Cluster Analysis for Medium Voltage Distribution Feeders

Jneid Jneid

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

September 2020

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Engineering Thesis option.

2020/09/21

Abstract

Different green energy technologies will be deployed in electricity distribution networks at a large scale in the near future, and technical problems will arise if adequate planning is not carried out to accommodate them. Therefore, it is crucial for distribution network operators and planners to study the impact these technologies will have and develop strategies to mitigate problems that may arise. However, with a large number of feeders and a wide variety of models, a full-scale impact analysis of different technologies is a computationally intensive and time-consuming process as the execution of multiple simulations are needed. Consequently, it is desirable to narrow down to only a few models that provide generic representations of the different models. This would benefit in rapidly evaluating a test case and studying the impact of new grid technologies.

This thesis focuses on identifying a set of representative distribution feeders by analyzing various electrical characteristics using unsupervised clustering methods. For this purpose, this thesis proposes an unsupervised clustering procedure for medium voltage distribution feeders that consists of using two dimensionality reduction techniques, three clustering algorithms as well as three indicators to determine the optimal number of clusters. A comparative analysis across the different techniques to find the best structure of the data is performed. Subsequently, the clustering results are systematically validated in terms of topology, demand and voltage profiles, a process that enabled the iterative improvement of cluster quality, contrary to conventional non-iterative approach. The work in this thesis provides a collective recipe, utilizing state of the art tools like principal component analysis and t-distributed stochastic neighbour embedding, for the best selection of clusters in the exercise of identifying representative feeders.

A case study is conducted on a dataset of 2975 feeders from a distribution utility covering a large area and different types of consumers. The case study confirms the effectiveness of the proposed procedure, where the whole dataset of feeders gets represented by 11 feeders each being of a very distinct type.

Abrégé

Différentes technologies d'énergie verte seront déployées à grande échelle dans les réseaux de distribution d'électricité dans un avenir proche, et des problèmes techniques se poseront si une planification adéquate n'est pas effectuée pour les accueillir. Il est donc crucial que les opérateurs et planificateurs de réseaux de distribution étudient l'impact que ces technologies auront et élaborent des stratégies pour mitiger les problèmes qui pourraient survenir. Cependant, avec un grand nombre de réseaux et une grande variété de modèles, une analyse d'impact à grande échelle des différentes technologies est un processus qui exige beaucoup de calculs et de temps car l'exécution de multiples simulations est nécessaire. Par conséquent, il est souhaitable de se limiter à quelques modèles qui fournissent des représentations génériques des différents modèles. Cela permettrait d'évaluer rapidement un cas d'essais et d'étudier l'impact des nouvelles technologies.

Cette thèse se concentre sur l'identification d'un ensemble de réseaux représentatifs en analysant diverses caractéristiques électriques à l'aide de méthodes de regroupement ou clustering non supervisées. À cette fin, cette thèse propose une procédure de clustering non supervisée pour les réseaux de distribution de moyenne tension qui consiste à utiliser deux techniques de réduction de la dimensionnalité, trois algorithmes de clustering ainsi que trois indicateurs pour déterminer le nombre optimal de clusters. Une analyse comparative entre les différentes techniques pour trouver la meilleure structure des données est effectuée. Par la suite, les résultats du clustering sont systématiquement validés en termes de topologie, de demande et de profils de tension, un processus qui a permis l'amélioration itérative de la qualité des clusters, contrairement à l'approche non itérative classique. Le travail de cette thèse fournit une recette collective, utilisant des outils de pointe comme l'analyse en composantes principales et l'algorithme t-distributed stochastic neighbor embedding, pour la meilleure sélection de clusters dans l'exercice d'identification de réseaux représentatifs.

Une étude de cas est menée sur un ensemble de données de 2975 réseaux d'une utilité de distribution d'électricité couvrant une large zone et différents types de consommateurs. L'étude de cas confirme l'efficacité de la procédure proposée, où l'ensemble de données des distributeurs est représenté par 11 réseaux , chacun étant d'un type très distinct.

Acknowledgments

I would like to express my sincere appreciation to Prof. Francois Bouffard and Prof. Géza Joós for their guidance and support throughout my work. They inspired me with thought-provoking discussions that helped me deepen my understanding of the research topic and motivated me to always walk the extra mile. They convincingly guided and encouraged me to be professional and do the right thing even when the road got tough. I also highly appreciate the chance they gave me to work with an industrial partner and to become an intern there.

I gratefully acknowledge the funding received towards my master's degree from NSERC and from Hydro-Quebec Industrial Research Chair on the Integration of Renewable Energies and Distributed Generation. I would like to further acknowledge the financial support provided by Fonds de recherche du Québec - Nature et technologies (FRQ-NT), B1X scholarship Group 1 Energy.

I would like to recognize the invaluable assistance that Hydro Quebec Distribution provided during my internship. On this note, I would like to thank Bruno Fazio and Angelo Giumento who believed in me, in my work and helped me transition from academia to the industry. I would also like to thank the rest of my colleagues at Hydro Quebec Distribution for their feedback on my work and for the great times that we shared during my internship.

I want to give thanks to my friend and colleague Laith Mubaslat, for the good memories we made and for the discussions about research and life. I want to also thank all my friends and colleagues in the lab for all the technical discussions and good memories.

Finally, I must express my very profound gratitude to my parents and to Elsa-Lynn Nassar for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis.

Preface

I was the major contributor for all the chapters in this dissertation where I was responsible for the major areas of proposing, designing and developing the solution, and analysing the results, as well as the totality of dissertation composition. Bruno Fazio from Hydro Quebec Distribution contributed to defining the problem as part of my internship mandate at Hydro Quebec Distribution.

All the text and equations that are taken from previously published articles were cited in the dissertation. The corresponding sources of the algorithms and data analysis tools that were adopted in this dissertation were cited where appropriate. I adapted all these algorithms and tools to my analysis procedure described in this dissertation with additional incidental recommendations from Bruno Fazio.

The validation technique described in Chapter 4 is of my own design with additional improvements based on the recommendations of Bruno Fazio and other members of his team.

The data analysis in Chapter 5 is my original work and the simulations in Section 5.4 were performed by me. The tools to perform the analysis and simulations including the computer with the software and its license were provided by Hydro Quebec Distribution.

Hydro Quebec Distribution also provided the data for the case study. The data collection was performed by me with assistance from Bruno Fazio and other members of his team.

Professor Bouffard and Professor Joós contributed by supervising me and providing feedback on the technical contents and results, as well as giving best practice guidelines in writing academic articles.

Contents

List of Figures vii

List of Tables x

1 Introduction 1

1.1 Background . . . 2

1.1.1 Feature Engineering . . . 2

1.1.2 Cluster Analysis . . . 3

1.1.3 Distribution Test Feeders . . . 4

1.2 Literature Review . . . 4

1.2.1 Integrating New Technologies . . . 5

1.2.2 Distribution Feeder Modelling . . . 6

1.2.3 Relevant Applications of Cluster Analysis . . . 7

1.3 Problem Definition . . . 12

1.3.1 Thesis Statement and Contributions . . . 12

1.4 Thesis Organization . . . 13

2 Feature Engineering 14

2.1 Introduction . . . 14

2.2 Proposed Features . . . 15

2.2.1 Data Quality . . . 17

2.3 Data Processing . . . 18

2.3.1 Outliers . . . 18

2.3.2 Correlation Study . . . 19

2.3.3 Scaling . . . 20

2.4 Categorical Features . . . 21

2.5 Conclusion . . . 21

3 Dimensionality Reduction **22**
3.1 Introduction . . . 22
3.2 PCA . . . 23
3.2.1 Mathematical Formulation (Shalev-Shwartz and Ben-David, [1]) . . . 23
3.2.2 Clustering via PCA . . . 24
3.3 t-SNE . . . 24
3.3.1 Mathematical Formulation . . . 24
3.3.2 t-SNE vs PCA . . . 26
3.4 Conclusion . . . 27

4 Cluster Analysis **28**
4.1 Introduction . . . 28
4.2 Clustering Algorithms . . . 29
4.2.1 K-means++ Algorithm . . . 29
4.2.2 Hierarchical Algorithm . . . 30
4.2.3 GMM Algorithm . . . 31
4.3 Optimal number of clusters . . . 31
4.3.1 Silhouette Score . . . 32
4.3.2 Calinski-Harabasz Score . . . 32
4.3.3 Davies-Bouldin Score . . . 33
4.4 Comparative Analysis . . . 34
4.4.1 Remark . . . 34
4.5 Validation Process . . . 34
4.6 Conclusion . . . 36

5 Case Study **37**
5.1 Introduction . . . 37
5.2 Data Processing . . . 39
5.2.1 Data Pre-Processing and Feature Engineering . . . 39
5.2.2 Dimensionality Reduction . . . 41
5.3 Cluster Analysis Results . . . 44
5.3.1 Cluster Analysis Results with PCA . . . 44
5.3.2 Cluster Analysis Results with t-SNE . . . 51
5.4 Validation of Results . . . 58
5.4.1 PCA . . . 58
5.4.2 t-SNE . . . 64
5.5 Conclusion . . . 67

6 Conclusions and Future Work **68**
6.1 Summary 68
6.2 Conclusions 69
6.3 Future Work 70

A Python Packages and Hyper-Parameters **73**

References **76**

List of Figures

1.1 A simple example of Cluster Analysis for two features x1 and x2, using the k-means algorithm. 3

2.1 Box-Whisker or Boxplot example. 19

5.1 Diagram of the overall procedure proposed in this thesis 38
5.2 Correlation matrix between the features. 40
5.3 Pairplots of features 7, 9 and 12 with the correlation coefficient, r. 41
5.4 Data in reduced dimensionality with the eigenvectors pointing in the direction of maximum variance. 42
5.5 Data in reduced dimensionality 1. 43
5.6 Data in reduced dimensionality 2. 44
5.7 The silhouette score with different number of clusters. 45
5.8 The VRC score with different number of clusters. 46
5.9 The Davies-Bouldin score with different number of clusters. 46
5.10 Results of the k-means++ clustering algorithm with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids. 47
5.11 Results of the hierarchical clustering algorithm (agglomeration technique) with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids. 48
5.12 Results of the GMM clustering algorithm with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids. 48
5.13 The silhouette score with different number of clusters. 51
5.14 Results of k-means++ clustering algorithm with 7 clusters and with t-SNE dimensional reduction. The feeders with regulators are represented in one cluster. 52

5.15 Results of k-means++ clustering algorithm with 10 clusters and with t-SNE dimensional reduction. The feeders with regulators are still represented in one cluster. 53
5.16 Results of k-means++ clustering algorithm with 11 clusters and with t-SNE dimensional reduction. The feeders with regulators are represented in two different clusters. 53
5.17 Silhouette profile with PCA dimensional reduction. 57
5.18 Silhouette profile with t-SNE dimensional reduction. 57
5.19 Topology of the center 58
5.20 Topology of the nearest neighbour 58
5.21 Topology of the farthest neighbour 58
5.22 Boxplots of the feature 1 (number of branches) outlining the range of values, namely, the median and IQR, in each cluster. 59
5.23 The normalized demand profile of an urban residential/industrial cluster. . . 60
5.24 Boxplots of the feature 1 (number of branches) outlining the range of values, namely, the median and IQR, in each cluster. 64

List of Tables

2.1 Proposed Features 17

5.1 Eigenvalues ratio of the principal components 42
5.2 Number of feeders with negative silhouette scores 49
5.3 Number of feeders per cluster 49
5.4 Representative feeders 50
5.4 Representative feeders (Continued) 50
5.5 Number of feeders having a negative silhouette score 54
5.6 Representative feeders with t-SNE 54
5.6 Representative feeders with t-SNE (Continued) 55
5.7 Number of feeders per cluster with t-SNE 55
5.8 MSE between the demand profile of the center, the nearest and farthest feeders 60
5.9 PV simulation results 63
5.10 MSE between the demand profile of the center, the nearest and farthest neighbor 65
5.11 PV simulation results with t-SNE 66

List of Acronyms

DG	Distributed Generation
DNO	Distribution Network Operator
EM	Expectation Maximization
EV	Electric Vehicle
FN	Farthest Neighbor
GMM	Gaussian Mixture Models
HV	High Voltage
HVAC	Heating, Ventilation and Air Conditioning
IEEE	Institute of Electrical and Electronics Engineers
IQR	Interquartile range
KNG	K-Neighbors Graph
LOF	Local Outlier Factor
LV	Low Voltage
MLOP	Maximum Load Operating Point
MSE	Mean Square Error
MV	Medium Voltage
NLP	Natural Language Processing
NN	Nearest Neighbor
NOP	Network Operating Point
NREL	National Renewable Energy Laboratory
ONC	Optimal Number of Clusters
PCA	Principal Component Analysis
PNNL	Pacific Northwest National Laboratory
PV	Photovoltaic
SVD	Singular Value Decomposition
TFWG	Test feeder Working Group
TOV	Temporary Overvoltage

t-SNE	t-Distributed Stochastic Neighbor Embedding
VVO	Voltage Violations and Overload

Chapter 1

Introduction

Identifying a set of representative feeders is of great interest for distribution network planners and operators [2]. The identified set of feeders can be used in performing a limited number of simulations to evaluate the impact of new grid technologies. Therefore, distribution network operators (DNOs) are capable of studying and understanding the impact of different new technologies while avoiding an exhaustive process of simulations on every feeder of the network. The process of finding the representative set of feeders typically involves analyzing different electric characteristics and using unsupervised cluster analysis methods.

Understanding how feeders are modelled is essential to determine the electric components or features to be considered. In parallel, due to the variety of methods used in cluster analysis [3], finding the optimal grouping of feeders can only be guaranteed by comparing the performance of the most used techniques in similar exercises. Consequently, using the recommended electric characteristics and the variety of cluster analysis methods, a procedure to find a set of representative feeders can be drafted for the distribution utilities. The procedure will represent a guideline for the DNOs of different utilities to apply to their feeders.

This thesis presents a cluster analysis procedure for medium voltage (MV) distribution feeders. The current chapter presents background knowledge and a literature review on cluster analysis and distribution feeder modelling as well as the problem definition and the thesis statement.

1.1 Background

1.1.1 Feature Engineering

Feature engineering or feature extraction is the process of refining data without losing any important information; it is part of any data analysis problem [4]. Accordingly, the purpose of feature engineering is to have a feature subset that:

- Increases model performance.
- Reduces time/memory requirements.
- Improves the interpretability of the results.

In [4], the authors provided a basic guideline on how to derive useful features from original dataset. In doing so, they revisited generic feature construction techniques covering basic linear transformations of the original features, variable ranking, which involves studying the correlation of the features used and matrix factorization or singular value decomposition (SVD). In fact, principal component analysis (PCA) is a linear dimensionality reduction that uses SVD to project data to a lower-dimensional space providing a low dimensional summary of the data [5, 6]. Alternatively, the authors in [7] went more in-depth and demonstrated the ability of PCA to extract features relevant to cluster structure. The advantages of PCA will be thoroughly discussed in the upcoming chapters.

Further, the authors in [7] addressed the topic "clustering stability" which denotes that the structure of clusters should not vary significantly when introducing perturbation to the data. Clustering stability covers issues that should be considered when using cluster analysis techniques [8]. These issues are:

- The choice of a clustering algorithm.
- Choice of a normalization and similarity/dissimilarity measure.
- Which variables/features to cluster.
- Which patterns to cluster.
- How many clusters to have.

The authors concluded that PCA supports clustering stability and improves the refinement of cluster structure.

This section highlighted the steps that should precede the clustering exercise. In this thesis, different feature selection and extraction techniques such as PCA and variable correlation will be used on original data before performing cluster analysis.

1.1.2 Cluster Analysis

The practice of cluster analysis consists of finding a convenient and valid grouping of objects where the objects of a group share similar characteristics. An object is described by a set of measurements (e.g.: coordinations in Euclidean space) or by a relationship with other objects (i.e. proximity).

One of the first problems in identifying clusters in data is to specify what proximity is and how to measure it [3]. The notion of proximity depends on the nature of the problem. Trivially, the proximity can be the difference of the Euclidean distance, the Manhattan distance, or any predefined metric between two objects. For instance, in [9], a pre-defined logarithmic metric on the Euclidean distance was introduced to emphasize the small differences between different objects or data points. In order to obtain a well-defined partitioning or grouping, it is necessary to:

- Minimize intra-class variance or the distance between the points of the same cluster.
- Maximize inter-class variance or the distance between different clusters.

A simple example of cluster analysis is presented in Fig. 1.1, where the data is partitioned into four groups.

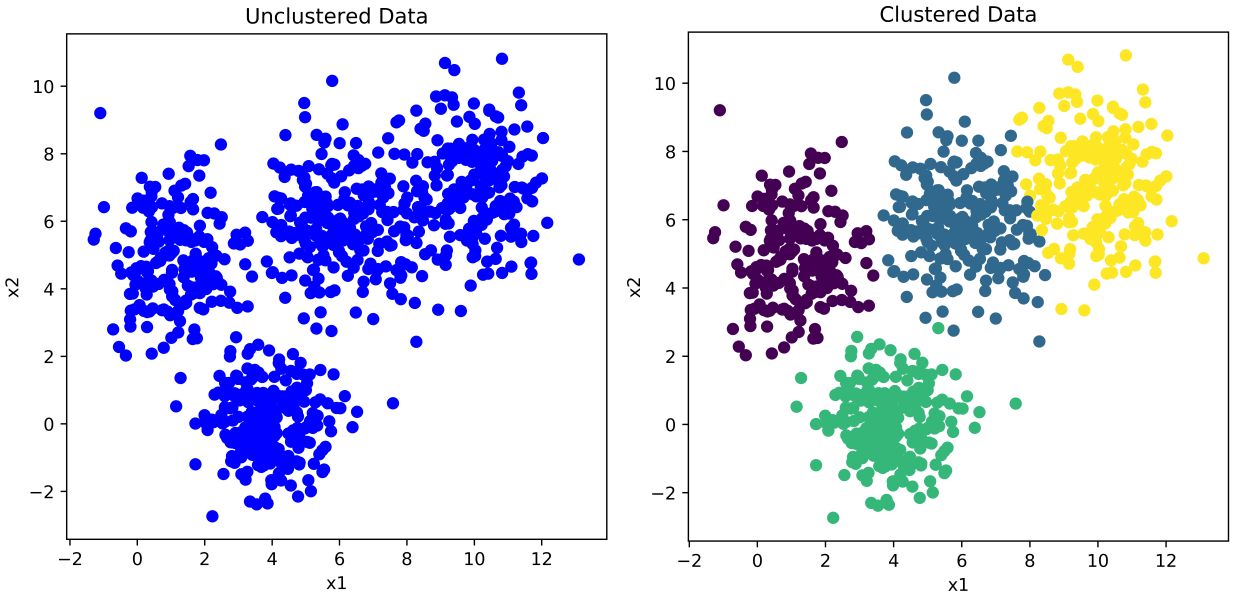


Fig. 1.1 A simple example of Cluster Analysis for two features x1 and x2, using the k-means algorithm.

It is further essential to distinguish cluster analysis from discriminant analysis, where categorizing data is based on ground truth or prior knowledge. Data clustering is an unsupervised classification method that is distinct from supervised classification, where data

points are already labelled. Reference [10] provides a detailed breakdown regarding the difference between supervised and unsupervised learning.

In this thesis, different clustering algorithms are used and their performance is compared in order to obtain the best feasible partitioning of data.

1.1.3 Distribution Test Feeders

Modelling feeders that researchers can use in their simulations was initially introduced in 1992 by the Test Feeder Working Group (TFWG) [11] of the IEEE Power and Energy Society Distribution System Analysis Subcommittee. Their work resulted in the development of four test feeder models. A more recent study was presented in 2001 [12] where a new test feeder, the IEEE Four Node Test Feeder, was added to the existing four feeders. This new test feeder considered three-phase transformer models. The strategy followed by [12] consisted of the following aspects:

1. Different types of constant unbalanced loads
2. Shunt capacitance
3. Conductor characteristics
4. Voltage regulation

The increasing widespread integration of renewable energy resources in distribution networks which caused the changes in how power is generated, delivered, as well as scheduled necessitated the revision of the existing test feeders [13]. After more than 15 years, the TFWG revisited its test feeders and highlighted underlying challenges. In fact, each feeder was purposely modelled for testing certain analysis techniques and could not be used otherwise. Accordingly, the subcommittee drafted a guideline for researchers outlining which test feeders were most appropriate for different types of analysis.

In this thesis, the selection of features that will be used in cluster analysis is based on the recommendations of distribution system operators as well as on the work in [12] and [13]. Precisely, this thesis considers the distribution elements that [13] deemed necessary to be modelled in the process of analyzing the electrical behaviour of the distribution feeders.

1.2 Literature Review

The current section presents a more detailed description of existing problems and a review of the literature regarding the challenges of integrating new technologies, distribution feeder modelling, and relevant applications of cluster analysis in distribution power systems.

1.2.1 Integrating New Technologies

This thesis is motivated by the importance of studying the impact of new technologies such as photovoltaic systems (PVs) and electric vehicles (EVs) in distribution networks. As such, this section introduces the problem statement and explicates how the different impacts of these technologies are studied.

One of the main aims of obtaining representative distribution feeders is to facilitate the impact analysis of new technologies, mainly in the form of distributed generation (DG). DG includes renewable resources like photovoltaic systems as well as various other technologies [14]. Since distribution feeders were originally designed and developed to operate unidirectionally [15], the integration of these technologies could pose a challenge to distribution utilities touching on several aspects of distribution systems planning and operations [14, 16]. In the work of [14], a handbook for distribution engineers was prepared where an investigation of the potential impacts caused by high-penetration PV scenarios was conducted. High-penetration PV is a scenario where there is a possibility of exceeding the voltage, thermal, and protection limits of some distribution feeders. A significant aspect of this investigation is the time-varying analysis, which, contrary to single-point analysis, can capture the interaction between all the distribution components. In fact, dynamic analysis was previously applied in [17] where an emphasis regarding the impacts of PV distribution generation was made, and mitigation measures were proposed.

