

IMPERIAL COLLEGE

Big Data Analytics in Power Systems

A thesis submitted to Imperial College London for the degree of Doctor of Philosophy

by

Mingyang Sun

November 2016

Department of Electrical and Electronic Engineering
Imperial College London

Declaration

The material contained within this thesis is my own work, except where other work is appropriately referenced. Any use of the first-person plural is for reasons of clarity.

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

With the increasing penetration of advanced sensor systems in power systems, an influx of extremely large datasets presents a valuable opportunity to gain insight for improving system operation and planning in the context of the large-scale integration of intermittent energy sources. To this end, it becomes imperative to implement big data methodologies to handle such complex datasets with the challenges of volume, velocity, variety, and veracity (4Vs) for further dealing with the problems in power systems.

The large-scale integration of intermittent energy sources, the introduction of shiftable load elements and the growing interconnection that characterizes electricity systems worldwide have led to a significant increase of operational uncertainty. The construction of suitable statistical models is a fundamental step towards building Monte Carlo analysis frameworks to be used for exploring the uncertainty state-space and supporting real-time decision-making. This thesis firstly proposes the novel composite modelling approaches that employ dimensionality reduction, clustering and parametric modelling techniques with a particular focus on the use of pair copula construction schemes. Large power system datasets are modelled using different combinations of the aforementioned techniques, and detailed comparisons are drawn on the basis of Kolmogorov-Smirnov tests, multivariate two-sample energy tests and visual data comparisons. The proposed methods are shown to be superior to alternative high-dimensional modelling approaches. In addition, the benefits of the proposed model are demonstrated through the applications on the calculation of a system's PNS profile and the security assessment based on decision trees.

Furthermore, this thesis presents a novel finite mixture modelling framework based on C-vine copulas (CVMM) for carrying out consumer categorization. The superiority of the proposed framework lies in the great flexibility of pair copulas towards identifying multi-dimensional dependency structures present in load profiling data. CVMM is compared to other classical methods by using real demand measurements recorded across 2,613 households in a London smart-metering trial. The superior performance of the proposed approach is demonstrated by analyzing four validity indicators. In addition, a decision tree classification module for partitioning new

consumers is developed and the improved predictive performance of CVMM compared to existing methods is highlighted. Further case studies are carried out based on different loading conditions and different sets of large numbers of households to demonstrate the advantages and to test the scalability of the proposed method.

In addition, in the interest of economic efficiency, design of distribution networks should be tailored to the demonstrated needs of its consumers. However, in the absence of detailed knowledge related to the characteristics of electricity consumption, planning has traditionally been carried out on the basis of empirical metrics; conservative estimates of individual peak consumption levels and of demand diversification across multiple consumers. Although such practices have served the industry well, the advent of smart metering opens up the possibility for gaining valuable insights on demand patterns, resulting in enhanced planning capabilities. This thesis is motivated by the collection of demand measurements across 2,613 households in London, as part of Low Carbon London project's smart-metering trial. Demand diversity and other metrics of interest are quantified for the entire dataset as well as across different customer classes, investigating the degree to which occupancy level and wealth can be used to infer peak demand behavior.

This thesis also presents a novel TNEP framework that accommodates multiple sources of operational stochasticity. Inter-spatial dependencies between loads in various locations and intermittent generation units' output are captured by using a multivariate Gaussian copula. This statistical model forms the basis of a Monte Carlo analysis framework for exploring the uncertainty state-space. Benders decomposition is applied to efficiently split the investment and operation problems. The advantages of the proposed model are demonstrated through a case study on the IEEE 118-bus system. By evaluating the confidence interval of the optimality gap, the advantages of the proposed approach over conventional techniques are clearly demonstrated.

Finally, this thesis proposes a novel scenarios selection framework for the transmission expansion problem to obtain an accurate solution in terms of operating costs and investment decisions with a significantly reduced number of operating states. Different classification variables and clustering techniques are considered and compared to determine the most appropriate combination for this specific problem.

Benders decomposition is applied to solve the TNEP problem by splitting the investment and operation problems. The superior performance of the proposed scenarios selection framework is demonstrated through a numerical case study on the modified IEEE 118-bus system.

Table of Contents

Acknowledgements.....	7
List of Publications	8
List of Symbols	9
1. Introduction.....	12
1.1 Background.....	12
1.2 Research Questions.....	14
1.3 Thesis Structure	18
1.4 Original Contributions	20
2. Background Methodology	23
2.1 Clustering Techniques	23
2.2 Dimensionality Reduction Techniques.....	26
2.3 Copulas and Pair-copula Construction.....	30
3. Evaluating Composite Approaches to Modelling High-Dimensional Stochastic Variables in Power Systems.....	44
3.1 Introduction.....	44
3.2 Proposed Composite Modelling Framework	46
3.3 Model Evaluation Techniques	49
3.4 Case-study: Evaluation Based on Real Data.....	50
3.5 Case-Study: PNS Calculation in IEEE 118-BUS System.....	56
3.6 Case-Study: System Security Assessment.....	61
3.7 Conclusions.....	64
4. C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data.....	66
4.1 Introduction.....	66
4.2 C-vine Mixture Model Clustering Algorithm.....	69
4.3 Clustering Quality Evaluation.....	73
4.4 Case Study Application.....	74
4.5 Conclusion	91
5. Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data	93
5.1 Introduction.....	93
5.2 Diversified Demand Quantification.....	96
5.3 Analysis of Smart-metering Data.....	98
5.4 Conclusions.....	104
6. Transmission Network Expansion Planning with Stochastic Multivariate Load and Wind Modelling.....	106

6.1	Introduction.....	106
6.2	Stochastic Multivariate Dependence Modelling Using Copulas	108
6.3	Proposed TNEP Model	112
6.4	Simulation Study and Results Analysis	115
6.5	Conclusions.....	119
7.	A Target-Based Scenarios Selection Approach for Transmission Network Expansion Planning....	120
7.1	Introduction.....	120
7.2	Clustering variables and clustering techniques.....	123
7.3	Proposed target-based scenarios selection framework	126
7.4	Simulation study and results analysis	129
7.5	Conclusions.....	142
8.	Conclusion and Future Works	144
8.6	Future Work.....	149
	References.....	151
	Appendix A.....	157

Acknowledgements

I am deeply indebted to my supervisor, Prof. Goran Strbac, for his crucial guidance and consistent encouragement in my doctoral work. I appreciate all his contributions of ideas, supports, and funding to provide me a fulfilling and challenging Ph. D. experience. His profound knowledge and innovative thinking of system operation and planning problems inspired me to make efforts in these areas with certain contributions.

Many thanks also to past and present colleagues in the Control and Power Group at Imperial College, especially Dr Ioannis Konstantelos, Dr Simon Tindemans, and Dr Fei Teng, for all their valuable discussions, assistants in the field of statistics, uncertainty modelling, transmission planning, and stochastic optimization. In particular, Dr Ioannis Konstantelos has guided me to be a qualified researcher from learning new knowledge to propose novel algorithms, evaluating results of experiments to finally write academic papers.

I am also very grateful to all the colleagues from RTE, National Grid, and UKPN for their important inputs over these years.

My deep appreciation goes out to my friends who always accompany with me and support me when I had difficulties in the process of study and life. Special thanks to Yilun Zhou, Boli Chen, Yujian Ye, Yanfei Chen, Ming Ge, Chen Wang, Fangjing Hu, Zongxiang Wu, Yujie Wu, Yunyang Wu, Yuanqi Hu, Zhengwei Wang, Hailing Fu, Yuan Qin, Dixi Liu, Jingye Sun, Cheng Cheng, Chengyuan Liu, Yang Wang, Ying Fan, and my dear girlfriend, Yixuan Chang.

Lastly, I would also like to express my heartfelt thanks to my parents for their sustained support and selfless love. It is impossible to complete this thesis without their encouragement and care.

List of Publications

Journal Papers

M. Sun, I. Konstantelos and G. Strbac, “C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data,” *IEEE Trans. Power Syst*, Accepted.

M. Sun, F. Teng, I. Konstantelos and G. Strbac, “A Target-Based Scenario Selection Approach for Transmission Network Expansion Planning,” *Energy*, Submitted.

I. Konstantelos, M. Sun, S. Tindemans, G. Strbac, S. Issad, P. Panciatici “Modelling of High-Dimensional Stochastic Variables Using Vine Copulas,” *IEEE Trans. Power Syst*, Submitted.

Conference Papers

M. Sun, I. Konstantelos, S. Tindemans and G. Strbac, “Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems,” PSCC '16, Genoa, pp. 1-8, 2016.

M. Sun, I. Konstantelos and G. Strbac, “Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data,” IEEE PES GM 2016, Boston, Jul. 2016.

M. Sun, I. Konstantelos and G. Strbac, “Transmission Network Expansion Planning with Stochastic Multivariate Load and Wind,” PMAPS 2016, Beijing, Oct. 2016.

Technical Report

I. Konstantelos, M. Sun, and G. Strbac, “Quantifying demand diversity of households,” Report for the 'Low Carbon London' LCNF project: Imperial College London, 2014.

I. Konstantelos, M. Sun, S. Tindemans, G. Strbac and S. Issad, “Methodology for Sampling of External Stochastic Variables,” Deliverable D4.1 Report for the 'iTesla' FP7 project: Imperial College London, 2013.

List of Symbols

Sets and indices

$t \in T$	Scenarios.
$i, j \in I$	Indices of the network buses.
$g \in G$	Indices of all generators.
$\gamma \in \Gamma$	Indices of thermal generators, $\Gamma \subset G$.
$w \in W$	Indices of wind generators, $W \subset G$.
$d \in Nd$	Indices of load.
$i, j \in \Lambda$	Existing transmission lines.
$i, j \in \Lambda_{new}$	Candidate transmission lines.

Input parameters

n_{ij}	Candidate lines built between bus i and j , $i, j \in \Lambda_{new}$.
C_{ij}	Cost of building a line from bus i to j (\$/year).
π_g	Generation cost of generator g (\$/MWh).
Υ	Value of lost load (\$/MWh).
τ	Duration of each scenario (hour).
p^t	Probability of scenario t .
B_{ij}	Susceptance of existing line from bus i to j (p.u.), $i, j \in \Lambda$.
B_{ij}^{new}	Susceptance of candidate line from bus i to j (p.u.), $i, j \in \Lambda_{new}$.
f_{ij}^{max}	Power flow limit of the transmission line from bus i to j (MW).

Decision variables

f_{ij}	Power flow of the transmission line from bus i to j (MW).
----------	---

θ_i	Phase angle at bus i (rad).
S	Branch-node incidence matrix with elements s_{ij} .
D	Matrix of sampled load (MW).
P_W^{max}	Matrix of sampled wind power (MW).
P_Γ^{max}	Matrix of maximum conventional generators capacities (MW).
U	Matrix of load curtailment with elements U_d (MW).
F	Matrix of power flow with elements f_{ij} (MW).
P_Γ	Matrix of conventional generator output (MW).
P_W	Matrix of wind generator output (MW).

Abbreviations

SOs	System Operators.
TNEP	Transmission Network Expansion Planning
CVMM	C-Vine Copulas Mixture Model
LCL	Low Carbon London
ADMD	After-Diversity Maximum Demand
OPF	Optimal Power Flow
DNOs	Distribution Network Operators
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis
ANN	Artificial Neural Networks
SOM	Self-Organizing Map
LPP	Locality Preserving Projections
NPE	Neighborhood Preserving Embedding
LLE	Local Linear Embedding

KNN	K Nearest Neighbors
DNOs	Distribution Network Operators
CF	Coincidence Factor
LCZ	London's Low Carbon Zones
LPN	London Power Networks
EIZs	Engineering Instrumentation Zones
PNS	Power Not Supplied
DT	Decision Tree
NPV	Negative Predictive Value
PPV	Positive Predictive Value
ACC	Accuracy

1. Introduction

1.1 Background

With the increasing penetration of advanced sensor systems in power systems, an influx of extremely large datasets presents a valuable opportunity to gain insight for improving system operation and planning in the context of the large-scale integration of intermittent energy sources. To this end, it becomes imperative to implement big data methodologies to handle such complex datasets with the challenges of volume, velocity, variety, and veracity (4Vs) [1] for further dealing with the problems of power systems. Specifically, volume, referred to the unprecedented massive amounts of data, brings about the challenges in data storage, data loading, and information analyzing. For real-time tasks (e.g. outage prevention, efficient condition-based asset management [2]), it is imperative to efficiently collect and process the data with sufficiently fast speed. In addition, different types of data sources such as social graphs, weather forecasting systems, and the web, result in the challenge of variety that require utilities to take into account the diversified information when solving planning, operation, and decision-making problems. Finally, veracity signifies the uncertainty of the data, including the accuracy and the quality of measured data.

Under this new reality, data analytics are facing significant challenges when employing conventional data analysis techniques in the applications of power systems. For the System Operators (SOs), it becomes necessary and difficult to deal with such amounts of variables and to predict stochastic outputs of wind and solar generation. At the same time, the SO needs to ensure that the large generation over/under-supply that could occur due to this intermittency is considered and the system is kept well within the security boundaries. As a result, the increasing uncertainties surrounding the system operating conditions have rendered the need for more advanced on-line stability analysis tools. Until now, this has mainly involved running dynamic studies (e.g. simulating line outages) for a set of credible contingencies close to real-time to ensure that no operational constraint is violated. However, there is a limit to the number of simulations that can be run in real-time, severely limiting the analysis scope and forcing

SOs to operate in a conservative manner, potentially foregoing actions that could lead to a safe yet more optimal use of the available energy resources.

Electricity market liberalization has largely unbundled the distribution and supply services in many jurisdictions, providing customers with the freedom to select their electricity supplier. In this competitive environment, retail companies can improve the commercial attractiveness of their product by formulating tariffs aimed at different customer types. An important part of the tariff design process is the identification of meaningful customer classes that exhibit different consumption patterns, enabling the development of diversifiable products. The advent of smart metering has led to large-scale availability of consumption data that render clustering analysis increasingly possible. However, it is another difficult task to carry out load profile clustering that aims to allocate consumers into a small number of homogeneous groups, ensuring that elements of the same cluster are similar between them, while being dissimilar to elements of different clusters.

In the context of distribution network design, peak electricity consumption patterns of the households that a network should serve are the key drivers for determining the ratings and location of the necessary distribution assets. The variational as well as inter-temporal characteristics of peak electricity demand are of paramount importance for achieving efficient and economic distribution network design. The design of distribution network should be tailored to the demonstrated needs of its consumers i.e. meeting the coincident winter peak demand. However, in the absence of detailed knowledge related to the characteristics of electricity consumption, as has historically been the norm, taking a fully-informed decision has not been possible. Planning engineers traditionally based their designs and investment decisions on empirical metrics and established rules that have been derived from experience and have been proven to serve the industry well over the past decades. Such approaches are based on historical estimates of individual peak consumption levels as well as on the fact that the coincident demand of a large number of households exhibits reduced sensitivity to the attributes of individual consumers (that may vary wildly across a group of households) due to the effect of demand diversification. The availability of a large number of actual residential demand measurements enable us to calculate such diversification metrics

based on real data. This presents an unprecedented opportunity to evaluate the applicability of empirical approaches to the new reality of electricity consumption habits and come up with updated values towards informing the distribution planning process.

From the view of transmission network expansion planning (TNEP), the ever-expanding spectrum of possible operating points necessitates the consideration of a very large number of scenarios within a cost-benefit framework, leading to computational issues. On the other hand, failure to adequately capture the behaviour of stochastic parameters can lead to inefficient expansion plans. It is imperative to accommodate multiple sources of operational stochasticity by considering inter-spatial dependencies between loads in various locations and intermittent generation units' output to form the basis of a Monte Carlo analysis framework for exploring the uncertainty state-space.

In addition, TNEP is facing a tremendous challenge to consider the system operational variability and uncertainty introduced by the increasing integration of renewable energy resources. To this end, a vast number of scenarios within a cost-benefit framework are conventionally used to handle this uncertainty, leading to computational issues, especially when it comes to a long-term TNEP problem with high intermittent generation penetration. Each scenario of demand-generation patterns with a certain probability is used to solve the TNEP problem. Therefore, it is imperative to reduce the amount of operational states by selecting the representative scenarios from the original dataset via clustering techniques.

1.2 Research Questions

This thesis focuses on developing and proposing advanced data analysis techniques to deal with the challenges of big data in power systems. The *Research Objectives* of this PhD thesis can be summarised as:

Q1: High-dimensional stochastic variables modelling: The large-scale integration of intermittent energy sources, the introduction of shiftable load elements and the growing interconnection that characterizes electricity systems worldwide have led to a significant increase of operational uncertainty. The vast number of stochastic variables beyond the operator's control are expanding the range of possible system states, thus

rendering experience-based operational practices inefficient or even obsolete [3]. Under this new reality, the use of historical data will be instrumental in informing operators of possible system states likely to occur in the near-term and providing useful information regarding the risk the system is likely to face.

There are several advantages to having a stochastic model matched to historic observations instead of relying solely on past measurements. Most importantly, a parametric model is capable of producing samples that are similar but not identical to what has already been encountered, thus interpolating and extrapolating the historical dataset. In addition, the model parameters themselves can be useful in extracting properties of a multivariate dataset that could be too complex to identify through direct examination. Another practical advantage is not having to deal with missing data. Beyond all the above, statistical models can provide versatility when combined with other sampling or data-mining techniques. For example, the ability to generate a sample population of arbitrarily large size can be extremely useful in generating unseen training and testing data sets for machine learning algorithms to subsequently produce efficient security rules. Despite the numerous applications, there has been little progress in developing a unified stochastic modelling framework. Particularly in the case of high-dimensional datasets, there are four distinct issues that render this task highly challenging; the non-Gaussian marginal probability distribution of individual variables, the non-linear dependency structure, the large number of observations to be fit as well as the vast number of variables, resulting in increased computational complexity.

Q2: Customer classification using smart-meter data: Electricity market liberalization has largely unbundled the distribution and supply services in many jurisdictions, providing customers with the freedom to select their electricity supplier. In this competitive environment, retail companies can improve the commercial attractiveness of their product by formulating tariffs aimed at different customer types. An important part of the tariff design process is the identification of meaningful customer classes that exhibit different consumption patterns, enabling the development of diversifiable products. Moreover, electrical customer classification can also play a crucial role in load forecasting [4],[5] and modelling [6], electricity market development [7], energy system planning and operation [8] and theft detection [9]. Naturally,

information on customer type (e.g. industrial, commercial, and residential) provides important information regarding the likely electricity consumption pattern and intensity. However, for further partitioning and exploratory analysis to be carried out effectively, high-frequency demand measurements are necessary [10]. As such, the advent of smart metering has led to large-scale availability of consumption data that render clustering analysis increasingly possible. To this end, it is imperative to propose a high-performance clustering algorithm to handle such complex data for the purpose of classifying the customers.

Q3: Analysis Demand Diversity: When designing a distribution substation and sizing assets such as transformers and cables, the peak coincident demand across the house-holds that each asset should serve is a key consideration. Underestimating the coincident peak demand, which in Europe typically occurs during low-temperature winter days, will result in undersized assets and an inability to service load during some periods. On the other hand, over-estimating the peak coincident demand can lead to substantially cost-inefficient decisions, given that the same level of reliability could be provided with less expensive assets of reduced rating. Given that Distribution Network Operators (DNOs) must meet strict reliability targets in combination with the inability to accurately infer the coincident peak demand of a group of consumers due to lack of instrumentation at the household level, DNOs have historically undertaken network design based on some pre-determined empirical metrics. These metrics are typically conservative to allow for worst case scenarios and have little customization related to the specific planning case at hand. The use of these general-purpose metrics is made possible because of the fact that the coincident demand of a large number of consumers' exhibits reduced sensitivity to the attributes of individual consumers, which may otherwise vary wildly when viewed in isolation. This effect is known as demand diversification and is a very important concept that pertains to electricity consumption. Demand diversity exists because the use of individual appliances in different households occurs at different times due to consumers' different schedules and preferences. It follows that accurately quantifying the effect of demand diversity is a quintessential aspect of efficient network design. Therefore, another problem comes up

with understanding demand diversity based on the high-resolution smart meter measurements and demographic data.

Q4: Scenarios Generation for TNEP Problem: With the increasing penetration of renewable energy resources, it is significant to carry out transmission network expansion planning (TNEP) with the consideration of uncertainties of uncertainties stemming from different sources. In particular, variations of loads and wind generators outputs are of interest to be investigated when performing planning. Conventionally, wind and load are assumed to be uncorrelated in [11]. In [12], uncertainties of load and wind are represented by scenarios generated using clustering method. However, entirely depending on real observations cannot provide a sufficient number of samples for Monte Carlo simulation. To generate scenarios that have similar dependence structure but not identical to what has already been encountered, a parametric model can be employed to interpolating and extrapolating the historical dataset. In [13], total load and available wind power are modelled via a bivariate Gaussian copula based on historical data. The samples generated from the constructed model are then used to carry out a two-stage stochastic program for transmission planning. However, although this method captures the dependence between total load and total wind output, the individual variables (electricity consumption at each bus and output of each wind farm) are assumed to be perfectly correlated. This is a substantial simplification undertaken in the interest of relieving the increased computational load that comes with modelling the stochastic variables at the level of individual variables instead of their aggregation. In this thesis, how to capture the interspatial correlations of load and renewables output at various places can be regarded as another important question for investment planning.

Q5: Scenarios Selection for TNEP problem: In the context of large-scale systems, transmission network expansion planning is encountered with various challenges especially when taking into account the uncertainties in demand and renewable power generation. In particular, it is essential to capture the temporal correlations and interspatial dependencies between loads in various locations and intermittent generation units' output. This can be achieved by representing the uncertainties via a set of historical operating points, each of them indicates a demand-wind pattern with an estimated probability that can be exploited in the proposed TNEP model. Nevertheless,

the vast number of historical data renders it indispensable to select the representative scenarios as less as possible on condition that an accurate planning result is ensured. As stated in [14], clustering the scenarios based on their similarity is one of the most broadly used methods to compress the size of input data for the transmission expansion problem. To this end, three major directions in terms of the clustering based scenarios selection method can be concluded as:

- 1) For the TNEP problem, the selection of clustering variables is the main topic of interest to be investigated;
- 2) Given a type of clustering variables, among the varieties of clustering methods, it is hard but imperative to choose the most appropriate one for this specific problem;
- 3) After obtaining the constructed clusters of data, the selection of representative profile in each cluster also has great influence on the final solution of the target problem.

In this thesis, we make efforts to mainly explore the first two issues of clustering variables and clustering methods, and then preliminarily analyze the effects of selecting different scenarios (e.g. mean value or actual median point) in each cluster.

1.3 Thesis Structure

This thesis consists of six technical chapters to deal with the above-mentioned research questions. Note that the relevant literature reviews for each topic are included in each chapter.

Chapter 2 introduces some fundamental data-mining and data analysis techniques that contain clustering methods, dimensionality reduction methods, and the definitions of copulas and pair-copula construction. Different combinations and means to use these tools for handling various types of data and solving power system problems are further explained in the following chapters.

Chapter 3 describes the proposed novel composite modelling approaches that employ dimensionality reduction, clustering and parametric modelling techniques with a particular focus on the use of pair copula construction schemes. Large power system datasets are modelled using different combinations of the aforementioned techniques, and detailed comparisons are drawn on the basis of Kolmogorov-Smirnov tests,

multivariate two-sample energy tests and visual data comparisons. The proposed methods are shown to be superior to alternative high-dimensional modelling approaches. This chapter is based on a journal paper that has been submitted to IEEE Transaction on Power System [15] and a conference paper that has been published and presented [16].

Chapter 4 introduces a new finite mixture modelling framework based on C-vine copulas (CVMM) for carrying out consumer categorization. The superiority of the proposed framework lies in the great flexibility of pair copulas towards identifying multi-dimensional dependency structures present in load profiling data. CVMM is compared to other classical methods by using real demand measurements recorded across 2,613 households in a London smart-metering trial. The superior performance of the proposed approach is demonstrated by analyzing four validity indicators. In addition, a decision tree classification module for partitioning new consumers is developed and the improved predictive performance of CVMM compared to existing methods is highlighted. Further case studies are carried out based on different loading conditions and different sets of large numbers of households to demonstrate the advantages and to test the scalability of the proposed method. This chapter is based on a journal paper that has been accepted by IEEE Transaction on Power System [17].

Chapter 5 is motivated by the collection of demand measurements across 2,613 households in London, as part of Low Carbon London project's smart-metering trial. Demand diversity and other metrics of interest are quantified for the entire dataset as well as across different customer classes, investigating the degree to which occupancy level and wealth can be used to infer peak demand behavior. This chapter is based on a conference paper that has been accepted and presented [18].

Chapter 6 presents a TNEP framework that accommodates multiple sources of operational stochasticity. Inter-spatial dependencies between loads in various locations and intermittent generation units' output are captured by using a multivariate Gaussian copula. This statistical model forms the basis of a Monte Carlo analysis framework for exploring the uncertainty state-space. Benders decomposition is applied to efficiently split the investment and operation problems. The advantages of the proposed model are demonstrated through a case study on the IEEE 118-bus system. By evaluating the

confidence interval of the optimality gap, the advantages of the proposed approach over conventional techniques are clearly demonstrated. This chapter is based on a conference paper that has been accepted and presented [19].

Chapter 7 proposes a novel scenarios selection framework for the transmission expansion problem to obtain an accurate solution in terms of operating costs and investment decisions with a significantly reduced number of operating states. Different classification variables and clustering techniques are considered and compared to determine the most appropriate combination for this specific problem. Benders decomposition is applied to solve the TNEP problem by splitting the investment and operation problems. The superior performance of the proposed scenarios selection framework is demonstrated through a numerical case study on the modified IEEE 118-bus system. This chapter is based on a journal paper that has been submitted to Applied Energy [20].

Chapter 8 contains the concluding remarks and the potential further work of these topics.

1.4 Original Contributions

This thesis proposes and applies various novel data analysing tools to handle the research questions stated in chapter 1.2. The main contributions of this thesis can be summarized as:

- Propose a composite modelling framework to deal with the challenges of high-dimensional stochastic variables. In the proposed framework, vine-copulas module aims to handle the non-Gaussian marginal probability distribution of individual variables and the non-linear dependency structure are handled; dimensionality reduction module and clustering module are employed to deal with the issues of the vast number of variables and observations, respectively. The main contribution of this part is an in-depth evaluation of potential synergies between these different components with the aim to identify well-performing composite modelling approaches.
- Propose a novel C-vine copula mixture model clustering algorithm to classify customers based on residential electrical load pattern data gathered via smart-

meters. One main advantage of distribution-based clustering is the ability to systematically select the number of clusters while penalizing model complexity to construct parsimonious models. Another advantage is the ability to obtain a probability measure regarding the classification of each consumer, as opposed to the ‘hard’ clustering that characterizes other methods. In addition, a practical advantage relates to reduced memory requirements; by relying on parametric functions it is possible to effectively compress large amounts of data in a reduced-sized model. Under the new reality of the highly non-linear dependence structures and non-standard marginal distributions observed in demand datasets, the proposed C-vine based algorithm shows superior performance than the other conventional distribution-based clustering methods. In addition, CVMM can efficiently manage huge amounts of data and handle the datasets under different loading conditions (i.e. days and seasons). Finally, efficient detection of outliers is another important feature of the proposed clustering scheme.

- Comprehensive analysis of the high-resolution residential demand measurements, provided by the LCL smart meter trail that enables us to gain valuable insights towards the evolution of residential electricity consumption patterns in the UK and evaluate how they are affected by consumer demographic characteristics such as household occupancy and wealth level. In addition, peak and diversified demand metrics (e.g. After-diversity maximum demand (ADMD)) have been analysed and quantified, aiming to inform future distribution planning processes.
- Development of a novel TNEP framework that accommodates multiple sources of operational stochasticity. Inter-spatial dependencies between loads in various locations and intermittent generation units’ output are captured accurately re-constructed by using a multivariate Gaussian copula. A substantial improvement, from independent to bivariate to multivariate, and from aggregated to disaggregated variables that pertain to the specific network locations, highlights the importance of capturing stochastic dependence at the

level of individual variables when performing transmission expansion planning.

- Propose a novel target-based scenarios selection framework that aims to significantly reduce the essential number of operating points for the TNEP problem with large penetration of renewable energy resources, by clustering the original scenarios directly based on their target outputs. Another vital contribution of this thesis is to present three major research directions in terms of the clustering based scenarios selection method that concluded as clustering variables selection, clustering methods selection and the representative scenario selection in each constructed cluster. Beyond the demand-wind patterns, OPF patterns of existing lines, OPF patterns of candidate lines, and investment costs, are proposed as three new types of clustering variables that used as clustering criteria for the operating points. Comprehensive comparisons among varieties of clustering methods and different clustering variables are proceeded to choose the most appropriate clustering technique and scenarios selection framework for TNEP problem.

2. Background Methodology

2.1 Clustering Techniques

In order to handle the problem of the vast number of historical observations and the complex dependency patterns, exploiting clustering techniques has been demonstrated as an efficient way to partition the large dataset into groups according to their different statistical behavior. In this thesis, the considered clustering methods are classified into four categories: centroid-based clustering (k-means, k-medoids clustering), connectivity-based clustering (hierarchical clustering), distribution-based clustering (Gaussian Mixture Model (GMM) clustering), and artificial neural networks (ANN) based clustering (the self-organizing map (SOM)).

2.1.1 K-means clustering and k-medoids clustering

As one of the most popular unsupervised clustering algorithm, the k-means clustering technique [21] is broadly used to classify the input data into K (user defined integer value) clusters by means of an iterative procedure. Let $X = [x_1, \dots, x_M]^T \in \mathbb{R}^{N \times M}$ denote the input data, the main target of the iterations is to find the K centroids $C = [c_1, \dots, c_K]^T \in \mathbb{R}^{K \times M}$ of clusters $X^k \subset X$, for $k = 1, \dots, K$ that minimize the within-cluster Euclidean distances between the sets of points, and their corresponding centroids. Mathematically, the objective function, within-cluster sum of squares (WCSS), can be written as follows:

$$\min \left(\sum_{k=1}^K \sum_{i \in I_k} \|X_i^k - c_k\|^2 \right) \quad (2.1)$$

where c_i is the centre of the cluster set X^k , and I_k represents the indices set of the cluster k . In details, the k-means clustering algorithm can be summarized in the following steps:
Step1. Randomly place K points (i.e. $C = [c_1, \dots, c_K]^T$) as an initial group of centroids in the historical data space.

Step2. Assign each observation X_m to the cluster dataset k with the closest centroid c_k for $m = 1, \dots, M$.

Step3. Re-calculate the centroids C by:

$$c_k = \frac{1}{n_k} \sum_{i \in I_k} X_i^k, \forall k \in [1, \dots, K] \quad (2.2)$$

where n_k is the number of observations placed in cluster k .

Step4. If the newly-calculated centroids are same as the previous iteration with a tolerance ε , stop running. Otherwise, go back **Step 2**;

Although k-means clustering is a fast, robust and easy implementing clustering technique, the quality of clustering highly depends on the initial centroids and the number of clusters, which is an unknown prior. Based on the k-means clustering and the medoid shift algorithm, k-medoids clustering method aims to minimize the sum of dissimilarities between the data points assigned in a cluster and its corresponding central point [22]. The partitioning around medoids (PAM) algorithm is one of the most widely used k-medoid clustering. The difference between k-means and k-medoids methods is that the mean value in each cluster is replaced with the actual median.

