Heidl, M.; Winter, C.; Hanek, D.; Abart, A. Evaluation of voltage control algorithms in smart grids: Results of the project: MorePV2grid. In Proceedings of the 29th European Photovoltaic Solar Energy Conference and Exhibition, Amsterdam, The Netherlands, 22–26 September 2014.
8. Stetz, T.; Marten, F.; Braun, M. Improved Low Voltage Grid-Integration of Photovoltaic Systems in Germany. *IEEE Trans. Sustain. Energy* **2013**, *4*, 534–542. [CrossRef]
9. Dehghani, F.; Nezami, H.; Dehghani, M.; Saremi, M. Distribution feeder classification based on self organized maps (case study: Lorestan province, Iran). In Proceedings of the 20th Conference on Electrical Power Distribution Networks Conference (EPDC), Zahedan, Iran, 28–29 April 2015; pp. 27–31.
10. Nijhuis, M.; Gibescu, M.; Cobben, S. Clustering of low voltage feeders from a network planning perspective. In Proceedings of the CIRED 23rd International Conference on Electricity Distribution, Lyon, France, 15–18 June 2015.
11. Willis, H.L.; Tram, H.N.; Powell, R.W. A Computerized, cluster based method of building representative models of distribution systems. *IEEE Trans. Power Appar. Syst.* **1985**, *12*, 3469–3474. [CrossRef]
12. Schneider, K.P.; Chen, Y.; Chassin, D.P.; Pratt, R.G.; Engel, D.W.; Thompson, S. *Modern Grid Initiative: Distribution Taxonomy Final Report*; Pacific Northwest National Laboratory: Richland, WA, USA, 2008.
13. Kerber, G. *Aufnahmefähigkeit von Niederspannungsverteilnetzen für die Einspeisung aus Photovoltaikkleinanlagen*; Technical University of Munich: Munich, Germany, 2011.
14. Lindner, M.; Aigner, C.; Witzmann, R.; Frings, R. Aktuelle Musternetze zur Untersuchung von Spannungsproblemen in der Niederspannung. In Proceedings of the 14 Symposium Energieinnovation, Graz, Austria, 10–12 February 2016.

15. DeNA. *DeNA-Verteilernetzstudie. Ausbau-und Innovationsbedarf der Stromverteilnetze in Deutschland bis 2030*; DeNA: Tokyo, Japan, 2012.
16. Dickert, J.; Domagk, M.; Schegner, P. Benchmark low voltage distribution networks based on cluster analysis of actual grid properties. In Proceedings of the 2013 IEEE Grenoble Conference on PowerTech (POWERTECH), Grenoble, France, 16–20 June 2013; pp. 1–6.
17. Broderick, R.J.; Williams, J.R. Clustering methodology for classifying distribution feeders. In Proceedings of the 2013 IEEE 39th Photovoltaic Specialists Conference (PVSC), Tampa, FL, USA, 16–21 June 2013; pp. 1706–1710.
18. Gust, G. *Analyse von Niederspannungsnetzen und Entwicklung von Referenznetzen*; KIT: Cornwall, UK, 2014.
19. Cale, J.; Palmintier, B.; Narang, D.; Carroll, K. Clustering distribution feeders in the Arizona Public Service territory. In Proceedings of the 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC), Denver, Colorado, 8–13 June 2014; pp. 2076–2081.
20. Li, Y.; Wolfs, P.J. Taxonomic description for western Australian distribution medium-voltage and low-voltage feeders. *IET Gener. Transm. Distrib.* **2014**, *8*, 104–113. [CrossRef]
21. Walker, G.; Nägele, H.; Kniehl, F.; Probst, A.; Brunner, M.; Tenbohlen, S. An application of cluster reference grids for an optimized grid simulation. In Proceedings of the CIRED 23rd International Conference on Electricity Distribution, Lyon, France, 15–18 June 2015.
22. Varela, J.; Hatziargyriou, N.; Puglisi, L.J.; Rossi, M.; Abart, A.; Bletterie, B. The IGREENGrid Project: Increasing Hosting Capacity in Distribution Grids. *IEEE Power Energy Mag.* **2017**, *15*, 30–40. [CrossRef]
23. Bletterie, B.; Kadam, S.; Abart, A.; Priewasser, R. Statistical analysis of the deployment potential of Smart Grids solutions to enhance the hosting capacity of LV networks. In Proceedings of the 14 Symposium Energieinnovation, Graz, Austria, 10–12 February 2016.
24. PowerFactory—DIgSILENT Germany. 2017. Available online: http://www.digsilent.de/index.php/products-powerfactory.html (accessed on 6 November 2015).
25. Python. 2017. Available online: https://www.python.org/ (accessed on 21 March 2016).
26. Kadam, S.; Bletterie, B.; Gawlik, W. A Large Scale Grid Data Analysis Platform for DSOs. *Energies* **2017**, *10*, 1099. [CrossRef]
27. Bletterie, B.; Abart, A.; Kadam, S.; Burnier, D.; Stifter, M.; Brunner, H. Characterising LV networks on the basis of smart meter data and accurate network models. In Proceedings of the Integration of Renewables into the Distribution Grid (CIRED 2012 Workshop), Lisbon, Portugal, 29–30 May 2012; pp. 1–4.
28. Xu, R.; Wunsch, D. Survey of Clustering Algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
29. Bletterie, B.; Gorsek, A.; Abart, A.; Heidl, M. Understanding the effects of unsymmetrical infeed on the voltage rise for the design of suitable voltage control algorithms with PV inverters. In Proceedings of the 26th European Photovoltaic Solar Energy Conference and Exhibition, Hamburg, Germany, 5–9 September 2011; pp. 4469–4478.
30. Jenkins, N.; Allan, R.; Crossley, P.; Kirschen, D.; Strbac, G. *Embedded Generation*; The Institution of Engineering and Technology: London, UK, 2000.
31. Sutton, C.D. Classification and Regression Trees, Bagging, and Boosting. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2005; Volume 24, pp. 303–329.
32. Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **1997**, *30*, 1145–1159. [CrossRef]
33. Walker, G.; Krauss, A.-K.; Eilenberger, S.; Schweinfort, W.; Tenbohlen, S. Entwicklung eines standardisierten Ansatzes zur Klassifizierung von Verteilnetzen. In *VDE-Kongress 2014*; VDE Verlag: Berlin, Germany, 2014.