Coupled with a time-varying analysis, the investigation in [14] highlighted the impacts of voltage fluctuations, especially under low load conditions. For example, with a high PV generation near the source substation of the feeder, it was reported that voltage regulation problems could arise if line drop compensation is used [18]. It was further found that PV generation could mask the actual load in the measurements which could lead to several problems throughout cold load pickup or fault current calculation.

There have been numerous studies to investigate the challenges of integrating new technologies. The challenges for integrating inverter-based resources like PVs were summarized by [19]. The presentation covers the issues of stability (transient and dynamic), frequency, and control of the reactive power and voltage regulation resources on the distribution system (also called voltage var control).

In addition, several mitigation techniques were proposed in [14]. These mitigation techniques exploit the inverter characteristics and specifications such as the voltage var setting. One of these techniques is accelerating the ability of inverters to detect temporary overvoltage (TOV). This aspect has been covered by the IEEE standard 1547 rev. 2018 where improved settings have been recommended [20]. Many of the aforementioned impacts and mitigation techniques were revisited by IEEE PES Task Force on Voltage Control for Smart

Grids. They extended their work to cover voltage control in transmission grids [21].

It is also important to mention that some simulation platforms and analysis tools are available for performing detailed full-scale simulations like the Integrated Grid Modeling System (IGMS) [22]. These platforms can help utilities in studying the challenges illustrated in this section. However, they usually have expensive and specific requirements that discourage their adoption by utilities.

On one hand, the IGMS framework models the transmission and distribution grid and also enables time-varying analysis of the whole electric grid. On the other hand, this framework is based on GridLaB-D [23], which is a power system simulation and modelling tool that adopts an agent-based approach to conduct a variety of studies like voltage var control and cost-benefit analysis. Utilities that use other simulators like CYMDIST [24] will need to make risky decisions and adopt GridLaB-D first. For such utilities, adopting a new simulator like GridLaB-D invovles substantial investments to buy licenses, train engineers and merge all existing models. Moreover, IGMS requires expensive high-performance computers.

1.2.2 Distribution Feeder Modelling

In this section, a comprehensive and in-depth review of test feeder modelling and load models is presented.

Despite the fact that power systems were developing concurrently worldwide, their evolution was fundamentally different due to design, structure, and regional differences [13]. The most relevant finding of the TFWG in [13] was the identification of the distribution elements that should be modelled for different types of analysis. For power flow analysis, single and three-phase cables should be considered [15]. The impedance should also be calculated accordingly. Similarly, for transformers, all types of connections should be represented. The list of elements stretches to cover shunt capacitors, voltage regulation, and load models.

While it is common practice to use constant load models, the authors in [25], expanded these models and developed time-variant load models. In their work, they considered end-user factors such as heating, ventilation and air conditioning (HVAC) consumption and TV quality. They based their constant load model on the ZIP model [15], which consists of determining a respective Z (impedance), I (current), and P (active power) values for end users. Additionally, ZIP model can include a reactive component. Later, the same authors proposed time-variant end-user models that account for the behaviour of the end-user under changes in electricity prices [26].

The distribution elements that are used in the modelling of the feeders characterize their electric behaviour. Therefore, these elements must be considered when solving for a set of representative feeders. The feature selection component of the process described in this

thesis will be based on the work done by the TFWG.

1.2.3 Relevant Applications of Cluster Analysis

This section exhibits some of the relevant uses of the cluster analysis techniques in distribution networks.

Load Profiles

Studying electricity consumption has gained increased attention over the years. Categorizing or characterizing consumers based on their consumption patterns can enable many options for electric utilities. Information on electricity consumption can help improve energy management strategies, tariffing, and load control [27]. Detecting consumption patterns could be based on probabilistic models [28] or simply load aggregation methods [29]. Alternatively, cluster analysis techniques can be used to detect consumption pattern similarities or representative patterns. The author in [27] proposed a procedure that uses clustering techniques to perform load pattern categorization. The procedure covered steps related to data preprocessing, outlier detection, and feature extraction, as well as performance assessment of different clustering techniques like k-means and hierarchical algorithms [3].

Similar work was conducted in [30] where the authors examined the topic of load profile clustering. A dimensionality reduction method was incorporated on smart meter data before performing clustering using the k-means algorithm. In fact, the first use of artificial intelligence technology to electrical power engineering was with smart metering data [31]. Clustering techniques can be used in a wide variety of applications, especially at the distribution level. Categorizing customers provides utilities with more insight on how to adjust their strategies and plans of resource dispatching and scheduling and electricity pricing.

Clustering Operating Points

Another interesting application of cluster analysis techniques in distribution power systems is the determination of relevant network operating points (NOP). These NOPs could be used not just in reliability assessment but in many other studies.

With the adoption and integration of new technologies, the reliability of the grid needs to be revisited more often. Currently, two main methods are used for reliability calculations in the distribution grid: Monte-Carlo and Analytical methods [32]. Analytical methods are deterministic methods that usually use the worst-case scenario or maximum load operating point (MLOP). However, with the increased integration of DG that is not highly affected by the load, using MLOP instead of a less strict operating point can lead to overly optimistic

reliability calculations [32]. Therefore, a clustering approach was proposed in [32] to determine relevant network operating points. That can be achieved by grouping similar operating points together and choosing the point with maximum load in each cluster as representative. The aforementioned approach expanded to adopt the criteria of representing the data in three dimensions as well as choosing the representative NOP of each cluster in such a way voltage violations and overload (VVO) of the NOPs are not neglected in the reliability calculations.

Clustering Distribution Feeders

Applying cluster analysis techniques to find representative distribution feeders was first proposed in [33]. The authors used the k-means algorithm [3] to find twelve feeders that represented an entire system of 1350 feeders. Their work was not only based on the statistical nature of the clustering but also on the electrical analysis of the results. In particular, comparisons of voltage drops, losses, and power factors were performed, and, at the end, a list of variables that are potentially significant for feeder clustering was formulated. The objective of [33] was to find an equivalent system that is less computationally expensive.

The Pacific Northwest National Laboratory (PNNL) then reintroduced the topic [2]. The authors in [2] used a different type of clustering method, the hierarchical algorithm [3], to find a set of 24 medium voltage (MV) representative feeders across the USA. Their approach was different from [33] in the sense that they first performed a global categorization based on climate and voltage levels. Next, they performed the cluster analysis considering 35 characteristics and operating variables.

In both [2] and [33], the only method used to determine the optimal number of groups or clusters (ONC) was the study of the sum of squared error (SSE). Using a method referred to as the *elbow criterion*, the slope of the SSE was examined, and the number of clusters was selected when the slope started decreasing (i.e., when the marginal value of the SSE stopped to improve).

A slightly different approach was used in [34] for Australian MV feeders, where more work was put into choosing the features, which covered more topology parameters. The primary goal of this work was to capture the topology or the construction of the distribution networks in Australia. The authors in [34] finalized 17 clusters describing best the Australian feeders. The optimal number of clusters for [34] and [35] was found using the silhouette method [8, 36]. Unlike the SSE method used by [33] and [2], the silhouette evaluates the average similarity and dissimilarity of the clusters.

Clustering of 280 22 kV distribution feeders in Western Australia was performed [37] using a combination of clustering and discriminant analysis techniques. The ONC was determined

using statistical parameters that are computed in a hierarchical fashion.

In [38], the authors used PCA to reduce the dimensionality of the problem. Their work resulted in twelve representative feeders covering over 3000 feeders in the Western USA, and it aimed at improving the screening process of PVs.

PCA was also used in [39], where 1295 distribution feeders in Arizona were re-grouped into nine clusters using the k-medoids algorithm [40].

PCA [5] is a data analysis method that consists of transforming variables related to each other ("correlated" in statistics) to new variables that are de-correlated from each other. These new variables are called principal components, or main axes. It allows the reduction in the number of variables as well as making the information less redundant.

However, PCA is a linear transformation that aims at maximizing the variance in the new components and preserving large pairwise distances. Thus, it cannot capture complex non-linear relationships between the features and it emphasizes the distant data points without accounting for close data points. This can sometimes lead to poor representation of the data, especially when the structure of the data has many significant variations [41]. This technique will be revisited in more detail later in Chapter 3.

The analysis of finding a set of representative distribution feeders was also conducted on 232 residential low voltage (LV) feeders of the North West of England [42]. The 232 feeders were partitioned into two data sets, one with PV and one without PV. The authors' work involved using four different algorithms, 19 features covering all the residential feeders' electric and customer characteristics. They chose gaussian mixture models (GMM) as the best performer in their study. In addition, the results were validated by assessing the capacity of the representative feeders to host PV generation (PV hosting capacity). PV hosting capacity, in this context, is defined as the capacity of PV generation that the feeders can accommodate without causing any problems.

In [43], the authors performed clustering using the k-means algorithm on over 8000 distribution feeders in California. The process resulted in eight clusters and the authors decided to select 214 feeders from these eight clusters and assessed their PV hosting capacity as a way to verify that the feeders of the same cluster have a similar hosting capacity. They found that the hosting capacity for feeders within the same cluster varies widely and thus, concluded that the accuracy of the clustering technique in terms of predicting PV hosting capacity is low. They also proposed various techniques to improve the accuracy of the PV hosting capacity via clustering. These techniques include adding new features, studying correlation, and weighting relevant features.

In [33, 34],[38],[39], and [43], the authors considered only using one indicator to determine the ONC (silhouette, variance ratio criterion or cubic criterion [36, 44]) with only one

algorithm (k-means, k-medoids or hierarchical algorithm [3, 40]).

The statistics community adopted the cubic criterion used in [2] and [38] to find the ONC, which is known to perform best in spherical clusters. However, the cubic criterion is an empirical method that lacks a rigorous mathematical proof.

The variance ratio criterion (VRC) [44] prefers convex clusters and it highly depends on the data. While the authors in [39] chose the optimal number of clusters based on the VRC, they did not comment on how decisive the indicator was in determining the ONC.

From an algorithmic perspective, k-means and k-medoids have the same objective, which is minimizing, over each cluster, the sum of the square of the distance between all the points and the centroid. Due to the random initiation of the k-means and k-medoids, it is difficult to reproduce the same results (which means with different runs, different results may be generated) [3, 40]. The hierarchical algorithm with its agglomerative approach, on the other hand, starts by taking each element as a cluster and merges the formed clusters recursively based on a linkage criterion. There are many linkage criteria like average linkage where the distance (or dissimilarity) between two clusters is represented by the average distances of all the elements of the two clusters. Another linkage criterion is called ward; it aims at minimizing the variance between the clusters that are being merged using the sum of squared distances within the clusters. As an objective function, the ward linkage criterion can be seen to be very similar to the other two algorithms, k-means and k-medoids.

Even though k-means is widely used and efficient with big datasets, it is not robust against local minima [45]. The same thing goes to the hierarchical algorithm, which will produce results that might significantly differ from k-means.

Another relevant algorithm is the GMM [46, 47]. It is based on a stochastic approach that is completely different from the partition and non-partition approaches of k-means and hierarchical algorithms. With such a variety of algorithms, one cannot be certain that an optimal clustering is achieved with just one algorithm. Different algorithms need to be used to validate the clustering results or to explore different possible results.

The majority of relevant references used only one algorithm with one indicator for the ONC. However, the authors in [38] used several indicators and clustering algorithms. They chose the GMM alongside the k-means++ algorithm as the best performers based on their highest silhouette score commenting that a deeper investigation must be conducted for finding the optimal results. The k-means++ algorithm differs from k-means in the initial step of choosing centroids, with k-means++ the initial centroids are chosen far apart from each other whereas, for k-means, they are chosen randomly. The silhouette score is a metric to evaluate the coherence of the clusters. By coherence, it is meant that the clusters are dense and distinctly isolated.

Indicator driven algorithm selection is not sufficient since the indicator only identifies the ONC. In other words, deciding which algorithm results in more coherent clustering involves performing a comparative analysis covering the indicator, how many points are mis-clustered (on the borders between two neighbouring clusters), the study of the characteristics of the centroids as well as the distribution of feeders among the clusters.

Moreover, references [33, 34], [39] and [43] used the assessment of PV hosting capacity as a validation tool for the cluster analysis. While this method is efficient in terms of interpreting the applicability of the cluster analysis, it does not help in improving the clustering model.

As mentioned above, the authors in [43] concluded that the clustering is not precise and proposed various techniques like studying correlation and weighting relevant features to improve the clustering.

In [38], the results showed that the majority of residential feeders are capable of hosting 100% of PV generation in terms of number of clients, meaning that all the clients recorded on the feeders can install PVs with a rated capacity of 3kW power on their roofs without causing any voltage violations and overloads (VVOs). Only the feeders with prior technical problems with PVs could not achieve 100% PV generation, which is almost a known fact for well-designed and planned networks.

If systematic validation of the clustering results is followed, then the cluster analysis could be transformed into an iterative process. Systematic validation is validation in terms of topology, followed by a validation of demand profiles and a validation of VVOs occurring on the feeders with different levels of PV penetration (similarly to the process performed in [38]). If the validation of topology fails then it means variables related to the architecture of the feeders need to be added or adjusted. Similarly, if the validation of the demand profile fails, then variables related to the types of clients must be revisited.

Moreover, studying the topology, the demand profile and the VVOs with PV penetration will be more informative than aiming directly for the hosting capacity. It is also worth mentioning that performing the three validations on the centre, closest neighbour and furthest neighbour of each cluster will determine if the variation within a cluster is substantial or not. The closest neighbour to the centre represents the case of the best similarity within a cluster, whereas, the furthest neighbour represents the feeder with the largest possible difference within a cluster [32].

An emerging non-linear technique, especially in the field of clustering, is the t-Distributed Stochastic Neighbor Embedding (t-SNE) [41], which Van der Maatens and Hinton developed in 2008. This method has a higher computational and memory complexity than PCA. It is capable of retaining the local structure of the data while at the same time displaying global structures. This means that in the case of overlap between two clusters using PCA,

the separation between these two clusters will be well defined with t-SNE. This is because t-SNE is purposed to interpret complex relationships in the feature space and to account not just for distant dissimilar data points but also for close similar ones. Thus, this thesis assesses the performance of t-SNE as a dimensionality reduction technique in parallel with the linear technique of PCA.

1.3 Problem Definition

Generally, distribution systems were designed to flow in one direction [15], from source to load. Recently, DG which includes renewable resources like PVs and other technologies are becoming more prevalent. DNOs face many upcoming hazards if no proper planning for these technologies is performed. The hazards relate to thermal overload, voltage fluctuations and statutory limit violations, and even protection miscoordination. The importance of analyzing the impact of these technologies in distribution systems has been addressed at length in the literature. However, with a large number of feeder models, full-scale impact analysis, in this case, is expensive and requires a substantial number of heavy simulations.

With a small set of feeders that characterize the whole distribution network, fewer simulations can be performed to understand the impact of any technology and eventually necessary mitigation measures can be set up. The improved procedure proposed in this thesis is expected to provide a guideline for the identification of a set of representative feeders that can be utilized by distribution utilities.

1.3.1 Thesis Statement and Contributions

This thesis builds on previous work and proposes an improved clustering procedure that utilities can use to find representative feeders of the totality of their feeders. These feeders can be used for the purpose of technology introduction studies and test runs. In particular, the proposed procedure first describes a guideline on feature selection and pre-processing. These features are further refined by using state-of-the-art dimensionality reduction techniques. Furthermore, to cover all the possible structures of the data, three clustering algorithms, and three indicators to determine the ONC are used and comparative analysis across these techniques is performed. Moreover, based on the nature of the problem, a validation technique of the results is proposed.

The contributions of this work consist of the following:

1. Detecting optimal structure of the data by performing a comparative analysis across different clustering techniques.

2. Proposing a systematic validation technique for the purpose of interpreting the applicability of the clustering techniques used and improving the clustering models for distribution feeders.

3. Introducing t-SNE into the domain of cluster analysis for distribution feeders and assessing the performance of this dimensionality reduction technique relative to PCA.

1.4 Thesis Organization

The remainder of this thesis is structured as follows:

Chapter 2 introduces the proposed features and how they can be selected. Moreover, it explains how the correlation between features is studied and how outliers can be detected. This is essential for achieving coherent and consistent clusters. Chapter 3 presents the techniques available to perform dimensionality reduction after refining the data in Chapter 2. It also describes mathematically the meaning of each technique. It covers a comparison between PCA and t-SNE. After selecting the features and reducing the complexity of the variation in Chapters 2 and 3, the cluster analysis step can be introduced. Chapter 4 revolves around the steps proposed in performing and analyzing clusters. It presents the different techniques and algorithms used and discusses the proposed approach to validate the resulting clusters. Chapter 5 exhibits the results of the case study on a large distribution system. It covers the performance assessment of different techniques to choose the optimal clusters. Chapter 6 concludes with closing thoughts on the work done in this thesis. It also describes future research improvements and perspectives of this work.

Chapter 2

Feature Engineering

2.1 Introduction

Most of the steps related to data processing fall under the label of Feature Engineering in the area of machine learning algorithms. Generally speaking, feature engineering involves two steps [4], [10]:

- Feature extraction
- Feature selection/construction

Prior to performing feature extraction in a dataset, some considerations in the form of a checklist need to be made [4]. The following considerations assist in understanding the nature of the features to use for model building.

1. Do we have a domain knowledge? If so, construct an adequate set of *ad hoc* features.
2. Are the features commensurate? If not, normalize them.
3. Are the features independent? If so, expand the feature space by constructing relevant features.
4. Suspecting outliers or noise in the data? If so, detect and remove them.

These questions define the scope of this chapter. In this chapter, we go through the very first steps to be taken in terms of data preparations and pre-processing for our procedure.

There is a grey zone between the contents of Chapters 2 and 3. These two chapters describe feature engineering. Specifically, Chapter 2 deals more with feature construction (better-said feature transformation) and Chapter 3 deals with dimensionality reduction.

2.2 Proposed Features

In the field of machine learning, constructing a set of *ad hoc* features is relatively uncomplicated. Usually, practitioners extract all relevant features, then, evaluate their relevance, which will be covered in future steps. The same idea applies to the clustering process in this thesis.

Alternatively, in [2], the authors used 35 different electric features in their clustering model. Although the authors did not give details on the complexity of the variation of the correlation between these features, the majority of them are relevant. Some of these features are overhead circuit length, underground circuit length, connected kVA (industrial, commercial and residential), impedance, etc.

In [38], the authors refined these features and added features related to the three-phase length of conductors, voltage regulation, capacitors, and winter-to-summer consumption ratio. The knowledge gained from the previous work combined with the understanding of how feeders are modelled (Chapter 1) enables the formulation of a list of variables or features that are the most relevant for the clustering approach proposed in this thesis. Moreover, the experience of the utilities' engineers helps in deciding if some of these features are useful or not. For example, for the case study in this thesis, the list of the features does not include features related to shunt capacitors. This is because the shunt capacitors are used for transmission support and can rarely affect the performance of the distribution feeders [48].

In general, the proposed features to be used must account for the topology of the feeders, the loads' behaviour and distribution, the voltage regulation, as well as the shunt capacitors along the feeders, if the latter is used in the distribution network. Grouping MV feeders at the distribution level and based on these aspects should be sufficient to capture the electric behaviour [13]. Therefore, a list of 16 features is proposed in this thesis. It can be found in Table 2.1. A systematic definition for all the features in order of numerical label is first provided below.

1- Number of three-phase branches The total number of three-phase connections present on the main three-phase conductor of the feeder.

2- Maximal Impedance The impedance per unit of the furthest point of the feeder, which corresponds to the maximal impedance of the feeder.

3- Total kVA consumed The total maximum consumption in kVA recorded on the feeder during the winter peak period.

4- % of kVA commercial The percentage of total maximum kVA that is consumed by commercial loads.

5- % of kVA industrial The percentage of total maximum kVA that is consumed by industrial loads.

6- % of kVA residential The percentage of total maximum kVA that is consumed by residential loads.

7- Single-phase overhead distance The total length of the single-phase overhead conductors, in km.

8- Single-phase underground distance The total length of the single-phase underground conductors, in km.

9- Three-phase overhead distance The total length of the three-phase overhead conductors, in km.

10- Three-phase underground distance The total length of the three-phase underground conductors, in km.

11- Number of voltage regulators The total number of three-phase voltage regulators present on the feeder. (Assumption that all voltage regulators have the same ratings)

12- Coefficient of variation of the distance between the load and the source Ratio of the standard deviation and the average of the distance between all the loads and the source of the feeder. It is referred to as the coefficient of variation and calculated using (2.1).

13- Average distance between the load and the source The average of the distance between all the loads and the source of the feeder.

14- Coefficient of variation of kVA Ratio of the standard deviation and the average of the kVA values of all the loads of the feeder. It is referred to as the coefficient of variation and calculated using (2.1).

15- Average kVA consumed The average of kVA values of all the loads of the feeder.

16- Number of shunt capacitors/compensators The total number of three-phase voltage capacitors present on the feeder. (Assumption that all capacitors have the same ratings)

The coefficient of variation is expressed by (2.1), and it is used as an indication of the level of variation around the mean for features 12 and 14 [49].

$$c_v = \frac{\sigma}{\mu} \tag{2.1}$$

where σ is the variance (can also be the standard deviation) and μ is the mean of all the loads of a feeder.

Feature 11 "Number of voltage regulators" is used not just as an indication of whether the feeders are long in distance but also to represent the voltage regulation aspect.

Feature 16 "Number of shunt capacitors" is added to cover the aspect of voltage control, which may be found in some distribution networks.

As can be seen below, Table 2.1 is divided into three parts: one part is dedicated to topology and another part is dedicated to capturing the distribution of the load and the type of the load along the feeder and the third part is considered between the topology and load (an example is the number of voltage regulators).