2.1.2 Hierarchical clustering

Hierarchical clustering method [23] can construct a hierarchy of clusters by employing a measure of similarity between groups of data measurements rather than a pre-defined number of clusters. Regarding the clustering strategies, it can be categorized as divisive clustering algorithm (top-down approach) and agglomerative hierarchical clustering algorithm (bottom-up approach), which is considered in this research. In general, agglomerative hierarchical clustering can be concluded in the following steps. Firstly, each observation of X is assigned to its own singleton group. Afterwards, a similarity matrix $D = \{d^{i,j}; i = 1, \dots, N, j = 1, \dots, N\} \in \mathbb{R}^{N \times N}$ is constructed for the input dataset X based on the Euclidean distance. On the basis of the calculated similarity, each pair of clusters that are closest to each other should be merged to a higher level. This procedure is successively repeated until there is only one cluster left. Note that each element $d^{i,j}$ represents the similarity between clusters X^i and X^j . In terms of the intergroup similarity, different linkage criteria have been proposed with varieties of properties that have been concretely investigated and compared in [24]. In this thesis, we implement average linkage, complete linkage, and ward linkage for the sake of evaluating and comparing their performance in the proposed scenarios selection

framework for the TNEP problem. From the statistical point of view, hierarchical clustering method has the deficiencies that the clustering result cannot be re-evaluated and further adjusted due to its deterministic nature. An example of the hierarchical tree obtained by grouping the real historical demand-wind data of size $1,4251 \times 128$ via using complete linkage is shown in Figure 2-1. In the dendrogram, data points are connected by a vast number of U-shaped lines, whose height represents the similarity between each pair of clusters. Also, the horizontal axis includes the demand-wind patterns identifiers. Eventually, the target clusters can be built by choosing the calculated distance according to a determined number of clusters.

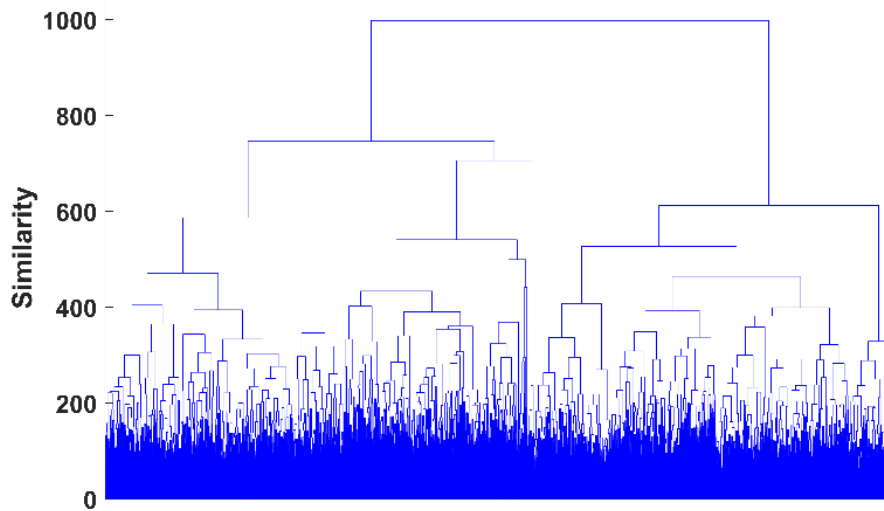


Figure 2-1: Dendrogram of the hierarchical clustering with complete linkage based on the demand-wind data in France, measured between Jan 2012 and Feb 2012. Horizontal axis: demand-wind pattern identifier. Vertical axis: similarity measure between clusters.

2.1.3 Gaussian mixture model

Beyond centroid and connectivity models, an alternative approach is the finite mixture model based clustering technique that renders it possible to tackle the issues of determining an optimal number of clusters by fitting the clusters of data to appropriate probability models [25]. This is a ‘soft’ clustering technique that can obtain a probability measure in terms of the classification of each data point. The whole dataset is described as a mixture of parametric distributions. In general, Gaussian distribution is commonly employed for fitting continuous data, which is suitable for this case. Mathematically, the Gaussian Mixture Models (GMM) [26] can be presented as follows.

Let $f_k(x_n|\mu_k, \Sigma_k) = \exp\left\{-\frac{1}{2}\left((x_n - \mu_k)^T \Sigma_k^{-1}(x_n - \mu_k)\right)\right\} / \sqrt{\det(2\pi \Sigma_k)}$ denote a multivariate Gaussian distribution of component k , for $k = 1, \dots, K$, where μ_k and Σ_k are the mean vector and the covariance matrix of f_k , respectively. The likelihood for a Gaussian mixture model with K components distributions is:

$$\mathcal{L}_{GMM}(\pi_1, \dots, \pi_K; \mu_1, \dots, \mu_K; \Sigma_1, \dots, \Sigma_K | X) = \prod_{n=1}^N \sum_{k=1}^K \pi_k f_k(x_n | \mu_k, \Sigma_k) \quad (2.3)$$

where $\sum_{k=1}^K \pi_k = 1$ and π_k is the proportion in $[0,1]$ for component f_k . Note that the optimal parameters that maximize the likelihood of the Gaussian mixture models are usually estimated by Expectation-Maximization (EM) algorithm [27].

2.1.4 Self-Organizing Map

Different from the previous clustering techniques, the self-organizing map (SOM) [28] is one of the most important artificial neural networks (ANN) based on competitive, unsupervised learning. It usually produces a two-dimensional arrangement of neurons, named the Kohonen layer, which maps input data from a high-dimensional space onto a plane. In this chapter, the SOM aims to perform a topology-preserving mapping from the historical data X to clusters X^k , for $k = 1, \dots, K$. Let $W = [w_1, \dots, w_K] \in \mathbb{R}^{K \times M}$ be the weight matrix, where $w^{k,i}$ is defined as the weight vector that represents the weight assigned to measurement x_n , for $n = 1, \dots, N$, to the cluster (unit) k . On the basis of the Euclidean distance between the training historical measurement x_n and all the weight vectors W , the winning unit is selected by finding the closest one. Afterwards, the weights of all the non-winning units remains unchanged and the weights of the winning unit is updated from step s to step $s + 1$ by following the Kohonen update rule:

$$W_k^{s+1} = W_k^s + \alpha^s (x_n - W_k^s) \quad (2.4)$$

where α is the learning rate. Generally, the SOM method has the benefit of dealing with large and complex data, however, it is difficult to determine the best number of clusters (units) K and the initial values of weight matrix W .

2.2 Dimensionality Reduction Techniques

For the measurements from the power system, high-dimensionality is another main challenge of modeling real-world data. Dimensionality reduction techniques can realize

the transformation of large data measurement sets from the original high-dimensional space to an appropriate lower-dimensional space while retaining the main properties of the original data sets as much as possible. In this thesis, we investigate three unsupervised linear dimensionality reductions techniques: Principal Components Analysis (PCA), Locality Preserving Projections (LPP), and Neighborhood Preserving Embedding (NPE). Among these techniques, we subdivide them into global structure-preserving methods (e.g. PCA) and neighborhood structure-preserving methods (e.g. LPP and NPE).

2.2.1 Principal Components Analysis

Principal Component Analysis (PCA) is a widely-used linear and full-spectral dimensionality reduction technique [29]. The basic concept of PCA is to perform an eigenvalue decomposition of the historical variables' covariance matrix. The resulting eigenvectors can be used to construct a lower dimensional representation of the original data set, named principal components (PCs), that keeps as much of the variance as possible. The corresponding eigenvalues can be seen as a metric of the information (more precisely the variance) contained within each dimension of this new domain (PCs domain). This way we can decide which eigenvectors are retained subject to retaining variance threshold VR^{target} . Let $X = [x_1, \dots, x_N]^T \in \mathbb{R}^{N \times M}$ denote the M-dimensional historical dataset with N observations and then calculate the covariance matrix Λ of X , which is positive semi-definite. After performing the eigenvalue decomposition of the

covariance matrix Λ , we define a diagonal matrix $D = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_M \end{bmatrix}$ whose diagonal

elements are the eigenvalues $(\lambda_1, \dots, \lambda_M)$ of the covariance matrix and an orthonormal matrix ψ that is composed of the eigenvectors of matrix Λ and an orthonormal matrix ψ . These three matrices satisfy the following expression: $\Lambda = \psi^T D \psi$. The detailed algorithm for performing PCA on X can be described below:

Step1. Calculate the zero-mean data X_c by subtracting the mean from the historical data:

$$X_c = X - \mu \quad (2.5)$$

where X is the historical data and $\mu = E(X)$ i.e. the vector of means;

Step2. Compute the covariance matrix of the mean-centered historical data X .

$$\Lambda = \text{cov}(X_c) \quad (2.6)$$

Step3. Perform the eigenvalue decomposition of the covariance matrix Λ as in:

$$\Lambda = \psi^T D \psi \quad (2.7)$$

where the orthonormal matrix ψ is composed of the eigenvectors of matrix Λ and the diagonal matrix D has eigenvalues $(e_1, \dots, e_M)$ of Λ as its diagonal elements;

Step4. Determine the number of retained PCs that satisfies the target variance retain criterion VR^{target} and let it to be q (i.e. $q \leq M$);

Step5. Keep the q eigenvectors associated to the largest q eigenvalues. Let ψ_q denote the matrix composed of the ordered q eigenvectors (as column vectors) and D_q denote the diagonal matrix containing the q largest eigenvalues;

Step6. Project the cloud of points in the principle components space:

$$Y_c = X_c \cdot \psi_q \quad (2.8)$$

Step7. Finally, in order to compute the reconstructed (back-projected) data $\hat{X}$ we have to project back to the original (known as actual) signal space and also add the means that were subtracted in Step 1:

$$\hat{X} = Y_c \psi_q^T + \mu \quad (2.9)$$

As mentioned in Step 4, the number of final retained principle components is computed using the ‘Variance Retained’ metric. Its computation is based on the number of retained dimensions and the corresponding eigenvalues. Assuming that the number of retained PCs is q (i.e. $q \leq M$), the variance retain criterion (VR) can be formally expressed in a mathematical way as below:

$$VR = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^n \lambda_i} \quad (2.10)$$

The above formula indicates that with an increasing q , more signal variance will be retained (variance can be regarded as an indicator of information content within each variable).

2.2.2 Locality Preserving Projections and Neighborhood Preserving Embedding

LPP and NPE are two sparse-spectral linear dimensionality reduction techniques that proposed in [30] and [31], designed to find a low-dimensional data set that preserves

the local neighborhood structure of the original data manifold. In fact, they can be regarded as the linear approximation to the popular non-linear Local Linear Embedding (LLE) method [32] and Laplacian Eigenmaps [33], respectively. Regarding the algorithmic procedure, LPP and NPE have some common steps:

Step1. Adjacency Graph Construction: Let Q denote a graph with M nodes for the data set $X = [x_1, \dots, x_N]^T \in \mathbb{R}^{N \times M}$. By using k nearest neighbors (KNN) method, node $i \in \{1, \dots, N\}$ (x_i) and node $j \in \{1, \dots, N\}$ (x_j) are connected by an edge if x_i is among k nearest neighbors of x_j and vice versa;

Step2. Weights Selection: In the constructed graph Q , each edge joining x_i and x_j is weighted by $w^{i,j}$, which is one element of the sparse symmetric weights matrix $W = \{w^{i,j}; i = 1, \dots, N, j = 1, \dots, N\} \in \mathbb{R}^{N \times N}$.

Up to this point, the constructed graph and the calculated weights can be used either in LPP or NPE. However, the final steps of LPP and NPE that aim to generate the linear projections are different. For LPP, we have:

Step3 (LPP). Eigenmaps: Solve the generalized eigenvector problem:

$$XLX^T \alpha = \lambda XDX^T \alpha \quad (2.11)$$

where D is a diagonal matrix with $D_{i,i} = \sum_j w^{j,i}$ and $L = D - W$ is the Laplacian matrix. Let q ($q < M$) denote the target reduced low dimensions, then the solution vectors $A = [\alpha_0, \dots, \alpha_{q-1}] \in \mathbb{R}^{M \times q}$, which ordered based on their eigenvalues, $\lambda_0 < \dots < \lambda_{q-1}$, are used to construct the q -dimensional data $Y = [y_1, \dots, y_N]^T \in \mathbb{R}^{N \times q}$ as follows :

$$Y = XA \quad (2.12)$$

For NPE, we have:

Step3 (NPE). Projections calculation: Solve the generalized eigenvector problem:

$$XMX^T \alpha = \lambda XX^T \alpha \quad (2.13)$$

where $M = (I - W)^T(I - W)$ is a symmetric and positive semidefinite matrix and $I = \text{diag}(1, \dots, 1) \in \mathbb{R}^{N \times N}$. Similar to LPP, the low-dimensional data $Y = XA$ can be directly obtained using the solution matrix A of (2.13).

2.3 Copulas and Pair-copula Construction

Given the highly complex dependence structure characterizing the inter-relationships between the variables of interest, it is imperative to capture higher-order statistics such as tail-to-tail dependencies of the data in power system. To this end, copulas are a type of multivariate joint distribution functions that are capable of modelling such non-Gaussian dependence structures. A copula is essentially a description of how two or more random variables relate to each other. To be more precise, copulas deal with the inter-dependency behavior of uniform random variables. However, given that any random variable may be transformed to a uniformly-distributed variable through its cumulative distribution function (CDF), copulas are essentially marginal-agnostic enabling the decoupling between marginals and dependence structure.

The basic concept of copulas can be demonstrated with the introduction of Sklar's theorem. Consider n random variables $X = (X_1, X_2, \dots, X_n)$ with marginal density functions $f_i(x)$. In case of independence, the joint density function is given by the product of all marginal densities:

$$f(x_1, x_2, \dots, x_n) = f_1(x_1)f_2(x_2) \dots f_n(x_n) \quad (2.14)$$

In case of dependence, we can write the joint density function as:

$$f(x_1, x_2, \dots, x_n) = c_{12\dots n}(F_1(x_1), F_2(x_2), \dots, F_n(x_n)) \times f_1(x_1)f_2(x_2) \dots f_n(x_n) \quad (2.15)$$

where we use the notation $x_{1:n}$ as a shorthand for the sequence $x_1, x_2, \dots, x_n$. The *copula density* $c_{12\dots n}: [0,1]^n \rightarrow \mathbb{R}$ describes the dependence between uniform random variables $\{U_1, \dots, U_n\} = \{F_1(X_1), \dots, F_n(X_n)\}$. The copula representation provides a convenient way to separate the marginal distributions of X_i from their dependency structure. The copula connecting all variables is unique if all marginal distribution functions are continuous, and for independent variables $c_{12\dots n} = 1$. The practical problem of interest is the identification of the best-fitting parametric copula to the random vector $\mathbf{X}^e$ defined by the empirical data.

Moving beyond the above definition, there are many different types of copulas; empirical, parametric, time copulas etc. For the purposes of this research, we will focus on parametric copulas, which are copula functions that can be fully described by a

family and corresponding parameters. The idea is to identify the family and parameters that give rise to a copula that best matches the dependency of historical variables. For the bivariate case there is a well-investigated and rich variety of copula families [34]. For example, five popular copula families that will form the fundamental modelling blocks for our approach are presented next.

2.3.1 Gaussian Copula

The Gaussian copula is also called normal copula and is usually signified by its defining distribution $N_n(0, R)$, where n is the number of dimensions ($n = 2$ for bivariate) and R is the correlation matrix. The analytical expressions for the Gaussian copula are presented in Appendix A.1. Gaussian copula is one kind of elliptical copula that only has mean dependence with no upper or lower tail dependencies [35]. The range of pairwise correlations is $\rho \in [-1, 1]$. These following figures give a visual illustration of the bivariate Gaussian copula:

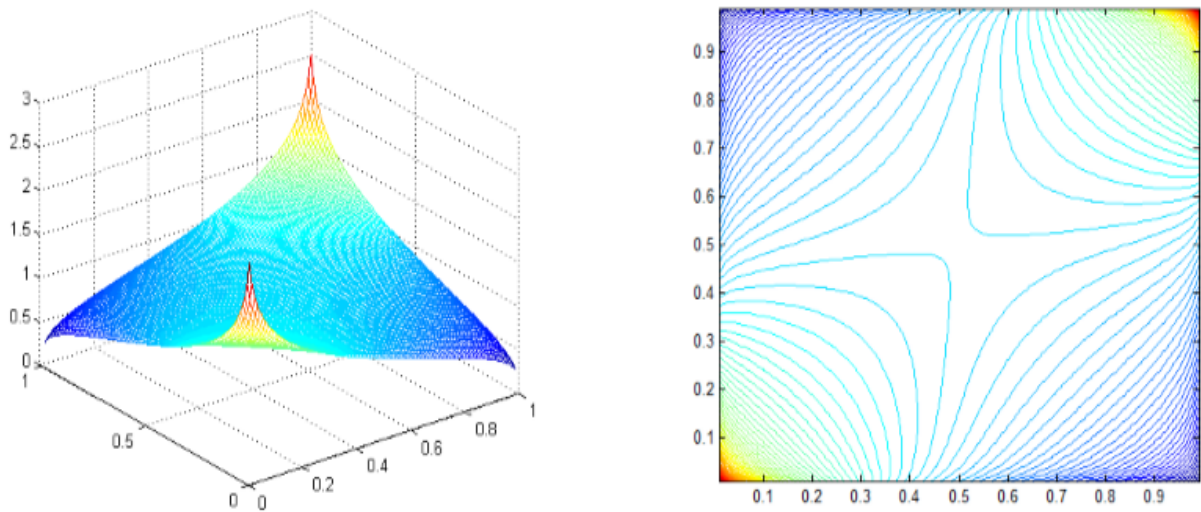


Figure 2-2: Bivariate Gaussian Copula distribution density ($\rho = 0.2$)

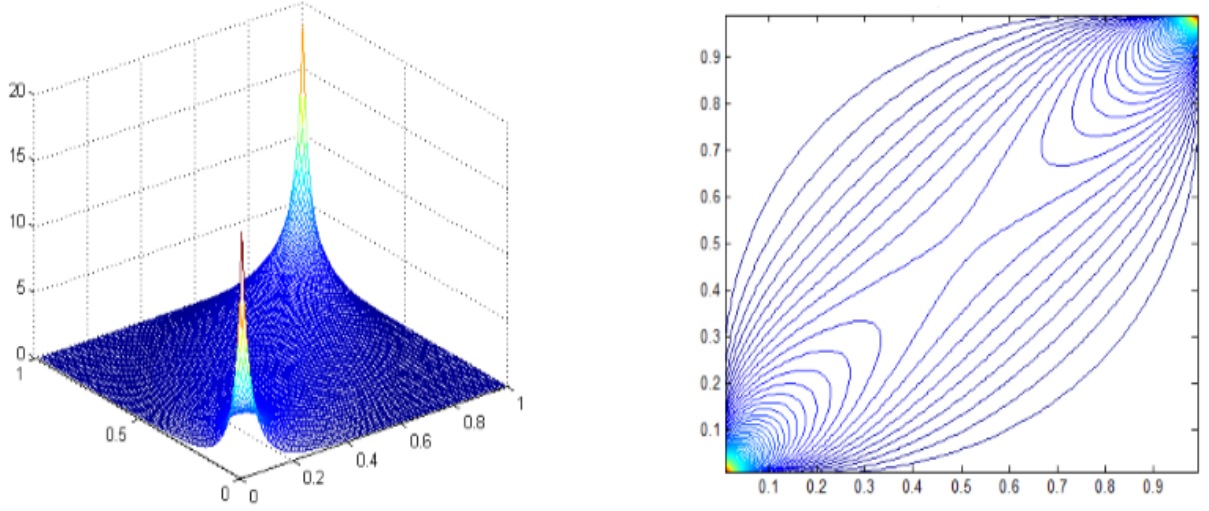


Figure 2-3: Bivariate Gaussian Copula distribution density ($\rho = 0.8$)

2.3.2 Student-t Copula

Student-t copula is also one kind of elliptical copula and is the copula defined by the random variable $X \sim t_d(\mu, \mathbf{R}, \vartheta)$, where $t_d(\mu, \mathbf{R}, \vartheta)$ is a d -dimensional t-distribution with mean vector $\mu = \mathbf{0}$, correlation matrix $\mathbf{R}$ and the ϑ degrees of freedom. The detailed formulas of t-copula are described in the Appendix A.2. This copula has symmetric upside and downside dependencies [35]. Two bivariate t-copulas with $\rho = 0.2$, $\nu = 6$ and $\rho = 0.8$ and $\nu = 4$ can be shown as below:

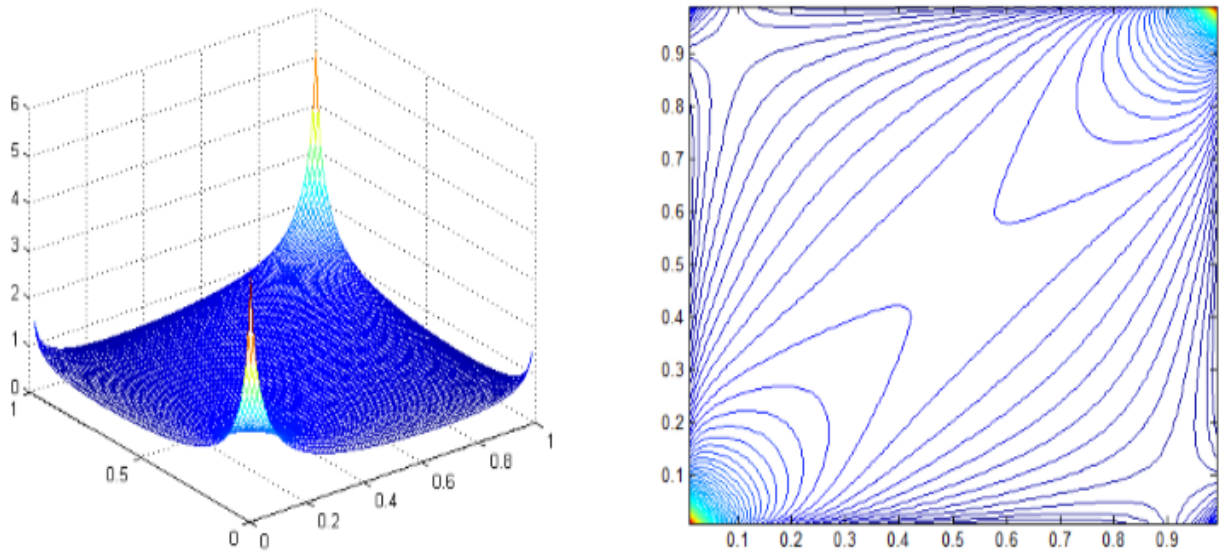


Figure 2-4: Bivariate Student's t-copula distribution density ($\rho = 0.2$ and $\nu = 6$)

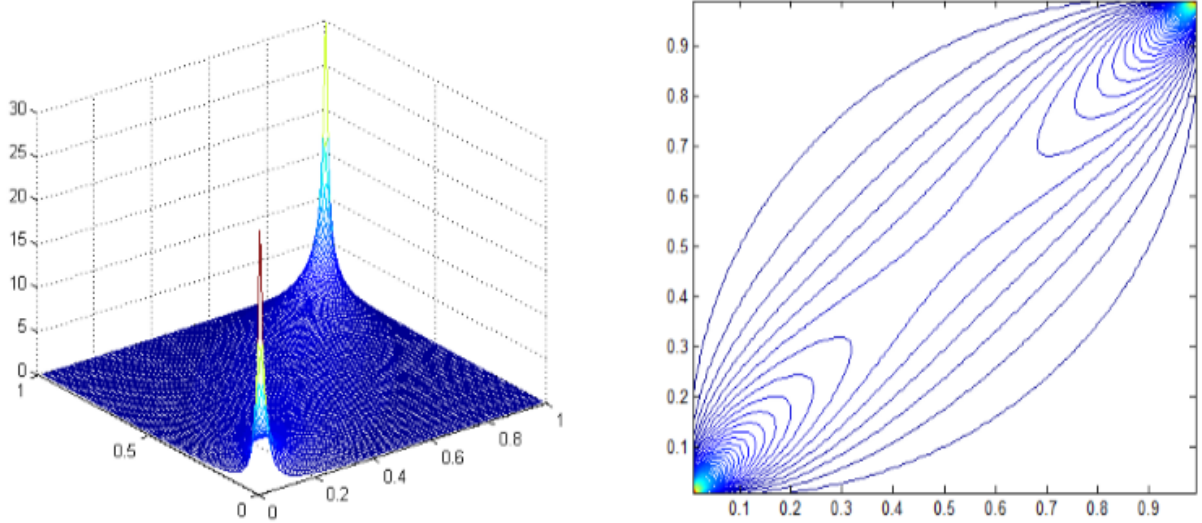


Figure 2-5: Bivariate Student's t-copula distribution density ($\rho = 0.8$ and $\nu = 4$)

2.3.3 Clayton Copula

Clayton is a type of Archimedean copula that was introduced by D.G. Clayton in 1978. The analytical expressions for the Clayton copula are presented in Appendix A.3. The Clayton copula has lower-tail dependence but no upper-tail dependence [35]. It is characterized by a single parameter ϑ and is defined on the range $(0, \infty)$. When $\vartheta \rightarrow 0$, it will tend to the independence copula ($C(u_1, u_2) = u_1 u_2$) while when $\vartheta \rightarrow \infty$, it tends to be the comonotonicity copula. Two bivariate Clayton copulas with $\vartheta = 0.2$ and $\vartheta = 0.8$ can be observed below:

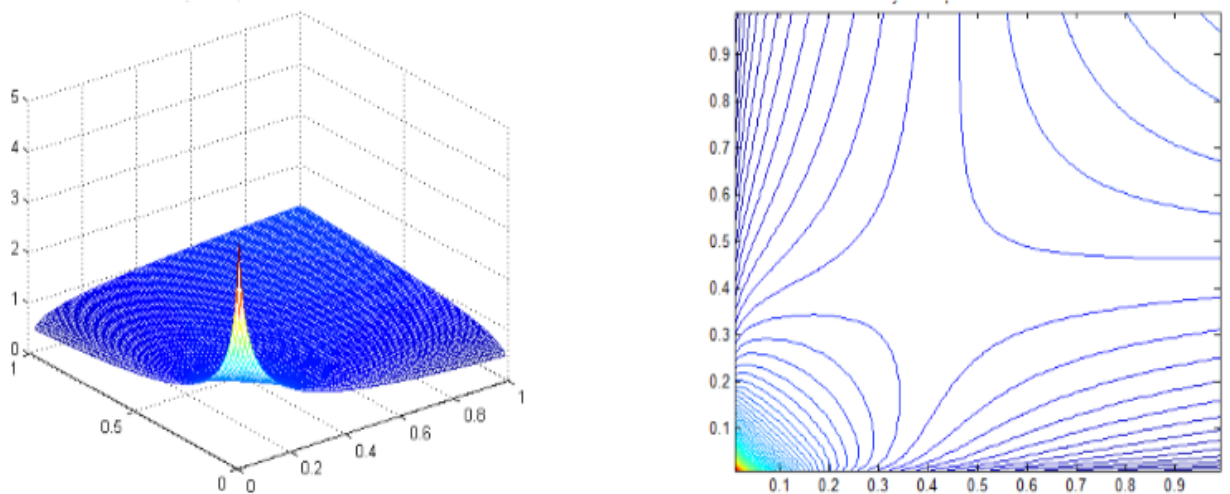


Figure 2-6: Bivariate Student-t copula distribution density ($\rho = 0.2$)

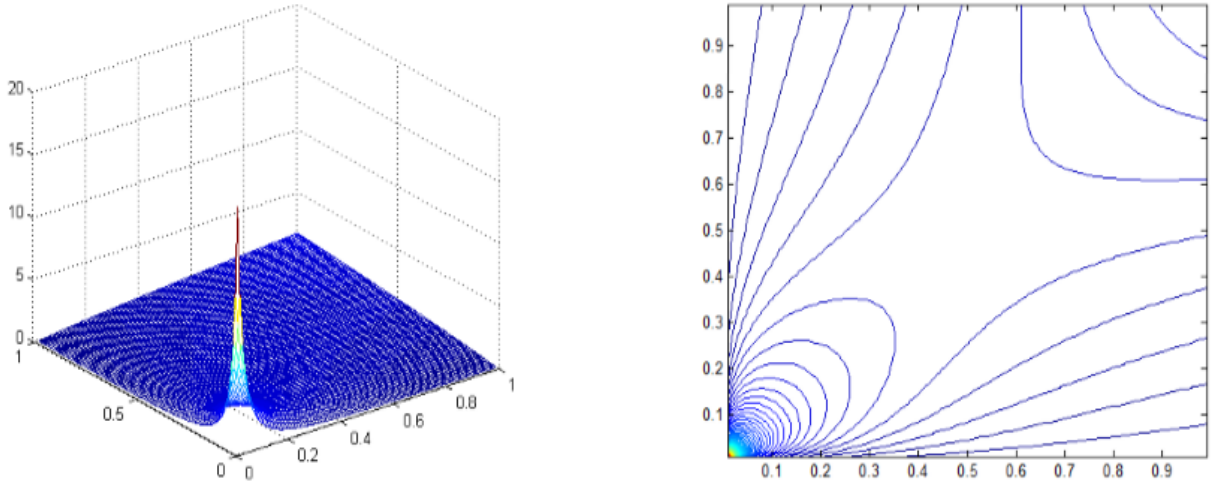


Figure 2-7: Bivariate Student-t copula distribution density ($\rho = 0.8$)

2.3.4 Frank Copula

Frank copula is one type of the Archimedean copulas. The expressions of Frank copula are described in Appendix A.4. The Frank copula does not have upper- and lower-tail dependence [35], while it is the only one type of Archimedean copula that is radially symmetric. The parameter of this type of copula ϑ has its range in $(0, \infty)$. Similar to the Clayton copula, when $\vartheta \rightarrow 0$, the Frank copula converges to the independence copula ($C(u_1, u_2) = u_1 u_2$) while when $\vartheta \rightarrow \infty$, it converges to be the comonotonicity copula. Two bivariate Frank copulas with $\vartheta = 0.2$ and $\vartheta = 0.8$ can be observed below:

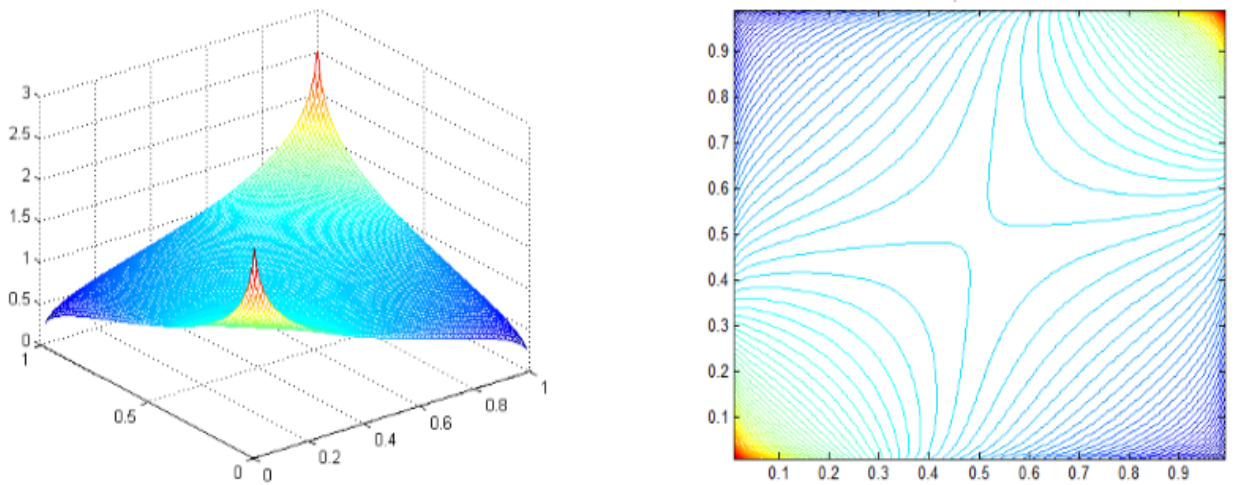


Figure 2-8: Bivariate Frank copula distribution density ($\vartheta = 0.2$)

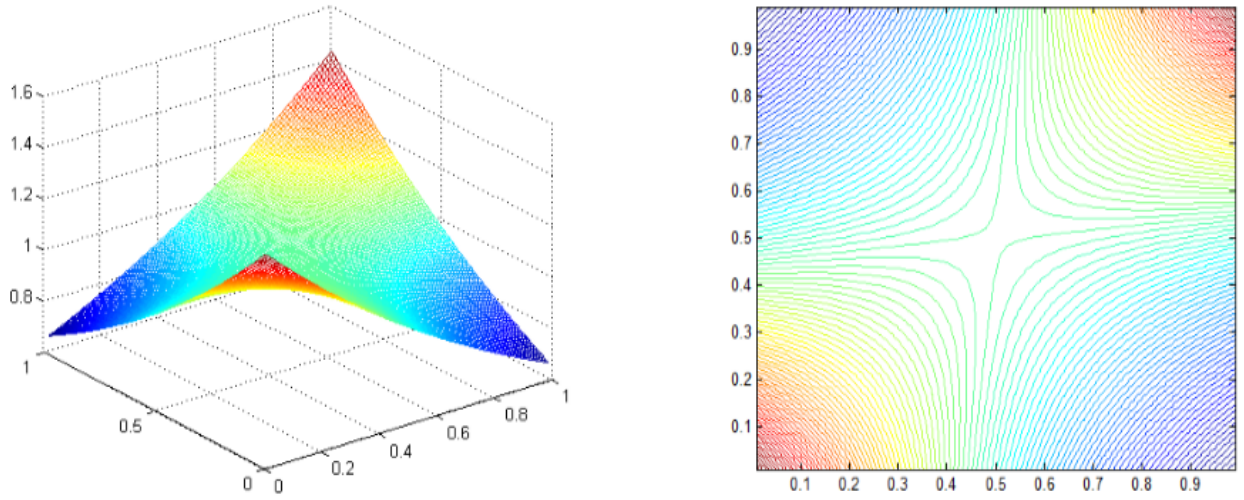


Figure 2-9: Bivariate Frank copula distribution density ($\vartheta = 0.8$)

2.3.5 Gumbel Copula

Gumbel is another Archimedean copula proposed by E.J. Gumbel in 1960. The analytical expressions for the Gumbel copula are presented in Appendix A.5. The Gumbel copula exhibits upper-tail dependence, while it does not have any lower-tail dependence [35]. The most significant property of the Gumbel copula is that it is the only member of the Archimedean family that is at the same time an extreme value copula, appropriate for modelling rare events. The parameter ϑ of this type of copula is defined on the range $(1, \infty)$. As all Archimedean copulas, when $\vartheta = 1$, Gumbel converges to the independence copula ($C(u_1, u_2) = u_1 u_2$) while when $\vartheta \rightarrow \infty$, it converges to be the comonotonicity copula. Two bivariate Gumbel copulas with $\vartheta = 1.5$ and $\vartheta = 4$ can be observed below:

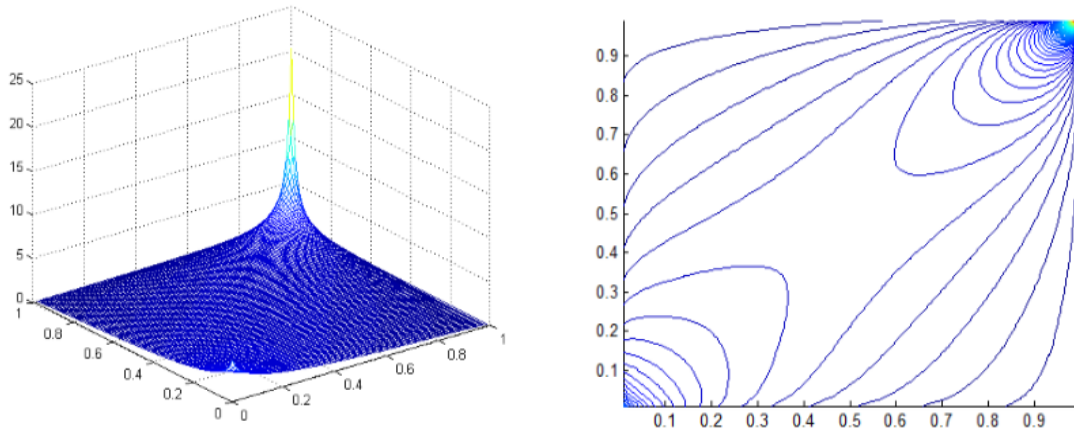


Figure 2-10: Bivariate Gumbel copula distribution density ($\vartheta = 1.5$)

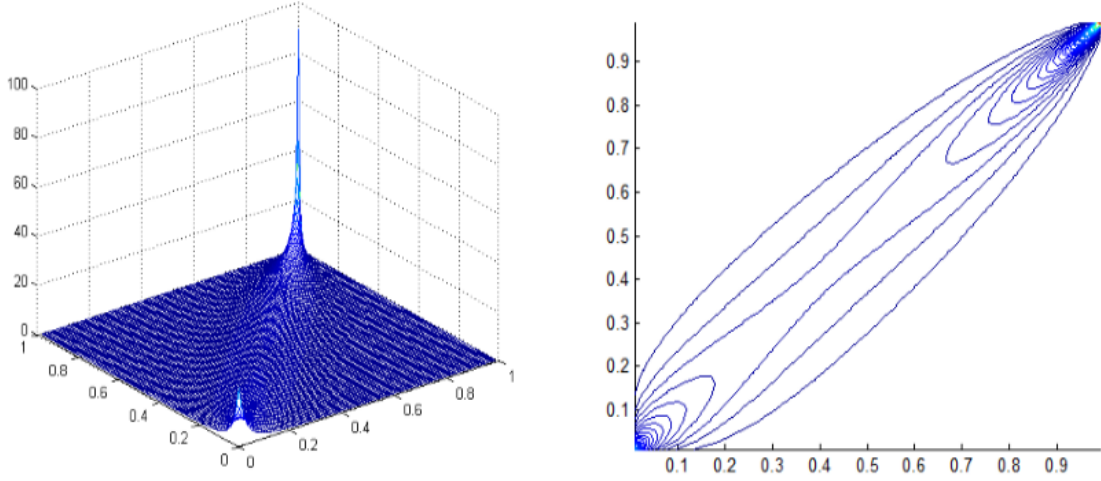


Figure 2-11: Bivariate Gumbel copula distribution density ($\theta = 4$)

2.3.6 Pair-copula construction

In the case of higher dimensions, the dependency patterns that may exist between large numbers of variables are far more complex than for the bivariate case. Attempting to capture this dependency structure using a single multivariate parametric copula can be restrictive and lack the flexibility required. The pair-copula construction (PCC) approach represents a novel way of constructing high-dimensional complex models of multivariate dependence, by extending bivariate theory to an arbitrary number of dimensions [35]. It was first introduced by Joe in [34] and developed in more detail in [36], [37] and [38]. The main idea of PCC is to decompose a multivariate distribution into a product of bivariate copulas by using recursive conditioning. A joint probability density can be recursively factorized as follows:

$$f(\mathbf{x}_{1:n}) = f_1(x_1) \cdot f_{2|1}(x_2|x_1) \cdot f_{3|21}(x_3|x_2, x_1) \cdots f_{m|m-1:1}(x_m|x_{m-1:1}) \cdot f_{n:m+1|m:1}(x_{n:m+1}|x_{m:1}) \quad (2.16)$$

Each of the univariate conditional probability distributions can be decomposed into (conditional) pairwise copulas using:

$$\begin{aligned} f_{j+i|j:1}(x_{j+i}|x_{j:1}) &= \frac{f_{j+i|j-1:1}(x_{j+i}, x_j|x_{j-1:1})}{f_{j|j-1:1}(x_j|x_{j-1:1})} \\ &= c_{j+i|j-1:1}(F_{j+i|j-1:1}, F_{j|j-1:1}) f_{j+i|j-1:1}(x_{j+i}|x_{j-1:1}) \end{aligned} \quad (2.17)$$

For clarity we have used the shorthand notation $F_{j+i|j-1:1}$ to represent the conditional distribution function $F_{j+i|j-1:1}(x_{j+i}|x_{j-1:1})$. This equation can be used recursively to

express all conditional densities for dimensions $1, \dots, m$ in Eq. (16) as products of bivariate copula densities and the corresponding marginal densities, resulting in the following decomposition.

$$f(x_{1:n}) = f_{n:m+1|m:1}(x_{n:m+1}|x_{m:1}) \left(\prod_{k=1}^m f_k(x_k) \right) \cdot \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{j+i,j|j-1:1}(F_{j+i|j-1:1}, F_{j|j-1:1}) \quad (2.18)$$

Note that the distribution in higher dimensions $m+1, \dots, n$ is left undetermined for now. Through this decomposition it becomes possible to retain the separation between marginals and dependency modeling, while at the same time bivariate copulas can be chosen independently of each other, allowing the construction of a wide range of different dependency structures. It is constructive to note that there are a vast number of possible pair-copula constructions for a high-dimensional distribution. To this end, a graphical structure called regular vine structure (R-vine) was proposed to organize the decomposition. As two special and simplest regular vines, Canonical vines (C-vines) and D-vines, were introduced in [36], [39]. For C-vine, each tree has a unique node that is connected to all the other nodes, whereas each tree is a path in D-vine structure. In other words, C-vine has a more hierarchical structure than D-vine that leads to synergizing with the ordered variables according to their variances obtained via dimensionality reduction technique (e.g. PCA). In general, the C-vine and the D-vine representations for the density function $f(x_1, \dots, x_m)$ can be expressed as follows:

$$f(x_1, \dots, x_m) = \left(\prod_{k=1}^m f_k(x_k) \right) \times \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{j,j+i|\omega_{i,j}}(F_{j|\omega_{i,j}}, F_{j+i|\omega_{i,j}}) \quad (C - vine) \quad (2.19)$$

$$f(x_1, \dots, x_m) = \left(\prod_{k=1}^m f_k(x_k) \right) \times \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{i,i+j|v_{i,j}}(F_{i|v_{i,j}}, F_{i+j|v_{i,j}}) \quad (D - vine) \quad (2.20)$$

where $\omega_{i,j} = \{1, \dots, j-1\}$ and $v_{i,j} = \{i+1, \dots, i+j-1\}$. Graphically, both of the C-vine and D-vine constructions can be represented as a sequence of dependency trees $T = [T_1, \dots, T_{d-1}]$. Examples of C-vine and D-vine for $d = 4$ are shown in the Figure 2-12.

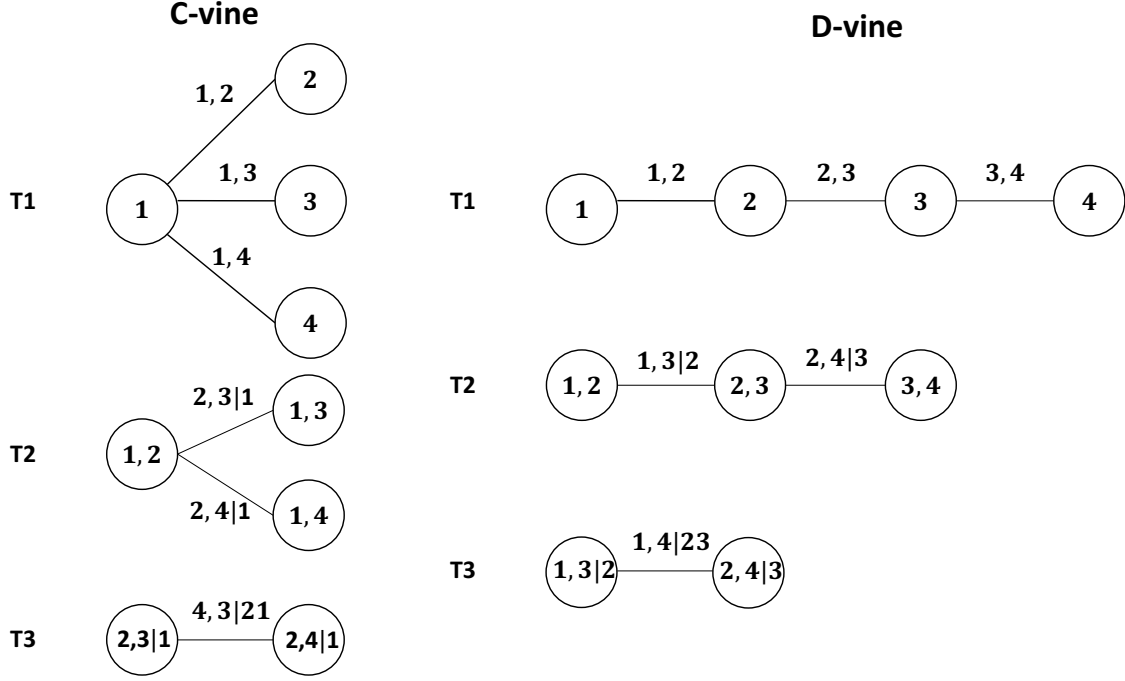


Figure 2-12. Example C-vine (left) and D-vine (right) structures for $d = 4$.

For either the C-vine or the D-vine, there are totally $d(d-1)/2$ edges, where d is the number of variables, in the whole structure. Each edge corresponds to a pair-copula density function fitted to the connected nodes. Another important procedure of constructing C-vine or D-vine is to determine the order of the variables when fitting C-vine and D-vine copulas. In details, the root nodes should be selected for each tree for C-vine and for D-vine, the order of the variables in T_1 needs to be determined as well. For convenience, we will use the concept of h -functions introduced in [37] to denote conditional distribution functions as in

$$F_{j+i|j:1}(x_{j+i}|x_{j:1}) = \int_{-\infty}^{x_{j+i}} \frac{f_{j+i,j|j-1:1}(x', x_j | x_{j-1:1})}{f_{j|j-1:1}(x_j | x_{j-1:1})} dx'$$

$$\equiv h_{j+i,j}(F_{j+i|j-1:1}(x_{j+i}|x_{j-1:1}), F_{j|j-1:1}(x_j | x_{j-1:1})), \quad (2.21)$$

with

$$h_{j+i,j}(u, v) = \int_0^u c_{j+i,j|j-1:1}(w, v) dw. \quad (2.22)$$

The h -functions therefore define the recursive relation between the conditioned unit random variables as

$$U_{j+i|j:1} = h_{j+i,j}(U_{j+i|j-1:1}, U_{j|j-1:1}). \quad (2.23)$$

Note that we focus on describing the parameterization, goodness-of-fit, and sampling procedures of the C-vine structure in the following part of this chapter.

2.3.7 C-Vine Model Parameterization

Returning to model construction, a C-vine copula model is to be constructed for the first m dimensions of the set of observations $\mathcal{X}$, recalling that this procedure is executed in an identical fashion for each cluster. In the following, we index the elements $x^{(s)}$ of $\mathcal{X}$ by the index $s \in \{1, \dots, N^c\}$.

As far as the univariate marginal densities are concerned, in general their complex nature does not permit approximation by parametric functions. Instead, for each variable index i we independently transform the empirical data to the unit domain using the empirical cumulative distribution functions F_i^e . The observations are transformed as $u_i^{(s)} = F_i^e(x_i^{(s)})$, resulting in a set $\mathcal{U}$ of points in the multivariate unit domain $[0,1]^n$.

In this domain we fit a sequence of bivariate copulas $\{\hat{c}_{j+i,j|j-1:1}(\cdot); 1 \leq j < j+i \leq m\}$ to $\mathcal{U}$. We restrict the fitting procedure to parametric copula functions that are defined by their family and family-dependent parameters. For example, the Clayton copula admits a single parameter and is used to model variables that exhibit lower-tail but no upper-tail dependency whereas the Student-t copula is an elliptical copula that admits two parameters and has both symmetric upside and downside dependencies.

Whereas the simple pairwise copulas $c_{i,1}(\cdot)$ in tree T_1 are computed with respect to the components $u_i^{(s)}$, the remaining copulas describe dependencies between variables $\hat{u}_{j+i|j:1}^{(s)} = \hat{F}_{j+i|j:1}(x_{j+i}^{(s)} | x_{j:1}^{(s)})$ that have been conditioned on fitted copulas in underlying trees. Using the h -function notation these values are computed recursively as $\hat{u}_{j+i|j:1}^{(s)} = \hat{h}_{j+i,j}(\hat{u}_{j+i|j-1:1}^{(s)}, \hat{u}_{j|j-1:1}^{(s)})$. The $\hat{h}$ -function corresponding to the fitted

copula is fully determined by the family $\hat{\phi}_{j+i,j}$ and the family-dependent parameter(s) $\hat{\theta}_{j+i,j}$: $\hat{h}_{j+i,i}(u, v) \equiv h(u, v; \hat{\phi}_{j+i,j}, \hat{\theta}_{j+i,j})$. Expressions for common families are available in [37].

For a given choice of family ϕ , the optimal parameter value $\hat{\theta}$ for each pair-copula is determined using Maximum Likelihood Estimation (MLE):

$$\hat{\theta}_{j+i,j}(\phi) = \arg \max_{\theta} \left\{ \sum_{s=1}^{N^c} \ln c_{\phi}(u_{j+i|j-1:1}^{(s)}, u_{j|j-1:1}^{(s)}; \theta) \right\} \quad (2.24)$$

Note that MLE provides a consistent way for selecting the best parameter(s) for a chosen parametric family, but is not useful for determining which of the candidate families provides the best fit.

2.3.8 Assessing C-Vine Goodness-of-Fit

For the purpose of determining which of the candidate families best describes the dependence between two variables, a goodness-of-fit (GOF) test must be performed. Before introducing the algorithm, it is important to appreciate the fact that GOF-testing is a non-trivial task. Considering that all family and parameter choices propagate and influence subsequent parameterization choices due to the inherently recursive structure of a C-Vine, it becomes apparent that determining the best-fitting model can quickly become a highly-complex task. An exhaustive examination of all combinations of families and pair-copulas is the only way to identify and verify the optimum family selection. However, this means that $d^{(m(m-1)/2)}$ GOF tests have to be performed, where d is the number of candidate copula families. It is evident that this approach can quickly become intractable and is not suitable for practical applications on high-dimensional data sets. A pragmatic alternative approach is to assume independence between our choices so as to reduce the number of combinations to be examined. In this step-by-step approach we perform d GOF tests to determine the best-fitting family for each pair-copula in the C-Vine structure. The number of GOF tests performed in this method $dm(m-1)/2$ scales quadratically with the number of dimensions m and linearly with the number of families d , making it a suitable compromise for high-dimensional applications.

We make use of the GOF test for bivariate copulas introduced in [37] that relies on a basic property of the C-vine construction. The copulas $c_{j+i,j|j-1:1}(\cdot)$ are constructed as part of tree T_j and are used to compute the conditioned input values $u_{j+i|j:1}^{(s)}$ for tree T_{j+1} . By construction, the resulting random variable $U_{j+i|j:1}$ (i.e. the set of equiprobable states indexed by s) should be independent of $U_{j|j-1:1}$. However, because the parametric copulas are typically not an exact match for the dependencies in tree T_j , the realized distributions $\hat{U}_{j+i|j:1}$ and $\hat{U}_{j|j-1:1}$ will have residual dependencies. A measure for the GOF of the bivariate copula is obtained by quantifying the independence of $\hat{U}_{j+i|j:1}$ and $\hat{U}_{j|j-1:1}$.

To this end we compute the realized values of the quantity

$$Z_{j+i,j} = \left(\Phi^{-1}(\hat{U}_{j+i|j:1}) \right)^2 + \left(\Phi^{-1}(\hat{U}_{j|j-1:1}) \right)^2, \quad (2.25)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard Gaussian distribution. If $\hat{U}_{j+i|j:1}$ and $\hat{U}_{j|j-1:1}$ are independent uniform random variables $Z_{j+i,j}$ should follow the chi-square distribution with 2 degrees of freedom. A p-value for the GOF can be computed using the Kolmogorov-Smirnov test. Within the C-Vine modelling process this test is performed for each family, and the family that results in the largest p-value is used for the copula in question.

The method described above can also be used to compute a single p-value for the entire C-Vine. The construction indicates that the following unit random variables (the nodes on the left of each tree) form a mutually independent set:

$$\begin{aligned} V_1 &= U_1 \\ V_2 &= U_{2|1} \\ &\vdots \\ V_m &= U_{m|m-1,\dots,1} \end{aligned} \quad (2.26)$$

Again, we establish a measure for the GOF by testing the independence of the fitted variables $\{\hat{V}_i\}_{i=1}^m$. If they are independent, the random variable $Z_m = \sum_{i=1}^m (\Phi^{-1}(\hat{V}_i))^2$ should be distributed according to the chi-square distribution with m degrees of freedom. A p-value for the goodness of fit may again be determined using the Kolmogorov-Smirnov test.

2.3.9 Sampling from a C-Vine model

Simulations from vines copulas were first briefly proposed in [36] and [39]. Then the explicitly programmable algorithms, which are used for sampling from C-vine and D-vine structures, were introduced in [37]. For the general regular vines (R-vines), the sampling algorithm was described in [40] and extended in [35]. In this project, we apply the sampling algorithm for C-vines. The method for sampling m dependent uniform variables on $[0,1]$ can be illustrated as below [37]:

C-Vine Sampling Algorithm

Step1. Sample m -dimensional independent uniform variables s_i on $[0,1]$ where $i = 1, \dots, m$;

Step2. Set the sampled dependent variables x_i as below:

$$\begin{aligned} x_1 &= F_1^{-1}(s_1) \\ x_2 &= F_{2|1}^{-1}(s_2|x_1) \\ &\vdots \\ x_i &= F_{i|1,\dots,i-1}^{-1}(s_i|x_1, \dots, x_{i-1}) \\ &\vdots \\ x_m &= F_{m|1,\dots,m-1}^{-1}(s_m|x_1, \dots, x_{m-1}) \end{aligned} \quad (2.27)$$

Given the hierarchical structure of a C-Vine, the conditional distribution function $F_{i|1,\dots,i-1}^{-1}(x_i|x_1, \dots, x_{i-1})$ can be expressed as:

$$F(x_i|x_1, \dots, x_{i-1}) = \frac{\partial C_{i,i-1|1,\dots,i-2}(F(x_i|x_1, \dots, x_{i-2}), F(x_{i-1}|x_1, \dots, x_{i-2}))}{\partial F(x_{i-1}|x_1, \dots, x_{i-2})} \quad (2.28)$$

The algorithm of the above procedure can be expressed as below:

C-Vine Sampling Algorithm

Sample s_i independent uniform on $[0,1]$ where $i = 1; \dots, m$

$$v_{1,1} = s_1$$

$$x_1 = F_1^{-1}(v_{1,1})$$

for $i \leftarrow 2, 3, \dots, m$

$$v_{i,1} = s_i$$

for $k \leftarrow i-1, i-2, \dots, 1$

```


$$v_{i,1} = h^{-1}(v_{i,1}, v_{k,k}, \Theta_{k,i-k})$$

end for


$$x_i = F_i^{-1}(v_{i,1})$$


if  $i == m$  then
    Stop
end if

for  $j \leftarrow 1, 2, \dots, i - 1$ 

$$v_{i,j+1} = h(v_{i,j}, v_{j,j}, \Theta_{j,i-j})$$

end for

end for

```

In the algorithm shown above, the outer for-loop runs over the number of variables n . This loop includes two further for-loops. The first one is used for sampling the i th variable and the other one is used for computing the conditional distribution functions which are required for sampling the $(i + 1)$ th variable. In the h function, $\Theta_{i,j}$ is the set of estimated parameters of the corresponding copula density $c_{j,j+i|1,\dots,j-1}(\cdot)$.

3. Evaluating Composite Approaches to Modelling High-Dimensional Stochastic Variables in Power Systems

3.1 Introduction

The large-scale integration of intermittent energy sources, the introduction of shiftable load elements and the growing interconnection that characterizes electricity systems worldwide have led to a significant increase of operational uncertainty. The vast number of stochastic variables beyond the operator's control are expanding the range of possible system states, thus rendering experience-based operational practices inefficient or even obsolete [3]. Under this new reality, the use of historical data will be instrumental in informing operators of possible system states likely to occur in the near-term and providing useful information regarding the risk the system is likely to face.

There are several advantages to having a stochastic model matched to historic observations instead of relying solely on past measurements. Most importantly, a parametric model is capable of producing samples that are similar but not identical to what has already been encountered, thus interpolating and extrapolating the historical dataset. In addition, the model parameters themselves can be useful in extracting properties of a multivariate dataset that could be too complex to identify through direct examination. Another practical advantage is not having to deal with missing data. Beyond all the above, statistical models can provide versatility when combined with other sampling or data-mining techniques. For example, the ability to generate a sample population of arbitrarily large size can be extremely useful in generating unseen training and testing data sets for machine learning algorithms to subsequently produce efficient security rules. Despite the numerous applications, there has been little progress in developing a unified stochastic modelling framework. Particularly in the case of high-dimensional datasets, there are four distinct issues that render this task highly challenging; the non-Gaussian marginal probability distribution of individual variables, the non-linear dependency structure, the large number of observations to be fit as well as the vast number of variables, resulting in increased computational complexity.

In recent years, copulas have been applied to capture the relation between different stochastic attributes in the context of electrical energy system operation. Their superiority lies in their ability to decouple marginal distribution from dependency

modelling, rendering them well-suited to capture statistical properties of datasets that do not belong to standard multivariate distributions. For example, in [41] a multivariate Gaussian copula has been applied to generate synthetic wind power output from 15 sites in the Netherlands. A statistical modelling framework based on a multivariate Gaussian copula is also proposed in [42] to obtain complete temporal dependence structure and actual distributions of wind speed time series. However, when moving to a high-dimensional setting, a single multivariate copula lacks the required flexibility for accurate dependence structure modelling. Different pair copula schemes have recently been proposed in the literature to address this issue. However, they entail significant computational load, and their computation can become intractable even for a moderate number of dimensions. The FP7 iTesla project has recently developed a composite modelling framework for high-dimensional variables, consisting of a combination of pair copula construction with dimensionality reduction and data clustering. In a forthcoming chapter, we explore the performance of a particular modelling scheme across datasets of different size and complexity. In addition, a case study demonstrated the method's clear advantages when applied to the construction of training-testing datasets for building proxies for a system's security boundary.

In the present chapter we investigate the performance of a number of alternative composite schemes. Quantitative comparisons of goodness-of-fit are used to identify synergies between different modelling stages with the aim of identifying well-performing composite modelling approaches. We mainly focus on the use of pair-copula construction schemes, namely C-vine and D-vine copulas models, which utilize cascades of two-dimensional copulas as building blocks for flexible high-dimensional models. Also, clustering techniques are used to create mixture models that achieve an overall better model fit. The effectiveness of different data clustering methods and relevant clustering validity indicators are examined. Finally, statistical models utilizing different dimensionality reduction techniques, are compared. Note that the proposed modelling approach is time-independent. The autocorrelations and temporal dependencies between variables of the historical population are captured in the sampled data set, but it is not possible to reference a particular sample in terms of time. On the other hand, spatial dependencies between variables are fully captured. Detailed

comparisons are drawn on the basis of Kolmogorov-Smirnov tests, multivariate two-sample energy tests and further supported by visual data comparisons. A large power system dataset is modelled using different combinations of the techniques above and the proposed composite modelling approaches are shown to considerably out-perform alternative methods. The computational performance of each modelling scheme is also discussed.

The outline of the remainder of this chapter is as follows: Chapter 3.2 explicitly introduces the structure of the proposed composite modelling framework and describes the main characteristics of stochastic data in the context of power systems. Chapter 3.3 introduces the model evaluation techniques used for performance comparison. Chapter 3.4 presents and compares the results of different modelling schemes when applied to historical data from the French power system. Chapter 3.5 demonstrates the use of the proposed statistical model in the calculation of a system's Power Not Supplied profile. Chapter 3.6 presents how to applied the proposed model to generate the database for offline training of decision trees (DTs). Finally, Chapter 3.7 summarizes and concludes the chapter.

3.2 Proposed Composite Modelling Framework

To illustrate some of the challenges related to modelling of power system stochastic variables such as active/reactive load and power injections, an example dataset is shown in Figure 3-1. It demonstrates that stochastic variables in power systems have highly non-standard marginal distributions (diagonal histograms) and non-linear dependencies (scatter plots) that limit the effectiveness of traditional parametric models. To tackle these challenges, a composite modelling approach has been developed in the iTesla project [43] that consists of K-means clustering method, Principal Components Analysis (PCA) and C-vine copulas. In the present chapter, another version of this modelling framework is introduced alternative clustering techniques (i.e. K-means, hierarchical clustering, and Gaussian mixture model clustering (GMM)), dimensionality reduction methods (PCA, Locality Preserving Projections (LPP)), and parametric modelling (C-vine and D-vine copulas) to identify well-performing combinations of modelling stages.

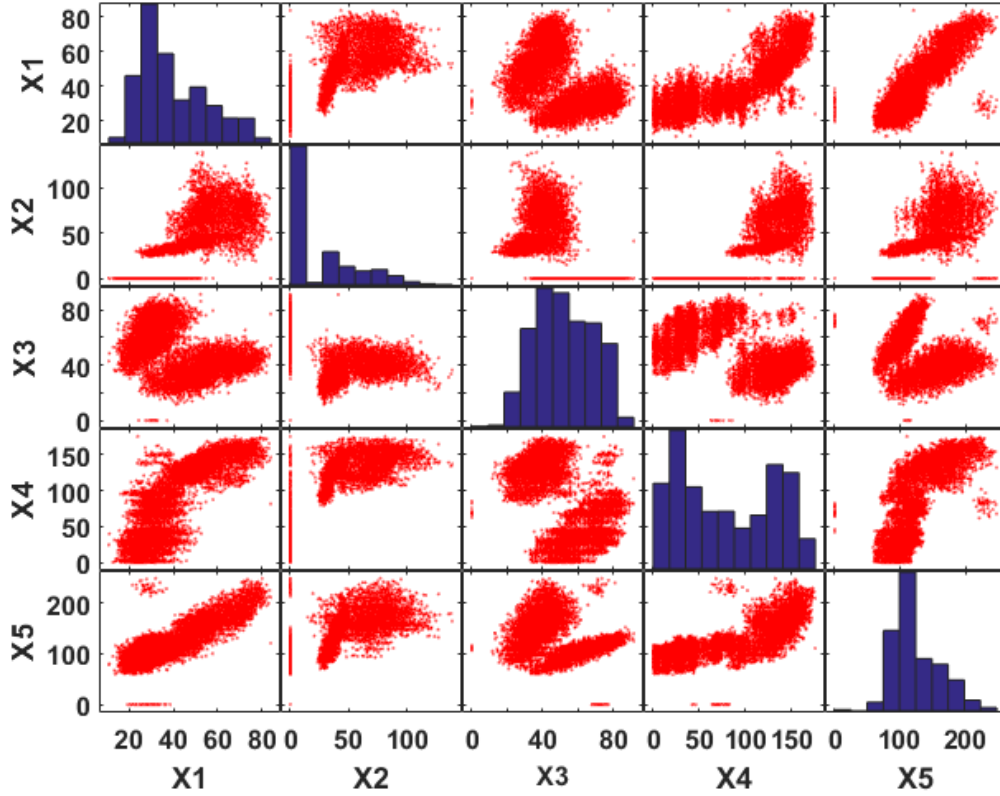


Figure 3-1. Histograms of marginals distributions and bivariate scatter plots of active loads (MW) for five randomly selected buses in France, measured between Jan 2012 and Feb 2012.