Table 2.1 Proposed Features

Topology	***Load***
1- Number of three-phase branches	3- Total kVA consumed
2- Maximal Impedance	4- % of kVA commercial
7- Single-phase overhead distance	5- % of kVA industrial
8- Single-phase underground distance	6- % of kVA residential
9- Three-phase overhead distance	14- Coefficient of variation of kVA
10- Three-phase underground distance	15- Average kVA consumed
11- Number of voltage regulators	
12- Coefficient of variation of the distance between the load and the source	
13- Average distance between the load and the source	
16- Number of shunt capacitors/compensators	

2.2.1 Data Quality

To have a dataset that contains consistent and correct values is something that is presumed in most of the topics concerning feature engineering. However, with the introduction of big data and the modernization of the grid, data management is a major challenge for utilities [10]. Therefore, data retrieval becomes challenging and errors are easily made.

Data quality in this sense is different from the data quality where the presence of outliers is suspected. It simply means that the data retrieved is correct and consistent. For illustration, if the data retrieved displays that the impedance of an urban feeder is as high as that of a rural feeder than there is a problem. This issue is sensitive since some erroneous data points

might not be detected as outliers and, hence, could affect the performance of the clustering model.

Unfortunately, this issue does not have a direct solution. Nevertheless, the elimination of outliers and the study of the features' correlation coupled with pair plots can attenuate the effects of possible errors.

2.3 Data Processing

Before starting any analysis or performing a transformation of the data, there are certain steps to follow after data extraction. These steps involve eliminating outliers, studying the correlation between the features, and scaling of the data.

2.3.1 Outliers

The statistical definition of an outlier is an observation that varies significantly from the rest of the observations [49]. An outlier can be caused by many errors, from data collection to measurement errors. These errors can introduce bias to the clustering model and thereby affect the coherence of the clusters. Furthermore, the presence of outliers directly affects the correlation coefficients [50], which in turn influence the results of data transformation techniques such as PCA [5].

A more concrete definition of an outlier would be a point that falls beyond the 1.5 times interquantile range (IQR) [49]. The interquantile range is a measure of statistical dispersion, being equal to the difference between upper and lower defined quantiles [49].

The steps to detect outliers using IQR can be summarized as follows:

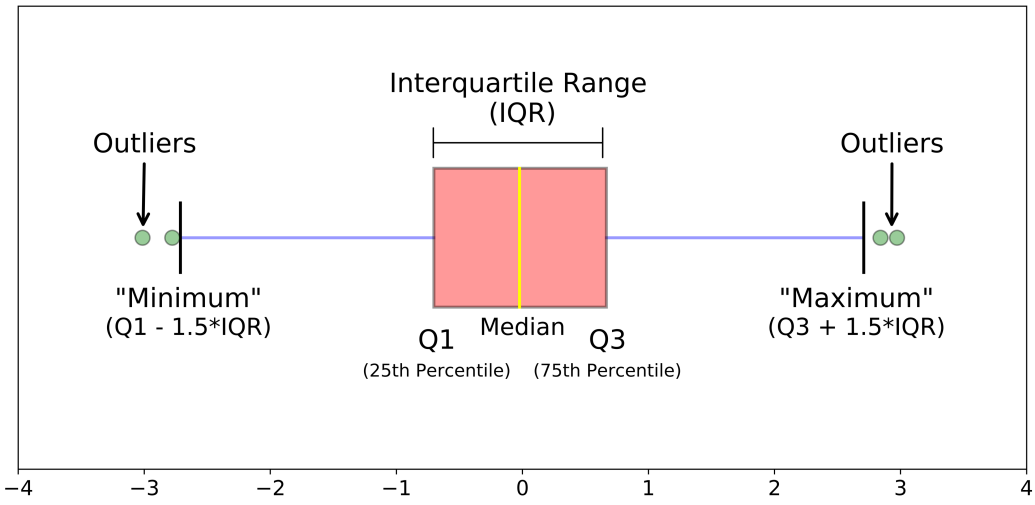
First, the data is sorted in an ascending manner. Then, lower and upper percentages are decided (e.g. q_1=25% and q_2=75%).

After that, the lower and upper values of the sorted data are identified at two positions $Q_1 = \lfloor q_1 \cdot (n+1) \rfloor$ and $Q_2 = \lfloor q_2 \cdot (n+1) \rfloor$, where n is the total number of points and Q_1 and Q_2 are the lower and upper values.

Once Q_1 and Q_2 are identified, (2.2) is used to find the outliers. As previously mentioned, the outliers are any value in the data beyond 1.5 times the interquantile range.

$$\begin{aligned} IQR &= Q_2 - Q_1 \\ outlier &= \begin{cases} > Q2 + 1.5IQR \\ < Q1 - 1.5IQR \end{cases} \end{aligned} \tag{2.2}$$

Fig. 2.1 illustrates the notion of an outlier using a boxplot where q is defined as a multiple of quarters and not a percentile. Boxplots are an efficient method to quickly test for outliers [49].



Source: towardsdatascience.com, Understanding Boxplots.

Fig. 2.1 Box-Whisker or Boxplot example.

There are many other methods for outlier elimination. For instance, the scaling technique that will be discussed in the following section is less affected by the presence of outliers compared to min-max scaling [51]. In addition, clustering algorithms that are designed to detect clusters based on the points' density like the DBSCAN algorithm can further be used to identify as well as eliminate outliers [52].

2.3.2 Correlation Study

Studying the correlation between features is part of feature selection and is a subsequent step to outlier elimination. In fact, it is an essential step to determine if the same information is presented by more than one feature.

One method for determining the correlation between features is the visualization of the covariance matrix of the data. A large covariance between two features indicates that the features are highly correlated, suggesting that they contain information that can be predicted or represented by a single feature, therefore implying redundancy.

The Pearson correlation coefficient is the type of correlation used in this thesis [49]. The Pearson correlation coefficient measures the extent of the linear relationship between two features. The goal of studying the correlation using the Pearson coefficient is to eliminate the features that are highly linearly correlated. Other coefficients, such as the Spearman

correlation coefficient [49], measure non-linear correlations. We are, however, less interested in these types of coefficients since they will indicate a relatively high correlation between the majority of the features. This is because, if the length of the conductors increases then, the maximum impedance, the number of three-phase branches, and many other features will most likely increase in typical feeders. Hence, the use of Pearson correlation is best for our purposes in this thesis [50].

According to the Cauchy–Schwarz inequality, the Pearson correlation coefficient has a value between $+1$ and -1, where 1 is a total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation.

It is calculated based on (2.3) where r is the Pearson coefficient value between two vectors X and Y, n is the number of points, $\bar{x}$ and $\bar{y}$ are respectively the mean values of vectors X and Y, σ is the standard deviation and $cov(\cdot,\cdot)$ is the covariance.

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{cov(X,Y)}{\sigma_X\sigma_Y} \end{aligned} \tag{2.3}$$

Deciding on a threshold for the correlation involves visualizing the features by pair plots to verify the relationship between the different features.

The statistics community considers a Pearson coefficient of ± 0.77 to be indicative of a high correlation, given the absence of outliers [53]. Since the outliers are eliminated, a general coefficient threshold of ± 0.77 is set, above which, the features are considered correlated.

Therefore, after visualizing the covariance matrix, any pair of features that have a correlation of ± 0.77 and more should be investigated with the goal of keeping only one of them. It is also advised to visualize pair plots of the features in question to verify the linear relationship between them.

2.3.3 Scaling

Centering the data or scaling is common jargon in the field of machine learning; it means putting the features on the same order of variance, magnitude, etc. in order to make them comparable. This is the final step prior to performing any transformation on the original data. There are many types of scaling. Features can be scaled by removing the mean and matching the variance like in (2.4) (known as z-score or standard scaling).

$$x'(i,j) = \frac{x(i,j)-\mu(F_i)}{\sigma(F_i)} \tag{2.4}$$

where x' is the new value of feature i denoted by (F_i) at data point j. μ and σ are respectively the mean and variance values of feature i.

Another scaling can be MinMax. Using this type of scaling, the features are compressed to the [0,1] range of values, (2.5).

$$x'(i,j) = \frac{x(i,j) - \min(F_i)}{\max(F_i) - \min(F_i)} \tag{2.5}$$

where x' is the new value of feature i denoted by (F_i) at data point j. min and max are respectively the minimum and maximum values of feature i.

Standard scaling is favoured over the MinMax scaling [51] especially when using PCA, which reduces the dimensionality based on the variance [5].

2.4 Categorical Features

This section explains an issue that is usually not encountered in the field of clustering. It relates to the presence of many categorical features among the discrete and continuous features (mixed data). Categorical features can be, for example, the number of regulators present or the number of shunt capacitors. These features have relatively a few numbers of values (typically the number of regulators varies between one and five). In the case where many categorical features are present, the scaling should exclude them and a different type of clustering algorithms is advised. An example of these algorithms is the k-prototype algorithm [54]. This algorithm takes into account the different categorical values or classes of the data while performing the clustering on the discrete or continuous features [54]. This section is outside the scope of this thesis. Nonetheless, it is imperative to briefly discuss the issue of mixed data in case practitioners choose to encode some discrete or continuous features into categorical ones.

2.5 Conclusion

This chapter discussed the very first steps in the procedure outlined in this thesis. The features to be used in the upcoming chapters were identified. This chapter further highlighted the processes that the raw data should pass through. All in all, the outliers should be detected and removed, the correlation of the features should be studied, and the resulting data should be scaled using an adequate scaling technique.

Chapter 3

Dimensionality Reduction

3.1 Introduction

Dimensionality reduction is a class of problems in the field of unsupervised learning. Dimensionality reduction is simply reducing the dimension of the feature set used whether by the elimination of some irrelevant features and the transformation of others.

Dimensionality reduction is used to tackle the curse of dimensionality. The curse of dimensionality is basically the case where the number of features is relatively big, compared with the number of data points. A machine learning model trained with a large number of features and fewer data points is at risk of overfitting and thus, increasing the odds of poor performance on new data.

Moreover, dimensionality reduction has many more advantages that motivate its incorporation into the approach proposed. Evidently, fewer features mean less computation and memory requirements. However, the most relevant advantage of dimensionality reduction is that the output is a simplified representation of the relative positioning of data points, thus obtaining a reduction in the number of explanatory factors necessary to model data variation. With less misleading data, the clustering model's ability to detect coherent clusters increases. There is also another specific advantage of PCA, which will be discussed in the next section.

This chapter explains how dimensionality reduction is performed after eliminating the outliers and scaling the data as seen in chapter 2. In particular, this chapter covers the two main dimensionality reduction techniques used in this thesis, PCA and t-SNE. While PCA is a linear technique, t-SNE is a non-linear stochastic technique that is computationally expensive. At the same time, t-SNE is capable of unfolding more complex variations in the data than PCA.

3.2 PCA

In general, PCA is a data-analytic technique that linearly transforms some features into fewer uncorrelated ones. It aims to maximize data variance by transforming the data using the eigenvectors that correspond to the highest eigenvalues of the correlation matrix, in other words, respecting the major structure of the original data [5, 6].

3.2.1 Mathematical Formulation (Shalev-Shwartz and Ben-David, [1])

Let X be a collection of vectors in $\mathcal{R}^m$. The aim is to reduce the dimensionality of X using a linear transformation.

Let the linear mapping $X \mapsto W^T X$, where W is called the compression matrix with dimensions $\mathcal{R}^{m \times m'}$, $m > m'$.

Another matrix U in $\mathcal{R}^{m' \times m}$ can be used to reconstruct the original data from its compressed version.

If $Y = W^T X$ is the compressed version of X, then $\hat{X} = UW^T X$ is the reconstructed version of X.

The goal is to minimize the error between X and its recovered version $\hat{X}$, thus, to solve 3.1.

$$\text{argmin}_{W,U} \left\| X - XWU^T \right\|^2 \tag{3.1}$$

Based on [1], the solution is given by the eigenvalue decomposition of the matrix XX^T where $W = U^T$ and U is the matrix whose columns are the eigenvectors of XX^T.

The principal components are the first m' eigenvectors of XX^T or the first m' left singular vectors of X. The vectors are sorted based on the value of the eigenvalues.

The matrix XX^T has many characteristics:

- It is symmetric which means that the eignevectors are orthogonal.
- It is a positive semidefinite matrix which means that the eigenvalues are non-negative.

By definition, the first principal component, which corresponds to the largest eigenvalue, is the axis directed along the component with the most variance in the data. Similarly, the second principal component corresponds to the second largest eigenvalue and is, at the same time, orthogonal to the first principal component. The orthogonality of the eigenvectors makes the new spanned features uncorrelated.

Evidently, m' should be chosen in a manner that preserves enough information from the original data. Therefore, one can decide on the new dimension of the data m' by analyzing

the variance of the projections. This can be done by analyzing the eigenvalues because they explain the total variance as described by each component.

Practitioners use the term "explained variance", which is the normalized value of the eigenvalues, to evaluate how much variance is preserved with different values of m' [47]. In this thesis, m' is chosen such that to have a total explained variance of over 85%, meaning that at least 85% of the information is preserved in the new data.

A final characteristic of the principal components is the ability to quantify how much each original feature contributes to the formation of the new feature space. In other words, it is possible to determine which features are represented the most in the new feature space by analyzing the weights of the principal components or eigenvectors.

3.2.2 Clustering via PCA

In this section, many advantages of PCA in the context of cluster analysis are demonstrated. It is shown that the clusters formed are better refined when PCA is used [7], concluding that PCA improves the extraction of cluster structure, with respect to stability, refinement, and coincidence with known ground truth. This means that PCA can partially unveil the structure of the data, especially if the structure is defined by the largest variances.

The work of Ding and He in [55] showed how the clustering algorithm k-means and PCA relate mathematically. They proved that PCA automatically performs data clustering according to the k-means objective function.

3.3 t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique for dimensionality reduction, particularly in two or three dimensions. The technique is a variation of Stochastic Neighbor Embedding. t-SNE is better than existing techniques at revealing structure at many different scales. This means that t-SNE is capable of revealing the local structure of the high-dimensional data, while also revealing its global structure [41].

3.3.1 Mathematical Formulation

For the sake of simplicity, a simple version of the mathematical formulation of t-SNE is presented in this section, while the full details can be found in [41].

Let $X = x_1, x_2, ..., x_n$ be a high dimensional dataset and $\mathscr{Y} = y_1, y_2, ..., y_n$ be a two or three dimensional representation of the dataset X.

We start by calculating the similarity between points in high dimensional space. The similarity between points is calculated as the conditional probability $p_{j|i}$ that a point x_i

would choose point x_j as its neighbour if neighbours were picked in proportion to their probability density under a Gaussian distribution centered at x_i. This step is expressed by (3.2):

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)} \ , \ p_{i|i} = 0 \tag{3.2}$$

where σ_i is the variance of the Gaussian that is centered at x_i, the calculation of this parameter is discussed later in this section.

The pairwise similarities in the high-dimensional space are defined in (3.3)

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{3.3}$$

In SNE, the pairwise similarities in the low-dimensional space are given by (3.4)

$$q_{ij} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq l} \exp\left(-\left\|y_k - y_l\right\|^2\right)} \tag{3.4}$$

However, in t-SNE, a Student t-distribution with one degree of freedom is used instead in the low-dimensional space for the pairwise similarities, which is defined as follows

$$q_{ij} = \frac{\left(1 + \left\|y_i - y_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \left\|y_k - y_l\right\|^2\right)^{-1}} \ , \ q_{ii} = 0 \tag{3.5}$$

The reason for using a Student t-distribution is to allow a small dissimilarity distance in the high-dimensional space to be represented by a relatively large dissimilarity distance in the low dimensional space.

The goal of t-SNE is to minimize a function, called Kullback-Leibler divergence (3.6), between a joint probability distribution, P, in the high-dimensional space and a joint probability distribution, Q, in the low-dimensional space.

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.6}$$

where p_{ii} and q_{ii} are zero and p_{ij} and q_{ij} are given by (3.3) and (3.5).

After calculating the similarities in the high-dimensional space using (3.2) and (3.3), we iterate T times and calculate:

- The similarities in the low-dimensional space using (3.5).
- Followed by the calculation of the gradient of the cost function, which is calculated

using (3.7)

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j=1}^{n} (p_{ij} - q_{ij}) (y_i - y_j) \left(1 + \|y_i - y_j\|^2\right)^{-1} \tag{3.7}$$

- Finally, calculating the new coordinates using (3.8) with pre-defined learning rate η and momentum for the gradient $\alpha(t)$. $\frac{\partial C}{\partial \mathscr{Y}}$ is calculated using (3.7).

$$\mathscr{Y}^{(t)} = \mathscr{Y}^{(t-1)} + \eta \frac{\partial C}{\partial \mathscr{Y}} + \alpha(t) \left(\mathscr{Y}^{(t-1)} - \mathscr{Y}^{(t-2)}\right) \tag{3.8}$$

The aforementioned pre-defined parameters are set with default values: $T = 100, \eta = 200$ and α can be any value in the range $[0.2, 0.8]$.

It is important to note that the variance of the Gaussian in (3.2), σ_i, is set based on a user-defined parameter called perplexity. However, t-SNE is not substantially affected by the perplexity [41]. The perplexity can be considered as a measure of the effective number of neighbours, its values are between 5 and 50 and the value to use in this thesis ranges between 20-25.

3.3.2 t-SNE vs PCA

In this section, the differences between the two techniques are discussed. The users are also advised on which one to use based on their specific application.

As mentioned before, PCA finds a linear transformation of the data that maximizes the variance or, as shown by [55] minimizes the sum of the squared errors between high-dimensional pairwise distances and their low-dimensional representatives. Since it is linear, it will not be able to represent more complex relationships between the features.

PCA focuses on representing the distances between widely separated data points (maximum variations) rather than on preserving the distances between nearby data points. t-SNE, on the other hand, can unveil global structures while taking into account the small differences between data points.

However, since PCA is a linear technique, the data points in the low-dimensional space preserve their relative closeness or similarity. t-SNE does not have this feature due to its stochastic nature. If two points are close to each other in the low dimensional space, using PCA, we can quantify how much these two points differ in each feature. But using t-SNE, it is indicated that these two data points are similar because they are placed close to each other.

t-SNE is a stochastic technique, which means different structures can be produced with different runs, even with the same parameters. This can be solved by producing different

runs and performing a visual assessment.

Evidently, t-SNE is computationally expensive, and running it for a relatively small dataset can take minutes and even hours, using a relatively powerful computer. PCA, on the other hand, will not exceed seconds or minutes, as PCA relies on well-established eigenvalue/eigenvector calculation libraries.

After highlighting the advantages and disadvantages of each technique, the conclusion is that it is advised to use PCA when the user is interested in studying the relative closeness between datapoints or clusters. However, if the aim is to have a better separation between clusters then t-SNE is the method to use.

3.4 Conclusion

In this chapter, the dimensionality reduction is explained and the advantages of adopting this step in the procedure of this thesis are presented. Two state-of-the-art techniques are used to perform dimensionality reduction. PCA is a linear data transformation technique based on the variance maximization. t-SNE is a non-linear stochastic approach that can reveal complex structures in the data. Each approach has its advantages and disadvantages and depending on the user's preference, the user can choose to use either of the two approaches in the process.

If users are more interested in the relative similarity between the data points after performing clustering, they are recommended to use PCA to reduce the dimensionality of the problem at hand. However, if it is more important to have better-defined clusters then t-SNE is recommended. In the case study presented in Chapter 5, the two techniques are applied and a comparison is made to illustrate the pros and cons of each method as discussed in this chapter.

Chapter 4

Cluster Analysis

4.1 Introduction

The different techniques available to perform unsupervised cluster analysis are discussed in this Chapter. In particular, different methods to find the ONC and the algorithms to use in the process are presented. Moreover, an approach to validate the results of various clustering techniques is proposed. This step involves studying the topology and load profiles of the distribution feeders as well as performing network studies on selected feeders.

The literature review in Chapter 1 revealed that there exist several approaches and algorithms to perform cluster analysis. These algorithms and techniques were used individually in most cases. Hence, a strategy to perform a comparison between the algorithms is proposed in this chapter. Performing a comparative analysis could assure that optimal results are achieved by having explored several possible solutions.

4.2 Clustering Algorithms

Clustering algorithms are in general based on a number of paradigms [3]. These include partitioning, hierarchical, density-based and distribution-based methods. In this thesis, the selection of algorithms was done so as to introduce variety in the manner by which clustering is achieved. As such, three algorithms, each belonging to a different clustering paradigm, are adopted and are as follows:

- Partitioning method: k-means++ algorithm
- Non-partitioning method: hierarchical algorithm
- Distribution-based method: GMM algorithm

These algorithms support unsupervised learning applications where no labeling of the data points is available which is the case for our application. Furthermore, it has been emperically established that these three algorithms perform well on different tasks [3, 47]. More details on each algorithm are provided in this section.

The metric or distance used in this thesis is the Euclidean distance and the representative feeder of each cluster is the feeder or neighbour closest to the fictitious center or centroid calculated by the algorithms. The fictitious centers are found automatically in the case of k-means++ algorithm; however, for the other algorithms, the centers are the averages of all the samples of each cluster.

4.2.1 K-means++ Algorithm

K-means++ [56] is a partitioning method that groups the data by attempting to separate the samples into n groups based on minimizing a criterion called inertia or sum of squares of the cluster criterion. Its objective function is described in (4.1).

$$\sum_{j=1}^{C} \sum_{i=1}^{n} d(x_i^{(j)} - \mu_j)^2 \tag{4.1}$$

where n is the number of samples or feeders, C is the number of clusters, μ_j is the mean of the samples in cluster j, and $d(x_i^{(j)} - \mu_j)$ is the Euclidean distance between sample i of cluster j and μ_j. As for the k-means algorithm, μ are the centers of the clusters, so-called fictitious centers. The representative feeder of each cluster is the one closest to this fictitious center.

The documentation of [47] best describes the concept of the k-means algorithm. The algorithm starts assigning centroids; then it loops in two steps. The first step consists of

assigning each sample to its nearest centroid (sometimes a distance threshold can be set), while the second step consists of computing new centroids through calculating the mean of the samples in each cluster formed. The iterations of these two steps continue until the centroids cease changing.

This method functions well with a large number of samples. Nonetheless, this method starts losing efficiency when the number of clusters is high or when the clusters have non-convex shapes. In addition, k-means++ starts with the random assignment of a cluster center, then searches for other centers based on the first one. The former approach is more effective than the arbitrary initialization of the n centroids by the k-means algorithm, where converging to a local minimum is more probable.

4.2.2 Hierarchical Algorithm

The hierarchical algorithm [3] is a method that merges or splits clusters in a successive way. It involves two techniques, namely: agglomerative and divisive.

The agglomeration approach performs a hierarchical classification using an ascending approach, whereas the divisive approach is a bottom down process. With the agglomeration algorithm, each observation starts as its own cluster, then the two closest clusters are merged together into one. Successively merging the closest clusters is repeated until there is only one cluster.

A reduction in the number of clusters is achieved by means of utilizing linkage criterion between the different clusters. Linkage criterion refers to the process of calculating the new distance between the clusters. Specifically, clusters with minimal linkage distance are paired into a new cluster. As such, the procedure of merging the closest clusters results in a minimal increase in the linkage distance between the resulting clusters. Linkage criteria or merging strategy, can be done in different ways [47].

An average linkage means that the algorithm will minimize the average distance between the samples of a cluster. Another linkage type is *ward*, which minimizes the sum of squared differences in all clusters [57]. It is an approach comparable to the objective function of k-means++ and it will be utilized in this thesis. Intuitively, the agglomerative algorithm leads to uneven cluster sizes since the larger a cluster is, the more samples it can group. Essentially, with this algorithm, the rich get richer and the linkage type ward is the best strategy to counteract this behaviour. In this thesis, the agglomerative approach is adopted since a linkage-type or strategy for merging that is similar to the objective function of the k-means++ algorithm could be used.

In order to avoid a computationally expensive problem, connectivity constraints can be added to the algorithm restraining it from merging very far apart samples. These constraints

can be added via a connectivity matrix [47]. The matrix can be generated using a graph of k-neighbours (KNG) [47]. A KNG generates a sparse binary matrix indicating which samples can be merged together. This matrix is generated based on a distance metric or number of neighbours. Both of these parameters are user-defined parameters.

4.2.3 GMM Algorithm

A Gaussian mixture model [46] is based on parameterizing Gaussian distributions. It involves a mixture or superposition of multiple Gaussian distributions and implements the expectation-maximization algorithm (EM) to fit the Gaussian mixing models by finding their corresponding means, co-variances and mixing probability. EM is a statistical algorithm that performs an iterative process similar to the k-means++. It starts with a random assignment of data points and then, assigns for each point a probability of being generated by a Gaussian model. It iterates over the various parameters (mean, variance and mixing probabilities of Gaussian models) with a maximization step (M) to maximize the likelihood (E) of the data [58]. Even though the aforementioned algorithm is fast in terms of fitting the mixture of models, it is susceptible to divergence, especially with small datasets.

Density-based algorithms like DBSCAN [52] could also be tested. This type of algorithm tends to group points into clusters of different densities. The density, in this context, is the average distance between a point and a defined number of neighbours. In the case where the density, for the majority of data points is high with respect to their neighbours, this type of algorithms would not be very effective in identifying relevant clusters. They would only highlight the samples that are scattered or that are considered outliers.

4.3 Optimal number of clusters

Cluster analysis is an unsupervised learning technique where prior knowledge about the structure of the data is not available. One of the most essential tasks in cluster analysis is determining the ONC. Therefore, to use the algorithms presented in the section above, the user must define the number of clusters to apply. In this thesis, three different indicators or scores are used to find the optimal number of clusters.

We note that in some cluster analysis applications, the optimal number of clusters is not required. For example in [9], the goal was to merely group similar measurements together so the optimal number of clusters was not needed. The authors were trying to detect recurrent events by comparing measurements and grouping similar ones together. In that case, the hierarchical algorithm was used with specific settings.

4.3.1 Silhouette Score

The Silhouette score [36] is one of the most used methods to determine the optimal number of clusters. It involves assigning a silhouette score, $s(i)$, ranging between -1 and 1 to each sample i where this score describes the similarity/dissimilarity of the sample in its cluster. The silhouette score is calculated using (4.2):

$$s(i) = \frac{a(i) - b(i)}{\max(a(i) - b(i))} \ , \ -1 \leq s(i) \leq 1 \tag{4.2}$$

where $a(i)$ is the average distance of sample i to all other samples of the same cluster A, and $b(i)$ is the average distance of sample i to all other samples of the closest cluster B. The average distance, $a(i)$, represents the average dissimilarity of sample i from the other samples in the same cluster A, and $b(i)$ represents the average dissimilarity of sample i from the other samples of the closest cluster B.

Once all the samples are assigned a silhouette score, an overall average silhouette coefficient also called silhouette width, denoted by $\bar{s}_k$, for all the samples is calculated for a number of clusters k to assess the global coherence of the clusters. One should choose the value of k that corresponds to the highest average silhouette coefficient.

A higher average silhouette coefficient indicates more coherent (dense and well separated) clusters. Similarly to the inertia criterion of the k-means algorithm, the overall average silhouette coefficient tends to be lower for clusters with irregular shapes than clusters with convex shapes.

4.3.2 Calinski-Harabasz Score

The Variance ratio criterion (VRC) or *Calinski-Harabsz* criterion [44] is another method of finding the ONC. It is the ratio between the within-cluster dispersion and the between-cluster dispersion. Its mathematical formulation is expressed by (4.3):

$$vrc_k = \frac{tr(B_k)}{tr(W_k)} \cdot \frac{n-k}{k-1} \tag{4.3}$$

where k is the number of clusters, n is the number of data samples, $tr(B_k)$ is the trace of between-cluster dispersion matrix and $tr(W_k)$ is the trace of within-cluster dispersion matrix.

These matrices are calculated as follows:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$
$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{4.4}$$

where C_q is the group of samples in cluster q, n_q is the number of samples in q, c_q is the center of cluster q, and c_E is the center of the whole dataset [47].

Based on the formulation above, the objective is to minimize $tr(W_k)$ because it contains information about the variance inside the clusters and we expect a limited dispersion around the centers of the clusters. At the same time, $tr(B_k)$ needs to be maximized in order to have centers of clusters that are distanced from the global center of the dataset.

Similarly to silhouette, a higher VRC score indicates more coherent clusters; however, calculating the VRC score takes less time than the silhouette score. It shares the same drawback with silhouette where VRC tends to be higher with convex clusters.

4.3.3 Davies-Bouldin Score

The Davies-Bouldin (DB) score [59], by definition, takes into account the average similarity of each cluster with its neighbouring cluster. Its formulation is a simpler version of the silhouette. It indicates the level of separation between clusters only, and it suffers from the same drawback of silhouette score, i.e., is usually higher for convex clusters. Mathematically, the DB score is the average similarity between each cluster and its closest neighboring cluster.

The mathematical formulation of the DB score is expressed by (4.5).

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij} \tag{4.5}$$

where R_{ij} is the similarity index between clusters i and j, it is defined by (4.6).

$$R_{ij} = \frac{a_i + a_j}{d_{ij}} \tag{4.6}$$

where a_i is the average distance between the points of cluster i and its centroid, and d_{ij} is the distance between the centroids of clusters i and j.

Based on this, a lower DB indicates a better separation between the clusters.

4.4 Comparative Analysis

Comparing the performance of the three clustering algorithms presented in Section 4.2 is done by analyzing the three indicators that assess the quality and coherence of clusters. In addition to that, evaluating the number of feeders that are considered mis-clustered by each algorithm is another step adopted in the comparative analysis. Feeders that are assigned a negative or close to zero silhouette score are described mis-clustered and they are usually situated at the borders of their cluster. Moreover, the distribution of the feeders or the total number of feeders in the clusters is compared to check if the clustering results of the three algorithms is comparable. These factors combined determine which clustering technique achieved better results.

4.4.1 Remark

It is important to note that since the used scores earlier in Section 4.3 tend to be lower for clusters with irregular shapes, they are less effective with the dimensionality reduction technique t-SNE because it can reveal more elaborate cluster shapes.

4.5 Validation Process

Considering that clustering analysis is an unsupervised learning technique where the ground truth labels are not known, validation of the results can be done by exploiting the nature of the problem. As the problem is to identify representative distribution feeders, the verification of the coherence of the clusters formed can be achieved by comparing the electric behaviour of a subset of feeders from each cluster.

Choosing the feeders to verify the coherence of the clusters is essential since these feeders should cover most of the variations in the clusters. In [32], the authors chose the farthest NOP from the center of each cluster to perform reliability assessment as the farthest point from the center represents the worst similarity. In this thesis, the choice of the following feeders from each cluster is proposed: the representative feeder or the center of the cluster, one of its nearest neighbours (NN) and the farthest neighbour (FN) from the center, similarly to [32].

With this selection, the variation in each cluster can be determined, the NN represents the highest similarity achieved by each cluster and the FN represents the worst similarity or highest variation. Therefore, if a certain validation is verified for these feeders, in other words, if these three selected feeders have similar electric behaviour, then this is an indication that the clusters formed are dense and well separated. The validation consists of comparing

the selected feeders in three steps.

1. **Validation of the topology**: All feeders of the same cluster must share similar topologies and at the same time, must have different topologies from feeders of different clusters. The validation of topology can be performed visually by the users. An additional step can be comparing the number of three-phase branches in each cluster as this feature is a good indicator of the topology. This can be done by plotting boxplots of the three-phase branches feature for each cluster.

   Through scanning all the clusters, it should be found that the topology of the center, the NN, and FN are comparable and consistent. The same thing applied for the number of three-phase branches. For example, clusters characterized with short distances and low impedance should have relatively low values for the number of three-phase branches compared to clusters that are characterized with long distances and high impedance.

2. **Validation of the demand profile**: Similarly, the demand profile should be consistent with the characteristics of each cluster. The demand profile can be studied during a period with high load consumption, excluding the weekends, as the consumption during weekends may be abnormal. The period to be selected can be the average of the weekdays of a month that is characterized by high consumption (e.g., during the winter season). With high consumption, the demand patterns for different types of clients can be readily distinguished. The demand profile is also normalized based on the absolute maximum value.

   This means that the demand profile is a 24 hours vector containing the hourly average of the weekdays. The metric used for the normalized demand profile is the mean square error or (MSE). The MSE between the normalized demand profile of the center, the NN and FN should be less than $0.4 - 0.5$. Otherwise, this means that the demand profile varies substantially between the three feeders.

3. **Validation of Voltage Violations and Overload**: The validation is conducted by considering VVOs with different levels of PV penetrations (defined as % of clients installing PVs), similarly to the work done in [42].

   Voltage violations are identified by verifying if the voltage value at the clients connection point is outside the permissible range set by the utility. Overloads are identified by checking if the value of the current exceeds the permissible threshold. VVOs can be flagged by the simulation software used during the simulations.

   For 50%, 75% and 100% penetration levels, an identification of VVOs is performed while maintaining the load at the average consumption during summer. To perform

the simulations, a constant demand value corresponding to the average consumption during summer is allocated to all the clients.

It is expected that the center, the NN and FN show consistency in terms of VVO identification with different levels of PV penetration.

If one of the three validations fails, meaning that if the topology, demand profile, or VVO validation varies significantly (e.g., MSE above 0.4 for the demand profile) within the selected feeders of the same cluster, the features selected must be revisited and this is how the notion of the iterative process is introduced. For example, in a first run, the topology validation for the dataset used failed due to the absence of features related to underground cable distances. Hence, features 8 and 10 were added in a second iteration. The objective of these validations is to test the coherence of the clusters and at the same time to verify the adequacy of the choice of features. This validation is clarified in Chapter 5.

4.6 Conclusion

The last steps to follow in the procedure of this thesis and the process of determining the ONC and which algorithms to use were discussed and explained in this chapter. In particular, three different scores to find the ONC and three clustering algorithms, each of different type were adopted in this thesis. Furthermore, comparative analysis across the algorithms was proposed to determine the optimal results.

An approach to validate the results of the cluster analysis was further proposed in this chapter. This approach exploits the nature of the problem where the electric behaviour of selected feeders is studied to verify that the feeders of the same cluster behave similarly.

Chapter 5

Case Study

5.1 Introduction

In this chapter, all the steps discussed in Chapters 2-4 are followed and a case study is performed on 2975 distribution feeders from a distribution system of a utility covering a large area. First, a demonstration is carried out on how data processing is performed, including the choice of features and the elimination of outliers. In the following section, the dimensionality reduction using PCA and t-SNE is performed. The results of the two techniques are analyzed, and the differences between the two are pointed out. Afterwards, cluster analysis using the different techniques explained in Chapter 4 is performed. Before providing a conclusion, the validation steps explained in Chapter 4 are executed and the corresponding results are displayed.

In any data analysis or machine learning project, visualization of results is not only a way of communication, but also an analysis strategy. It is specifically of value to the field of cluster analysis. Therefore, this chapter extensively presents results graphically with detailed analysis. Moreover, the clusters in the different graphical representations of the results are identified using different colors. Moreover, information on the operating conditions and the hyperparameters of the algorithms used and their respective packages are presented in the appendix to allow the reader to reproduce the results. The diagram that illustrates the steps described in the previous chapters is shown in Fig. 5.1.

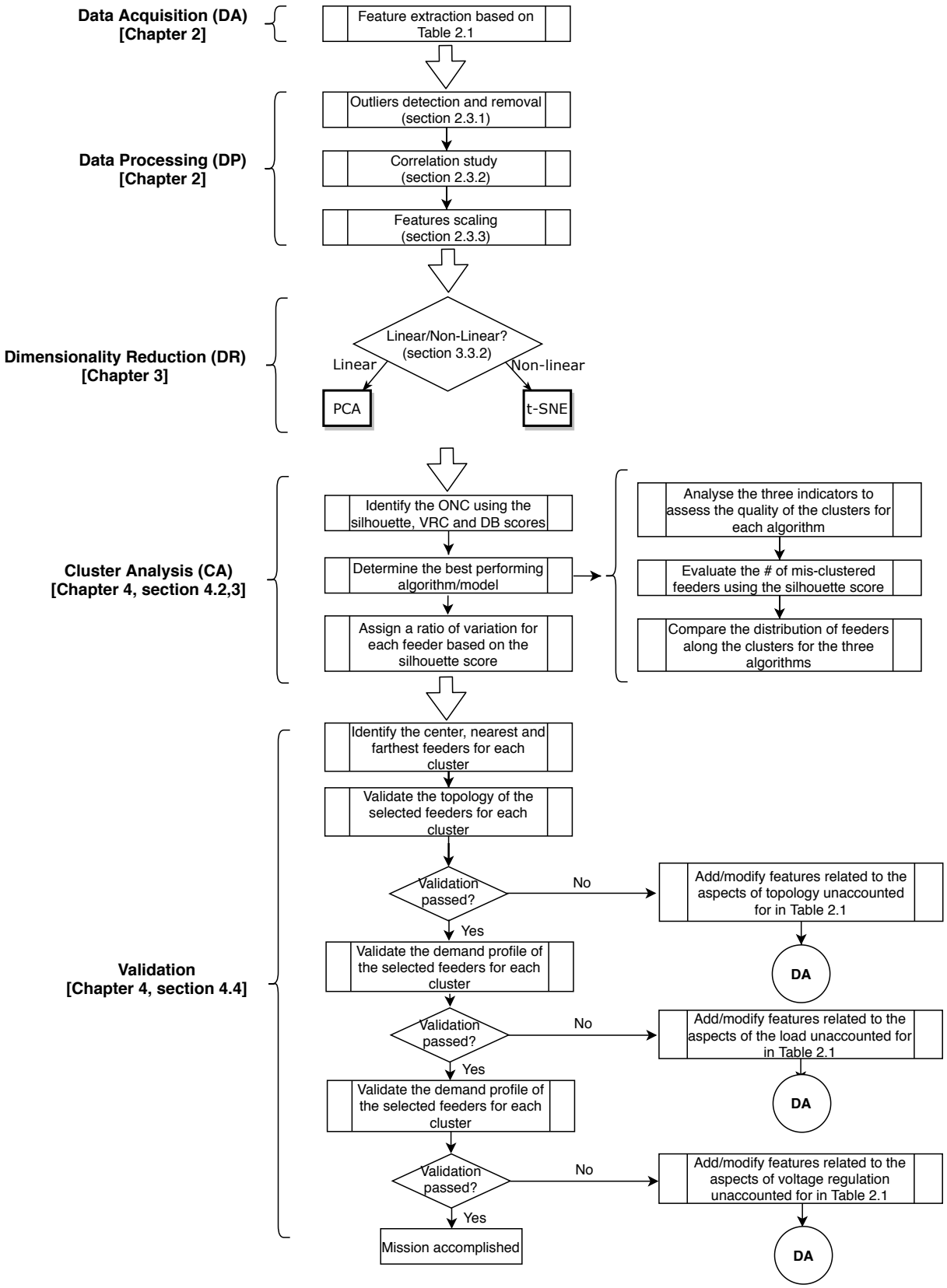


Fig. 5.1 Diagram of the overall procedure proposed in this thesis

5.2 Data Processing

This section covers the initial steps of the overall process (Chapters 2 and 3). It is important to note that the data provided is consistent and of good quality, meaning, the data is clean of errors. As mentioned in the introduction of this chapter, the dataset used for this case study consists of 2975 distribution feeders from a utility covering a large area.

5.2.1 Data Pre-Processing and Feature Engineering

The following work consists of eliminating outliers, studying the correlation between the features and performing the data scaling. The features used in this case study are presented in Table 2.1 except feature 16 (shunt capacitors presence). These features have also been recommended by the distribution system operators of the utility that provided the data. As discussed in Chapter 2, these features characterize the topology, the distribution of load and the voltage regulation along the feeders. Feature 16 is not included in this case study since the operation of distribution feeders does not involve shunt capacitors.

An important step in pre-processing is the elimination of outliers. Many techniques discussed in Chapter 2, like the interquartile range (IQR) can be used. For this study, since the data provided was consistent and of good quality, feeders with a total kVA of less than 500 kVA and more than 25 000 kVA, and networks with a nominal voltage other than 25 kV were considered outliers, and, as a result, the final number of feeders was reduced by 15%.

Before scaling the data based on (2.4), the correlation between the features used is studied by visualizing its correlation matrix. In Fig. 5.2, the correlation between the features is visualized using a heatmap, where dark colours indicate a high correlation. Fig. 5.2 indicates some high correlation between specific features like features 6 (% of kVA residential), 9 (three-phase overhead distance) and 13 (Average distance between the load and the source).

Thus, in the final covariance matrix, one eliminates the highly-correlated features (like 9 and 6 in Fig. 5.2) . Feature 6 corresponds to the % of residential kVA, which can be inferred from features 4 and 5, the percentages of commercial and industrial kVA, respectively.

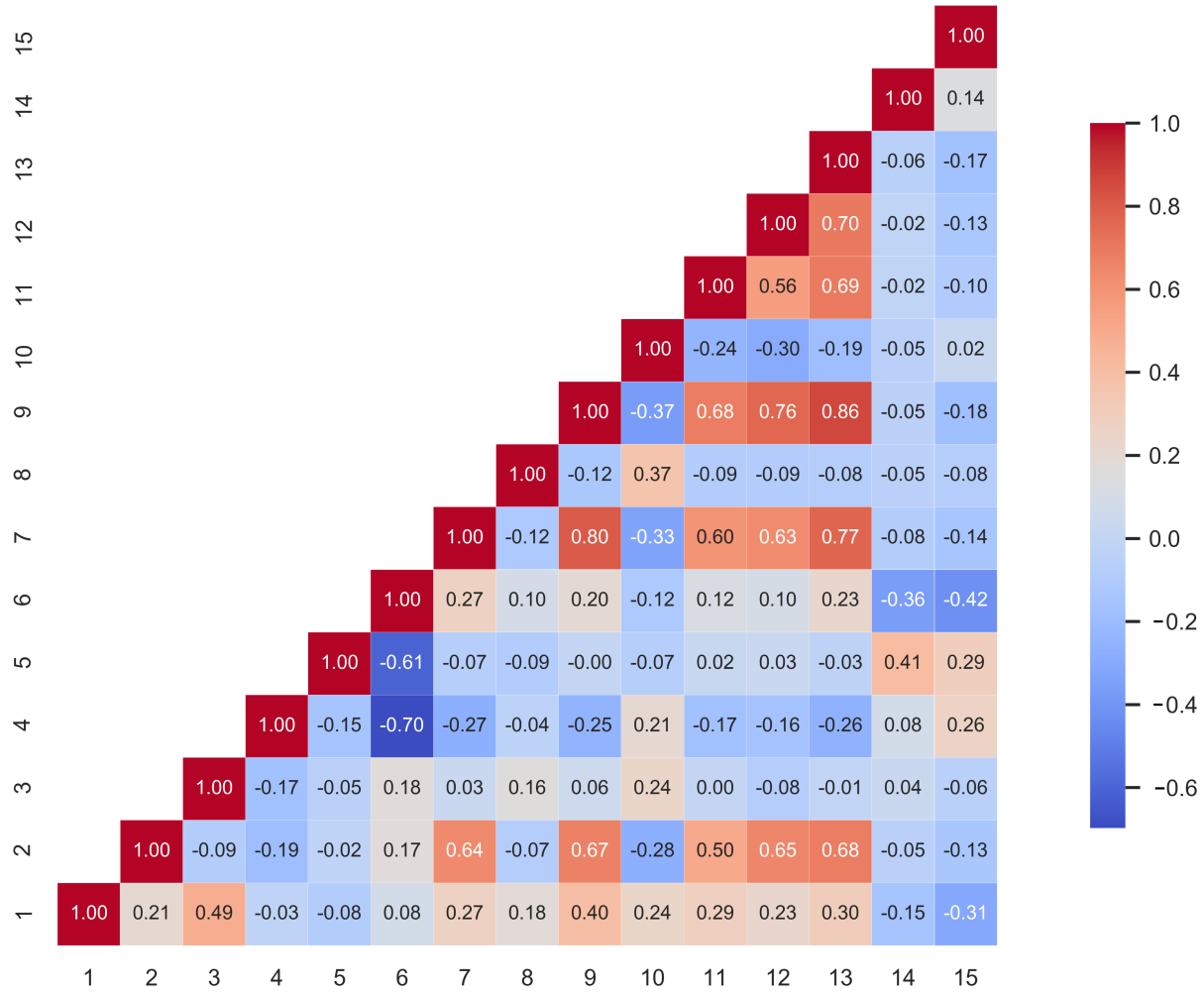


Fig. 5.2 Correlation matrix between the features.