The high-level structure of the composite high-dimensional modelling framework is illustrated in Figure 3-2. In general, the whole workflow can be split in two stages;

a) **Modelling Stage:** Given the input historical data, construct the composite model via the *Clustering Module*, *Dimensionality Reduction Module*, and *Vine Copulas Module*;

b) **Sampling Stage:** Generate samples through the *Vine Copulas Simulation Module* and the *Reconstruction Module*.

In particular, the aim of each module is as follows.

1) Clustering Module: Let d denote the total number of inter-dependent injections and loads in the power system to be modelled, the whole set of historical data set consists of N measurements and is denoted as $\mathcal{Z} = [z_1, \dots, z_N]^T \in \mathbb{R}^{N \times d}$. Clustering is performed to partition the observations into K clusters, where $\mathcal{Z}^k \subset \mathcal{Z}$, for $k = 1, \dots, K$. The weight of each cluster is calculated as the share of observations classified in cluster k .

- 2) **Dimensionality Reduction Module:** This step can be considered as a pre-processing procedure principally aimed at decreasing the computational cost of the following *Vine Couplas Module*. For each cluster k , we first need to transform the observations from the original domain to the rank-uniform domain via the corresponding empirical cumulative distribution functions (ECDF). As such, each clustered dataset $\mathcal{Z}^k$ is transformed to $\mathcal{Y}^k$ in the $[0,1]^d$ domain. Carrying out this transformation step is highly recommended as dimension reduction techniques are typically highly sensitive to data skewness. Let $q < d$ define the reduced target number of variables. Dimension reduction is performed on each $\mathcal{Y}^k$ to get the low-dimensional data set $\mathcal{X}^k \in \mathbb{R}^{N^k \times q}$, where N^k represents the number of observations in cluster k . This module also outputs the corresponding eigenvectors to be used as an input in the *Reconstruction Module*.
- 3) **Vine Copulas Module:** For vine copulas, the parametric model should be fitted to the data set in the uniform domain. Therefore, the first step of this module is performing the uniform transformation through the ECDFs of $\mathcal{X}^k$ to obtain $\mathcal{U}^k$. By using the vine copula modelling technique, described in the next section, a different parametric model is constructed for each cluster k .
- 4) **Vine Copulas Simulation Module:** Given the target number of samples N_S , each parametric cluster-model generates samples $\hat{\mathcal{U}}^k$ of size $N_S^k \times q$, where $N_S^k = N_S \times W^k$, and W^k is the weight of cluster k .
- 5) **Reconstruction Module:** The goal of this module is to transform the generated samples back to the original domain. Firstly, samples of each cluster $\hat{\mathcal{U}}^k$ are transformed back to the domain of $\mathcal{X}^k$; we obtain $\hat{\mathcal{X}}^k$ of size $N_S^k \times q$, by transforming $\hat{\mathcal{U}}^k$ through the inverse empirical distribution functions ($ECDF^{-1}$) of the original dataset $\mathcal{U}^k$. Subsequently, $\hat{\mathcal{X}}^k$ is back-projected to the original high-dimensional space $\mathbb{R}^{N_S^k \times d}$, denoted by the data set $\hat{\mathcal{Y}}^k$. Note that in this step it is possible to add some suitable noise to specific variables which may have exhibited low variability in the original dataset and are thus reconstructed as largely stationary. Finally, each $\hat{\mathcal{Y}}^k$ is transformed via the $ECDF^{-1}$ of original data set $\mathcal{Z}^k$.

Let $\hat{\mathcal{Z}}^k$ be the sampled data for the cluster k ; the final high-dimensional sampled data set is $\hat{\mathcal{Z}} = \{\hat{\mathcal{Z}}^1, \dots, \hat{\mathcal{Z}}^K\} \in \mathbb{R}^{N_S \times d}$.

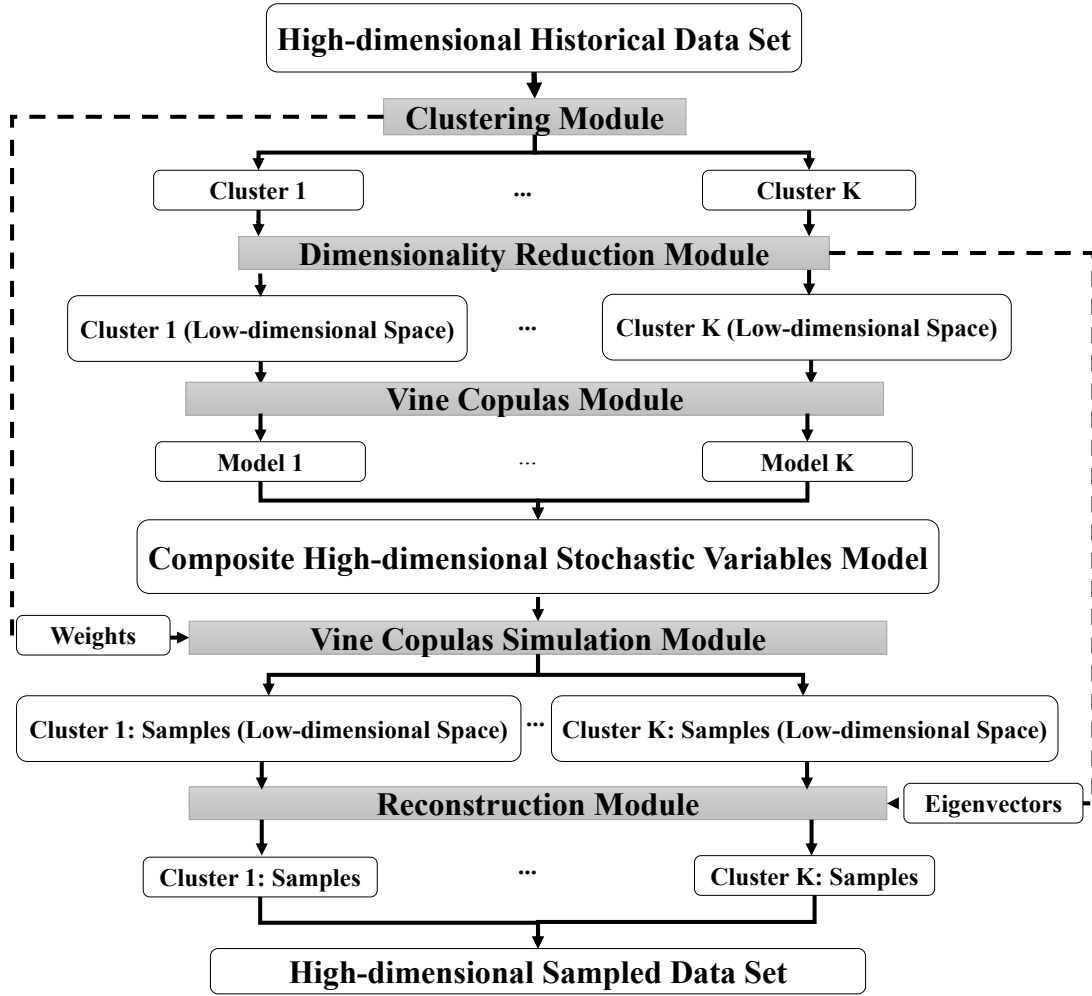


Figure 3-2. The structure of composite high-dimensional modelling framework.

3.3 Model Evaluation Techniques

In the context of high-dimensional data modelling, another problem that constitutes an open research question is quantifying the modelling accuracy and determining whether a particular data model successfully captures the main characteristics of the data being modelled. As a statistical analysis method, two-sample hypothesis tests are designed to determine if two samples are drawn from same distributions. In order to evaluate the performance of each proposed modelling schemes, test statistics of interest can be calculated between the historical data and the samples that are generated with different combinations of the introduced data mining techniques and vine copulas. In this chapter, we investigate two types of two-sample tests: two-sample Kolmogorov-

Smirnov tests and multivariate two-sample energy tests, to examine the marginal distributions individually as well as the entire multidimensional structure, respectively.

3.3.1 Two-sample Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov test (two-sample K-S test) [44] is a popular non-parametric test used to determine whether two underlying one-dimensional sample population have been drawn from the same probability distribution. Given historical data set $\mathcal{Z} = [z_1, \dots, z_d] \in \mathbb{R}^{N \times d}$ and its corresponding samples $\hat{\mathcal{Z}} = [\hat{z}_1, \dots, \hat{z}_d] \in \mathbb{R}^{N_S \times d}$, the two-sample K-S test is performed between each pair of margins under the null hypothesis that each pair of z_i and $\hat{z}_i$, for $i = 1, \dots, d$, are from the same distribution. The Kolmogorov-Smirnov statistic D_{KS} is given by:

$$D_{KS}^i = \sup_x |F_i(x) - \hat{F}_i(x)|, \quad (3.1)$$

where $F_i(x)$ and $\hat{F}_i(x)$ represent the empirical distribution functions of z_i and $\hat{z}_i$. Subsequently, a p-value can be obtained on the basis of this statistic.

3.3.2 Multivariate Two-sample Energy Test

Two-sample energy test [45] is a powerful non-parametric multivariate two-sample test based on a statistical quantity defined as energy to determine whether two multivariate samples are from the same distributions. The test statistic D is given by:

$$D_E = \frac{N \times N_S}{N + N_S} \left(\frac{2}{N \times N_S} \sum_{i=1}^N \sum_{j=1}^{N_S} \|z_i - \hat{z}_j\| - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \|z_i - z_j\| - \frac{1}{N_S^2} \sum_{i=1}^{N_S} \sum_{j=1}^{N_S} \|\hat{z}_i - \hat{z}_j\| \right) \quad (3.2)$$

where $\|\cdot\|$ denotes the Euclidean norm. The statistic D_E and its p-value are computed using the R package *energy*.

3.4 Case-study: Evaluation Based on Real Data

In this section, the performance of the proposed composite modelling schemes is illustrated with a large library of historical measurements of load that was provided by RTE, the French Transmission System Operator. The original library includes 14251 measurements at 15-minute intervals over the course of two months, from January to February 2012, and spans over 7,000 load buses and 200 wind plants. In this chapter

we focus on a subset; a 100-dimensional dataset consisting of 90 demand buses and 10 wind generators chosen at random. This dataset was used as the basis for comparing between different statistical modelling techniques.

3.4.1 Clustering Validity Assessment

The first step is to compare the performance of the clustering techniques via a set of clustering validity indicators. The Cluster Dispersion Indicator (CDI) and the Mean Index Adequacy (MIA) are indicators that have been presented in detail in [46] and used in this research; better clustering quality corresponds to lower indicator values. The clustering methods examined in this section are K-means, hierarchical clustering (average linkage), hierarchical clustering (Ward linkage) and GMM. Most of the clustering techniques need a pre-specified number of clusters. Figure 3-3 shows the results of adequacy comparisons among the tested clustering techniques by using the CDI and MIA validity indicators for numbers of clusters K from 2 to 20.

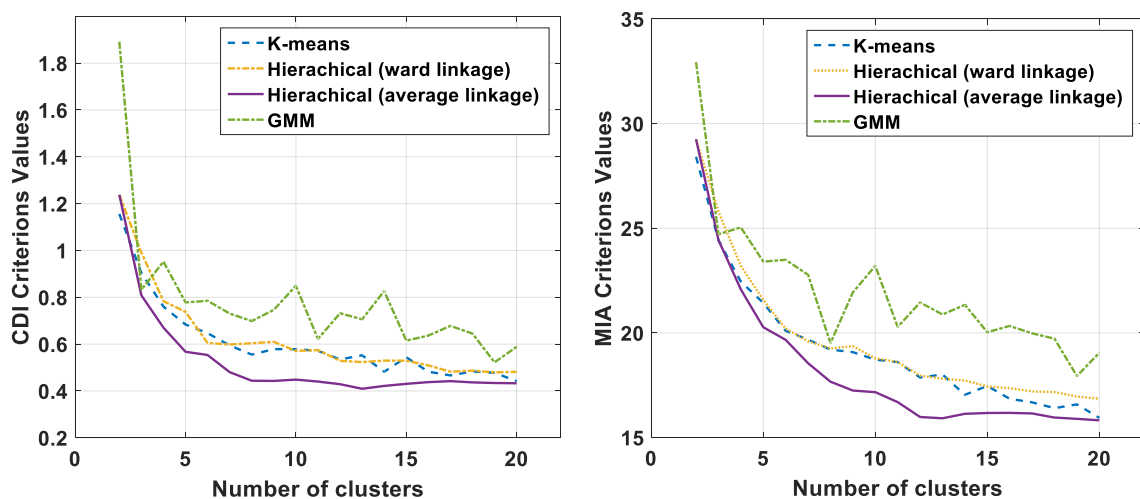


Figure 3-3. Comparisons among the clustering techniques by using (left) the CDI indicator and (right) the MIA indicator.

As discussed in [46], the optimal number of clusters corresponds to the knee of each curve. From Figure 3-3, it is clear that most of the validation indicators imply that the optimal number of clusters is equal to three ($K_{opt} = 3$) across all methods. Regarding the clustering method choice, as can be seen in Figure 3-3, the hierarchical clustering with the average distance linkage criterion exhibits the lower CDI and MIA criterion values, suggesting it is the most suitable clustering technique for the examined dataset. Accordingly, K is set to 3.

The next step is to generate a large sample population ($N_s=50,000$) from the historical dataset of 14,251 100-dimensional observations. Regarding the Dimensionality Reduction Module, the tested methods were PCA and LPP. The input parameter of this module is a metric, named information retainment (IR), which is the sum of eigenvalues of the retained PCs divided by the total eigenvalue sum of the non-reduced dataset. This metric was set to 97.5% and the average reduced dimensions for each cluster are 56 and 70 for PCA and LPP, respectively. In the Vine Copulas Module and the Vine Copulas Simulation Module, C- and D-vine copulas are implemented with bivariate copula families Gaussian, Student-t, Clayton, Frank, Gumbel, as well as their 90°, 180° and 270° rotated versions. In addition, we examine two conventional modelling techniques: multivariate Gaussian distribution (MGD) and multivariate Gaussian copula (MGC). All the methods were implemented in MATLAB and run on an Intel Xeon E5-2690 PC with 8 cores. CPU times for these six methods are shown in Table I.

Table 3-1. Computation Times for Different Methods

Modelling Methods	CPU Times (s)
Hierarchical(average linkage)+PCA+C-vine	3595.07
Hierarchical(average linkage)+PCA+D-vine	3602.12
Hierarchical(average linkage)+LPP+C-vine	4100.23
Hierarchical(average linkage)+LPP+D-vine	4125.15
MGD	3.25
MGC	5.62

As shown above, the MGD and MGC methods are far less computationally expensive but as will be shown later, underperform in terms of modelling accuracy. On the other hand, there is little difference between the four composite methods; most CPU time is expended in pair-copula fitting; 30,800 (20 candidate families across 1540 pair copulas) copulas have been fit in the case of PCA-based methods, while 43800 copulas are fit in the LPP case. Note that the load can be significantly reduced by choosing a lower IR value.

3.4.2 Visual Comparison

Considering the difficulties in fully visualizing such a high-dimensional data set, in Figure 3-4 we compare historical and sampled data generated by the six different

methods via the scatter plot of a single load variable and a single wind infeed variable chosen at random. In Figure 3-4 (a), the point-cloud of historical data is shown to be composed of several obvious sub-groups of highly non-linear dependence and complex data artifacts. As expected, the superiority of the proposed composite modelling schemes, shown in Figure 3-4 (b)-(e), can be observed in their reconstructed dependency structures. Among these schemes, the method with LPP performs relatively better than the methods with PCA. For the methods with C-vine and D-vine copulas, no significant differences can be found by this visual comparison. In Figure 3-4 (f), the MGD method is shown to have severely limited modelling capabilities due to the inherent constraints imposed on the marginal distributions and dependency structure. Also, as shown in Figure 3-4 (g), the MGC model cannot well accommodate the signal characteristics as a result of its limited type of dependence modelling.

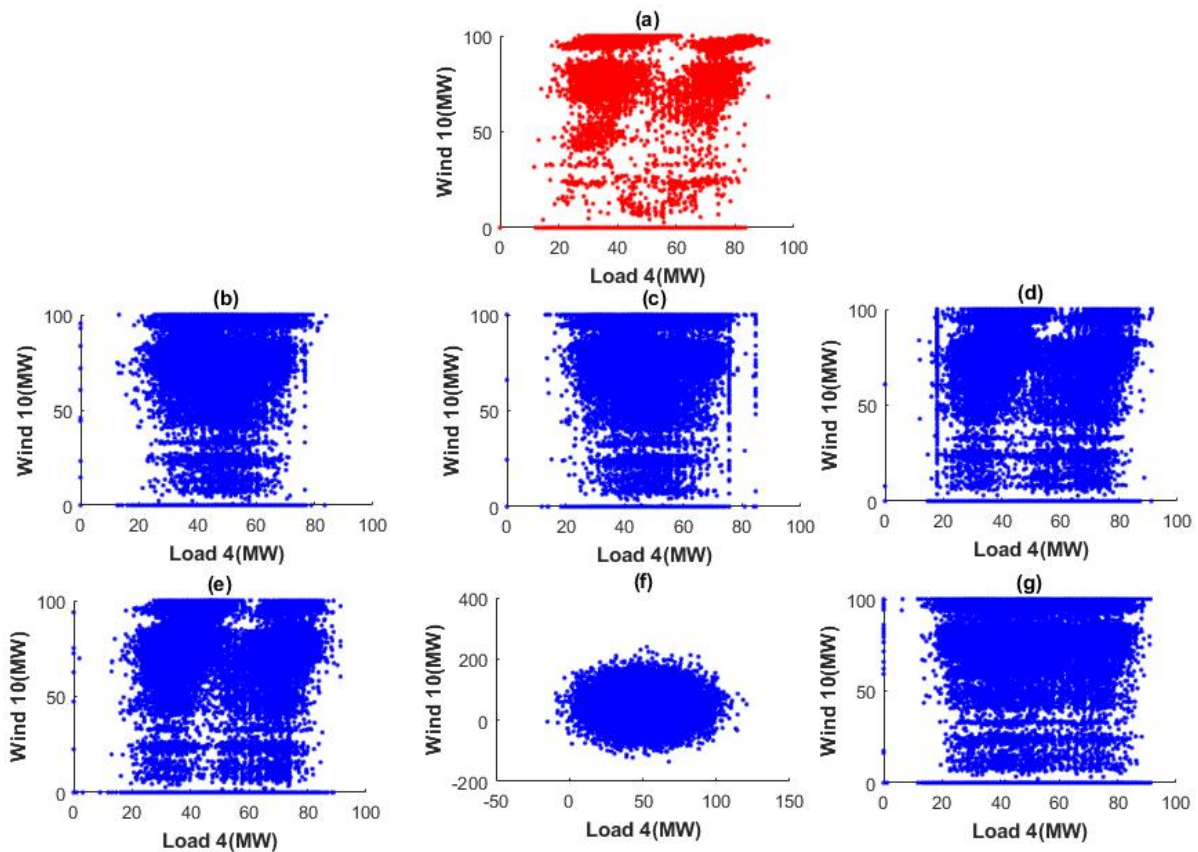


Figure 3-4. Scatter plots of historical (a) and data generated using Hierarchical(average linkage)+PCA+C-vine (b), Hierarchical (average linkage)+PCA+D-vine (c), Hierarchical (average linkage)+LPP+C-vine (d), Hierarchical (average linkage)+LPP+D-vine(e), MGD (f), and MGC (g) .

3.4.3 Two-sample Kolmogorov-Smirnov Test

In order to evaluate the performance of each modelling scheme on the basis of two-sample tests, a resampling method is employed to randomly generate comparison samples from the historical and simulated datasets. Sets of 100 (historical) and 200 (simulations) points were drawn at random, and the process was repeated 1000 times for each of the 100 variables. In this way, the distribution of test results can be obtained to increase the effectiveness of our analysis. For the null hypothesis to hold true (in this case meaning that historical and sampled data have been drawn from the same distribution), the p-values must be uniformly distributed. Recall that modelling high-dimensional stochastic signal has two main challenges of non-Gaussian marginal probability distributions and complex dependence structure. In terms of the first issue, the capability of reconstructing the original marginal distribution is illustrated in Figure 3-5 by comparing the marginal distributions of sampled data set to the corresponding historical margins, for each modelling method, via the two-sample K-S test. Note that the total number of the K-S tests performed for all variable pairs is 10^5 (100 dimensions times 1000 samples).

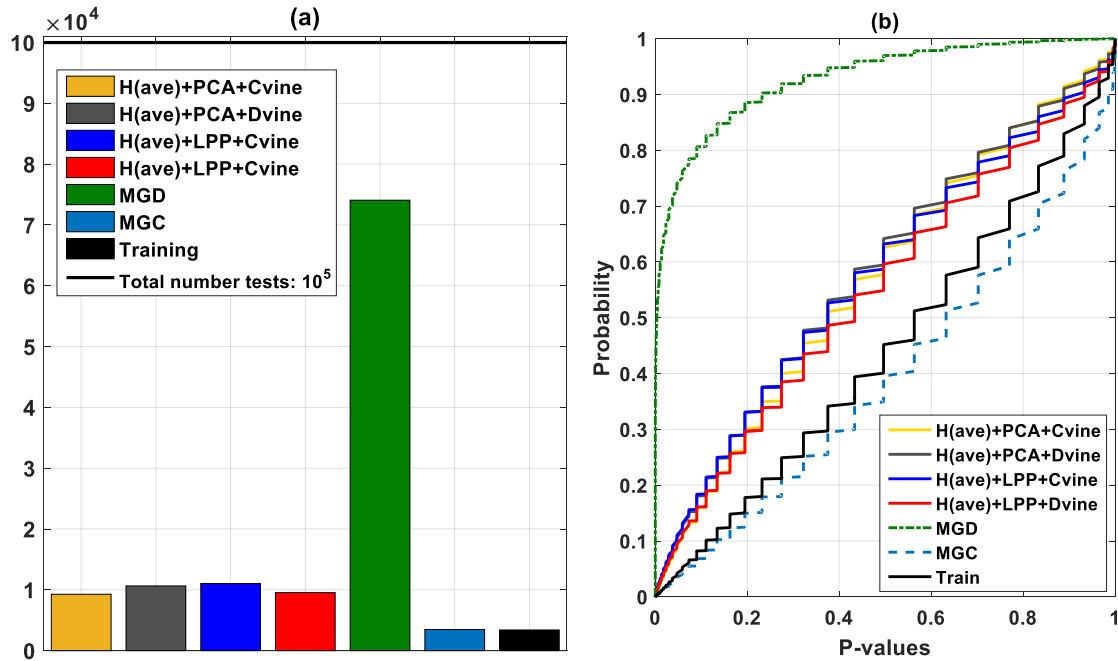


Figure 3-5. Results of the two-sample K-S tests: Number of rejected tests for each method (a) and ecdfs plot of test p-values for each method (b).

As can be seen in Figure 3-5 (a), which shows the total number of rejected sampling marginal distributions (i.e. p-value of K-S test less than 0.05), MGC performs best among all the methods because there is no information loss arising from the dimensionality reduction procedure. However, the proposed modelling schemes with PCA, LPP as well as C- and D-vine copulas (Methods 1-4) also exhibit low number of rejections (i.e. average about 10% in total tests). Among the proposed modelling schemes, the one combined with LPP and D-vine copulas has a slightly better performance than the other. In contrast, MGD is shown to achieve the worst performance (i.e. nearly 75% rejections), reflecting the non-normal margins of the original stochastic variables. The same pattern can be observed from the ECDFs of p-values of the K-S tests, as shown in Figure 3-5 (b). The marginal modelling capability of the examined composite modelling approaches is very close to the MGC which does not suffer from information reduction, while the traditional MGD apparently suffers from the imposed Gaussian marginals. The p-values of the two-sample K-S tests are almost perfectly uniformly distributed when the historical dataset is compared to itself (i.e. training dataset).

3.4.4 Multivariate Two-sample Energy Test

In this section, the results of multivariate two-sample energy tests are used to examine how the original dependency structure and marginal distributions are captured across different methods. The null hypothesis under investigation is whether historical and sampled data have been drawn from the same multivariate distribution. As before, 1000 random datasets have been generated and the distribution of p-values is computed in each case. A total of 1000 energy tests are carried out for each method. For the null hypothesis to hold, the p-values should be uniformly distributed, essentially indicating that the two datasets being compared are statistically indistinguishable.

As expected, the composite modelling schemes exhibit considerably more powerful ability to capture the original dependence patterns when compared to MGC and MGD. From Figure 3-6 (a) and (b), it can be observed that the results of composite modelling schemes show higher average p-values, and the distributions of their p-values are much closer to a uniform distribution. The test results of the schemes with C- and D-vine copulas do not exhibit a big difference when combined with PCA. However, the

combination of LPP and D-vine copulas results in the best performance overall. Although intuitively one would expect the variable ordering emerging from dimensionality reduction to synergize with the hierarchical structure of the C-vine, this is not the case. We suspect this is caused by two factors: 1) because high-variance variables are prioritized and detecting a well-fitting copula becomes a very challenging task; 2) model inaccuracies due to ill-fitting pair copulas are propagated throughout the vine structure. In contrast, the increase in modelling accuracy suggests that the D-vine may be a more robust to ill-fitting pair copulas and error propagation.

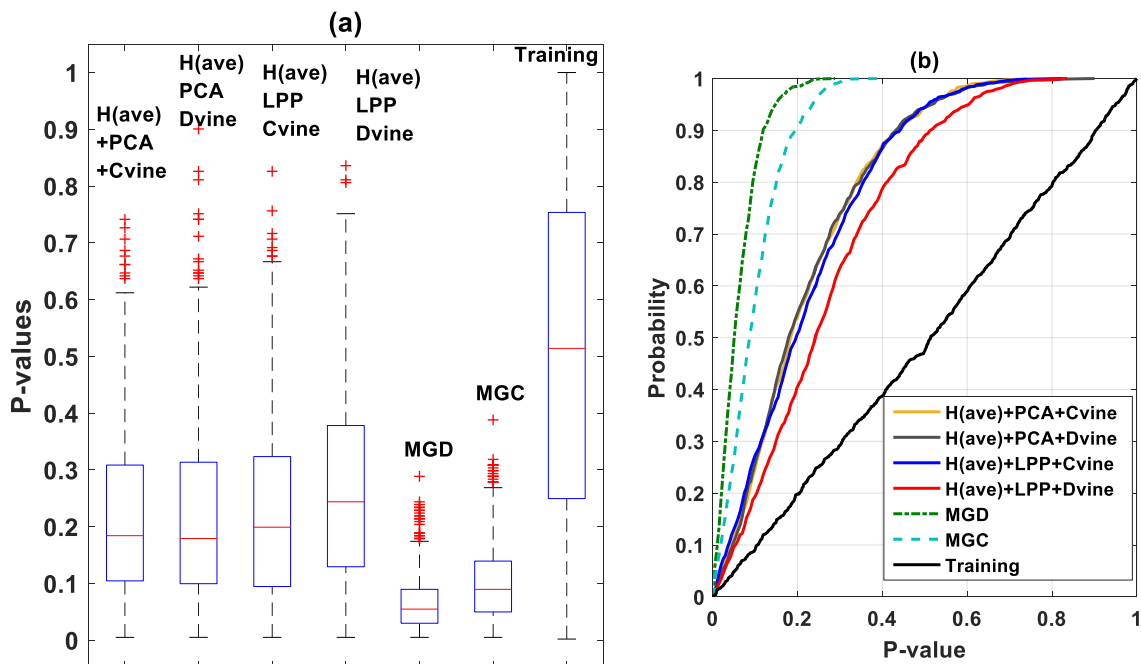


Figure 3-6. Results of the multivariate two-sample energy test: boxplot of test p-values for each method (a), ECDF plot of test p-values for each method (b).

3.5 Case-Study: PNS Calculation in IEEE 118-BUS System

In this chapter we provide an example use of the proposed modeling approach with the combination of C-Vine copulas, PCA, and K-means, applied to quantifying the risk profile of a large energy system. We are concerned solely with steady-state operation and more specifically with the loss of load that can occur following a generator outage due to plant inadequacy or due to transmission constraints. For this illustrative example we compute the Power Not Supplied (unmet demand) as a result of a single generator outage. One common approach for estimating such metrics is to perform system simulations over a population of historical operating points. In the presented study, the proposed high-dimensional sampling algorithm is used to generate new samples of

loads and wind infeeds. The aim is to demonstrate how the latter approach enables us to explore a larger area of the stochastic state-space than if relying solely on historical observations. This way we can carry out a more thorough analysis of the system's configuration space by interpolating the historical dataset and thus including stochastic realizations that may have not been encountered before, but may occur in the future.

3.5.1 Historical and Sampled Data

To demonstrate the aforementioned concept, we use a modified version of the IEEE 118-bus system, as shown. The original system consisting of 54 generators and 186 lines, as shown in Figure 3-7, has been modified to also include 10 wind farms of size 100MW.

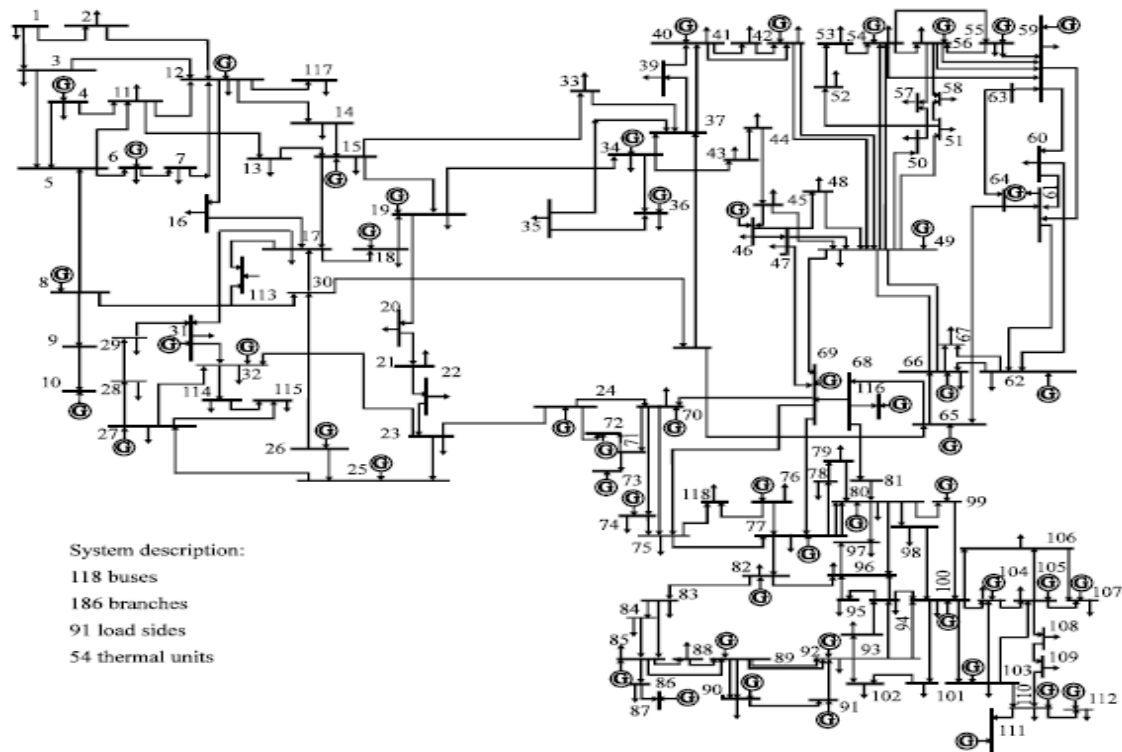


Figure 3-7. IEEE 118-Bus System

For the historical database of stochastic variables, it is worth noting that synthetic data (e.g. drawn from a parametric joint distribution function) exhibit a simple dependence structure and constitute an inherently good fit for parametric functions, rendering them of inadequate complexity. For this reason, a large library of historical anonymized measurements of load and wind power injections was provided by RTE, the French Transmission System Operator (TSO). The library consists of measurements

at 15-minute intervals over the course of January-March 2012 and spans over 7,000 load buses and 200 wind plants. This data can be ‘mapped’ onto the test system, creating a library of 14,251 historical observations. More specifically, 118 demand buses and 10 wind generators from the area of Nancy, France have been chosen at random and each variable scaled according to the maximum value defined in the coincident peak snapshot, while ensuring that the complex dependence structure of stochastic variables remains intact.

The first step is to use the proposed model to generate 10^5 new samples from the historical dataset of 14,251 128-dimensional observations. The number of clusters k was set to 50. A 97.5% information retainment threshold was used to define the choice of m for each cluster. For example, the cluster highlighted red in Fig. 4(a) consists of 315 observations and was modeled with $m = 33$, resulting in a truncated C-vine model of 528 bivariate copulas and a single Gaussian copula. The bivariate copula families modelled are Clayton, Frank, Gaussian, Gumbel, Student-t as well as their 90° , 180° and 270° rotated versions. The algorithm was implemented in MATLAB and run on an Intel Xeon PC with 8 cores. The overall C-Vine model parameterization and sampling was carried out in parallel for each cluster and took 34 minutes. Detailed CPU times are shown in Table 3-2.

Table 3-2. Computation times for parameterization and sampling

<i>Stage</i>	<i>CPU Time (s)</i>
Data Clustering	1.95
PCA	0.02
C-Vine Parameterization	2,051.45
Gaussian Copula Parameterization	2.56
Total time for model construction	2,055.98
Time to generate a single realization	0.02

It is not possible to fully visualize a multivariate dataset in a single graph. Thus, in Figure 3-8 (a) and (b) we present the scatter plot of total wind infeed vs. total load to compare between historical and generated data. Visual comparison of the plots confirms that all major signal characteristics are well captured with the proposed model. Furthermore, as expected, the point-cloud of sampled data covers a larger area, reflecting the enhanced state-space exploration aspect of the presented methodology. In

order to highlight the differences between the proposed method and other alternatives we focus on a single cluster of the historical data highlighted red in Figure 3-8 (a) and shown in detail in Figure 3-8 (c). In Figure 3-8 (d)-(f) we present the sampled data obtained when modelling the data dependency structure using a joint Gaussian distribution, a multivariate Gaussian copula and the truncated C-Vine method.

The joint distribution method is shown to have severely limited modeling capabilities due to the inherent constraints imposed on the marginal distributions and dependency structure. The benefit of decoupling dependency and marginal distribution modelling is evident in the considerably improved performance of the two copula methods. However, the multivariate copula model is not flexible enough to accommodate all signal characteristics, particularly in the tails. The truncated C-Vine approach is visibly more successful in this task.

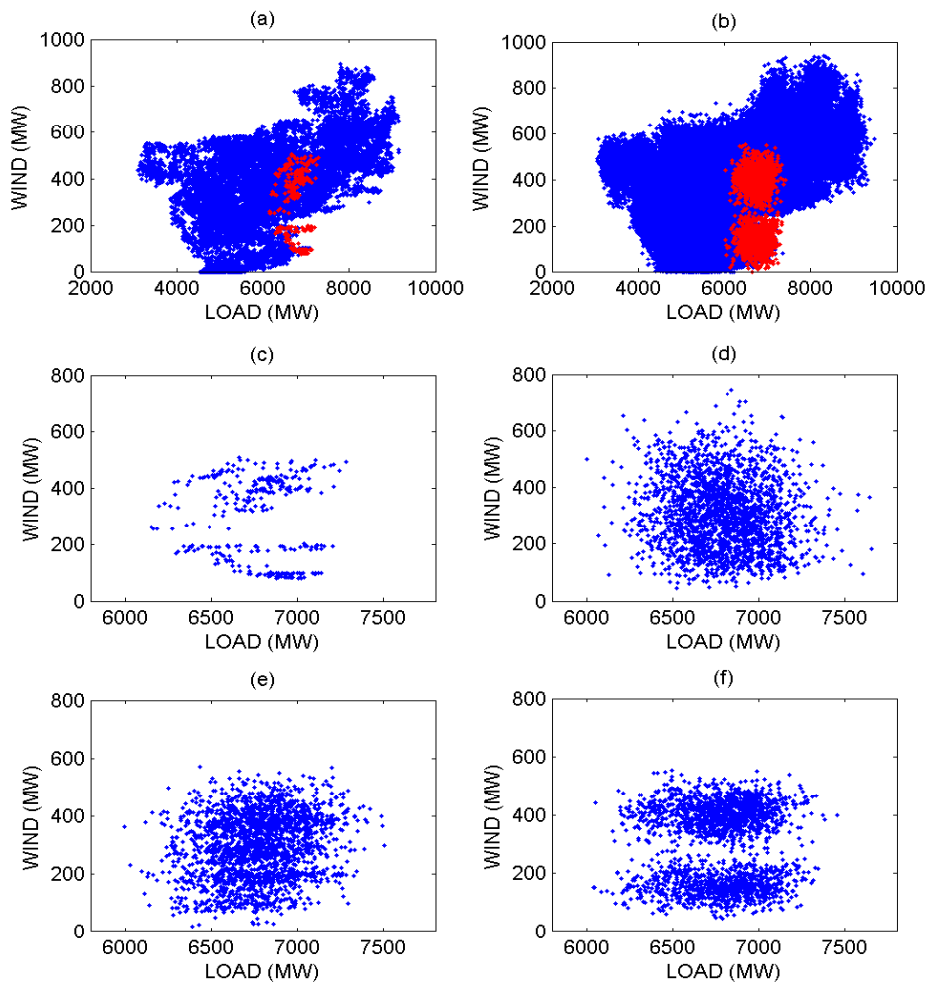


Figure 3-8. Scatter plots of historical (a) and data generated using the proposed method (b) ; target cluster is in red. Historical data in target cluster are shown in isolation in (c). Scatter plots of generated data using a joint Gaussian distribution (d), multidimensional Gaussian copula (e) and the proposed truncated C-Vine method (f).

3.5.2 Simulation Results

Given N operating points we are able to construct the system's Power Not Supplied (PNS) profile, conditional on a generator outage by simulating post-fault performance over a set of N_C credible contingencies. For this study, we only consider individual failures of the 10 largest generators and assume they are equiprobable; a total of 10^6 power flow simulations are run.

The PNS profiles obtained from historical and sampled data are shown in Figure 3-9 as complementary cumulative distribution functions. Note that we are plotting the 'tail' of the risk profile, excluding operating points that do not give rise to curtailed energy under any contingency. As can be seen below, the distribution functions of PNS obtained from the post-fault simulations have some important differences.

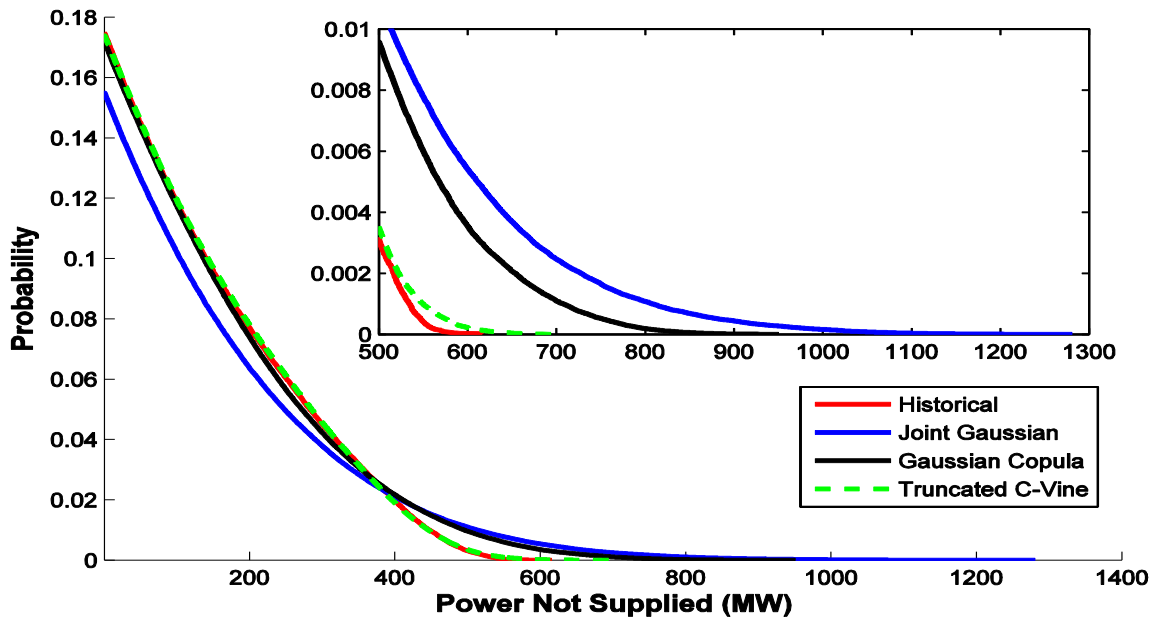


Figure 3-9. PNS profiles obtained using historical and generated data using different sampling methods. Inset focuses on tail of distributions.

The main observation is that the PNS curve resulting from the C-Vine method resembles very closely that obtained using the historical data. This is as expected, since we are interpolating between historical observations and generating thousands of unseen realizations that still follow the historical joint distribution. Regarding the multivariate Gaussian copula method, although the curve closely matches the historical in the first part, the latter part of the curve has a radically different shape with a significant tail overshoot. In the case of the joint Gaussian method, the overshoot is

even more pronounced. This highlights the shortcomings of other approaches when it comes to high-dimensional applications; methods that fit generic multivariate parametric distributions have limited accuracy when it comes to modelling tail-to-tail data dependencies.

3.6 Case-Study: System Security Assessment

In this subchapter, we provide an example use of the proposed modeling approach, applied to the database generation for offline training of decision trees for the purpose of system security assessment. In particular, we predict following the outage of the four tie lines connecting the west and east sections, namely lines 30, 44, 45, and 54 as denoted in Table 3-3.

Table 3-3. Selected Outage Lines for the IEEE 118-bus System

<i>No.</i>	<i>Lines</i>
30	(23 - 24)
44	(15 - 33)
45	(19 - 34)
54	(30 - 38)

In this case, the number of clusters k was set to 10, and an 85% information retainment threshold was used to define the choice of m for each cluster. A decision tree (DT) model is trained on the optimal power infeeds on each bus obtained by solving the pre-fault optimal dispatching problem. For the purpose of evaluating the constructed DT model, ten-fold cross-validation is used to get an estimate of classification accuracy. To this end, the original optimal power infeeds data set is randomly partitioned into 10 test datasets. Each test data set accounts for 10% of the total operating points, and the remaining 90% scenarios are attributed to the training data set. For each fold, the DT model is trained using one of the training data sets and then tested with the corresponding test data set. The security status of each operating scenario, defined as a binary variable (i.e. 1-safe, 0-unsafe), can be estimated by simulating the post-fault operation with line outage following the $N-1$ contingency criterion. Specifically, the unsafe operating point refers to its simulation result with unserved demand, and the safe status is denoted by the fully served operating point. By this way, each scenario is assigned with 0 or 1 for both test and training datasets. For each candidate outage line, No. 30, No. 44, No. 45, and No. 54, the Negative Predictive Value (NPV), Positive

Predictive Value (PPV), and Accuracy (ACC) of the constructed DT models with different modeling approaches are listed in Table 3-4, Table 3-5, Table 3-6, and Table 3-7, respectively. Also, the averaged errors are summarized and compared in Table 3-8. In this case, ACC is defined as the ratio of correctly classified operating points, averaged across all ten test data set. PPV and NPV indicate the proportions of correctly predicted security conditions that correspond to the conditions of unsafe (0) and safe (1), respectively. It is constructive to highlight that PPV is of most interest in the context of system security assessment because it is very dangerous to consider an unsafe scenario as a safe one.

Table 3-4. DT Test Errors (Outage Line: No. 30)

<i>Method</i>	<i>PPV</i>	<i>NPV</i>	<i>Accuracy</i>
Training Data	<u>82.84%</u>	96.33%	93.90%
MGD	<u>66.18%</u>	94.74%	89.52%
MGC	<u>78.90%</u>	96.30%	93.14%
Truncated C-Vine	<u>81.82%</u>	96.88%	<u>94.15%</u>

Table 3-5. DT Test Errors (Outage Line: No. 44)

<i>Method</i>	<i>PPV</i>	<i>NPV</i>	<i>Accuracy</i>
Training Data	<u>68.19%</u>	97.56%	95.06%
MGD	<u>43.33%</u>	98.76%	94.06%
MGC	<u>65.55%</u>	97.51%	94.80%
Truncated C-Vine	<u>68.25%</u>	97.62%	<u>95.12%</u>

Table 3-6. DT Test Errors (Outage Line: No. 45)

<i>Method</i>	<i>PPV</i>	<i>NPV</i>	<i>Accuracy</i>
Training Data	<u>68.67%</u>	97.95%	95.50%
MGD	<u>43.77%</u>	98.61%	94.04%
MGC	<u>67.31%</u>	97.71%	95.17%
Truncated C-Vine	<u>69.65%</u>	97.86%	<u>95.51%</u>

Table 3-7. DT Test Errors (Outage Line: No. 54)

<i>Method</i>	<i>PPV</i>	<i>NPV</i>	<i>Accuracy</i>
Training Data	<u>59.61%</u>	99.21%	98.27%
MGD	<u>30.02%</u>	98.90%	97.24%
MGC	<u>60.36%</u>	98.84%	97.92%
Truncated C-Vine	<u>65.25%</u>	99.19%	<u>98.39%</u>

Table 3-8. Overall DT Test Errors for Different Training Databases

<i>Method</i>	<i>PPV</i>	<i>NPV</i>	<i>Accuracy</i>
Training Data	69.83%	97.76%	95.68%
MGD	45.92%	97.75%	93.72%
MGC	68.03%	97.59%	95.25%
Truncated C-Vine	<u>71.24%</u>	97.89%	95.79%

As before, the proposed composite C-Vine method is compared against MGD, MGC and against the training data itself. Firstly, it is shown that all the tested methods are able to accomplish the accurate prediction of safe operation states, indicated by their high NPVs, whose minimum value reaches roughly 94.74% obtained via MGD when Line No. 30 is in outage. Afterwards, most of the ACC values also exceed 90%, whereas the most important metric, PPV, exhibits lower values but with larger difference among different training databases than NPV and ACC.

Overall, we clearly demonstrate that DTs trained on data generated via the composite C-Vine model exhibit the best performance for all the evaluation metrics, especially in terms of their negative-predictive values. Also, it illustrates the benefits of fitting the historical data into a statistical model as more non-existent but crucial samples could be generated via the well-constructed model, which is the proposed composite C-Vine model in this case. On the contrary, it is evident that the poorly constructed models such as MGD and MGC lead to lower NPV, PPV, and ACC values than the historical data itself.

It is also important to note that the order of the DTs' performance across the different modeling approaches is highly consistent with the results of the PNS calculation shown in Figure 3-9. For example, MGC is the second-best performing modeling algorithm expect the historical data itself, whereas the PPV of MGD is only 45.92%, which is the worst performing method as shown in Table 3-4.

Finally, one out of ten partitions is shown in Figure 3-10 as an example to visually inspect the advantage of the constructed DT via the proposed composite C-Vine model. Beyond the comparable reconstructed structure, the most important contribution of the proposed model is to generate 'smart' samples for carrying out the effective security assessment by identifying the accurate unsafe areas (i.e. red points areas), as shown in Figure 3-10 (e), which is similar but more concrete than using the historical data (i.e.

Figure 3-10 (a) and (b)). By comparison, the undesirable results of MGC and MGD are shown in Figure 3-10 (c) and (d), respectively, with significant difference between the samples and historical data regarding the dependency structure and the security areas.

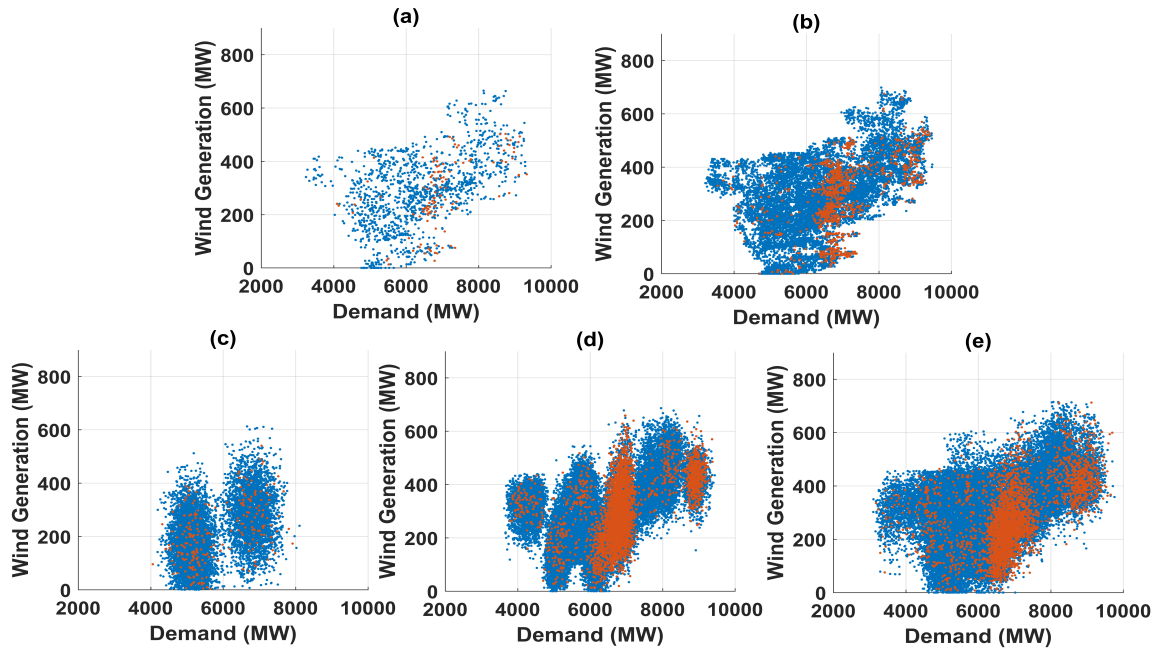


Figure 3-10. Scatter plots of test dataset 1 (a), training dataset 1(b), and data generated using a joint Gaussian distribution (c), multidimensional Gaussian copula (d) and the proposed truncated C-Vine method (e). Note that the safe and unsafe operating points are marked in blue and red, respectively.

3.7 Conclusions

The main contribution of the present chapter is an in-depth evaluation of potential synergies between these different components with the aim to identify well-performing composite modelling approaches. Among varieties of data mining techniques and vine copulas types, we choose the unsupervised clustering techniques: k-means, hierarchical clustering and GMM, the dimensionality reduction techniques: PCA and LPP as well as C-vine, D-vine copulas. Before assessing combinations of the above techniques, clustering validation indicators CDI and MIA are used to determine the optimal number of clusters and the optimal clustering method. On the basis of a 100-dimensional measurement dataset from the French transmission system, the calculation of these indicators for each clustering method shows that the most appropriate clustering technique is the hierarchical clustering with average linkage. To evaluate the

performance of the proposed composite modelling schemes with different combinations of the aforementioned techniques, visual comparisons, two-sample K-S tests, and multivariate two-sample tests are performed to test the reconstructed marginal distributions and dependency patterns of the sampled measurements. The results of the comparisons suggest the use of a modelling scheme that combines hierarchical clustering (average linkage), LPP, and D-vine copulas. Also, the results show that the performance of the proposed modelling schemes mostly depends on the choice of dimensionality reduction techniques. Overall, the composite modelling approach with LPP performs better than with PCA; in addition, D-vine copulas are more suitable for modelling the high-dimensional dependence encountered in the examined dataset.

Furthermore, case study applications on the calculation of a system's PNS profile and security assessment based on DTs demonstrate the benefits of integrating the developed statistical technique within a Monte Carlo simulation framework in the context of system security. The modeling approach presented here is set to become increasingly relevant with the advent of the smart grid and increasing penetration of intermittent generation. On one hand, many TSOs are bound to reduce controllability of a large part of their generation fleet and are already relying more and more on statistical techniques to forecast available resources and ensure robustness. On the other hand, measurement devices are being deployed across the network, gathering data that can be useful in characterizing in detail the statistical properties of stochastic parameters. All the above entail a fundamental shift from deterministic analysis to Monte Carlo simulation frameworks used to explore the uncertainty state-space. The present research, with its suitability for high-dimensional applications, enables operators to move beyond the selection bias of methods that rely exclusively on past observations.

Further research will focus on applying the proposed modelling approach to a broad range of problems that involve decision-making under high-dimensional uncertainty and require the generation of large populations of system states. There is also benefit in carrying out further tests on larger datasets spanning thousands of variables, as in large-area power systems. In terms of the model evaluation, additional metrics (e.g. [47]) could be considered to further assess the quality of the dependency modelling.

4. C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data

4.1 Introduction

Electricity market liberalization has largely unbundled the distribution and supply services in many jurisdictions, providing customers with the freedom to select their electricity supplier. In this competitive environment, retail companies can improve the commercial attractiveness of their product by formulating tariffs aimed at different customer types. An important part of the tariff design process is the identification of meaningful customer classes that exhibit different consumption patterns, enabling the development of diversifiable products. Moreover, electrical customer classification can also play a crucial role in load forecasting [4], [5] and modelling [6], electricity market development [7], energy system planning and operation [8] and theft detection [9]. Naturally, information on customer type (e.g. industrial, commercial, and residential) provides important information regarding the likely electricity consumption pattern and intensity. However, for further partitioning and exploratory analysis to be carried out effectively, high-frequency demand measurements are necessary [10]. As such, the advent of smart metering has led to large-scale availability of consumption data that render clustering analysis increasingly possible.

Load profile clustering aims to allocate consumers into a small number of homogeneous groups, ensuring that elements of the same cluster are similar between them, while being dissimilar to elements of different clusters. A large number of clustering techniques have been proposed in the past and applied to electrical load data. Examples of centroid-based approaches include k-means [48], [49], fuzzy k-means [50], k-medoids [51], and modified follow-the-leader [46]. Connectivity-based approaches have also been applied, focusing on hierarchical clustering methods with different linkages [46], [50]. Other studied techniques include ant colony clustering [52], self-organizing maps [53] and neural networks [54]. Typically, the number of clusters is determined using clustering validity indices as discussed in [8] and [55].

An alternative approach, beyond centroid and connectivity models, which is starting to gain interest among practitioners, is the use of distribution mixture models. One main

advantage of distribution-based clustering is the ability to systematically select the number of clusters while penalizing model complexity to construct parsimonious models [56]. Another advantage is the ability to obtain a probability measure regarding the classification of each consumer, as opposed to the ‘hard’ clustering that characterizes other methods. Another practical advantage relates to reduced memory requirements; by relying on parametric functions it is possible to effectively compress large amounts of data in a reduced-sized model.

In light of the above, a number of model-based clustering techniques have been used in the past to tackle customer clustering. Gaussian mixture model (GMM) is one of the most widely used model-based clustering approaches. It has been applied for time-of-use tariff design in [57] and clustering households in [58]. GMM has also been used in [59] to identify suitable data clusters for training Markov demand models. Other model types can also be deployed; for example, a multivariate Dirichlet process mixture model has been used to cluster electricity profiles in [60]. However, all such techniques suffer from the inherent limitations of standard multivariate functions; all mixture model components follow a pre-specified marginal distribution function and dependence structure. Given the highly non-linear dependence structures and non-standard marginal distributions observed in demand datasets, developing high-quality mixture models for clustering purposes is a substantial challenge.

In response to this, copulas can be utilized as a powerful tool to capture more complex dependency structures between variables. Gaussian Mixture Copula Models (GMCs) were proposed in [61], where a number of multivariate Gaussian copulas are fitted to a range of data sets. In [62], a Gaussian copula mixture model was developed for dependency-seeking clustering tasks for both synthetic and real data found in biological systems. In [63], the copulas-based mixture model clustering algorithm was extended beyond the Gaussian paradigm to also accommodate other copula families such as Gumbel and Clayton. Although the existing copula-based mixture model clustering techniques have demonstrated good performance, accurately capturing the complicated dependency structures exhibited by electrical system variables, such as loads, cannot be adequately described by solely relying on multivariate copula building blocks; more flexible modelling structures are required. Vine copula models make use

of pair copula construction (PCC) schemes to decompose a high-dimensional copula into a cascade of bivariate copula functions [64]. This substantially increases the flexibility of the model by being able to capture complex dependencies across a large number of variables. In addition, vine copula models have been shown not to suffer from the curse of dimensionality that characterizes other high-dimensional models [65]. Authors in [66] have shown that vine copulas are a class of density functions whose convergence rate does not depend on the number of dimensions. Given that a high-dimensional vine copula model is a cascade of bivariate functions, its convergence rate is equal to the rate of a two-dimensional estimator, thus evading the curse of dimensionality.

In this chapter, a specific pair-copula construction scheme known as C-Vine is used to perform a novel type of distribution-based clustering. It is important to note that an indirect clustering approach is followed; a transformation to a lower-dimensional feature space is first applied instead of the model being fit directly to the consumption data set. Many variants of indirect clustering have been applied in the past. For example, authors in [67] perform clustering on household occupancy states which have been inferred using a Hidden Markov model. In a similar vein, authors in [68] apply Fast Search and Find of Density Peaks [69] a novel density-based clustering method, on occupancy state transition matrices. Locality- sensitive hashing is used in [70] to substantially speed up subsequent similarity comparisons for clustering. Indirect clustering is also combined with GMM in [71]. A discussion on the selection of possible features takes place in [72]. In our work, by combining vine copulas with dimension reduction, we are capable of addressing the drawback of increased computational burden while also harvesting the synergy between the C-Vine's hierarchical structure and the ordered variables, as discussed in [73].

This chapter proposes a C-vine copulas based mixture model clustering (CVMM) algorithm for grouping the load pattern data. For CVMM, the parameters of the constructed mixture model are estimated via the expectation-maximization (EM) algorithm. The clustering quality of CVMM is evaluated and compared to other classical methods by using a set of selected clustering validation indicators based on real demand measurements. In addition, a decision tree based classification module for

assigning new consumers to the existing classes is developed to further assess the results of CVMM clustering.

This chapter is structured as follows. Chapter 4.2 illustrates the procedure of load pattern data processing and defines the proposed C-vine Mixture Model Clustering (CVMM) algorithm with EM estimation. Chapter 4.3 introduces the selected clustering validation indicators and their expressions. In Chapter 4.4, based on real residential load pattern data from London, the performance of the proposed CVMM algorithm is assessed and compared with other widely-used load pattern clustering algorithms via comparing various relevant metrics. Furthermore, the clustering results are analyzed using demographic metadata. In addition, the accuracy of a decision-tree based customer classification model is used as an additional performance metric. The clustering performance of CVMM is shown to be superior to existing methods when applied to the autumn dataset and combined weekday/weekend dataset. Finally, a scaling analysis for different sets of large numbers of households is presented. Chapter 4.5 contains the concluding remarks.

4.2 C-vine Mixture Model Clustering Algorithm

4.2.1 Load Pattern Data Processing

Given a set of N customers, the electrical consumption of each customer is typically represented via a daily load pattern, calculated by averaging the measured load data over a period of time (typically in the order of a few months or a year). Subsequently, the representative load pattern (RLP) defined in [46] can be obtained as the normalized daily load pattern in $[0,1]$ with regards to a reference value (i.e. peak value of the daily load pattern). In most of the existing literature on load pattern clustering, it has been demonstrated that it is effective to group customers into different classes via some appropriate clustering techniques on the basis of their RLPs.

Mathematically, let $X = [x_1, \dots, x_N] \in \mathbb{R}^{P \times N}$ denote the whole set of historical load data measurements, where P is the total number of data points measured during a determined period of time, each vector $x_n = [x_{n,1}, \dots, x_{n,P}] \in \mathbb{R}^P$ represents the monitored load data for the n^{th} customer, for $n = 1, \dots, N$. Assuming that m is the number of samples characterizing each daily load pattern, for each customer, the daily

load pattern $\bar{x}_n = [\bar{x}_{n,1}, \dots, \bar{x}_{n,m}] \in \mathbb{R}^m$ can be calculated by averaging every m data points of x_n . Furthermore, the data set D of the N RLPs is represented by $D = [d_1, \dots, d_N]^T \in \mathbb{R}^{N \times m}$ in which each vector $d_n = [d_{n,1}, \dots, d_{n,p}] \in \mathbb{R}^p$ can be calculated as:

$$d_n = \bar{x}_n / \max(x_n) \quad (4.1)$$

After building the N RLPs via the above data processing procedure, the next step is to group the N RLPS into a pre-determined number of clusters by using the proposed CVMM clustering algorithm.

4.2.2 CVMM Clustering Algorithm

Let K denote the number of customer classes to be partitioned; the density of the C-vine mixture model (CVMM) for the data set D of the N RLPs is given by:

$$f(D|\Theta) = \sum_{k=1}^K \pi_k f_k(D|\Theta_k) \quad (4.2)$$

where $\pi_k \in [0,1]$ with $\sum_{k=1}^K \pi_k = 1$ and $\Theta = [\Theta_1^T, \dots, \Theta_K^T]^T$ represent the mixing proportions and the parameters of the component densities $f_k(D|\Theta_k)$, for $k = 1, \dots, K$, respectively. Note that each parameter set Θ_k of Θ contains all the pair-copula parameters for the k^{th} component in the case of C-vine representation. According to equation (2), the multivariate distribution $f(D|\Theta)$ can be expressed as:

$$\sum_{k=1}^K \pi_k \prod_{l=1}^m f_k(d_{1:N,l}) \prod_{j=1}^{m-1} \prod_{i=1}^{m-j} c_{j,j+i|v_{i,j}}^k \left(F_{j|v_{i,j}}^k, F_{j+i|v_{i,j}}^k \right) \quad (4.3)$$

where $v_{i,j} \in \{1, \dots, j-1\}$, f_k and $c_{\dots}^k$ represent the empirical marginal distribution and a pair-copula of the k^{th} component, respectively. Let $S = [s_1, \dots, s_N]^T$ denote the latent random variables whose element s_n indicates the label for d_n , the n^{th} observation of data set D , for $n = 1, \dots, N$, the main aim of the proposed unsupervised CVMM clustering algorithm is to maximize the complete data log-likelihood function:

$$L = \sum_{n=1}^N \sum_{k=1}^K w_{n,k} \times \left(\sum_{j=1}^{m-1} \sum_{i=1}^{m-j} \ln c_{j,j+i|v_{i,j}}^k \left(F_{j|v_{i,j}}^k, F_{j+i|v_{i,j}}^k | \theta_{j,i}^k \right) + \sum_{l=1}^{m-1} \ln f_k(d_{S=k,l}) + \ln \pi_k \right) \quad (4.4)$$

where $w_{n,k}$ represents the posterior probability of the n^{th} measurement that was drawn from the k^{th} component, calculated by:

$$w_{n,k} = \frac{\pi_k f_k(d_n | \Theta_k)}{\sum_{i=1}^K \pi_i f_i(d_n | \Theta_i)} \quad (k = 1, \dots, K). \quad (4.5)$$

In addition, the prior probability (mixing proportion) of the k^{th} component is denoted by π_k :

$$\pi_k = \frac{\sum_{i=1}^N w_{i,k}}{N} \quad (k = 1, \dots, K). \quad (4.6)$$

The expectation-maximization (EM) algorithm is used to estimate the maximum likelihood parameters Θ and the latent variables S of equation (4.4), given measurement data D . The conventional idea of the EM algorithm can be deciphered as an alternate between estimating the latent variables and the parameters. Beyond that, a more comprehensive understanding is to consider the EM algorithm as a lower bound maximization procedure [74]. Accordingly, the expectation step (E-step) can be regarded as a process to build a local lower bound to the posterior distribution, while the maximization step (M-step) aims to optimize the constructed bound. In general, after initializing the EM algorithm, the t^{th} iteration of the EM algorithm for CVMM can be summarized as:

E-step: Calculate $w_{n,k}^{(t)}$ using equation (4.5), for $n = 1, \dots, N, k = 1, \dots, K$.