On a side note, the threshold of ±0.77 set in chapter 2 could be verified by examining the pairplot of selected features in Fig. 5.3. In particular, the correlation value between features 9 and 12, has a value of +0.76, which is close to +0.77, the general threshold set.

Investigating the relationship between features 9 and 12 using pair plots in Fig. 5.3 shows that a correlation value of 0.76 implies a linear relationship between the two. This observation is not valid in the case of features 7 and 12, where the correlation is below the threshold chosen and is equal to 0.63. Therefore, the features 6, 9 and 13 were removed from the list of features as they relatively have a high correlation with other features.

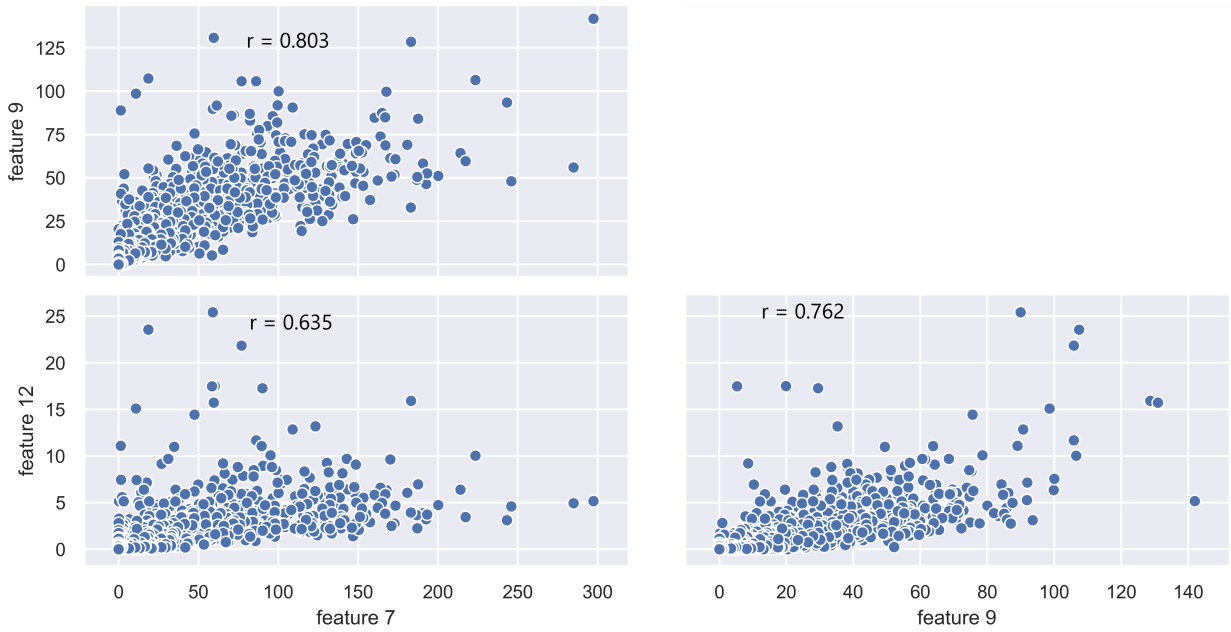


Fig. 5.3 Pairplots of features 7, 9 and 12 with the correlation coefficient, r.

5.2.2 Dimensionality Reduction

PCA

As discussed in Chapter 3, PCA performs a linear mapping of the data to a lower-dimensional space where the variance of the original data is maximized using the eigenvectors of the covariance matrix. The mapped data is presented in Fig. 5.4, where each point represents a feeder and the vectors represent the eigenvectors pointing in the direction of the maximum variance. As PCA can unveil partially the structure of the data, Fig. 5.4 shows the potential of using clustering methods on the mapped data where feeders are distributed in specific regions. This is a preliminary indicator that there is no overly significant variation in the data and that a coherent clustering can be achieved. Indeed, more components could be added if the total ratio of the explained variance of the selected components is less than 85% as explained in Chapter 3.

Moreover, the top five features that contributed to the PCA components or principal axis are features 2, 7, 11, 12 and 13 for component 1. For component 2, the major contributing features are 1, 3, 8, 10 and 15, and for component 3, these features are 1, 3, 5, 14 and 15. With these numbers, it could be confirmed that almost all the features presented in Table 5.1 contributed to the principal axes.

The dimensionality of the data is reduced by taking the first three components or eigenvectors. Table 5.1 presents the ratio of the explained variance of each of the three first

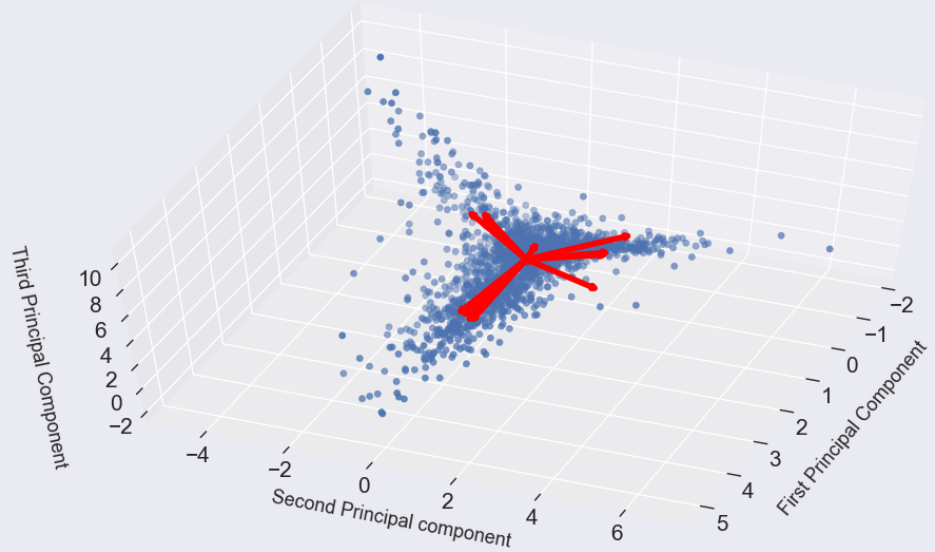


Fig. 5.4 Data in reduced dimensionality with the eigenvectors pointing in the direction of maximum variance.

components. The numbers in the table indicate that 99% of the total variance of the data is preserved.

Table 5.1 Eigenvalues ratio of the principal components

	Eigenvalues' ratio
Component 1	0.48
Component 2	0.30
Component 3	0.21
Total	0.99

t-SNE

In the previous section, PCA was used to reduce the dimensionality of the data. A specific global structure was observed; however, the data points or feeders were condensed, meaning that the local structure was not well represented.

In the case of t-SNE, the global and the local structures are represented in Figs. 5.5 and 5.6. By performing a visual comparison between the results of PCA and t-SNE, it can be seen that t-SNE stretched (i.e. increased the spread) the distance between the local points. This means that better-defined borders or more coherent clusters can be achieved with t-SNE than PCA. The group of points to the left of Fig. 5.5 cover the feeders that have regulators. Plainly, this group is easily identified in Fig. 5.5, which is not the case in Fig. 5.4. In the next section, a comparison between the clustering results of PCA and t-SNE is performed to validate the aforementioned statement.

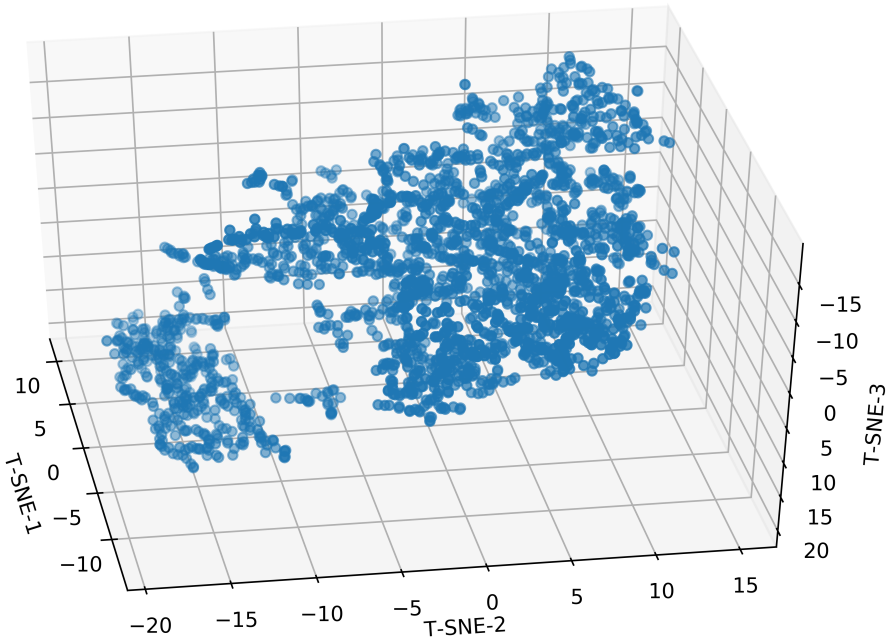


Fig. 5.5 Data in reduced dimensionality 1.

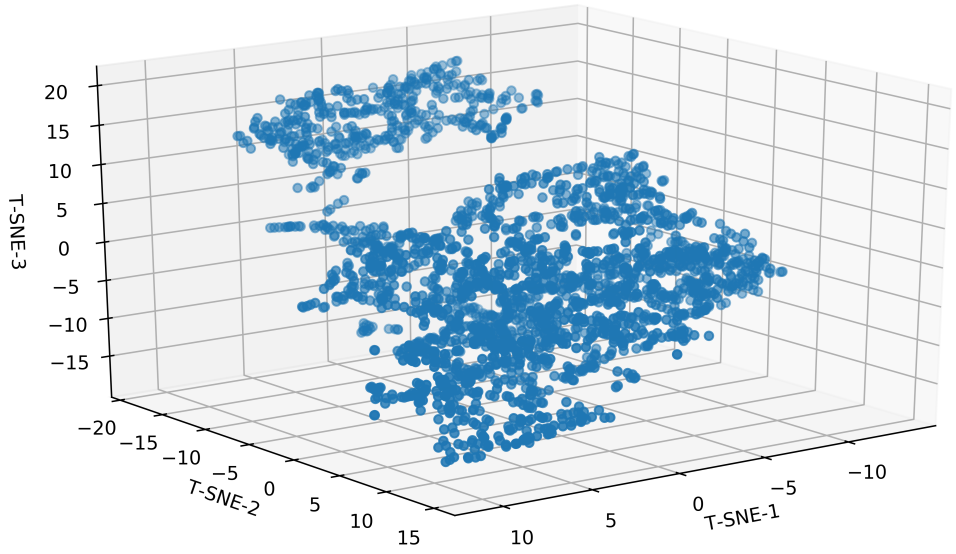


Fig. 5.6 Data in reduced dimensionality 2.

5.3 Cluster Analysis Results

After preparing the data, the cluster analysis is carried forward. As explained in Chapter 4, the steps are to find the ONC using different indicators and then analyze the results of the k-means++, hierarchical and GMM algorithms to choose optimal clusters. In this section, the results associated with the data transformation of PCA and t-SNE are displayed in two separate sections.

5.3.1 Cluster Analysis Results with PCA

To find the ONC, three indicators are used:

- Silhouette score in Fig. 5.7.
- VRC score in Fig. 5.8.
- Davies-Bouldin score in Fig. 5.9.

Evidently, the ONC cannot be 2, 3, 4 or 5 even if the results of the scores indicate otherwise. The cluster analysis goal is to find representative feeders that show a reasonable level of distinction in terms of characteristics. Moreover, the distribution utility categorizes their

feeders in five different databases based on the geographical area covered. It is a high-level categorization and it infers that an adequate ONC should be above five.

Comparing the results of the three indicators, two pieces of information can be extracted. The first one is that the ONC is 10. This statement is best expressed by the silhouette score, as seen in Fig. 5.7. The other two indicators were not as clear in identifying the ONC.

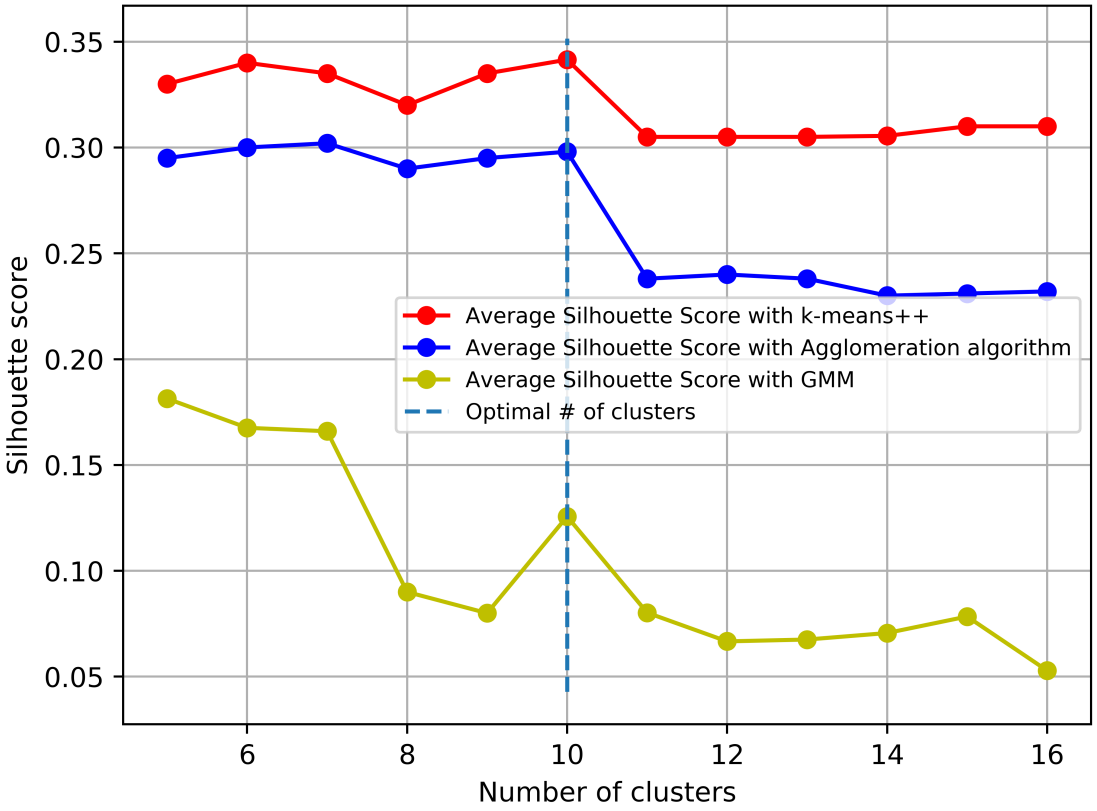


Fig. 5.7 The silhouette score with different number of clusters.

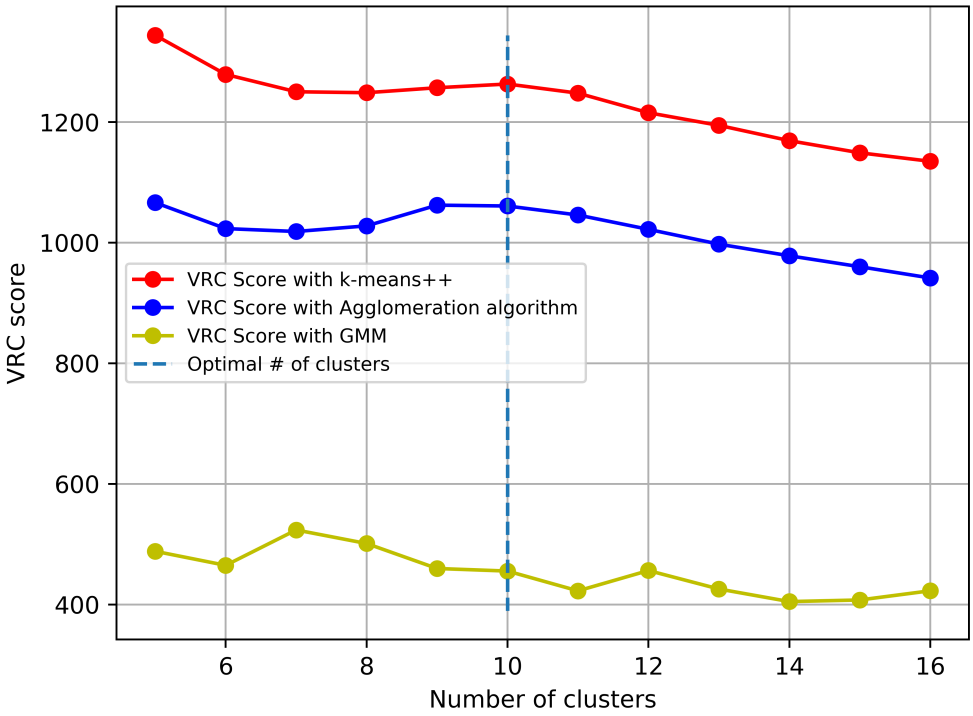


Fig. 5.8 The VRC score with different number of clusters.

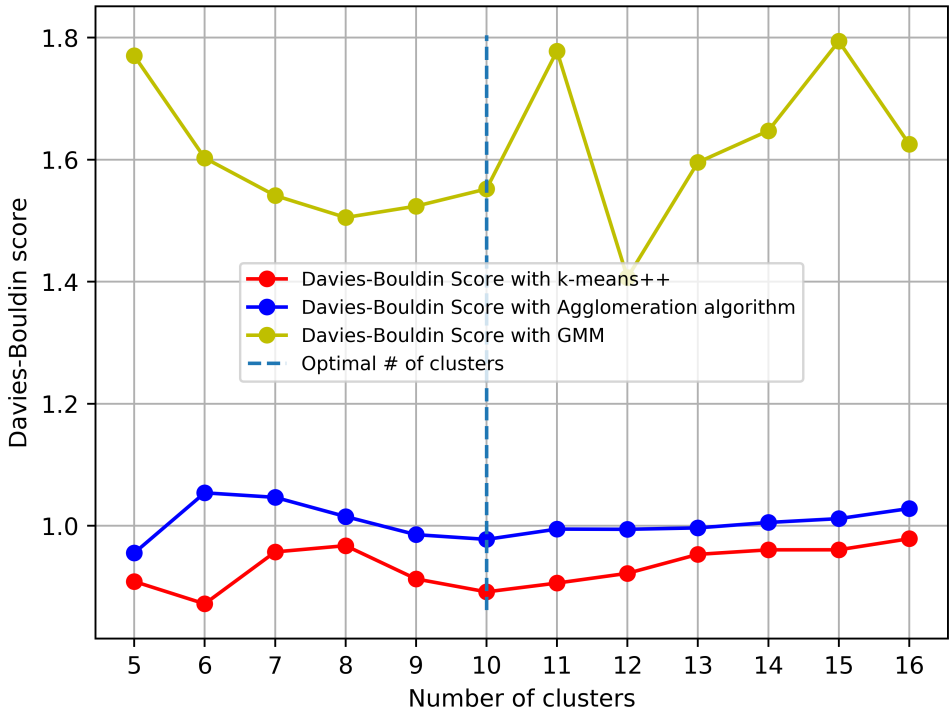


Fig. 5.9 The Davies-Bouldin score with different number of clusters.

The second information is that the best performing algorithm is k-means++, followed by the agglomeration algorithm, and finally the GMM algorithm. This statement will be further validated in the next phase of the analysis.

After finding the optimal number of clusters, the clustering results for the three algorithms are displayed in Figs. 5.10, 5.11 and 5.12. A visual comparison between figures 5.10 and 5.11 indicates that k-means++ and the hierarchical algorithm reveal similar structures or results. This observation is expected since the k-means++ and the ward linkage type of the hierarchical algorithm minimize similar objective functions.

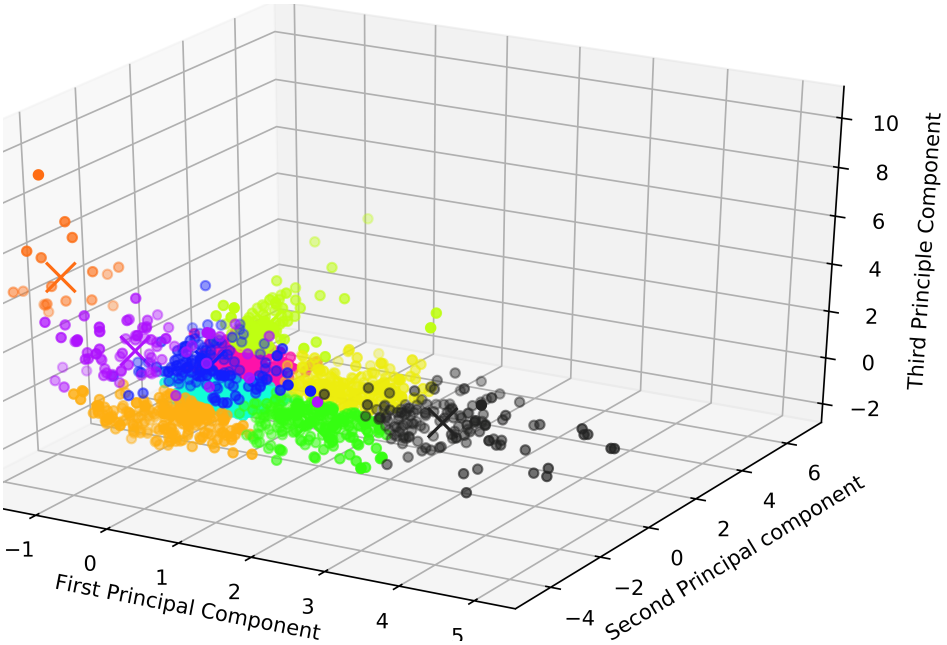


Fig. 5.10 Results of the k-means++ clustering algorithm with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids.