M-step: Update $\pi_k^{(t)}$ according to equation (4.6). Maximize the likelihood function $L^{(t)}$ to get the parameters $\Theta^{(t+1)}$ and the labels $S^{(t)}$.

A detailed explanation of the above EM estimation for CVMM is given as follows. Firstly, regarding the initialization step, the superiority of the hierarchical clustering with average distance linkage criterion has been demonstrated in [46] when compared to k-means, fuzzy k-means, modified follow-the-leader, hierarchical clustering with other types of linkages, and self-organizing maps. Therefore, in this case, the first step

is to initialize the EM algorithm by pre-grouping the input data D into K clusters via the hierarchical clustering with average distance linkage criterion. Note that the optimal number of clusters K is determined according to the clustering validity indicators, detailed in Section IV. Subsequently, the E-step of EM algorithm consists of the calculation of the posterior probabilities $W = [w_{1:N,1}, \dots, w_{1:N,K}] \in \mathbb{R}^{N \times K}$ and the update of labels S for N customers. The label s_n for each customer is set to the index of the maximum value in $[w_{n,1}, \dots, w_{n,K}]$, for $n = 1, \dots, N$. After updating the prior probabilities $\Pi = [\pi_1, \dots, \pi_K] \in \mathbb{R}^{K \times 1}$ using equation (10), the M-step maximize the likelihood function L by estimating the copula parameters set $\Xi = [\Theta_1, \dots, \Theta_K]$ for all the K C-vine models. During the procedure of C-vine construction, the first step is to transform the historical data to uniform margins through their corresponding empirical cumulative functions (ECDFs). Afterwards, model selection (via AIC) and parameter estimation (via MLE) are performed on each pair-copula to identify the best-fitting function. In each iteration, a total of $Km(m-1)/2$ pair copulas are fitted. The E-step and M-step are carried out iteratively until the relative change e in log-likelihood value L is less than a predefined threshold ε . The proposed algorithm is outlined below.

Algorithm 1 CVMM Clustering

Input: data set $D = [d_1, \dots, d_N]^T$, number of clusters K .

Output: clustering labels $S = [s_1, \dots, s_N]^T$.

- 1: Initialize customer labels $S^{(0)} = [s_1^{(0)}, \dots, s_N^{(0)}]^T$ via an initial clustering technique (e.g. hierarchical clustering).
- 2: Initialize the prior probabilities $\Pi^{(0)} = [\pi_1^{(0)}, \dots, \pi_K^{(0)}] \in \mathbb{R}^{K \times 1}$ and the posterior probabilities $W^{(0)} = [w_{1:N,1}^{(0)}, \dots, w_{1:N,K}^{(0)}] \in \mathbb{R}^{N \times K}$.
- 3: $L^{(0)} = -\infty$
- 4: $t = 1$
- 5: **repeat**
- 6: **for** $n \leftarrow 1, \dots, N$
- 7: Update $w_{n,k}^{(t)} \forall k$ using equation (4.5)
- 8: $s_n^{(t)} = \arg \max\{[w_{n,1}^{(t)}, \dots, w_{n,K}^{(t)}]\}$
- 9: **end for**

```

10:      Update  $\pi_k^{(t)} \forall k$  using equation (4.6)
11:      Transform  $f_k^{(t)}(d_{S^{(t)}=k,l})$  to  $F_k^{(t)}(d_{S^{(t)}=k,l})$  through its ECDF,  $\forall l \in 1, \dots, m$ 

12:      for  $j \leftarrow 1, \dots, m - 1$ 
13:          for  $i \leftarrow 1, \dots, m - j$ 
14:              Perform family selection and parameter estimation on
              the pair-copula
               $c_{j,j+i|v_{i,j}}^k{}^{(t)}(F_{j|v_{i,j}}^k{}^{(t)}, F_{j+i|v_{i,j}}^k{}^{(t)} | \theta_{j,i}^k{}^{(t)})$ 

15:          end for
16:      end for
17:      Calculate the log-likelihood  $L^{(t)}$  using equation (4.4)

18:       $e^{(t)} = L^{(t)} - L^{(t-1)}$ 
19:       $t = t + 1$ 
20:      until  $e^{(t-1)} < \varepsilon$ 

```

4.3 Clustering Quality Evaluation

Different clustering validity indicators have been proposed in the past as a comprehensive tool to quantitatively analyze the quality of clustering schemes [4], [8], [46], [55]. As such, they can substantially inform the choice on the number of clusters to be applied in each case and enable the comparison between different clustering methods. Consider that the clustering results, K partitioned customer classes with computed centroids $C = \{c_1, \dots, c_K\}$, are obtained by applying the proposed CVMM algorithm on the processed data set D of the N RLPs. Let D_k denote the customers that belongs to k^{th} class, for $k = 1, \dots, K$, the four indicators taken into account in this chapter are calculated as follows. Note that all distances are Euclidean.

1) The clustering dispersion indicator (CDI), defined as the ratio of the mean intraset distance between the RLPs in the same cluster and the intraset distance between the centroids of the K clusters [46], [50]:

$$CDI(D, K) = \widehat{d}(C)^{-1} \sqrt{K^{-1} \sum_{k=1}^K \widehat{d}^2(D_k)} \quad (4.7)$$

2) The modified Dunn index (MDI), is computed as follows, for $i, j = 1, \dots, K$:

$$MDI(D, K) = \max_{1 \leq k \leq K} \{\widehat{d}(D_k)\} / \min_{i \neq j} \{d(c_i, c_j)\} \quad (4.8)$$

where $d(.,.)$, $\hat{d}(.)$ represent cluster-to-cluster distance and intraset distance, respectively, as defined in [46].

3) The Davies-Bouldin index (DBI) [46], [50] is expressed as the system-wide average of the maximum ratio of the within cluster scatter to the between cluster separation, for $i, j = 1, \dots, K$:

$$DBI(D, K) = \frac{1}{K} \sum_{k=1}^K \max_{i \neq j} \left\{ \frac{\hat{d}(D_i) + \hat{d}(D_j)}{d(c_i, c_j)} \right\} \quad (4.9)$$

4) The mean index adequacy (MIA) [75] represents the average of the distances between the centroid of a cluster and each RLP in this cluster:

$$MIA(D, K) = K^{-1} \sum_{k=1}^K d^2(c_k, D_k) \quad (4.10)$$

For all four indicators, a lower value indicates better clustering performance. By applying a clustering algorithm a number of times, we can obtain the curve that describes the relationship between the indicator values and the number of clusters K . Subsequently, the optimal number of clusters can be obtained either by seeking the first knee of this curve, considering the Bayesian information criterion (BIC), or the informational complexity criterion (ICOMP) [46]. In this chapter, we determine the optimal value of K by utilizing the AIC. This way, the performance of the proposed CVMM clustering algorithm can be evaluated and compared to other existing technique (e.g. k-means, hierarchical clustering, etc.) through the comparison of their respective indicator values.

4.4 Case Study Application

4.4.1 LCL Load Dataset and Data Processing

The Low Carbon London (LCL) smart meter trials were designed to characterize the residential consumer demand of London and to assess the benefits from employing smart metering for distribution network operation. In the project, Landis and Gyr (L+G) E470 electricity meters were installed in 2,613 residential homes across the Mayor of London's Low Carbon Zones and the London Power Networks distribution network license area operated by UK Power Networks. Specifically, the Engineering Instrumentation Zones of the LCL trial contain Queen's Park, Merton, and Brixton. In this chapter, the LCL load dataset consists of 17,520 half-hourly electrical load

consumption measurements across 2,613 customers in kW for a full calendar year from 1st January 2013 to 31st December 2013. The RLPs of all $N = 2,613$ customers were calculated according to the data processing procedure described in Chapter 4.2.1. All RLPs are shown in Figure 4-1, each RLP consisting of 48 half-hourly normalized demand values. The mean of all RLPs, indicated in Figure 4-1 by a bold black line, follows the typical ‘duck curve’ shape indicating increased consumption levels during afternoon hours with a peak at around 7:00 pm. However, individual RLPs exhibit very high variation, with some consumers having a profile very different to the average RLP pattern.

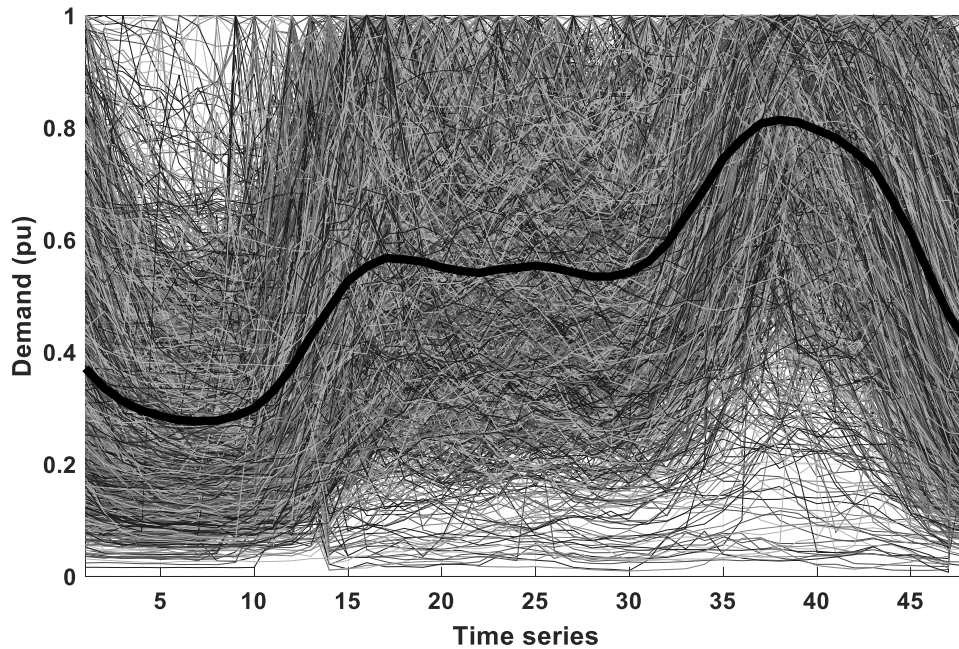


Figure 4-1. RLPs for all 2,613 customers of the LCL smart meter trials. Thick black line shows the mean RLP.

In this part, the proposed CVMM algorithm is applied to group the customers in the LCL dataset. For the proposed CVMM clustering algorithm, it is required to fit $(48 \times 47) / 2 \times t \times \varphi = 1128\varphi$ pair-copulas, where φ is the number of candidate bivariate copula families. The candidate families considered in this application are Gaussian, Student-t, Clayton, Frank, Gumbel as well as their 90° , 180° and 270° rotated versions i.e. $\varphi = 20$. As such, the presented modelling task will involve fitting $22,560 \times K$ copulas, repeated a number of times until convergence is achieved. Given that a single copula parameterization may take a few milliseconds, it is clear that the proposed procedure can involve a high computational burden.

To accelerate this procedure, a widely-used linear dimensionality reduction technique, principal component analysis (PCA), is used to construct a lower dimensional representation of the 48-dimensional RLPs named principal components (PCs), that retains as much of the variance as possible. The principal components are listed in descending order of the eigenvalues of the RLPs' covariance matrix. Note that the eigenvalues can be treated as a metric of the information contained within each principal component. To this end, we use a user-defined information retainment threshold to determine the number of PCs to be retained. In this case, we set the information retainment threshold to 97.5%, which corresponds to 12 PCs being retained. The size of the reduced dataset is 2,613 observations of 12 variables. The convergence threshold ε has been set to 10^{-5} . We indicatively mention that, for the CVMM algorithm with $K = 8$, convergence was achieved within 3 iterations, totaling CPU time of 3 minutes when running on 8 cores. Note that the CVMM algorithm can be parallelized so that each candidate copula family fit is carried out independently, resulting in significant speed-ups. In addition, a smaller number of retained PCs for the CVMM methods can also reduce the computational burden. A more extensive analysis of the method's computational performance and how it compares with other methods is shown in Section V.C.

4.4.2 Determining the Optimal Number of Clusters for CVMM

As stated in [77], one of the most important advantages of model-based clustering methods is rendering the issue of determining the optimal number of clusters to a statistical model selection problem based on information criteria such as AIC and BIC. For the proposed CVMM algorithm, the calculated AIC and BIC values for $K = 5$ to 20 are shown in Figure 4-2. Accordingly, the optimal number of mixture models (clusters) for the tested dataset is equal to eight ($K_{opt} = 8$) indicated by the first local maximum in both of the AIC and BIC curves. Although AIC and BIC differ in the way they are not guaranteed to agree in their selection, in this case both methods indicate the same optimal number of clusters.

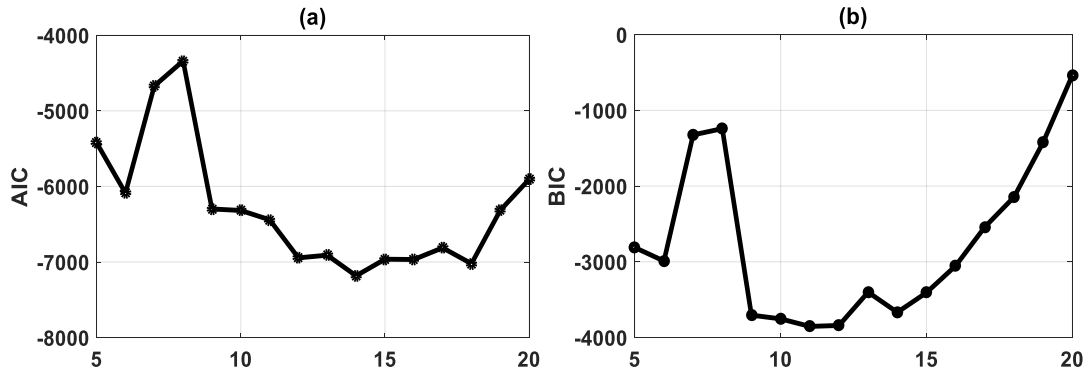


Figure 4-2. Calculated AIC (a) and BIC (b) values for $K = 5$ to 20.

4.4.3 Clustering Validity Assessment

In this section the performance of CVMM is evaluated and compared with other well-known clustering techniques used in the past; k-means, hierarchical clustering (complete linkage), hierarchical clustering (average linkage), hierarchical clustering (weighted linkage), hierarchical clustering (Ward linkage), and Gaussian mixture model clustering (GMM). Note that all the above approaches are applied to the original dataset of RLPs (48 variables), whereas the CVMM algorithm includes a dimension reduction stage that carries out clustering on a lower-dimensional dataset (12 variables in this case). As discussed in Section IV, four types of indicators, CDI, MDI, DBI and MIA, are chosen for carrying out the clustering validity assessments for different numbers of clusters K , ranging from 5 to 20.

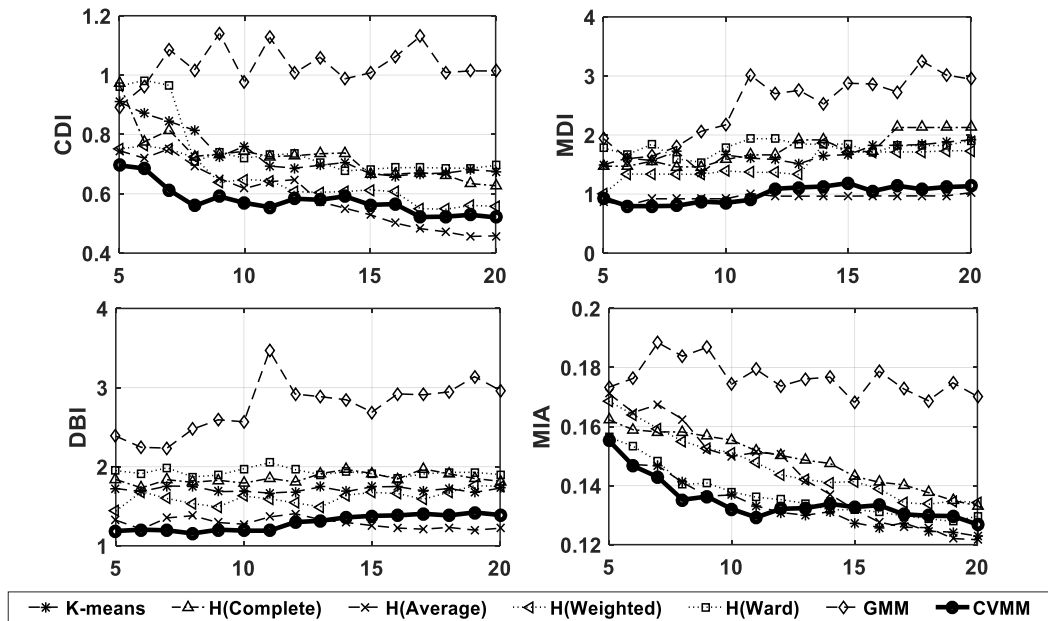


Figure 4-3. Comparisons among the clustering techniques by using (upper-left) the CDI indicator, (upper-right) the MDI indicator, (lower-left) the DBI indicator, and (lower-right) the MIA indicator for $K = 5$ to 20.

The results illustrated in Figure 4-3 indicate that, for the tested dataset, the proposed CVMM algorithm exhibits superior performance to the other techniques. This is evidenced by the fact that classification via CVMM consistently results in lower values across all four indicators in the range up to 12 clusters. Although the H(Average) hierarchical method does result in lower indicator values for larger numbers of K we are interested in comparing values in the range where the knee for each method occurs. Overall it is constructive to note that the hierarchical methods H(Average) and H(Weighted) perform relatively better than H(Complete), k-means and H(Ward). The limitations of GMM is indicated by the large values across all four indicators, highlighting the concern with standard multivariate model-based methods. As such, the proposed CVMM clustering algorithm is shown to form well-separated customer classes and to detect outlier load patterns.

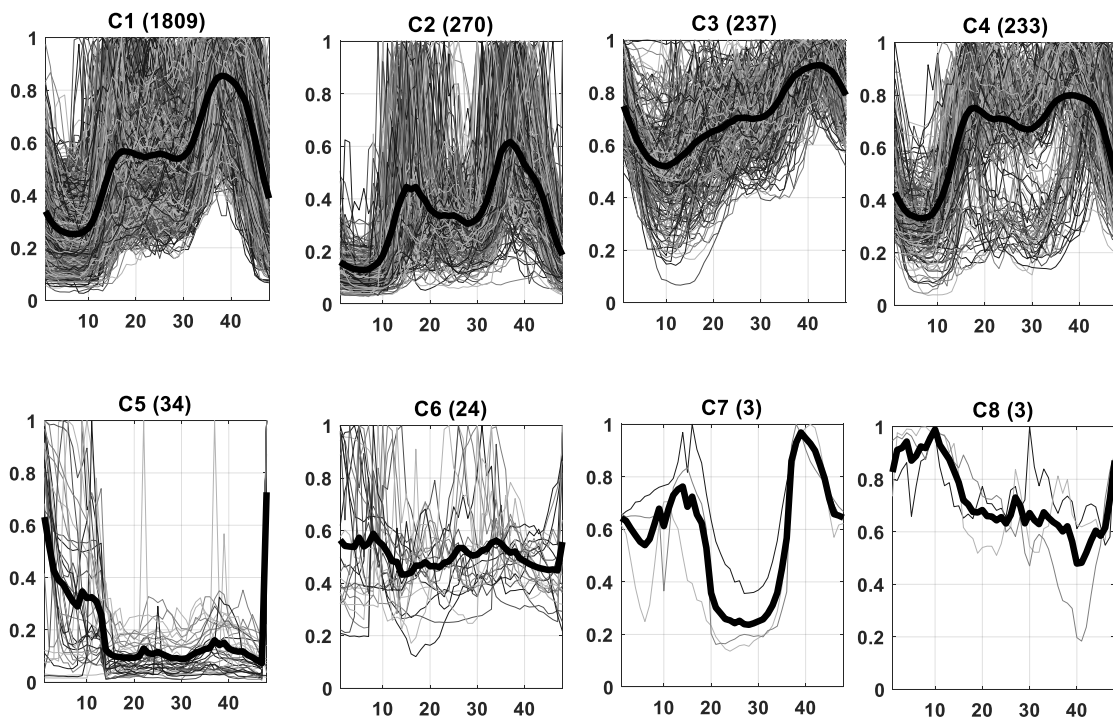


Figure 4-4. Clustering results for the proposed CVMM clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

Figure 4-4 shows the load patterns groups that were classified by the proposed CVMM clustering algorithm with $K = 8$. The number of consumers categorized in each cluster is shown in brackets next to each panel's title. Within each subplot, the average load pattern of all the customers in this class is plotted using a thick black line. From visually inspecting the plots, it is evident that the average load patterns of different

classes have radically different shapes. In addition, there are clusters that have a very large number of customers, while some clusters include outlying consumption behavior.

Cluster 1 (denoted C1) is the most populous cluster with over 69% of all customers. The mean profile exhibits the typical ‘duck curve’ shape, with peak consumption occurring in the afternoon hours. Also, mean profiles of C2 (10% of all consumers) and C4 (9%) have the similar shape as C1, but with slightly lower and higher consumption levels respectively. C3 (9%) and C8 (0.1%) contain the residents with relatively high consumption especially in the evening hours and the midnight hours, respectively, whereas most of the residents in C5 (1%) exhibit low electrical energy consumption during the daytime and have a sudden increase in midnight hours. In C6 (0.9%), mean consumption has less fluctuation than other clusters. On the contrary, large volatility is shown in C7 (0.1%) with abruptly low RLP values between 10:00 am and 17:00 pm. C8 exhibits high overall consumption with particularly high values during nighttime.

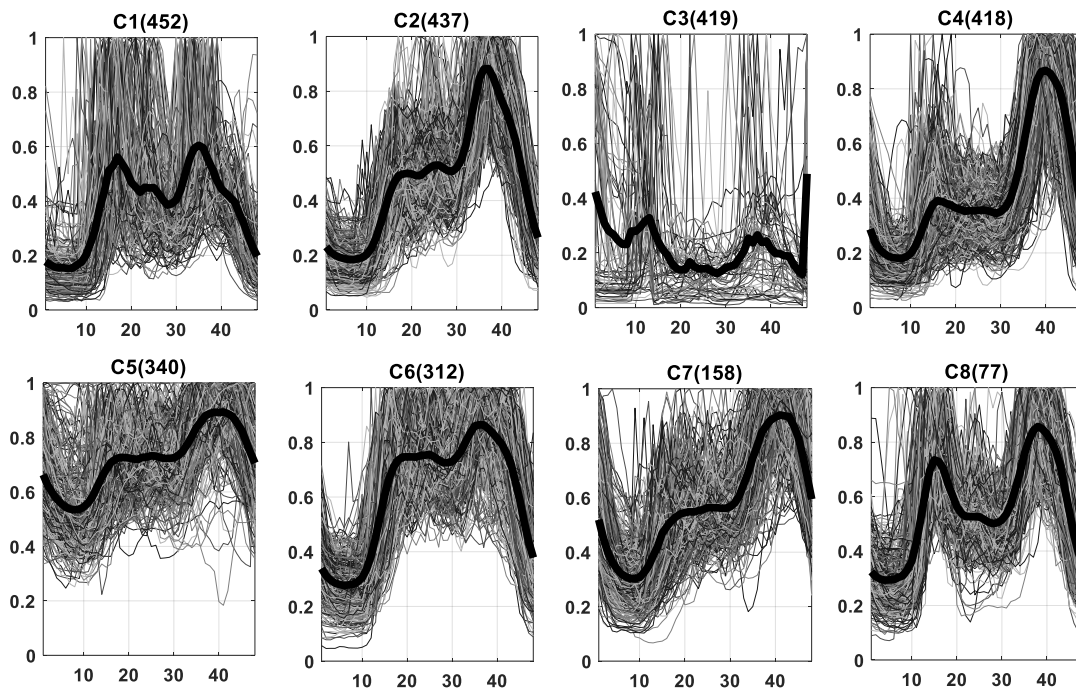


Figure 4-5. Clustering results for k-means with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

It is constructive to highlight that only 3 consumers have been placed in C7 and C8. By closer inspection of the individual patterns clustered in C7 and C8 we see that these profiles are indeed very similar between them and yet substantially different to other groups. Efficient detection of outliers is an important feature of well-performing clustering schemes; CVMM clearly succeeds in this task, presenting superior outlier detection capability. In addition, an example of the clustering results of k-means method is shown in Figure 4-5. It can be observed that the simpler method, k-means, fails to correctly identify the outliers identified in C7 and C8, clustered via CVMM.

A way to further analyze the average load profiles of each cluster is to calculate the normalized load shape indices that involve the information of the daily load curve shape during different periods of time. We calculate three attributes of interest as presented in [75]; load factor (d1), night impact (d3), and lunch impact (d5). In particular, d1 represents the full day consumption behavior, while d3 and d5 capture the load shape from 11:00 pm to 7:00 am (8 hours) and from 12:00 pm to 15:00 pm (3 hours), respectively. For the different clusters obtained by CVMM algorithm, the calculated d1, d3 and d5 values of the mean RLP of each cluster are shown in Table 4-1.

In terms of d1, C3 and C8 have relatively high values compared to other clusters. C5 exhibits the lowest d1 value but also the highest d3 value, which means it has a very high consumption during the night. In addition, C7 exhibits the lowest d5 value, which means severely reduced electricity usage around noon, as also evidenced in Figure 4-4.

Table 4-1. Load Attributes For Different Clusters

	<i>C1</i>	<i>C2</i>	<i>C3</i>	<i>C4</i>	<i>C5</i>	<i>C6</i>	<i>C7</i>	<i>C8</i>
d1	0.534	0.345	0.706	0.625	0.193	0.499	0.566	0.723
d3	0.197	0.190	0.287	0.221	0.622	0.345	0.379	0.413
d5	0.127	0.115	0.124	0.135	0.062	0.128	0.052	0.116

4.4.4 Clustering Results Analysis Using Metadata

In the LCL smart meter trials, in addition to consumption measurements, various socio-economic conditions of the participating household were also recorded. In this chapter, we focus on the data pertaining to household occupancy and wealth level. The former relates to the number of people living in the property. The latter has been drawn on the basis of mapping all participating households to ACORN groups [76]. Subsequently, three wealth classes have been defined: Adverse, Comfortable, and Affluent in increasing order. A demographic group is defined as a combination of

occupancy and wealth level. The number of customers belonging to each of the nine groups is shown in Table 4-2. The clusters generated via the CVMM algorithm are analyzed in Figure 4-6 using the metadata in Table 4-2.

Table 4-2. Number of Households Across Demographic Groups

	<i>1 occupant</i>	<i>2 occupants</i>	<i>3+ occupants</i>
Adverse	312	275	233
Comfortable	237	300	209
Affluent	428	398	221

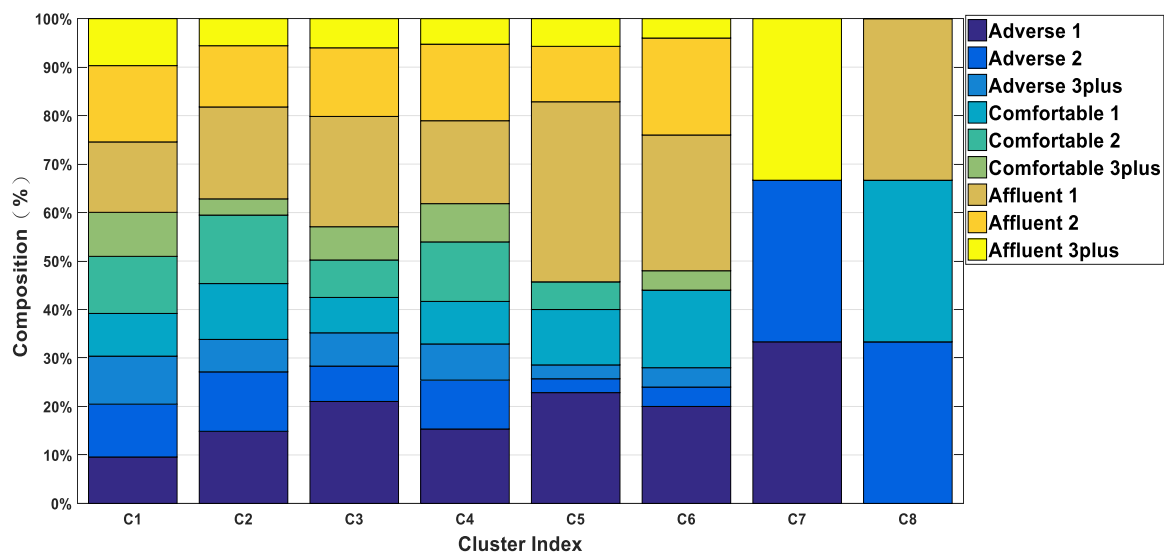


Figure 4-6. Cluster composition analysis using demographic metadata.

For C2, C5, and C6, there is no wealth category that dominates; these three clusters consist of Adverse, Comfortable, and Affluent households in almost equal distribution. The number of people living in the property also cannot be evidently distinguished in this case. On the other hand, in C6, households with one occupant account for the highest percentage of about 46%. C4 consists mostly of adverse wealth households (67%), with the rest of customers belonging to the Affluent group. In C6, groups of 1 and 2 occupants account for 67% and 23%, respectively. In addition, approximately 50%, 43%, and 53% of the RLPs in C1, C3, and C8 are in the Affluent group respectively. Meanwhile, households with 1 occupant account for 67%, 50% and 74% of households, respectively. The above results demonstrate that there are no clear nor consistent discrepancies across different wealth levels and number of occupants when comparing residential RLPs. This highlights the importance of performing customer partitioning on the basis of actual consumption measurements.

4.4.5 New Customer Characterization

An electrical customer characterization framework was proposed in [75] to generate customers classes based on the existing consumers' RLPs (*Clustering Module*) and subsequently build a classification model using decision trees (DTs) that assist system operators to assign new consumers to the previously constructed classes (*Classification Module*). In this chapter, customer clustering is carried out using the proposed CVMM clustering algorithm. In the Classification Module, d1, d3, and d5 are chosen as classifying variables. A DT model is trained on the load shape indices (d1, d3, and d5) and the corresponding cluster indices. Using the constructed DT model, we can accomplish the assignment of new customers to the classes obtained by the CVMM algorithm.

For the purpose of evaluating the constructed DT model, ten-fold cross-validation is used to get an estimate of classification accuracy. To this end, the calculated load shape indexes are randomly partitioned into 10 test datasets. In fact, the customers in the test dataset can be treated as new customers but with known labels. Each test data set accounts for 10% of the total customers, and the remaining 90% customers are attributed to the training data set. For each fold, the DT model is trained using one of the training data sets and then tested with the corresponding test data set. This process is presented in Figure 4-7.

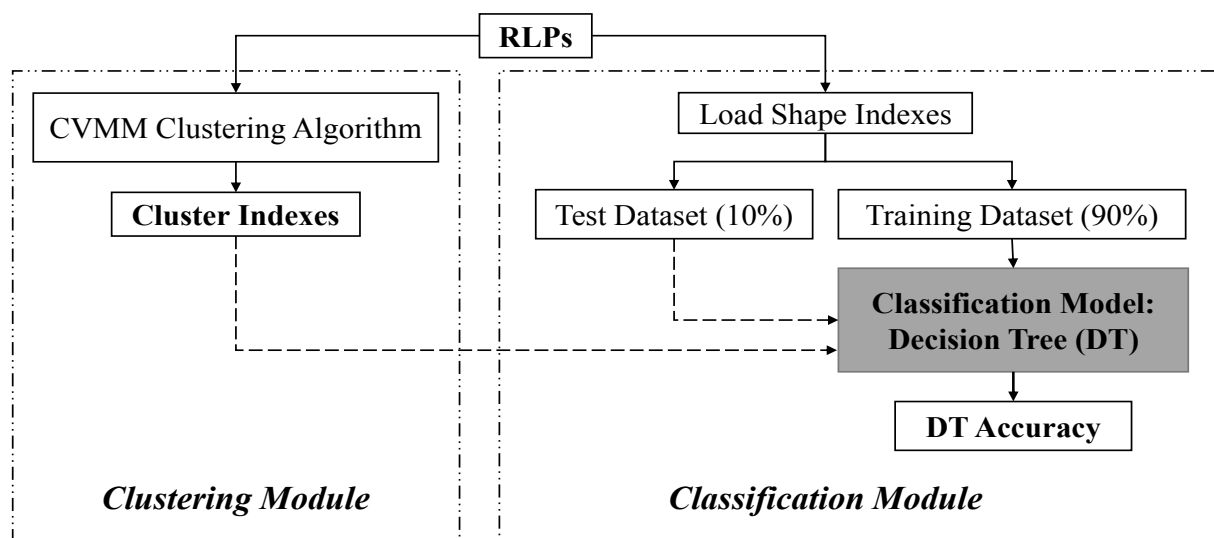


Figure 4-7. The structure of the modified customer characterization framework with the proposed CVMM clustering algorithm.

Table 4-3. Classification Module DT Accuracies

Method	Accuracy
CVMM	93.65%
Hierarchical(Average)	89.82%
Hierarchical(Weighted)	87.83%
Hierarchical(Complete)	62.31%
K-means	56.90%
Hierarchical(Ward)	55.42%
GMM	51.93%

The overall accuracies of the constructed DT models with different clustering methods are listed in Table 4-3 . In this case, accuracy is defined as the ratio of correctly classified households, averaged across all ten test sets. As can be seen below, the DTs trained on the classes generated via the proposed CVMM algorithm have the highest accuracy followed by hierarchical clustering. This analysis simulates a challenge that suppliers have to tackle in practice; classification of a new customer for which they may have limited information (in this case we assume we have an approximation of their overall consumption level, as well as an estimate of consumption intensity at night and noon times). The fact that CVMM exhibits the best predictive capability means that the proposed algorithm has partitioned customer profiles in such a way so that accurate conclusions about detailed usage patterns can be drawn in the absence of full information. As shown in Table 4-3, other methods such as k-means and GMM perform poorly in this respect, rendering them unfit for classifying new consumers. The hierarchical clustering method with average linkage, which was used to initialize the CVMM algorithm performs well, but it is evident that the proposed method enhances clustering performance.

It is also important to note that the order of the DTs' performance across the different clustering methods is highly consistent with the results of the clustering validation assessment shown in Figure 4-3. For example, H(Average) is the second-best performing algorithm after CVMM, whereas the DT accuracy of GMM is only 51.93%, which is the worst performing method as shown in Figure 4-3.

Note that when including ACORN demographic data in the training attribute set, the classification error improves only marginally. This highlights the significance of d1, d3, d5 in terms of clustering new customers and shows that the availability of demographic

data may convey little information regarding daily demand consumption patterns of a prospective user.

4.4.6 Clustering Across Different Seasons and Day Types

In all the preceding analysis we have analysed the full dataset corresponding to an entire year without differentiating between different calendar seasons nor between weekends and weekdays. Naturally, the season and day type has a substantial impact on consumer behavior as shown in Fig. 10 (note that seasons are defined in accordance to the specification of the UK market operator [78]). In Figure 4-8 (a), average residential electricity consumption begins later in the morning and leads to a small mid-day peak instead of the valley experienced in weekdays when residents are typically away from their home. The impact of season is also substantial as expected; for example, the 7pm consumption peak is higher in winter than the other seasons. To this end, it is of interest to evaluate the performance of CVMM when dealing with the datasets under different loading conditions.

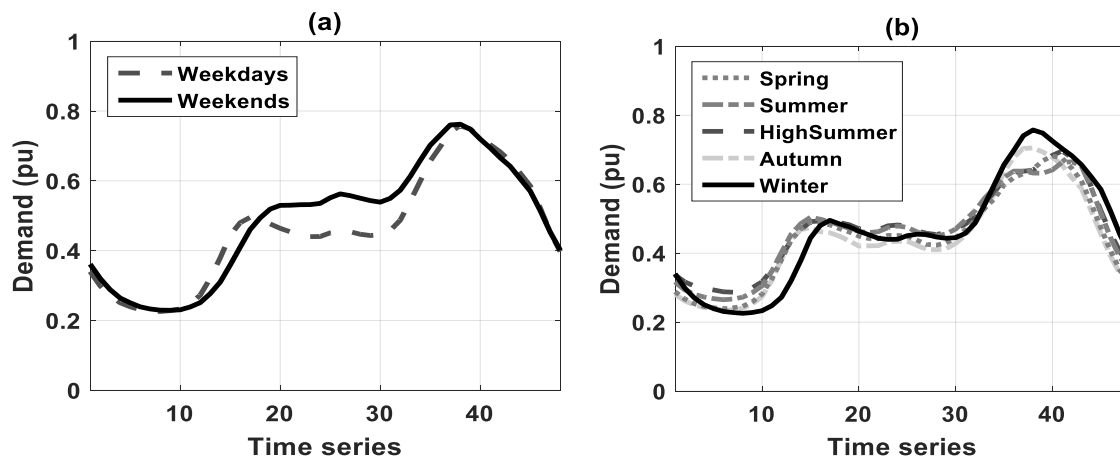


Figure 4-8. Average RLPs for all 2,613 customers of the LCL smart meter trials under the loading conditions: (a) Weekdays/Weekends; (b) Seasons.

For the purpose of determining tariffs for different customer types, the consumption behaviors of each customer in weekdays and weekends could be considered simultaneously. Therefore, we combine the calculated RLPs under loading conditions, weekdays and weekends, together to construct a 96-dimensional dataset. Based on this dataset, the performance of the proposed CVMM method is evaluated and compared with the other candidate methods, using the above-mentioned four types of indicators,

the CDI, the MDI, the DBI, and the MIA, for different numbers of clusters, ranging from 5 to 20.

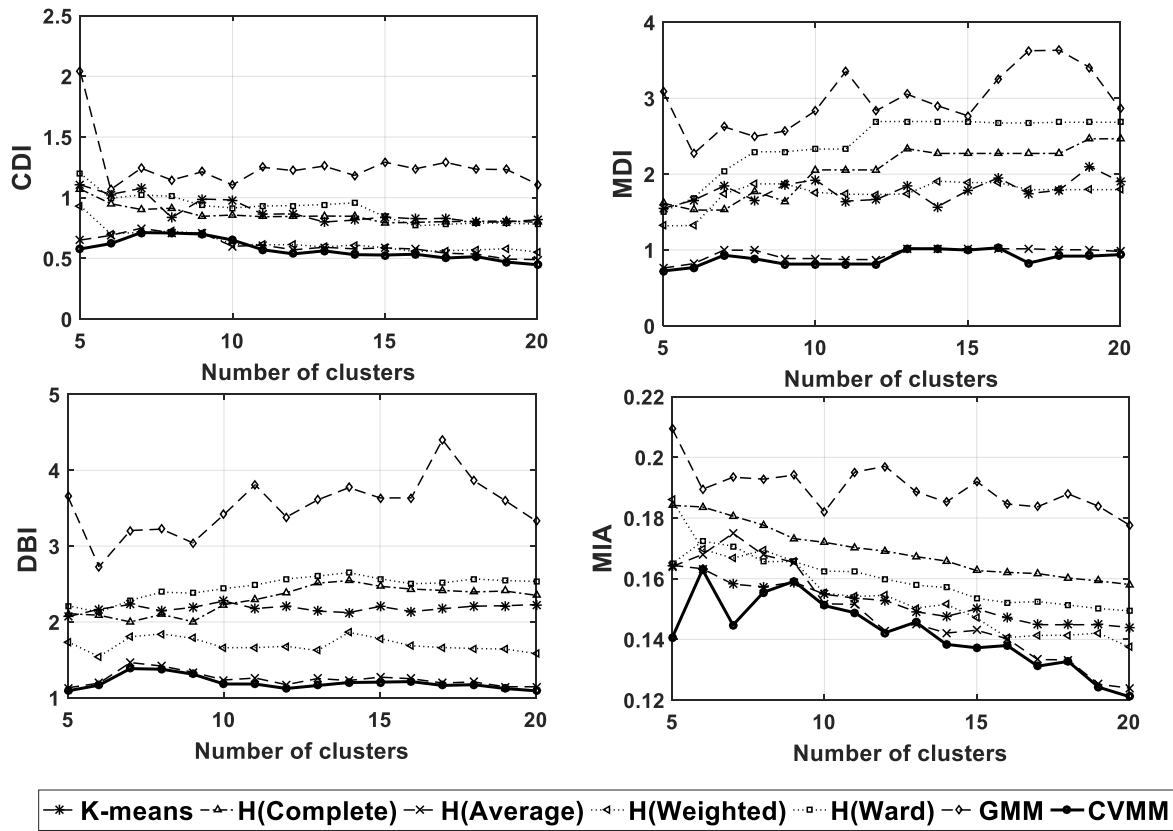


Figure 4-9. Comparison among the clustering techniques based on weekdays and weekends RLPs using CDI, MDI, DBI and MIA indicators for $K = 5$ to 20.

Similar to Figure 4-3, the results illustrated in Figure 4-9 also show better performance of the proposed CVMM algorithm comparing to the other methods according to their indicator values. It is imperative to note that the proposed CVMM method has the advantage of detecting outlier load patterns, although H(Average) method has a competitive performance in terms of the indicator values. In this case, the optimal number of classes is equal to twelve ($K_{opt}^{week} = 12$) according to the calculated AIC and BIC values.

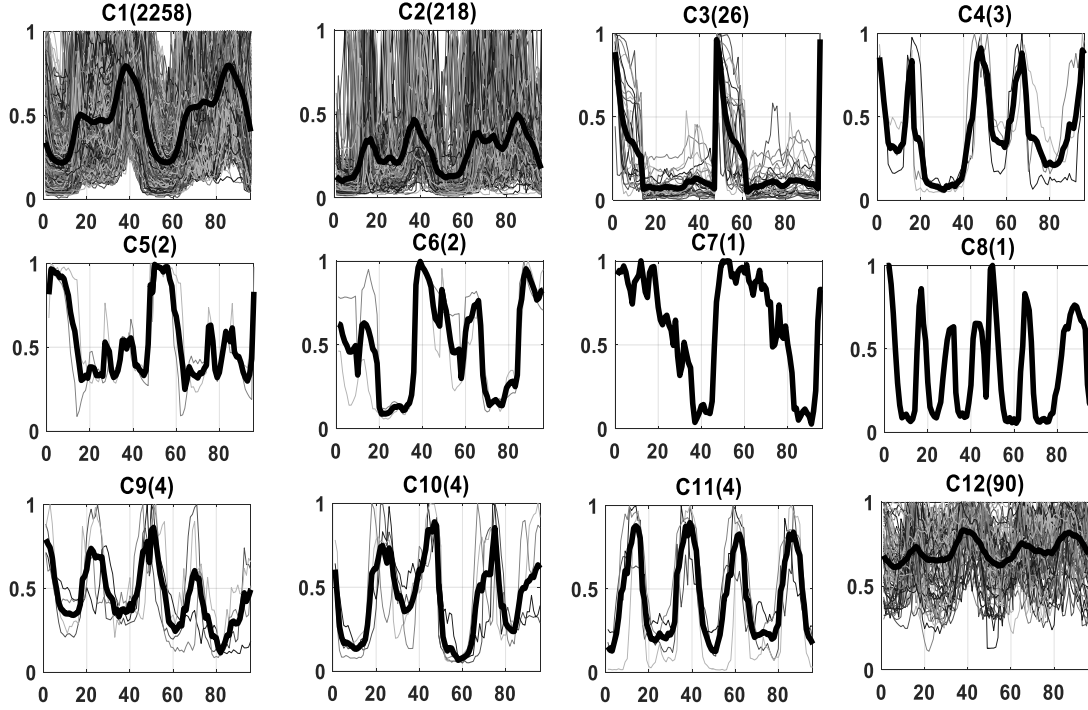


Figure 4-10. Clustering results of the combined weekdays and weekends dataset for the proposed CVMM clustering algorithm with 12 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

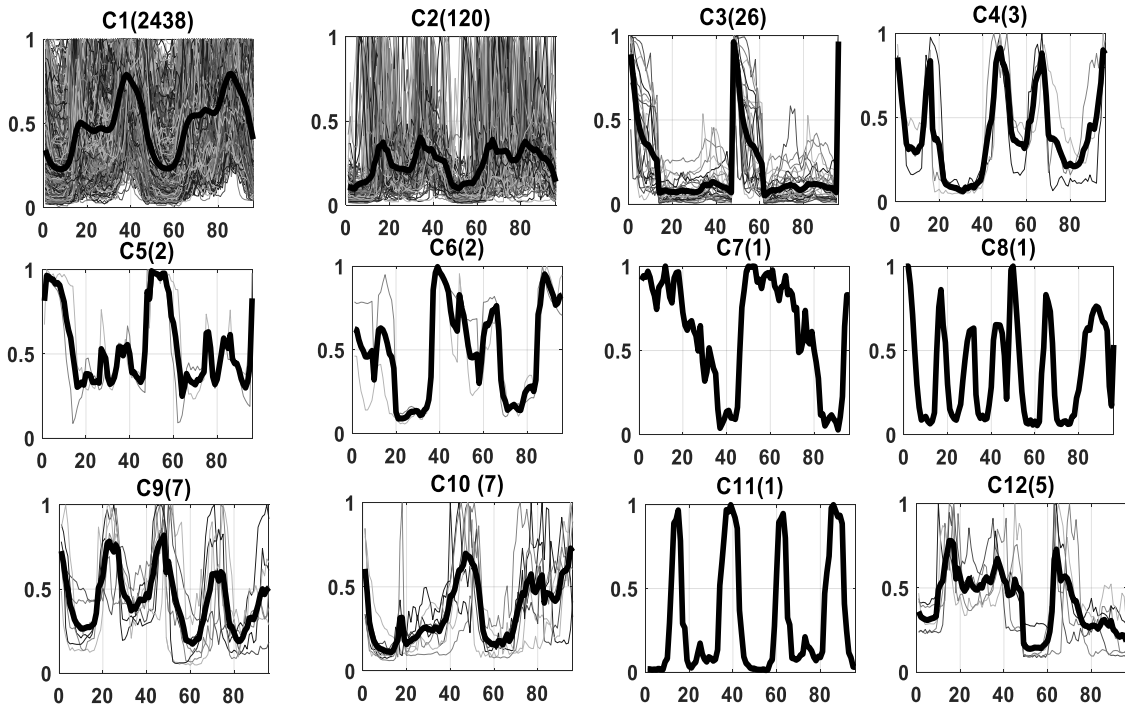


Figure 4-11. Clustering results of the combined weekdays and weekends dataset for the H(Average) clustering algorithm with 12 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series index (hour); and vertical axis indicates demand (pu).

Figure 4-10 and Figure 4-11 show the load patterns groups that were classified by the proposed CVMM clustering algorithm and H(Average) algorithm with $K = 12$, respectively. The average load pattern of all the customers in this class is plotted using a red bold line within each subplot. Visually inspecting indicates that both of these methods have the capability to identify outlying consumption behaviors. As H(Average) is the initial clustering algorithm of the proposed CVMM method, it is reasonable that the first nine clusters in Figure 4-10 and Figure 4-11 have similar average load patterns. However, comparing the last three clusters, more important load patterns are successfully detected by the proposed CVMM algorithm. In particular, the customers in C12 of Figure 4-10 exhibit consistent high consumptions during the whole week, which is important to be distinguished for the purpose of designing the tariffs.

Besides the influence of days, the impact of different calendar seasons can also be a significant consideration. As shown in Figure 4-8 (b), during the peak time between 17:00 pm and 22:00 pm, the average RLP of winter exhibits the highest normalized peak demand, whereas customers in high-summer include relatively low consumption. In addition, the average consumption autumn, spring, and summer decrease progressively. According to the previous results, H(Average) method has the most competitive performance compared with the proposed CVMM algorithm. Therefore, for different seasons' datasets, we only perform these two clustering methods with the optimal number of clusters $K=8$. The results of CVMM and H(Average) are evaluated by the aforementioned four indicators, given in Table 4-4 and Table 4-5, respectively. Comparing the criteria values across all seasons and all types of indicators, the proposed CVMM method always exhibit lower values which mean better clustering results than the H(Average) method.

Table 4-4. Performance of CVMM Across Different Seasons

	<i>CDI</i>	<i>MDI</i>	<i>DBI</i>	<i>MIA</i>
Winter	0.5976	0.8383	1.0912	0.1470
Spring	0.7173	0.8916	1.4450	0.1465
Summer	0.5486	0.8387	1.2494	0.1329
High Summer	0.5287	0.8952	1.2183	0.1257
Autumn	0.5388	0.8516	1.3003	0.1364

Table 4-5. Performance of H(Average) Across Different Seasons

	<i>CDI</i>	<i>MDI</i>	<i>DBI</i>	<i>MIA</i>
Winter	0.6322	0.8384	1.2380	0.1557
Spring	0.7212	0.9187	1.4571	0.1598
Summer	0.5601	0.8954	1.2870	0.1381
High Summer	0.5450	0.9052	1.2227	0.1340
Autumn	0.6104	0.9483	1.3942	0.1487

An example of the Autumn season is given to visually demonstrate the superior performance of the proposed CVMM method. In Figure 4-12 and Figure 4-13, the first five clusters have the similar load patterns. However, more customers within C3 and C5 are detected by the CVMM method. In addition, another type of important load pattern is identified by the proposed CVMM method, which exhibits a low consumption during the daytime and a sudden high consumption in the midnight. Furthermore, C7 in Figure 4-12 is also an outlier which is not contained in the results of H(Average). On the other hand, the outliers detected by H(Average) method include some similar shapes such as C1 and C6, C3 and C7, C4 and C8.

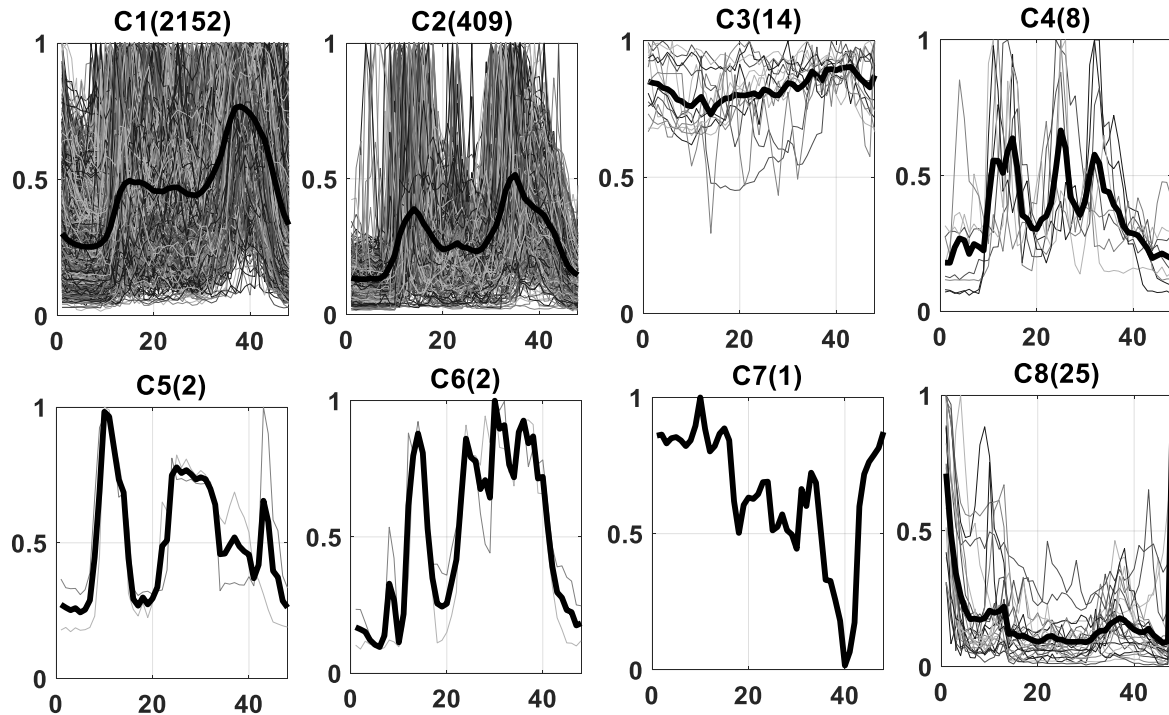


Figure 4-12. Clustering results of the Autumn dataset for the CVMM clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series (hour); and vertical axis indicates demand (pu).

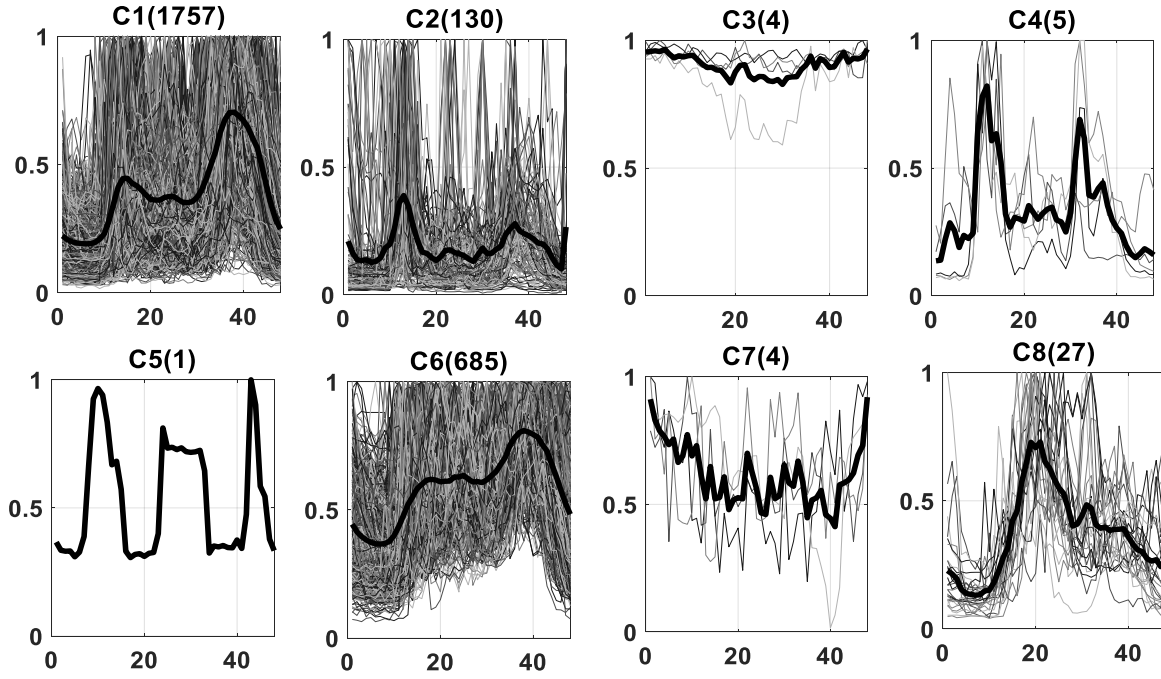


Figure 4-13. Clustering results of the Autumn dataset for the H(Average) clustering algorithm with 8 clusters. Thick black lines show mean RLP for each cluster. For each subplot, horizontal axis represents time series (hour); and vertical axis indicates demand (pu).

4.4.7 Performance for Large Data Sets

As mentioned in [8], one major direction that warrants further exploration is the development of well-performing clustering techniques capable of dealing with large amounts of smart metering data. In this section, we explore the proposed method's computational and performance scaling when dealing with a large number of customers, well above the size of the original dataset of 2,613 households. For this aim, synthetic data sets of 5,000, 10,000, 20,000, and 50,000 customers have been constructed by sampling the initial population of households. Gaussian noise has been added to the measurements to approximate the variability that would exist in larger datasets. The CPU times and CDI for all different methods are shown in Table 4-6 and

Table 4-7 respectively. Note that in the case of CVMM method two variants were carried out; 6 and 12 retained PCs to illustrate the trade-off between the computation time the clustering performance. Each method was carried out in parallel on 8 cores.

Table 4-6. Computation Times (seconds) For Different Methods

	<i>Customer Population</i>				
	2,613	5,000	10,000	20,000	50,000
CVMM-12PCs	178.55	310.71	680.26	1436.21	4027.38
CVMM-6PCs	24.50	38.85	176.95	358.6	1374.29
H(Average)	0.02	0.65	2.57	12.26	156.45
H(Weighted)	0.11	0.56	2.33	12.96	142.33
H(Complete)	0.17	0.58	2.41	13.81	143.44
H(Ward)	0.10	0.54	2.35	13.08	135.46
K-means	0.02	0.05	0.14	0.47	1.91
GMM	0.11	2.58	3.18	3.375	8.09

Table 4-7. CDI for Different Methods

	<i>Customer Population</i>				
	2,613	5,000	10,000	20,000	50,000
CVMM-12PCs	0.56	0.32	0.33	0.41	0.43
CVMM-6PCs	0.62	0.38	0.36	0.45	0.47
H(Average)	0.69	0.38	0.37	0.47	0.51
H(Weighted)	0.72	0.48	0.81	0.68	0.93
H(Complete)	0.73	0.62	0.99	1.05	1.10
H(Ward)	0.71	1.07	0.97	1.15	1.09
K-means	0.81	0.96	0.96	0.96	0.97
GMM	1.01	1.21	1.87	1.42	1.21

In terms of computation times, k-means and GMM are the fastest methods by a large margin. However, as evidenced by the high CDI criteria values shown in

Table 4-7, these conventional techniques exhibit poor clustering performance. The four hierarchical methods were found to scale well computationally with little CPU time variation between them. However, H(Average) is the best performer in terms of clustering quality, as indicated by the lower indicator values.

On the other hand, CVMM is clearly the most computationally expensive method. When 12 PCs are retained, the clustering of 50,000 customers takes more than one hour. However, the clustering quality of CVMM is substantially higher as suggested by the lower CDI values. CVMM's superior performance persists even when only 6 PCs are retained, which involves a substantial decrease in computation times. It is also constructive to highlight that CVMM's computational burden increases linearly as a function of households, which indicates good scaling performance for practical applications. We finally note that the fully-parallelizable nature of the algorithm could

be exploited further; deploying more CPU resources will result in a linear decrease of computation time, potentially allowing the clustering of millions of consumers within a few hours.

4.5 Conclusion

This chapter has proposed a novel copula-based mixture model clustering algorithm based on C-vine representations for grouping electrical customers according to consumption data. C-vine copulas, which can capture complex dependency structures by employing cascades of bivariate distribution, have been, for the first time, integrated into a mixture model-based clustering method. In order to address the high computational burden of the proposed scheme, model fitting takes place in a lower-dimension space where we exploit the synergy between PCA and the C-Vine structure's hierarchy. To evaluate the performance of the proposed CVMM algorithm, a set of well-established clustering validation indicators (CDI, MDI, DBI and MIA) have been used.

One major conclusion stemming from the analysis performed on a large dataset containing half-hourly smart meter recordings is that the CVMM method exhibits superior performance when compared to existing methods, indicated by the lower validation indicators' values. We also highlight the method's ability to detect outliers, in contrast to conventional techniques. The classification result of different techniques was further assessed according to a decision-tree based classification module. The result of the DTs' performance is highly consistent with the outcome of the clustering validation assessment. This result highlights the superiority of the proposed approach in a practical context when assigning new consumers to existing classes. In addition, the proposed method is used to analyze load partitioning behavior under different loading conditions such as different calendar seasons and days. This increases the dimensionality of the clustering problem but could uncover particular customer classes that have markedly daily and seasonal changes in their behavior, opening up the possibility for more tailored tariff structures. The superior performance of the proposed method is also well demonstrated under different conditions in terms of the indicator values and the visual comparisons.

Future research will focus on improving the proposed CVMM method by employing more possible types of vine constructions (e.g. R-vine), and exploring alternative dimension reduction techniques to further relieve the method's computational burden. Another topic of interest is the extension of the proposed model to mixed data sets so as to accommodate additional information such as load conditions; this is important to render the algorithm suitable for real-time pricing applications.

5. Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data

5.1 Introduction

When designing a distribution substation and sizing assets such as transformers and cables, the peak coincident demand across the house-holds that each asset should serve is a key consideration. Underestimating the coincident peak demand, which in Europe typically occurs during low-temperature winter days, will result in undersized assets and an inability to service load during some periods. On the other hand, over-estimating the peak coincident demand can lead to substantially cost-inefficient decisions, given that the same level of reliability could be provided with less expensive assets of reduced rating. Given that Distribution Network Operators (DNOs) have to meet strict reliability targets in combination with the inability to accurately infer the coincident peak demand of a group of consumers due to lack of instrumentation at the household level, DNOs have historically undertaken network design based on some pre-determined empirical metrics. These metrics are typically conservative to allow for worst case scenarios and have little customization related to the specific planning case at hand. The use of these general-purpose metrics is made possible because of the fact that the coincident demand of a large number of consumers' exhibits reduced sensitivity to the attributes of individual consumers, which may otherwise vary wildly when viewed in isolation. This effect is known as demand diversification and is a very important concept that pertains to electricity consumption. Demand diversity exists because the use of individual appliances in different households occurs at different times due to consumers' different schedules and preferences. It follows that accurately quantifying the effect of demand diversity is a quintessential aspect of efficient network design.

In order to measure demand diversity, several metrics have been proposed in the literature. After-diversity maximum demand (ADMD) is one of the most commonly used metrics defined as the coincident peak demand attributed to each customer, as the number of customers connected to the network approaches infinity [79]. This metric can be applied to distribution network design in a straightforward manner in order to estimate the peak coincident load for a large number of consumers [80], [81], [82] and sizing network assets [83], [84]. Another metric of interest is the coincidence factor

(CF), defined as the ratio of the coincident peak demand of a group of customers to the sum of individual peak demands in this group [85]. CF has also been widely used in the past for planning studies as in [86] and [87]. Other metrics such as the diversity factor and conversion factor have also been proposed in the past for planning purposes as showcased in [88], [89], [90].