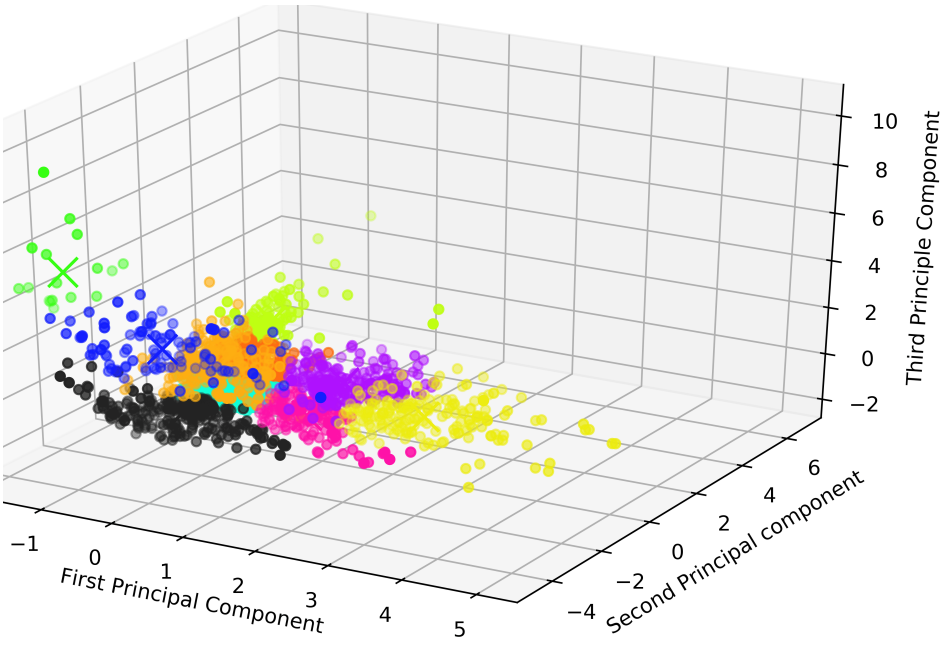


Fig. 5.11 Results of the hierarchical clustering algorithm (agglomeration technique) with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids.

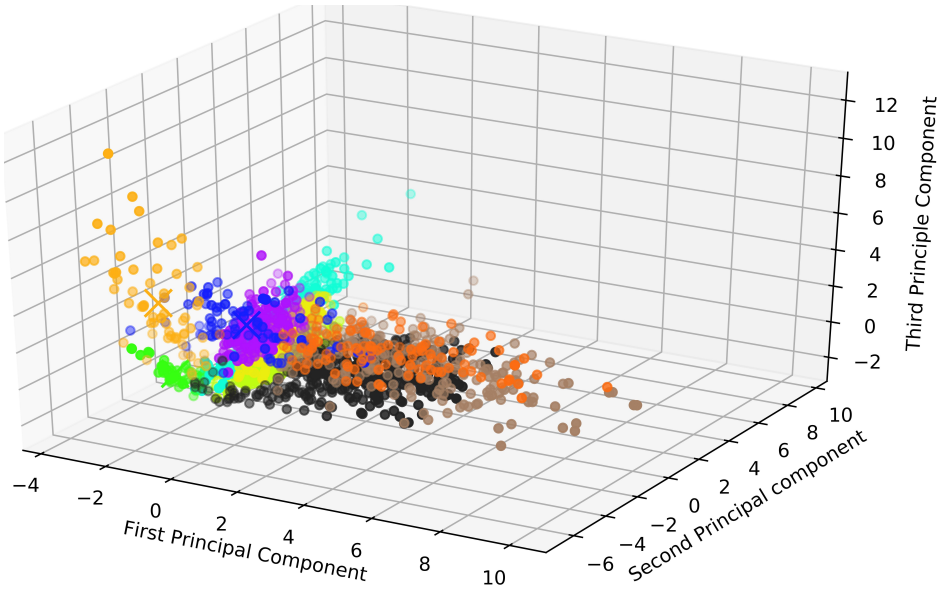


Fig. 5.12 Results of the GMM clustering algorithm with 10 clusters and with PCA dimensional reduction. Crosses indicate cluster centroids.

Moreover, Fig. 5.12 shows how the GMM algorithm is not revealing coherent clusters. The fact that the k-means++ algorithm is the best performer is further explained by Table 5.2, where the number of feeders having a negative silhouette score is minimal in the case of k-means++.

Table 5.2 Number of feeders with negative silhouette scores

	# of feeders forming the borders
GMM	784
Hierarchical algorithm	362
k-means++	100

Table 5.3 displays the distribution of feeders by the k-means++ and hierarchical algorithms. The numbers in Table 5.3 show that the distribution of the two algorithms is comparable.

Based on Tables 5.3 and 5.4, feeders of cluster 6, that are characterized with about 80% residential kVA, medium length urban topology and fairly distributed load, represent the majority of the feeders used for this case study.

The cluster descriptions in Table 5.3 relate to general key characteristics and by observing the representative feeder of each cluster in Table 5.4, a full description can be drafted.

Table 5.3 Number of feeders per cluster

Cluster description	**k-means++**	**Hierarchical algorithm**
Extreme rural	137	186
Rural with high kVA	212	220
Industrial	19	18
Short urban with low kVA	308	337
Urban/Underground	115	111
Rural	236	173
Urban/Residential	775	592
Urban/Industrial	143	99
Very short Industrial/Urban /Underground	75	212
Urban/Underground with concentrated load	492	564

Table 5.4 Representative feeders

feature / feeder	1	2	3	4	5	7	Cluster
feeder_0	37	6.73	14703.71	0.05	0	132.9	0
feeder_1	57	3.66	17053.93	0.24	0.1	62.6	1
feeder_2	3	0.3	10516.38	0	1	0	2
feeder_3	18	1.04	4895.99	0.23	0.11	2.3	3
feeder_4	61	0.79	20550.38	0.14	0.01	5	4
feeder_5	27	5.07	10190.67	0.13	0.08	89	5
feeder_6	28	1.47	13378.35	0.18	0.01	15.3	6
feeder_7	35	1.53	11947.61	0.13	0.43	11.3	7
feeder_8	14	0.31	9427.65	0	1	0	8
feeder_9	35	0.94	15778.93	0.15	0.01	13.5	9

Table 5.4 Representative feeders (Continued)

feature / feeder	8	10	11	12	14	15	Cluster
feeder_0	0.2	0.3	1	5.45	16.52	19.02	0
feeder_1	0.5	0.2	1	1.24	76.01	27.37	1
feeder_2	0	0.6	0	0.03	10508.6	5258.19	2
feeder_3	0.1	0.68	0	0.1	45.43	68	3
feeder_4	5.9	9.3	0	0.23	0.23	143.71	4
feeder_5	0	0.3	0	4.03	49.33	19.23	5
feeder_6	0.8	0.8	0	0.23	31.25	55.28	6
feeder_7	1.3	0.6	0	0.35	960.2	70.28	7
feeder_8	0.1	2.2	0	0.05	2920.44	1178.46	8
feeder_9	0.7	6	0	0.03	50.2	69.21	9

It is important to mention that, for the hierarchical algorithm, the number of neighbours for the KNG is set to an arbitrary value of 200. Therefore, the agglomerative algorithm will allow merging a maximum of 200 neighbours. However, to respect the ONC, the clusters formed contain more than 200 neighbours in Table 5.3. We conclude from this that the connectivity constraints added through the KNG can help reduce the computational complexity if the number of neighbours is set to 592, which is the maximum number of feeders grouped in one cluster (refer to Table 5.3).

On a seperate note, given the dense structure of the data in Fig. 5.4, density-based algorithm DBSCAN would focus on isolating the dispersed feeders positioned to the left of figures 5.10, 5.11 and 5.12.and would group all the dense feeders together.

Remark

Incrementing the number of clusters does not provide relevant information as the extreme cases are separated into unique clusters like in clusters 0 (Extreme rural) and 2 (Industrial) of the k-means++, which correspond to the points floating at the right and upper left in Fig. 5.10.

5.3.2 Cluster Analysis Results with t-SNE

In this section, the same steps performed in Section 5.3.1 will be followed using the data transformation of t-SNE. As k-means++ was the best performer in the previous section, it will be used in this section, and its results will be compared with those of Section 5.3.1.

Regarding the ONC, it was shown in Chapter 4 that the indicators used in this thesis will not be effective in determining the optimal number of clusters when using t-SNE because the structure of the data is complex.

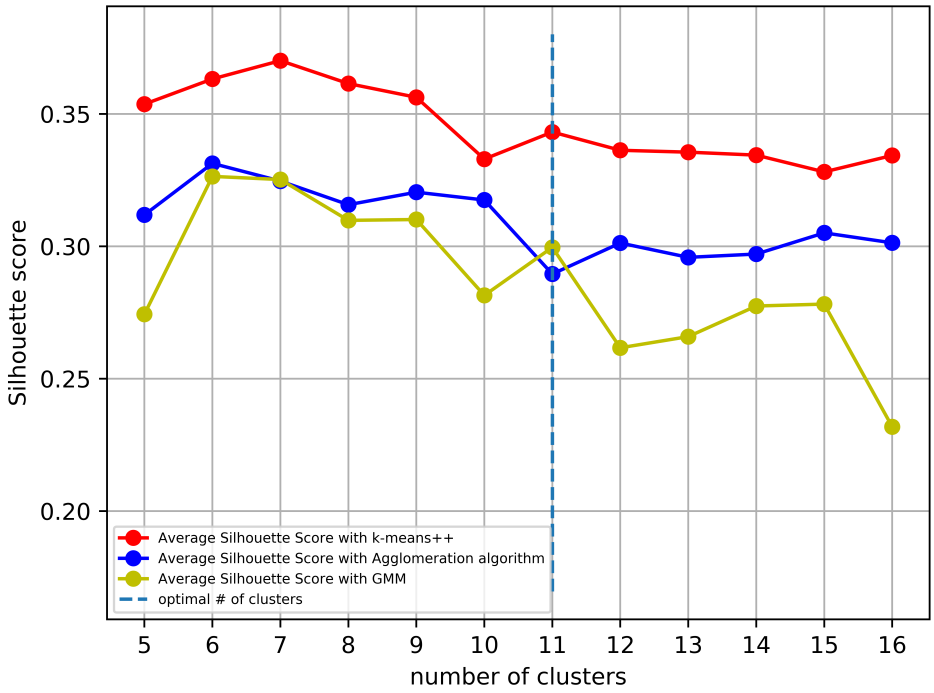


Fig. 5.13 The silhouette score with different number of clusters.

By inspection, Fig. 5.13 is not very informative about the ONC. As it can be seen, the ONC indicated in Fig. 5.13 is 7, which corresponds to grouping all the feeders with

regulators. The results of performing the cluster analysis with 7 clusters are presented in Fig. 5.14 where all the feeders with regulators are grouped in one cluster as indicated by the red pointer. Even with 10 clusters, all the feeders with regulators are still grouped in one cluster as shown in Fig. 5.15.

Therefore, for the sake of having detailed clusters in terms of presence of regulators, the ONC is chosen to be 11 and the results are displayed in Fig. 5.16.

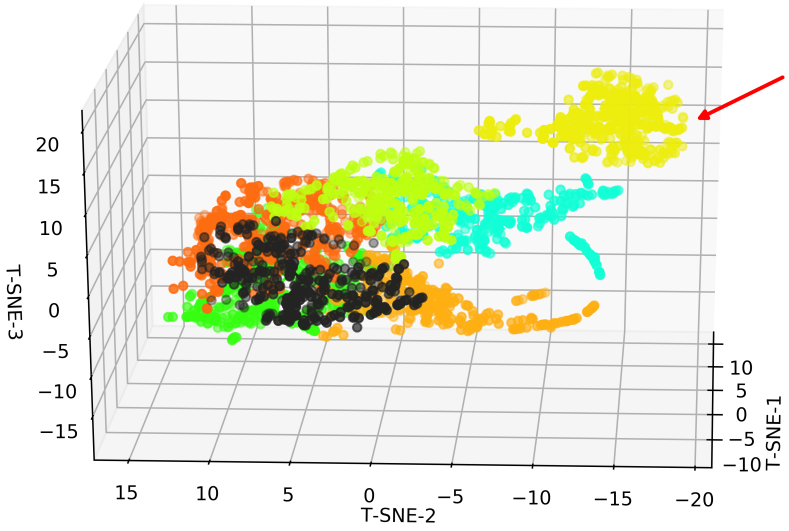


Fig. 5.14 Results of k-means++ clustering algorithm with 7 clusters and with t-SNE dimensional reduction. The feeders with regulators are represented in one cluster.

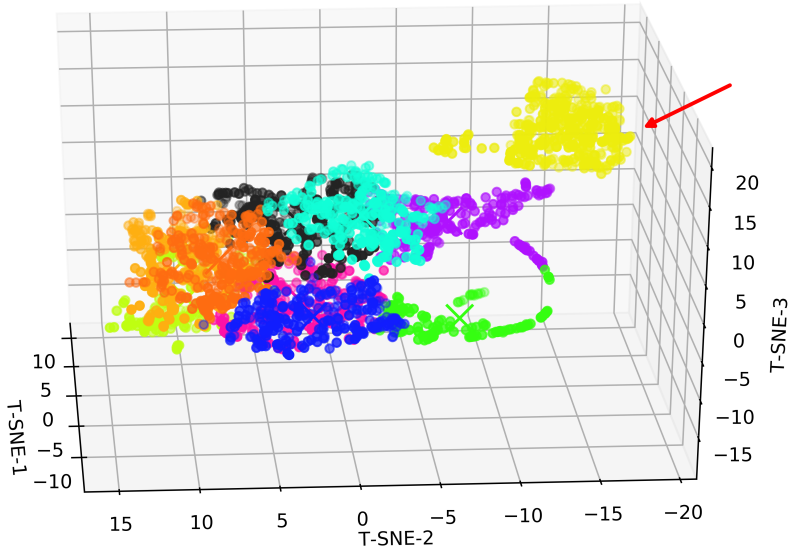


Fig. 5.15 Results of k-means++ clustering algorithm with 10 clusters and with t-SNE dimensional reduction. The feeders with regulators are still represented in one cluster.

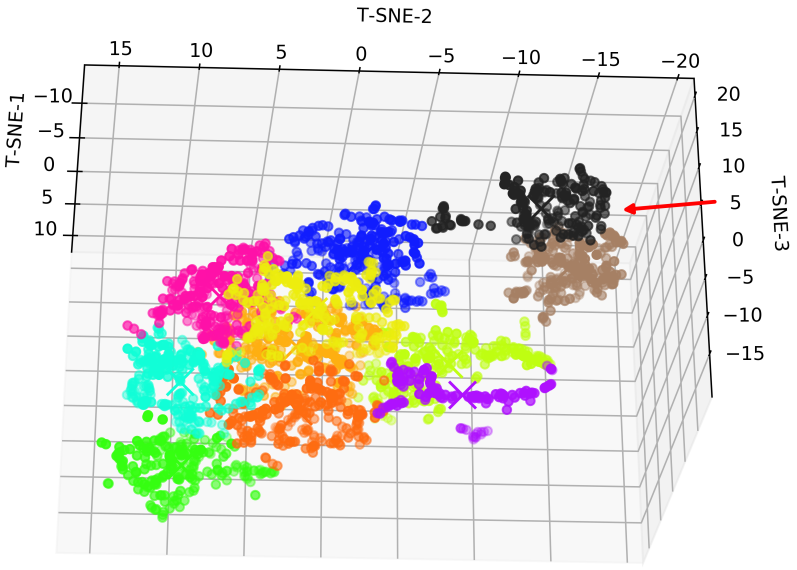


Fig. 5.16 Results of k-means++ clustering algorithm with 11 clusters and with t-SNE dimensional reduction. The feeders with regulators are represented in two different clusters.

Visibly, it can be noticed from figures 5.14, 5.15 and 5.16 above how the clusters are well defined in the case of t-SNE. Moreover, Table 5.5 shows the number of feeders that define the borders of the clusters. Comparing this number with Table 5.2, it is affirmative that the clusters achieved with t-SNE are more coherent than with PCA.

In addition to that, it can be observed in Table 5.7 that the number of feeders in each cluster is not extreme. Using PCA, over 700 feeders were grouped in one cluster (cluster 6) in Table 5.3.

On the other hand, with t-SNE, the number of feeders grouped in one clusters did not exceed 304. The distribution between the clusters is not as extreme as in the case of PCA.

Table 5.5 Number of feeders having a negative silhouette score

	# of feeders forming the borders
k-means++ with t-SNE	29

Table 5.6 Representative feeders with t-SNE

feature / **feeder**	**1**	**2**	**3**	**4**	**5**	**7**	**Cluster**
feeder_x	37	0.72	13615.02	0.11	0.00	4	0
feeder_y	29	0.94	11005.33	0.18	0.52	7.1	1
feeder_z	43	3.80	9685.88	0.15	0.01	63.8	2
feeder_m	22	0.70	7801.75	0.14	0.00	3.4	3
feeder_n	35	6.10	13462.35	0.16	0.17	121.8	4
feeder_o	34	1.88	12960.38	0.10	0.00	19.5	5
feeder_p	46	1.03	13536.33	0.36	0.13	9	6
feeder_q	10	0.38	10987.34	1.00	0.00	0	7
feeder_r	39	1.13	13550.02	0.42	0.05	5.5	8
feeder_s	55	4.08	10974.73	0.14	0.13	44.9	9
feeder_t	23	0.58	11313.09	0.20	0.02	1	10

Table 5.6 Representative feeders with t-SNE (Continued)

feature / feeder	8	10	11	12	14	15	Cluster
feeder_x	10.7	9.5	0	0.12	0.12	95.88	0
feeder_y	1.5	3.6	0	0.06	452.21	78.61	1
feeder_z	0	0	0	1.572	47.79	20.09	2
feeder_m	0	0.9	0	0.24	43.105	62.91	3
feeder_n	0.1	2.5	1	5.33	352.18	18.80	4
feeder_o	0.3	0.3	0	0.81	28.23	53.78	5
feeder_p	0.3	5.8	0	0.18	162.36	83.56	6
feeder_q	0	3.2	0	0.18	0.18	2746.83	7
feeder_r	0.6	5.5	0	0.17	448.37	79.71	8
feeder_s	0.2	0.4	1	1.65	116.35	22.08	9
feeder_t	0.7	3.3	0	0.11	0.11	168.85	10

The description of clusters in Table 5.7 can be derived from Table 5.6.

Table 5.7 Number of feeders per cluster with t-SNE

Cluster number	Cluster description	Number of feeders
0	Urban/Underground with high kVA	190
1	Industrial/Commercial with concentrated charge	292
2	Rural	241
3	Short Urban/Residential with fairly-distributed charge	270
4	Extreme rural	177
5	Medium length Urban/Residential with distributed charge	304
6	Urban/Residential/Commercial	194
7	Short Commercial Underground	130
8	Urban/Residential/Commercial with high variation in kVA values	245
9	Rural with regulators	241
10	Very short Urban/Residential /Commercial	228

Remark

As discussed in Chapter 4, a ratio of variation, which is the ratio of the silhouette scores of feeders in each cluster and the corresponding representative feeder, is assigned to each feeder in order to provide information about the variation within each cluster. It is also worth mentioning that information about the size and the silhouette scores of the feeders in each cluster as well as the average silhouette score of all the clusters can be extracted from the silhouette profile in figures 5.17 and 5.18. The left-hand side of figures 5.17 and 5.18 displays the average silhouette score, the silhouette scores in an ascending order covered in each cluster as well as the size of each cluster (this is reflected by the area each cluster color occupies). The right-hand side of figures 5.17 and 5.18 represents a two-dimensional projection of the three dimensional data of PCA where each center is identified by a white circle.

Comparing the left-hand side of both figures, it can be seen how the sizes of the clusters vary, in the case of PCA, some clusters have relatively a high number of feeders, however, for the case of t-SNE, the size of the clusters does not vary substantially. This information is validated by Tables 5.3 and 5.7 which describe the distribution of feeders among the clusters. Moreover, there exist more feeders with negative silhouette score in Fig. 5.17 (left side) than in Fig. 5.18 (left side), which matches the numbers in Tables 5.2 and 5.5.

The right side of 5.17 and 5.18 displays the relative position of the clusters in a two-dimensional projection. It can be seen that, in general, the distance between any two feeders (represented as a point) is higher in the case of t-SNE. This observation is expected since t-SNE accounts for local dissimilarity as well as for the global one. For the case of PCA, the majority of points are concentrated like in the case of clusters 3 and 6 to the right of Fig. 5.17, the rest are dispersed like in the case of clusters 4 and 8. This analysis confirms the comparison in performance between t-SNE and PCA in Section 3.3.2.

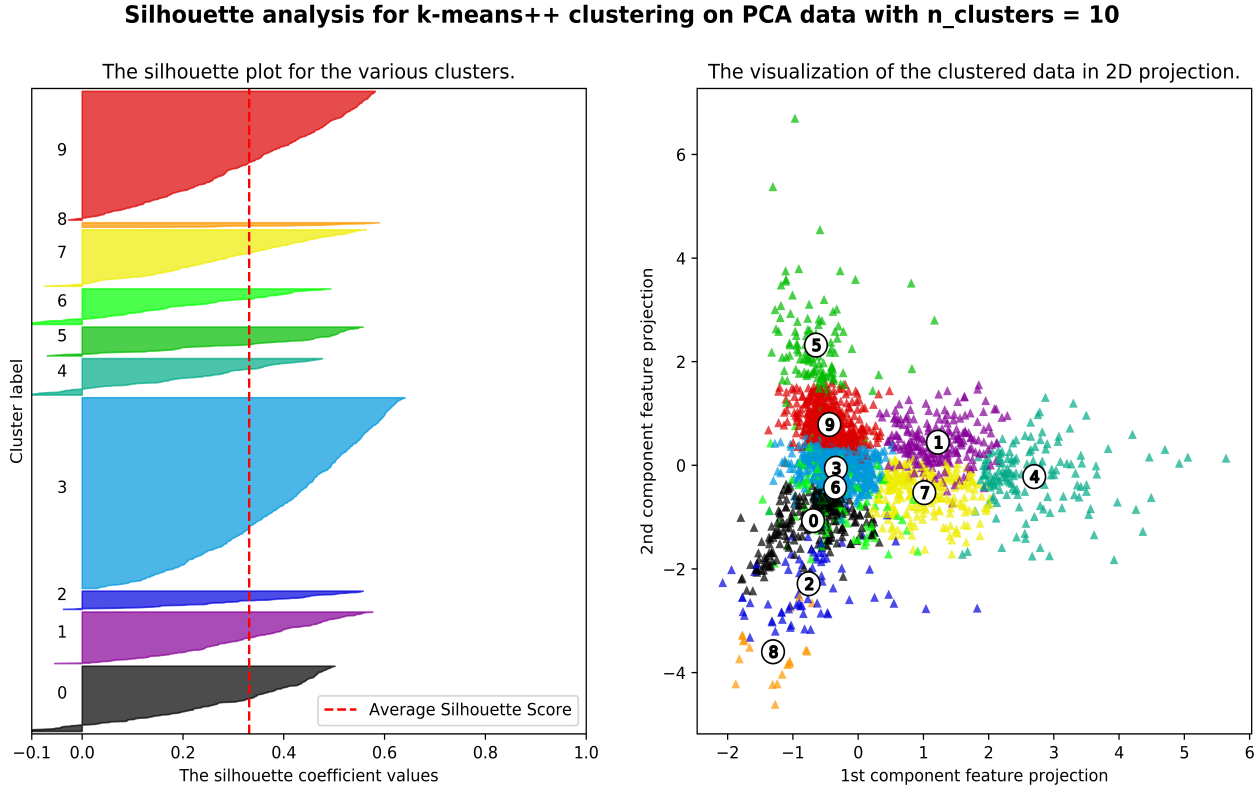


Fig. 5.17 Silhouette profile with PCA dimensional reduction.

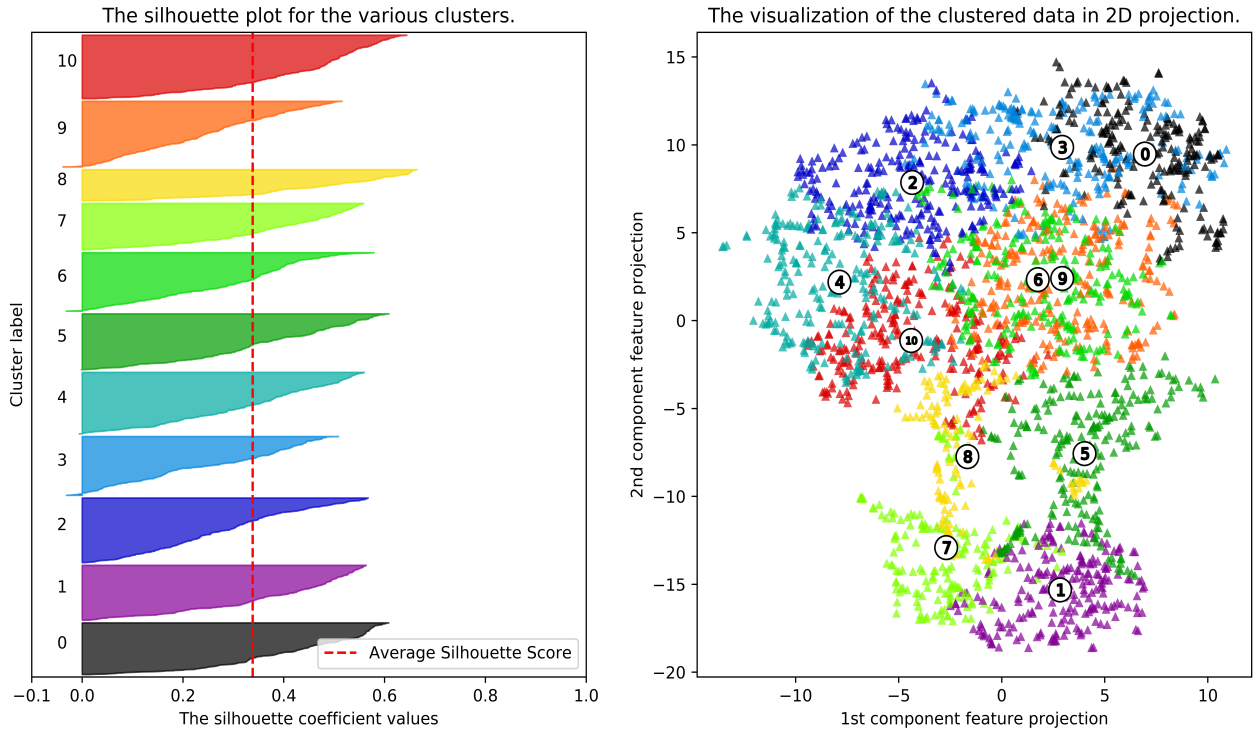


Fig. 5.18 Silhouette profile with t-SNE dimensional reduction.

5.4 Validation of Results

In this section, the results of the k-means++ algorithm with PCA and t-SNE are validated following the steps described in Chapter 4.

5.4.1 PCA

Validation of Topology

The validation of the topology can be simply performed visually or by comparing the number of three-phase branches between the three selected feeders. An example of this validation is presented in Figs. 5.19, 5.20 and 5.21 for cluster 9. These figures reveal the similarity in the topology between the three feeders. They are characterized by a concentrated load and an important underground cabling distance as indicated in Table 5.4. The same process is performed visually for the 10 clusters.

In Fig. 5.22, a boxplot of feature 1, which represents the number of three-phase branches, is displayed for each cluster. Based on Table 5.4, cluster 2 is characterized by a low number of three-phase branches and an industrial and concentrated load, for this reason, the IQR range for cluster 2 in Fig. 5.22 is short. However, for cluster 0 (extreme rural) and cluster 1 (rural) that represent feeders with long cable distances, the IQR is long. Similar observations can be drawn for the rest of the clusters.

Fig. 5.19 Topology of the center

Fig. 5.20 Topology of the nearest neighbour

Fig. 5.21 Topology of the farthest neighbour

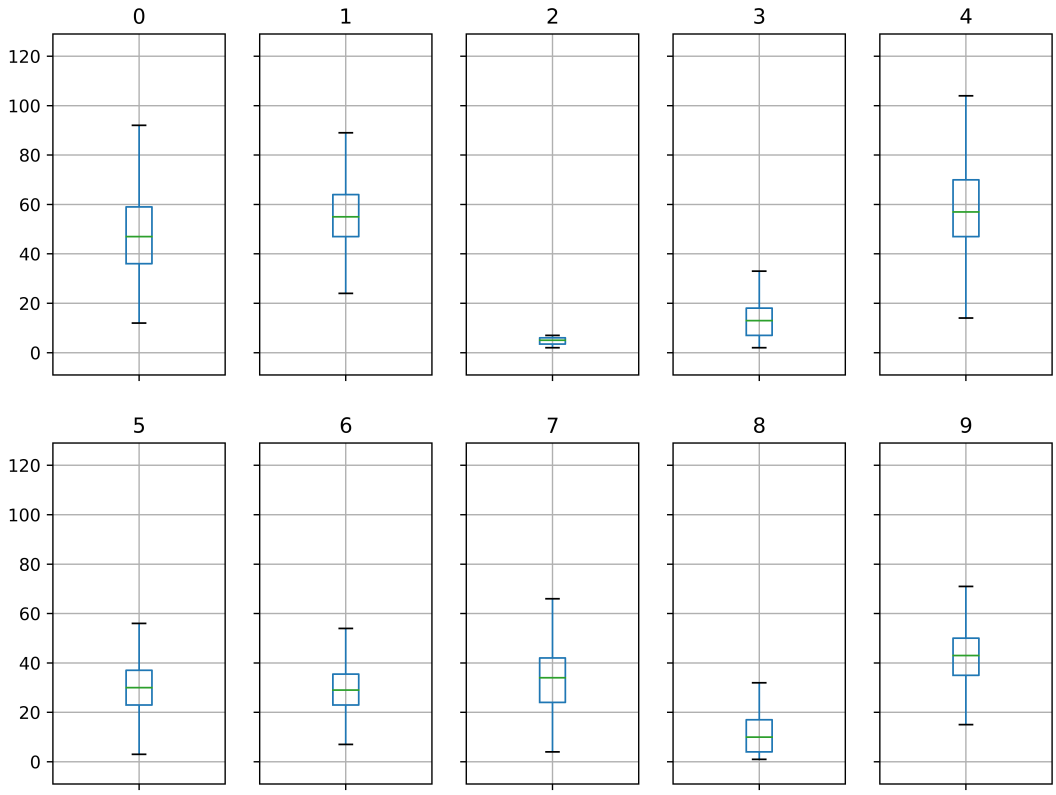


Fig. 5.22 Boxplots of the feature 1 (number of branches) outlining the range of values, namely, the median and IQR, in each cluster.

It is important to mention that in the first iteration of the cluster analysis, features 8 and 10 that represent the underground cabling distance were missing. This issue allowed the grouping of feeders with similar concentration of load as well as a slightly different topology in terms of overhead and underground distance. For this reason, the features related to the underground and overhead cabling distance were added.

Validation of Demand Profile

As discussed in Chapter 4, the metric used for the normalized demand profile is the MSE. The demand profile is chosen to be the average of the weekdays of the month of January 2018 and it is normalized based on the absolute maximum value. As example of the normalized demand profile of a cluster center, the nearest and the farthest feeders of cluster 7 which is characterized by about 45% residential load is presented in Fig. 5.23.

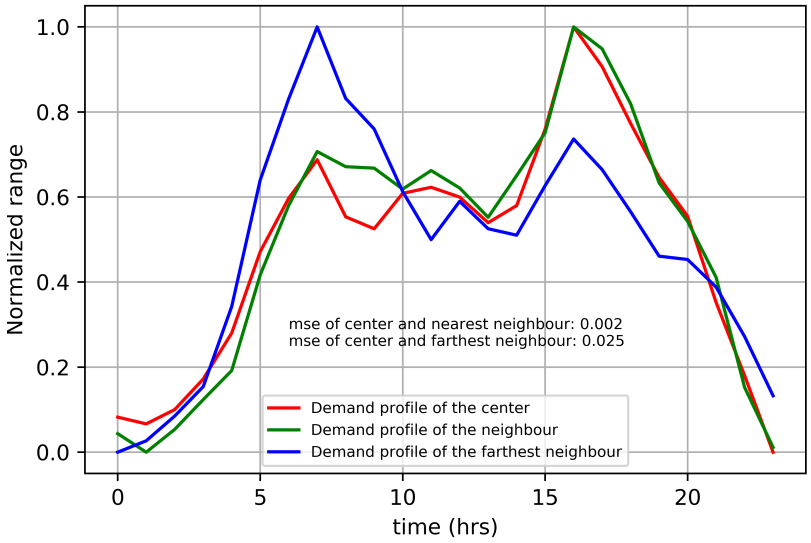


Fig. 5.23 The normalized demand profile of an urban residential/industrial cluster.

In Table 5.9, the MSE between the center, the nearest and the farthest feeder of each cluster is displayed. In the case of cluster 2, an elevated value of MSE is observed with around 40% difference in the demand profile between the center and farthest feeders. This is because cluster 2 is is characterized by industrial and commercial customers and the demand profile for the feeders in this cluster varies based on the type of the industry.

Table 5.8 MSE between the demand profile of the center, the nearest and farthest feeders

Cluster	MSE of the nearest feeder	MSE of the farthest feeder
0	0.071	0.126
1	0.033	0.220
2	0.119	0.402
3	0.010	0.077
4	0.010	0.044
5	0.045	0.026
6	0.184	0.103
7	0.002	0.025
8	0.102	0.184
9	0.002	0.025

It is also worth mentioning that there is a difference between the demand profile of commercial and industrial customers even though they share similar topologies (short distance, concentrated load). For industrial customers, their demand profile is almost flat around the clock. However, for commercial customers, their demand profile shows some periodicity.

Validation of Voltage Violations and Overload

The final step is assessing with different percentages of random clients installing PVs, whether VVOs are caused. The simulations are performed using the commercial software package for distribution system analysis, CymDist version 8.0 developed by CYME International [24]. For each candidate, a simulation with 50%, 75% and 100% of clients installing: PVs with capacities of 4kW for residential clients. For commercial clients, the PV capacity was defined as being 75% of the peak consumption recorded in kW. For these simulations, the load demand is maintained at the average consumption during summer months. When VVOs are detected during the simulations, a flag is raised by CymDist. The results of the simulations are presented in Table 5.9 where a value of 1 indicates that the feeder had VVO problems with PV integration. The feeders of cluster 9 that are presented in Figs. 5.19,5.20 and 5.21 are characterized by a very concentrated load (feature 12 in Table 5.4 indicates a 3% variation around the mean distance between the loads and the source of the feeder). This causes issues if the majority of the clients install PVs. Hence, with 50%, 75% and evidently, 100% of clients installing PVs, the feeders in this cluster are highly susceptible to have problems with PV integration as can be seen in Table 5.9.

Cluster 4 is an urban/underground cluster that shares similar characteristics with cluster 9. However, it is characterized by a higher coefficient of variation for the distance between the loads and the source, meaning that the load is not as concentrated as in the case of cluster 9. Hence, this cluster is less likely to have problems with PVs as indicated in Table 5.9, where the three candidates of the cluster did not have any VVOs. Based on the earlier results concerning topology and demand profile, it is more likely that problematic feeders in regards to PV integration can be found in rural feeders which are characterized by long distances and high impedance and presence of voltage regulators.

Moving away from the source, the voltage drops, therefore, rural feeders usually have voltage regulators to support the voltage drop. However, with increased PV generation on the feeder, which results in an increase in the voltage as well as a reverse power flow in certain cases, voltage regulation gets more and more difficult to achieve. It is important to note that voltage regulators operate under different modes and the bi-directional mode relates to supporting reverse power flow. The methods adopted to control the voltage under this mode may be optimized as the existing power flow algorithms for these methods may diverge [60]. Moreover, physically operating the voltage regulators in bi-directional mode requires utilities to consider setting up additional equipment like protection relays and sensors.

Clusters 0 and 1 are rural with regulators and based on Table 5.9, these clusters have VVO problems with PVs. Only the representative feeder of cluster 0 showed problems with 100% PV penetration, whereas in cluster 1, VVOs with PVs started with 50% of PV

penetration. This is expected since cluster 1 is characterized by a higher kVA consumption and concentration of load.

As mentioned in Chapter 4, VVOs are defined by the utility's operators, based on voltage and current limits they set on the transformer of each load point.

It is important to note that the majority of the feeders were capable of handling 100% of PV penetration as seen in Table 5.9. This is because the feeders of the utility are designed and planned to be able to handle a high volume of load under abnormal conditions, when there is a fault or a failed equipment on the feeder that might cause extended outages. Therefore, even when there is a high level of PV generation on the feeder where the excess power is injected back, it is less likely to have technical problems.

In particular, the feeders are planned for an N-1 contingency and designed to feed their loads under cold load pick-up. Planning the feeder for an N-1 contingency means that the feeder is able to feed or power its loads even when a transformer failure or a circuit reconfiguration due to a fault or work activities take place [61]. And cold load pick-up is the situation when the feeder is re-energized after an extended outage period which results in significantly higher than normal load levels [62].

Table 5.9 PV simulation results

Cluster	Feeder	50% PV	75% PV	100% PV
	feeder_0	0	0	1
0	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_1	0	1	1
1	nearest feeder	1	1	1
	furthest feeder	0	1	0
	feeder_2	0	0	0
2	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_3	0	0	0
3	furthest feeder	0	0	0
	feeder_4	0	0	0
4	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_5	0	0	0
5	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_6	0	0	0
6	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_7	0	0	0
7	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_8	0	0	0
8	nearest feeder	0	0	0
	furthest feeder	0	0	0
	feeder_9	1	1	1
9	nearest feeder	0	1	1
	furthest feeder	0	0	1

5.4.2 t-SNE

Similarly for the case of t-SNE, the topology of the clusters was consistent. Cluster 7 is characterized by short and underground cable distances and a prevalence of commercial load, hence it is distinguished by a low number of three-phase branches as indicated in Fig. 5.24. Clusters 4 and 9 are rural and are characterized by a relatively high number of three-phase branches.

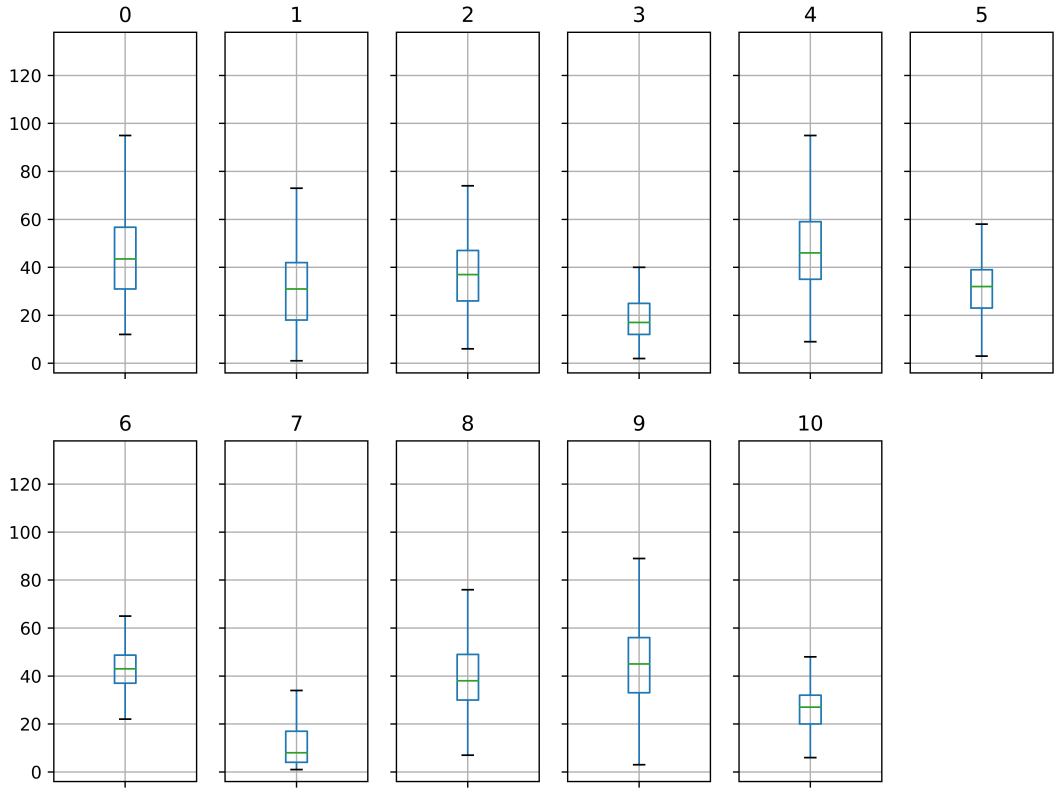


Fig. 5.24 Boxplots of the feature 1 (number of branches) outlining the range of values, namely, the median and IQR, in each cluster.

As for the demand profile, some elevated MSE values in Table 5.10 can be detected. In particular, the farthest feeder deviates by more than 30% from the center of clusters 7 and 8. This is because these clusters are characterized by commercial and industrial load.

Table 5.10 MSE between the demand profile of the center, the nearest and farthest neighbor

Cluster	MSE of the nearest feeder	MSE of the farthest feeder
0	0.050	0.019
1	0.053	0.047
2	0.021	0.030
3	0.034	0.068
4	0.040	0.263
5	0.012	0.014
6	0.009	0.016
7	0.054	0.318
8	0.058	0.417
9	0.335	0.049
10	0.014	0.029

As for the PV validation, we observe a similarity in the type of cluster that are problematic in regards to PV penetration between PCA and t-SNE. Based on Table 5.11, clusters 2,4 and 9 which represent rural feeders, indicated VVOs with 50%.

We can see from Table 5.6 that the urban feeders that are grouped in cluster 6, are characterized with a residential load, underground cable distances and relatively a concentrated load. The feeders of cluster 6 are among the problematic feeders in regards to PV integration, as seen in Table 5.11. This is equivalent to the observations drawn in Section 4.4.1 for feeders in cluster 9.

Table 5.11 PV simulation results with t-SNE

Cluster	Feeder	50% PV	75% PV	100% PV
0	feeder_x	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
1	feeder_y	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
2	feeder_z	0	0	1
	nearest feeder	0	1	1
	furthest feeder	1	1	1
3	feeder_m	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
4	feeder_n	0	1	1
	nearest feeder	1	1	1
	furthest feeder	1	1	1
5	feeder_o	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
6	feeder_p	0	1	1
	nearest feeder	0	1	1
	furthest feeder	1	1	1
7	feeder_q	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
8	feeder_r	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0
9	feeder_s	0	1	1
	nearest feeder	0	0	1
	furthest feeder	0	1	1
10	feeder_t	0	0	0
	nearest feeder	0	0	0
	furthest feeder	0	0	0

5.5 Conclusion

In this chapter, all the steps described in the procedure proposed in this thesis were showcased and applied on a dataset of 2975 distribution feeders. It has been also shown how reducing dimensionality can be advantageous for cluster analysis. Moreover, the performances of PCA and t-SNE were compared.

After conducting the comparative analysis across the different clustering techniques, the k-means++ algorithm was identified as the best performer on the dataset. The ONC was 10 for the data transformed using PCA and 11 using t-SNE. Moreover, in this chapter, t-SNE was found to be more effective than PCA in reducing the dimensionality of the data for the following reasons:

- t-SNE can reveal global and local structures of the data (refer to figures 5.16 and 5.18)
- The number of mis-clustered feeders is lower in the case of t-SNE (refer to Table 5.5)
- t-SNE featured a better silhouette profile (refer to figures 5.17 and 5.18)
- With t-SNE, a richer variety of types of clusters was obtained (refer to Table 5.7 and Section 5.4.2)

As discussed in chapter 3, users can choose to use PCA or t-SNE based on their requirements.

It has been underlined in Section 5.4 that the topology in the clusters is consistent as well as the demand profile with an exception for commercial and industrial feeders. As for the validation of VVOs with different percentages of PV integration, rural clusters that are distinguished by high impedance and presence of regulators were flagged in addition to the urban feeders with concentrated load. DNOs need to pay a close attention to these groups of feeders when conducting studies that involve distributed generation.

More importantly, this chapter showed that the proposed procedure can result in an adequate and reasonable representation of all the feeders. The representative feeders could be used to avoid a full-scale simulation that requires an expensive investment from the distribution utilities.

Chapter 6

Conclusions and Future Work

6.1 Summary

The research presented in this thesis built on previous work and proposed an improved procedure to identify representative feeders of the ensemble of distribution feeders of an electricity distribution utility. As indicated by previous studies in the literature, severe technical problems could arise if suitable planning to accommodate new technologies is not conducted. In addition, carrying out a full-scale study of the impact of these technologies would require substantial efforts. By using representative feeders in simulations and test runs, distribution utilities could avoid having to perform a prohibitive number of simulations. The proposed procedure in this thesis was applied to a dataset of 2975 feeders. In the proposed method, the relevant features to be used were identified based on the work of IEEE TFWG and the recommendations of the distribution engineers at the utility. In particular, we listed 16 features that contain information about the topology, load and voltage.

In order to properly use these features, specific pre-processing steps should be followed. We described how outliers affect the cluster analysis and how they are eliminated. In addition, we highlighted the importance of studying the correlation between features because in certain cases, the same information can be present in two different features. This requires the elimination of one of the two highly-correlated features. Moreover, in order to put the features on the same order of magnitude and variance, standard scaling should be done. In the case study, the data pre-processing resulted in eliminating certain features that showed high correlation with others.

Afterwards, two dimensionality reduction techniques, PCA and t-SNE, were applied on the remaining features. One of the advantages of reducing the dimensionality is the reduction in the complexity of the data variation. PCA is a linear technique that aims at maximizing the variance while reducing the dimensionality and t-SNE is a non-linear stochastic approach

that aims at maximizing the likelihood of having two neighbouring points. While PCA highlights the separation between the distant points, t-SNE can account for the distant as well as the close points. All in all, PCA unveils the global structure of the data, while t-SNE focuses on the local and global structures simultaneously.

Cluster analysis was performed on two different three-dimensional representations of the data. A variety of clustering algorithms can serve our purpose. Hence, k-meanss++, hiearchical and GMM clustering algorithms were adopted. Moreover, three different scores to find the ONC were used in the analysis. These scores indicated that the ONC is 10 for PCA data and 11 for t-SNE data. In order to pick the best results between the three clustering algorithms, a comparative analysis was conducted. The comparative analysis indicated that the k-means++ algorithm was the best performing algorithms, hence, its results were adopted before proceeding to the validation process. Moreover, by comparing the cluster analysis between the results of PCA and t-SNE, it was shown how t-SNE can unveil more elaborate cluster structure.

In order to validate the cluster analysis results, three feeders were picked from each cluster. The three feeders were the representative, its nearest and farthest neighbour. The validation was carried out in three steps that involved a comparison of the topology, the demand profile and simulations of different levels of PV penetration. It was observed that these three aspects were consistent among the three selected feeders.

While validating the demand profile, the feeders characterized with commercial and industrial load were distinguished since their demand profile varies by the type of industry or business. Moreover, we identified the rural feeders that are characterized with high impedance and presence of regulators as problematic feeders with PV penetration. In addition to these feeders, the urban feeders with concentrated load exhibited voltage and loading problems with PV penetration. Therefore, DNOs need to pay close attention to these feeders when conducting network studies.

6.2 Conclusions

The procedure proposed for the identification of representative feeders is capable of serving its purpose. Through the case study, it was demonstrated that unique representative feeders were highlighted. Results of the case study in Chapter 5 showed that the choice of features was adequate. Feeders were grouped based on topology, load distribution and voltage regulation aspects.

Furthermore, it was demonstrated that relevant structures of the data can be unveiled using PCA and t-SNE, however, with t-SNE, more detailed shapes of clusters were revealed.

Some limitations in terms of finding the ONC were faced with t-SNE, this is due to the fact that t-SNE is a non-linear technique and the scores to find the ONC used in this thesis were designed for convex clusters with relatively simple shapes. Moreover, the values of the hyper-parameters of t-SNE described in Chapter 3 were set by default. This issue is not considered as a limitation however, an incidental enhancement to the t-SNE algorithm can be achieved by fine tuning these hyper-parameters.

Through the comparative analysis, it was shown that k-means++ was more advantageous in achieving dense and well separated clusters while the GMM algorithm showed poor performance. The hierarchical algorithm achieved acceptable results even so the use of KNG to reduce the computational complexity was not successful. This is because the hierarchical algorithm grouped a higher number of neighbours from what was specified through the connectivity constraint matrix generated by KNG in order to respect the ONC.

Furthermore, it was shown that the systematic validation helped in improving the cluster analysis and in interpreting the results. The systematic validation resulted in adding new features, identifying feeders with varying demand profile and feeders that are problematic with PV integration. Two categories of feeders had problems with PV integration. One group was described as rural with high impedance and presence of regulators, while the other part was described as urban with concentrated load.

6.3 Future Work

In this thesis, a procedure to identify representative feeders is proposed. The procedure involved using dimensionality reduction techniques, that map the original dataset to a lower dimension, and a variety of clustering algorithms. This may result in certain cases to grouping some feeders in clusters that do not share similar characteristics. Feeders that diverge substantially from their cluster in terms of characteristics are considered as anomalies.

Thesis work can be extended to cover anomaly detection in each cluster. Unsupervised anomaly detection can be done using Local Outlier Factor (LOF) algorithm [63]. LOF can identify the feeders that are considered anomalies by evaluating how isolated these feeders are with respect to their surrounding feeders in the cluster.

Another potential extension is to investigate the problem that the rural feeders which are characterized with the presence of voltage regulators are having when different percentages of clients install PVs. As discussed in Chapter 5, there is a room for improving the voltage control algorithms of regulators. Moreover, the model of the regulators may also be involved in causing the problems.

In addition, thesis work can be extended by exploiting the linearity of PCA in helping

the utility planners design their feeders. Some clusters in Chapter 5 that were subjected to VVOs due to the concentration of load were detected. Their relative position with respect to the clusters with no VVOs in the three-dimensional feature space formed using PCA could be analyzed for design purposes. The changes required in the three-dimensional space to move the feeders with VVOs of this cluster to the neighbouring cluster where no VVOs were found could be computed. Since PCA is a linear technique, the change required in terms of features could be quantified. Thus, utility planners could use the results of the clustering using PCA and design their feeders in a way to avoid having the feeder placed in a cluster that was subjected to VVOs.

Finally, a list of interesting applications can be further explored in future work:

1. In Chapter 2, the application of the IQR method to eliminate outliers was considered. Alternatively, the DBSCAN algorithm might provide for a more efficient alternative for the purposes of outlier elimination. DBSCAN is a density-based algorithm that can be purposed to detect outliers by inspecting the density of a point and its neighbours. The density can be the average distance between a point and a pre-defined number of neighbours. A point is, then, marked as an outlier if its density is low with respect to its neighbors.

2. The algorithms presented in this thesis, especially the hierarchical algorithm, consider the distance or similarity matrix (Euclidean distance) and perform the clustering based on an the objective function. Alternatively, clustering with adaptive neighbours [64] modifies the two aforementioned steps. Thus, this could be examined on the dataset used in this thesis to assess its performance. The authors in [64] introduced a rank constraint that forces optimal neighbours to belong to the same cluster. This rank constraint stresses on the local distances, while the clustering is performed. Theoretically, this implies that the quality of the clusters achieved with this clustering model could be similar to that of this thesis.

3. Autoencoders [65] are neural networks whose outputs are their own inputs. The encoder compresses the data to a low-dimensional space. The decoder, on the other end, attempts the recovery of the original data with minimal reconstruction loss. As such, the hidden layer (i.e. the bottleneck) could be used as a low dimensional representation of the original data. This technique can be considered as a nonlinear PCA that improves the representation of the data in the low dimensional space of PCA and t-SNE.

   Autoencoders can be used as a dimensionality reduction technique in this thesis. The

autoencoder can be trained and the hidden layer can be used as a low dimensional representation of the original data, similarly to what was carried out in Chapter 3.

Appendix A

Python Packages and Hyper-Parameters

In this appendix, the different packages and hyper-parameters associated with the algorithms are presented for the purpose of allowing the user to reproduce the results presented in this thesis.

Different Scikit-Learn packages for visualization and cluster analysis were used in the programming language Python [47].

- The package **seaborn**, which is a Python data visualization library, was used to visualize the co-variance matrix and scatter-plots in Chapter 5.
- From **Scikit-learn** pre-processing library, **StandardScaler** was used to remove the mean and scale to unit variance, as explained in Chapter 2.
- From **Scikit-learn** decomposition library, **PCA** was used as a linear dimensionality reduction technique in Chapter 5, with the following parameters:

  n_components=3

  whiten=True
- To visualize the three-dimensional data in Chapter 5, the following classes were used:

  Axes3D from **mpl_toolkits.mpl**

  proj3d from **mpl_toolkits.mpl**

  FancyArrowPatch from **matplotlib.patches**

- From **Scikit-learn** manifold library, **TSNE** was used as a dimensionality reduction technique in Chapter 5, with the following parameters:

  n_components=3

  perplexity=25

  learning_rate=200

  n_iter=1000

  min_grad_norm=$1e-07$

  metric='euclidean'

- From **Scikit-learn** cluster library, **KMeans** was used for the k-means clustering with the following parameters:

  n_clusters=10,11

  init='k-means++'

  max_iter=500

  n_init=500

  n_jobs=1

  tol=0.0001

- From **Scikit-learn** metrics library, **silhouette_samples**, **silhouette_score**, **calinski_harabaz_score** and **davies_bouldin_score** were used to evaluate the ONC.

- From **Scikit-learn** cluster library, **AgglomerativeClustering** was used for the hierarchical clustering with the following parameters:

  n_clusters=10,11

  affinity='euclidean'

  linkage='ward'

  connectivity= connectivity matrix generated by kneighbors_graph

- From **Scikit-learn** neighbors library, **kneighbors_graph** was used to generate the connectvity constraint matrix for the agglomerative algorithm, with the following parameters:

  n_neighbors=200

- From **Scikit-learn** mixture library, **GaussianMixture** was used for the GMM clustering algorithm, with the following parameters:

  covariance_type='full'

  n_components=10,11

  max_iter=100

  n_init=100

- From **Scikit-learn** metrics library, **pairwise_distances** and **pairwise_distances_argmin_min** were used to find the actual centers, their nearest and farthest neighbours.

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[2] K. P. Schneider, Y. Chen, D. Engle, and D. Chassin, "A taxonomy of North American radial distribution feeders," in *Proc. of 2009 IEEE Power Energy Society General Meeting*, pp. 1–6, July 2009.

[3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[5] I. Jolliffe, *Principal Component Analysis*, pp. 1094–1096. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[6] J. E. Jackson., *A User's Guide to Principal Components*. John Willey and Sons, Inc., 2003.

[7] A. Ben-Hur and I. Guyon, "Detecting stable clusters using Principal Component Analysis," *Methods in Molecular Biology (Clifton, N.J.)*, vol. 224, pp. 159–82, February 2003.

[8] G. W. Milligan, "Clustering validation: results and implications for applied analyses," in *Clustering and Classification*, pp. 341–375, World Scientific, 1996.

[9] K. Manivinnan, C. L. Benner, B. D. Russell, and J. A. Wischkaemper, "Automatic identification, clustering and reporting of recurrent faults in electric distribution feeders," in *Proc. of 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pp. 1–6, 2017.

[10] R. Arghandeh and Y. Zhou, *Big data application in power systems*. Elsevier, 2017.

[11] W. H. Kersting, "Radial distribution test feeders," *IEEE Transactions on Power Systems*, vol. 6, pp. 975–985, Aug 1991.

[12] W. H. Kersting, "Radial distribution test feeders," in *Proc. of 2001 IEEE Power Engineering Society Winter Meeting*, vol. 2, pp. 908–912 vol.2, Jan 2001.

[13] K. P. Schneider and B. A. Mather et al., "Analytic considerations and design basis for the IEEE distribution test feeders," *IEEE Transactions on Power Systems*, vol. 33, pp. 3181–3188, May 2018.

[14] R. Seguin, J. Woyak, D. Costyk, J. Hambrick, and B. Mather, "High-penetration PV integration handbook for distribution engineers," tech. rep., National Renewable Energy Lab (NREL), Golden, CO (United States), 2016.

[15] W. Kersting, *Distribution System Modeling and Analysis*. CRC Press, 2006.

[16] P. Denholm, K. Clark, and M. O'Connell, "On the path to sunshot-emerging issues and challenges in integrating high levels of solar into the electrical generation and transmission system," tech. rep., EERE Publication and Product Library, 2016.

[17] J. R. Agüero and S. J. Steffel, "Integration challenges of photovoltaic distributed generation on power distribution systems," in *Proc. of 2011 IEEE Power and Energy Society General Meeting*, pp. 1–6, July 2011.

[18] M. McGranaghan, T. Ortmeyer, D. Crudele, T. Key, J. Smith, and P. Barker, "Renewable systems interconnection study: Advanced grid planning and operations," tech. rep., Sandia Report, 2008.

[19] B. D. Kroposki, "Summarizing the technical challenges of high levels of inverter-based resources in power grids," tech. rep., National Renewable Energy Lab.(NREL), Golden, CO (United States), 2019.

[20] "IEEE standard for interconnection and interoperability of distributed energy resources with associated electric power systems interfaces," *IEEE Std 1547-2018 (Revision of IEEE Std 1547-2003)*, pp. 1–138, April 2018.

[21] H. Sun, Q. Guo, J. Qi, V. Ajjarapu, R. Bravo, J. Chow, Z. Li, R. Moghe, E. Nasr-Azadani, U. Tamrakar, G. N. Taranto, R. Tonkoski, G. Valverde, Q. Wu, and G. Yang, "Review of Challenges and Research Opportunities for Voltage Control in Smart Grids," *IEEE Transactions on Power Systems*, vol. 34, pp. 2790–2801, July 2019.

[22] B. Palmintier, E. Hale, T. M. Hansen, W. Jones, D. Biagioni, H. Sorensen, H. Wu, and B. Hodge, "IGMS: An integrated ISO-to-Appliance scale grid modeling system," *IEEE Transactions on Smart Grid*, vol. 8, pp. 1525–1534, May 2017.

[23] D. P. Chassin, J. C. Fuller, and N. Djilali, "GridLAB-D: An Agent-Based simulation framework for smart grids," *Journal of Applied Mathematics*, vol. 2014, Jun 2014.

[24] CYME International, "CymDist v.8.0 software package," 2019. `http://www.cyme.com/software/#dist`.

[25] K. P. Schneider and J. C. Fuller, "Detailed end use load modeling for distribution system analysis," in *Proc. of IEEE Power Energy Society General Meeting*, pp. 1–7, July 2010.

[26] K. P. Schneider, J. C. Fuller, and D. P. Chassin, "Multi-State load models for distribution system analysis," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2425–2433, 2011.

[27] G. Chicco, "Overview and performance assessment of the clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68 – 80, 2012.

[28] A. K. Ghosh, D. L. Lubkeman, and R. H. Jones, "Load modeling for distribution circuit state estimation," *IEEE Transactions on Power Delivery*, vol. 12, pp. 999–1005, April 1997.

[29] J. Wang, X. Zhu, D. Lubkeman, N. Lu, N. Samaan, and B. Werts, "Load aggregation methods for Quasi-Static power flow analysis on high PV penetration feeders," in *Proc. of 2018 IEEE/PES Transmission and Distribution Conference and Exposition*, pp. 1–5, April 2018.

[30] M. Emmanuel and J. Giraldez, "Net electricity clustering at different temporal resolutions using a SAX-Based method for integrated distribution system planning," *IEEE Access*, vol. 7, pp. 123689–123697, 2019.

[31] T. C. Akinci, "Applications of Big Data and AI in electric power systems engineering," in *AI and Big Data's Potential for Disruptive Innovation*, pp. 240–260, IGI Global, 2020.

[32] K. Möhrke, Fabianand Kamps, M. Zdrallek, P. Awater, and M. Schwan, "Clustering and determination of relevant network operating points in analytical reliability calculations," *in Proc. of CIRED 2019 25th International Conference on Electricity Distribution*, vol. AIM, p. 754, 2019.

[33] H. L. Willis, H. N. Tram, and R. W. Powell, "A computerized, cluster based method of building representative models of distribution systems," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, pp. 3469–3474, Dec 1985.

[34] A. Berry, T. Moore, J. Ward, S. Lindsay, and K. Proctor, "National feeder taxonomy: describing a representative feeder set for Australian electricity distribution networks," *Newcastle, N.S.W.: CSIRO*, 2013.

[35] A. K. Jain and B. Mather, "Clustering methods and validation of representative distribution feeders," in *Proc. of 2018 IEEE/PES Transmission and Distribution Conference and Exposition*, pp. 1–9, April 2018.

[36] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[37] Y. Li and P. Wolfs, "Statistical discriminant analysis of high voltage feeders in Western Australia distribution networks," in *Proc. of 2011 IEEE Power and Energy Society General Meeting*, pp. 1–8, July 2011.

[38] R. J. Broderick and J. R. Williams, "Clustering methodology for classifying distribution feeders," in *2013 IEEE 39th Photovoltaic Specialists Conference (PVSC)*, pp. 1706–1710, June 2013.

[39] J. Cale, B. Palmintier, D. Narang, and K. Carroll, "Clustering distribution feeders in the Arizona public service territory," in *2014 IEEE 40th Photovoltaic Specialist Conference (PVSC)*, pp. 2076–2081, June 2014.

[40] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[41] L. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[42] V. Rigoni, L. F. Ochoa, G. Chicco, A. Navarro-Espinosa, and T. Gozel, "Representative residential LV feeders: A case study for the North West of England," in *2016 IEEE Power and Energy Society General Meeting (PESGM)*, pp. 1–1, July 2016.

[43] R. J. Broderick, K. Munoz-Ramos, and M. J. Reno, "Accuracy of clustering as a method to group distribution feeders by PV hosting capacity," in *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T D)*, pp. 1–5, May 2016.

[44] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[45] K. A. Nazeer and M. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in *Proceedings of the World Congress on Engineering*, vol. 1, pp. 1–3, Association of Engineers London, 2009.

[46] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing 3rd Edition*. New York, NY, USA: Cambridge University Press, 3 ed., 2007.

[47] F. Pedregosa, G. Varoquaux, and Gramfort et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[48] J. C. Das, *Application of Shunt Capacitor Banks*, pp. 453–502. 2015.

[49] G. Upton and I. Cook, *Understanding Statistics*. Oxford University Press, 1996.

[50] K. Ho, "Ten caveats of interpreting correlation coefficient in anaesthesia and intensive care research," *Anaesthesia and intensive care*, vol. 40, p. 595—597, July 2012.

[51] G. Milligan and M. Cooper, "A study of standardization of variables in cluster analysis," *Journal of Classification*, vol. 5, pp. 181–204, September 1988.

[52] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, July 2017.

[53] S. Boslaugh, *Statistics in a nutshell: A desktop quick reference*. O'Reilly Media, Inc., 2012.

[54] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[55] C. Ding and X. He, "K-means clustering via Principal Component Analysis," in *Proceedings of the Twenty-First International Conference on Machine Learning*, (New York, NY, USA), p. 29, Association for Computing Machinery, 2004.

[56] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, (USA), p. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[57] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion?," *Journal of Classification*, vol. 31, pp. 274–295, Oct 2014.

[58] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[59] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, pp. 224–227, April 1979.

[60] J. A. D. Massignan, B. R. Pereira, and J. B. A. London, "Load flow calculation with voltage regulators bidirectional mode and distributed generation," *IEEE Transactions on Power Systems*, vol. 32, no. 2, pp. 1576–1577, 2017.

[61] A. Mendes, N. Boland, P. Guiney, and C. Riveros, "(N-1) contingency planning in radial distribution networks using genetic algorithms," in *2010 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America (T D-LA)*, pp. 290–297, 2010.

[62] F. Friend, "Cold load pickup issues," in *2009 62nd Annual Conference for Protective Relay Engineers*, pp. 176–187, 2009.

[63] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying Density-Based local outliers," *SIGMOD Rec.*, vol. 29, p. 93–104, May 2000.

[64] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, (New York, NY, USA), p. 977–986, Association for Computing Machinery, 2014.

[65] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pp. 37–49, 2012.