In the UK, ADMD is the metric of choice, having been fully incorporated in DNOs' planning guidelines for asset sizing. In particular, for properties with up to 4 bedrooms and gas heating, an ADMD value of 2kW is used [91]. In case of no gas heating being available on the property, ADMD is increased to 3kW. In addition, the recommendation states that ADMD should be increased by 0.5kW for each additional bedroom. The effect of demand diversification is further taken into account by subsequently multiplying by a scaling factor to compute the coincident peak demand across n households (denoted by D_C^n):

$$D_C^n = 0.7 \times n \times ADMD \times \left(1 + \frac{12}{ADMD \times n} \right) \quad (5.1)$$

Accordingly, assuming up to four-bedroom households without electric heating, the effective ADMD value defined as $ADMD^n = D_C^n / n$ follows the curve shown in Figure 5-1.

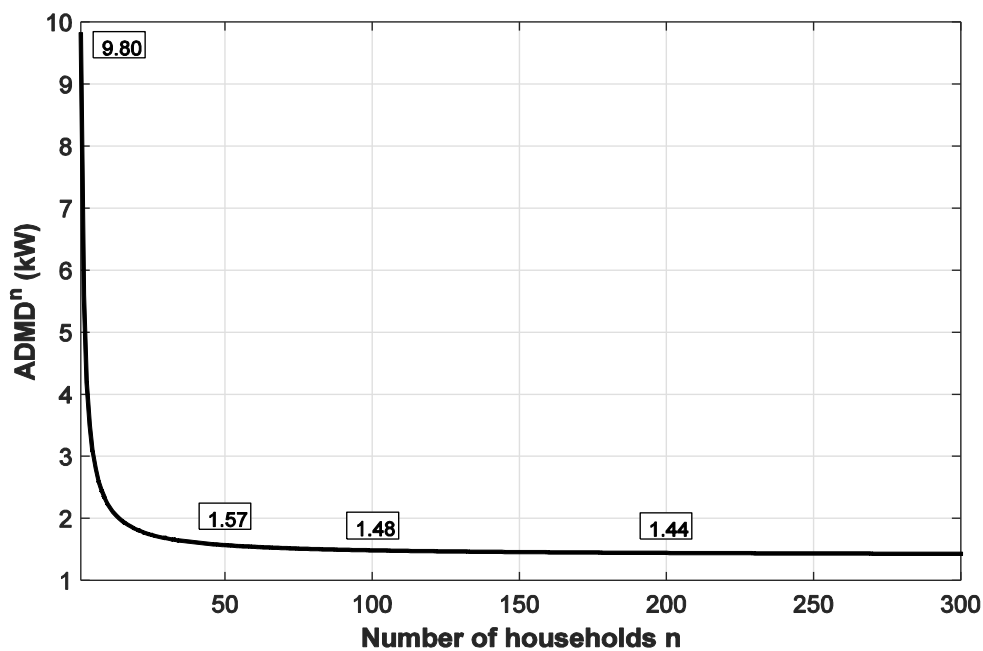


Figure 5-1. ADMD as a function of households as suggested by current UK planning methodology. For visual inspection, boxed numbers are provided for 1, 50, 100 and 200 households.

As can be seen in Figure 5-1 above, the maximum demand for a single household is set at 9.8kW; ADMD subsequently follows an exponential decrease pattern, reducing considerably in the range of 1-50 households and thereafter achieving a steady-state value that converges to 1.4kW per household. One crucial aspect to be highlighted in the overall calculation methodology is that there is little consideration of the particular details of individual households, (beyond heating type and an approximation of size) leaving little room for differentiation between areas of fundamentally different demographic makeup. It follows that this may be leading to highly conservative metrics to ensure the most adverse of possibilities are well covered, potentially resulting in asset overinvestment. The largest barrier to a transition to more custom and cost-efficient design methodologies has been the unavailability of data to experimentally calculate and validate diversified demand across different consumer groups. This reality is now being changed by the advent of smart meters creating new information streams and raising the opportunity for in-depth exploration and quantification of residential demand. To this end, data measurements from smart meter trials setup in anticipation of the UK-wide rollout target of 2020 constitute a source of valuable information.

The Low Carbon London (LCL) smart meter trial project [92] was setup as part of the UK regulator's Low Carbon Network Fund to enable DNOs to improve the understanding of the electricity consumption in the UK. Along with the collection of demand measurements, an integral part of the trial was the classification of all participating households in terms of occupancy and wealth level, allowing for a further analysis of how consumption is affected by economic factors and the number of people residing in a given property. The idea of making use of demographic data has already been proposed in the past based on the fact that different consumer categories exhibit significantly different levels of demand diversification and electric consumption intensity. The availability of this experimental data presents an unprecedented opportunity to evaluate the applicability of traditional approaches to the modern reality of electricity consumption habits and to design novel distribution network design methodologies on the basis of new information derived from high-resolution smart meter measurements and demographic data.

The chapter structure is as follows: Chapter 5.2 introduces the proposed methodology for quantifying demand diversity from experimental data. Chapter 5.3 presents the LCL dataset consisting of smart meter and demographic data; fundamental demand analyses are performed to show the main patterns of non-diversified and diversified demand. In addition, we calculate ADMD for the segregated subsets and investigate the relation wealth classes and occupancy levels. Chapter 5.4 summarizes and concludes the analysis, highlighting future potential uses of the derived metrics and insights.

5.2 Diversified Demand Quantification

Estimating the maximum coincident demand of a group of consumers is at the heart of designing efficient distribution networks. To this end, as already mentioned, ADMD is a standard metric used for planning purposes in the UK and other jurisdictions worldwide. Formally, it is defined as the coincident peak electrical demand per customer as the number of customers approaches infinity. Given that the number of customers connected to a distribution asset can vary considerably, we extend the concept of ADMD to be a function of customers connected to the network. As a result, $ADMD^n$ denotes the After Diversity Maximum Demand for n customers. It is constructive to note that, as shown in Figure 5-1, ADMD typically reaches a steady state value after a large number of households is analyzed; we denote this steady state value by $ADMD^\infty$. Its value and the number of households at which it is reached are of great interest and one of the important outputs of the present research, along with analyzing how $ADMD^n$ evolves as a function of connected households. Given a data matrix D consisting of M observations of electricity demand measurements across N consumers (size denoted $[M \times N]$), the coincident maximum demand D_C^n across n consumers is calculated according to equation (5.2). Subsequently, $ADMD^n$ is calculated by dividing the coincident peak demand by the number of customers, as shown in (5.3).

$$D_C^n = \max_{m=1,2,\dots,M} \left\{ \sum_{i=1}^n D_{m,i} \right\} \quad (5.2)$$

$$ADMD^n = D_C^{nmax} / n \quad (5.3)$$

In the cases that $n \in \{1, N\}$, the evaluation of equation (5.3) is trivial. However, in the cases that $1 < n < N$, the large number of possible ways to choose n households

among a total of N gives rise to a distribution of $ADMD^n$ values instead of a single value; in fact, there exist $n!/n! (N - n)!$ possible household combinations. Naturally, which houses are actually sampled gives rise to some variability; we communicate this variability through the explicit calculation of upper and lower bounds. To this end, in this analysis we are interested in computing the mean, minimum and maximum values, denoted by $\overline{ADMD}^n$, $ADMD_{min}^n$ and $ADMD_{max}^n$ respectively. Naturally, the maximum value is of most importance for practical planning purposes, but average and minimum values serve in gaining further insight towards diversity's degree of variability and convergence behavior. Furthermore, given that the number of combinations grows extremely fast (for example there are over 10 billion ways to choose 10 out of 50 households), a statistical approximation is necessary for its computation. To this end, random sampling techniques with replacement such as bootstrapping can be used to approximate the values of interest. In essence, instead of exhaustive enumeration we can perform the relevant calculations while only considering a smaller subset of $k \ll n!/n! (N - n)!$ combinations, where k is the number of sampled combinations. In order to compute the distribution of $ADMD^n$, a sampling algorithm is proposed as follows.

Step 1. Randomly select n households from the whole dataset D . Repeat this process k times with replacement to construct k subsets $D_C^{n(1)}$, $D_C^{n(2)}$, ..., $D_C^{n(k)}$.

Step 2. Calculate the maximum coincident demand of n household for each subset and construct a new vector $D_C^n = [D_C^{nmax}_1, D_C^{nmax}_2, \dots, D_C^{nmax}_k]$ representing the sampled population of coincident peak demand levels.

Step 3. Compute the expected, minimum and maximum value of $ADMD^n$.

The above procedure is shown in detail in Figure 5-2.

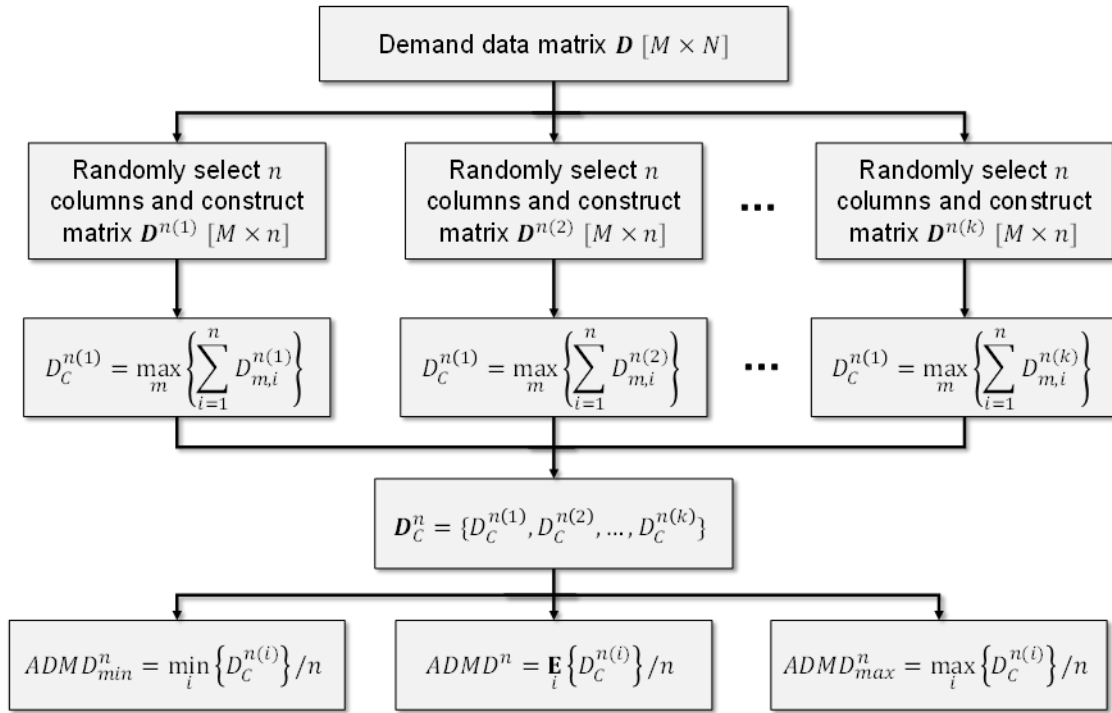


Figure 5-2. Sampling methodology for calculating the maximum, minimum and average $ADMD^n$ values from a large dataset. $E\{\}$ represents the expectation operator.

We continue with applying the presented calculation procedure to the analysis of the LCL dataset, where a k parameter of $k^* = 100,000$ has been used for the cases where $n!/n!(N - n)! > k^*$.

5.3 Analysis of Smart-metering Data

In this section, we present in detail the electricity consumption data measurements recorded by the smart meters and demonstrate how the customers have been split into categories based on property size and wealth. The peak demand distribution for all customer categories is shown and analyzed for the purpose of investigating wealth and occupancy level impact on electricity consumption in the absence of diversification. Subsequently, two analysis of ADMD are undertaken, one utilizing the entire LCL dataset and one focusing on segregated subsets, investigating the evolution of diversified demand as a function of connected households and customer classification.

5.3.1 Demand Dataset

The Low Carbon London smart meter trials aim to characterize the residential consumer demand of London and to evaluate the benefits from exploiting smart metering for distribution network design. Within the scope of the LCL project, Landis

and Gyr (L+G) E470 electricity meters were installed in 2,639 residential homes across the Mayor of London’s Low Carbon Zones (LCZ) and the London Power Networks (LPN) distribution network license area operated by UK Power Networks [92]. In particular, the Engineering Instrumentation Zones (EIZs) of the LCL trial included the areas of Brixton, Merton and Queen’s Park. The LCL demand dataset consists of half-hourly load consumption data for a full calendar year from 1st January 2013 to 31st December 2013. As such, the dataset contains 17,520 half-hourly measurements of demand across 2,639 customers in kW. In addition, various socio-economic conditions of the participated household were recorded. In this chapter we focus on the data pertaining to household occupancy and wealth level. The former relates to the number of people living in the property. The latter has been drawn on the basis of mapping all participating households to ACORN groups [93]. Subsequently, three wealth classes have been defined: Adverse, Comfortable and Affluent in increasing order. A customer category or class is defined as a combination of occupancy and wealth level. The number of customers belonging to each of the nine categories is shown in Table 5-1.

Table 5-1. Number of Participating Households Across Customer Categories

	1 occupant	2 occupants	3+ occupants
Adverse	315	278	234
Comfortable	240	304	214
Affluent	431	400	223

As can be seen above, the LCL smart-metering trials have a good representation across all household classes with several hundred customers in each of the nine categories.

5.3.2 Household peak demand analysis

The LCL dataset is summarized in Figure 5-3 where we present how the peak demand varies within each of the nine categories. We plot the probability density of peak demand of individual households in order to extract some characteristics of interest as well as explore the variability of consumption patterns among households belonging to the same class.

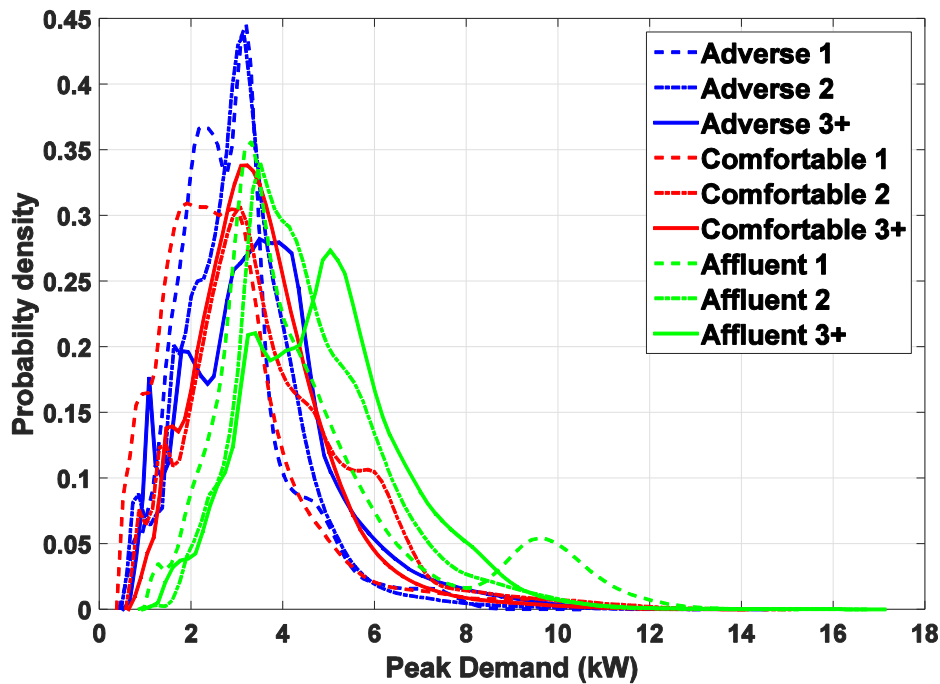


Figure 5-3. Peak demand probability distribution across the nine customer categories.

The maximum demand measurement for each customer class, equivalent to $ADMD^1$ is shown in Table 5-2. As can be seen below, the maximum demand recorded throughout the entire dataset is of 15.10kW from an Affluent 3+ household, while the group exhibiting the lowest peak demand of 6.26 kW is Adverse 1. It is imperative to highlight that these experimentally derived values do not agree with the current planning guidelines in UK, where the recommended sizing estimations for a single household of up to four bedrooms are 9.80kW and 10.50kW for properties with and without electrical heating respectively. More specifically, in the cases of Adverse 2 and all categories with 3+ occupants the data gathered from the LCL trials suggest that the guideline values are above the suggested numbers, meaning there is a real risk of asset undersizing in some cases. This may be indicating that consumer habits are changing over time, increasing the number of electrical appliances in households. In general, the patterns of maximum peak demand are as expected, with larger and wealthier households displaying increased values. Notably, there are exceptions in the case of Comfortable 2, Comfortable 3+ and Affluent 2 households, where they are slightly lower than their Adverse counterparts. These differences are relatively small and can be attributed to a single marginal case in the Adverse 2 class. These reversals indicate that worst case scenarios in the absence of diversification may be difficult to infer on the

basis of wealth data. On the other hand, household size is confirmed to be clearly correlated with maximum peak demand and can be a useful proxy for sizing connections of individual households. This is especially important in light of the ease with which household data can be obtained compared to wealth level information as well as their longer-term persistence.

Table 5-2. Maximum Peak Demand ($ADMD^1$) for Different Wealth and Occupancy Classes

	<i>1 occupant</i>	<i>2 occupants</i>	<i>3+ occupants</i>
Adverse	6.26 kW	10.53 kW	12.60 kW
Comfortable	8.98 kW	9.17 kW	11.28 kW
Affluent	9.06 kW	10.26 kW	15.10 kW

Regarding the population distributions, as can be seen in Figure 5-3, the peak demand curves exhibit a unimodal concentration towards larger values for the customers in the categories characterized by increased occupancy and wealth. There is a notable exception for ‘Affluent 1’ households, where two distinct peaks are observed. This in combination with the large range of values occupied highlights the fact that households belonging to this classification have significantly more variability among them and can potentially exhibit very high consumption levels. Further information obtained by observing the probability density plot is that the peak demands are mostly concentrated in relatively lower values ranging from 2 kW to 8 kW. Notably, there is a very low probability that an extreme high value of peak demand occurs (e.g. above 12 kW) across all categories. It is imperative to highlight that the above analysis only takes into account the peak demand of a single customer, while most of networks are designed to supply multiple customers. In this case the diversification effect becomes paramount, as shown in the following sections.

5.3.3 ADMD analysis across all participating households

In the previous subsection, we analyzed how the individual peak demand varies with occupancy and economic factors. In this section we investigate the diversified peak demand behavior as a function of connected customers. We compute $ADMD^n$ across all customers participating in the trial according to the methodology outlined in Section II and show the results in Figure 5-4. As mentioned earlier, up to 100,000 combinations have been sampled and three curves are provided. The black curve shows the maximum

$ADMD^n$ observed across the sampled combinations, denoted $ADMD_{max}^n$. Dark grey and light grey curves show the average and minimum values respectively. Note that in order to facilitate visual inspection, a plot inset has been provided zooming in the range of 1 to 500 households. This lower range can be of particular interest to distribution planners, enabling to better understand the effect of diversification when the number of customers is low and $ADMD^n$ is far from convergence. In addition, the values of $ADMD_{max}^n$ are shown in boxes for $n = 1, 50, 100, 200, 300, 400$ (inset), 500, 1000, 1500, 2000 and 2500 households.

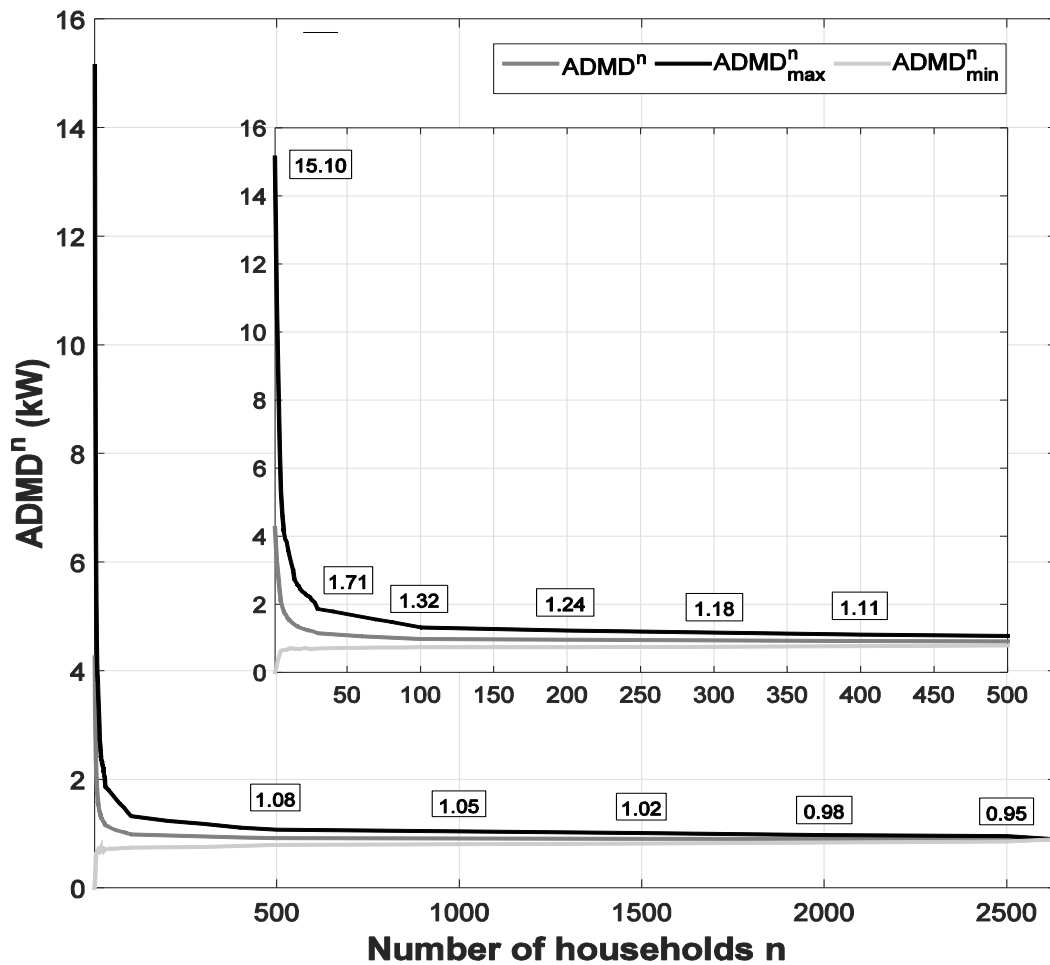


Figure 5-4. $ADMD^n$ across all participating households.

5.3.4 ADMD analysis across customer categories

In this section we investigate the diversified peak demand behavior as a function of connected customers. We compute $ADMD^n$ across the nine customer categories according to the methodology outlined in Chapter 5.2 and show the results in Figure 5-5. As before, three curves are shown to describe the evolution of upper, lower and

mean $ADMD^n$ values as a function of households; specific numbers are provided on the plots for $n = 1, 50, 100$ and 200 households. Note that $ADMD^1$ values shown on the plot are same as in Table 5-2. As before, the lower bound is the first to reach a steady state value, followed by the mean and then the upper bound. In addition, it is important to note that convergence of the upper bound to a steady-state value occurs after a few hundred households. The terminal $ADMD^\infty$ values for the nine customer categories are shown in Table 5-3.

Table 5-3. $ADMD^\infty$ for different wealth and occupancy classes based on smart-metering measurements

	<i>1 occupant</i>	<i>2 occupants</i>	<i>3+ occupants</i>
Adverse	0.51 kW	0.81 kW	1.06 kW
Comfortable	0.57 kW	0.91 kW	1.28 kW
Affluent	0.75 kW	1.07 kW	1.72 kW

Comparing $ADMD^\infty$ across different wealth and occupancy levels, it is apparent that the pattern observed in Table 5-2 does not persist. Whereas in the case of maximum peak demand, there were instances of Adverse and Affluent households exhibiting higher peak consumption than their Adverse counterparts, it is evident that diversified demand is clearly correlated with customer category. Note that this trend reversal occurs in the range of < 50 households. This is an important insight showing that coincident demand across an increased number of households exhibits reduced sensitivity to the attributes of individual consumers (that may vary wildly across a group of households) due to the effect of demand diversification. As a result, we can conclude that wealth and occupancy information can both be useful proxies in inferring diversified demand as long as the number of customers is above the order of a few tens of households. Another important observation is that houses of increasing wealth and occupancy level have a less pronounced diversification effect. This can be inferred from the curve shapes that continue to reduce considerably past the 100 households range, suggesting increased variability of consumption patterns.

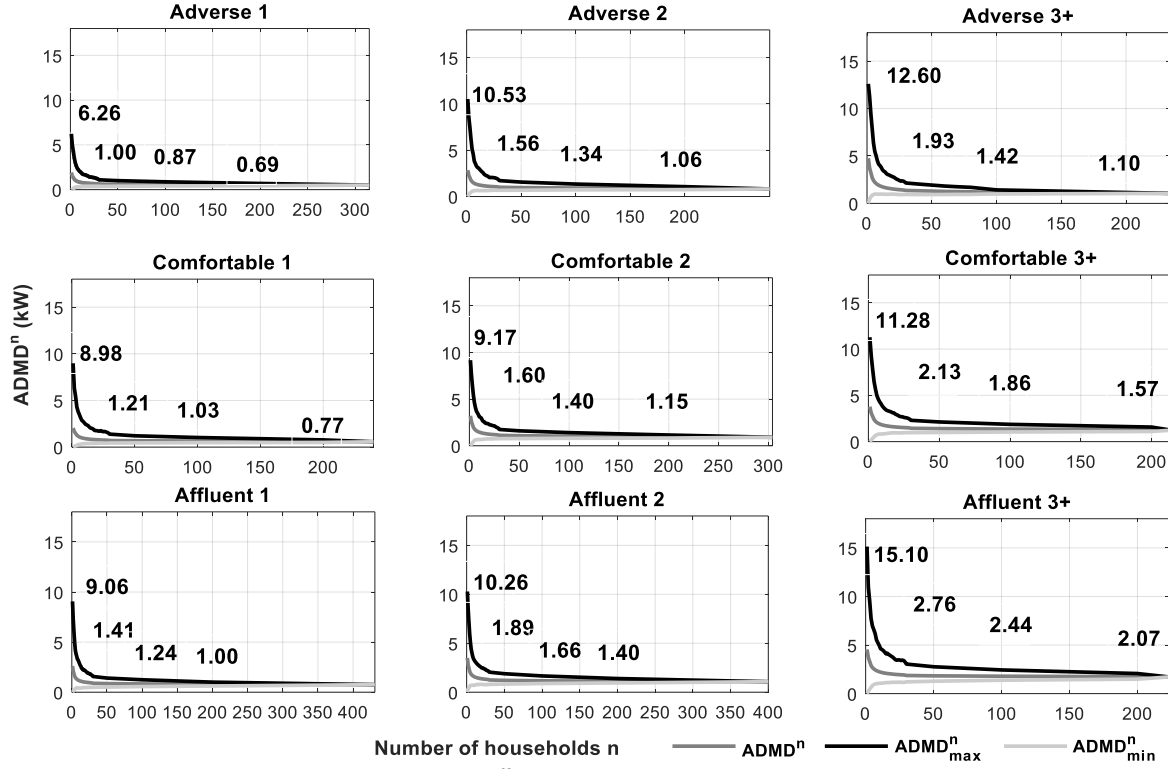


Figure 5-5. $ADMD^n$ for all nine customer categories.

5.4 Conclusions

The LCL smart meter trial provides a vast number of high-resolution residential demand measurements across 2,639 customers. The analysis of this dataset has enabled us to gain valuable insights towards the evolution of residential electricity consumption patterns in the UK and evaluate how they are affected by consumer demographic characteristics such as household occupancy and wealth level. In addition, peak and diversified demand metrics have been analysed and quantified, aiming to inform future distribution planning processes.

The chapter first introduces distribution planning practices currently applied in the UK, based on the concept of ADMD, highlighting the practical significance of accurate diversified demand metrics. Subsequently, the LCL dataset is presented and analysed. Major conclusions stemming from this analysis are that the ADMD value calculated across all customers suggests that currently adopted values may be overestimating residential demand while also overestimating the degree with which ADMD converges as a function of connected customers. In addition, household occupancy is shown to be a primary driver of peak demand of individual households. Finally, both household

occupancy and wealth level are shown to drive diversified demand, suggesting that there may be merit in novel planning methods that consider customers' demographic makeup.

Further research should propose novel methods for using the information extracted from the smart-meter data to design and operate distribution network in the smart grid era. Potential applications include the assessment of new connections for different mixes of consumers based on the detailed data presented in this chapter. In addition, further temporal analysis of the dataset can provide useful insights towards electricity consumption patterns across different calendar seasons, days and hours. The outputs of such an analysis can be used for the optimization of outage management processes to minimize service disruptions.

6. Transmission Network Expansion Planning with Stochastic Multivariate Load and Wind Modelling

6.1 Introduction

The increasing integration of large-scale intermittent power plants such as wind farms, the introduction of shiftable load elements, and the growing interconnection that characterizes electricity systems have led to significant uncertainty in planning, operation, and evaluation of transmission networks [3]. Under this new reality, transmission network expansion planning (TNEP) needs to be carried out with the consideration of uncertainties of loads and wind generators. TNEP is a nonlinear, nonconvex, and mixed integer optimization problem (MIP) problem. For practical purposes, it can be approximated via the linearized direct current (DC) power flow [94]. The main purpose of TNEP is to determine the best time, the most appropriate locations, and the optimal number of new lines that are required to be installed in an electrical network while minimizing the sum of investment and operation costs [95].

In order to solve the TNEP problem, deterministic approaches have been applied in the past [96], [97]. However, the stochastic variables beyond the operator's control (e.g. wind power) introduce substantial power flow fluctuations to the network. Consequently, it is not sufficient to perform the planning with stochastic variables only considering one power system operating profile. In recent years, a series of probabilistic TNEP models (e.g. fuzzy method [98], Monte Carlo simulation [99], and two-point estimation method [100], etc.) have been proposed to take into account the uncertainties stemming from various sources such as forced outage rates of lines [101], liberalized electricity markets [102], and wind farms [103], [104]. Furthermore, the uncertainty of both load and wind have been considered in [105], [106], [107], [94]. Nevertheless, wind and load are assumed to be uncorrelated in [105]. In [106], uncertainties of load and wind are represented by scenarios generated using clustering method. As stated in [99], entirely depending on real observations cannot provide a sufficient number of samples for Monte Carlo simulation.

A parametric model can generate samples that have similar dependence structure but not identical to what has already been encountered, thus interpolating and extrapolating the historical dataset. Another practical advantage is not having to deal with missing

data [16]. To this end, total load and available wind power are modeled via a bivariate Gaussian copula based on historical data in [99]. The samples generated from the constructed model are then used to carry out a two-stage stochastic program for transmission planning. However, although this method captures the dependence between total load and total wind output, the individual variables (electricity consumption at each bus and output of each wind farm) are assumed to be perfectly correlated. This is a substantial simplification undertaken in the interest of relieving the increased computational load that comes with modelling the stochastic variables at the level of individual variables instead of their aggregation. However, capturing the interspatial correlations of load and renewables output at various places can be an important factor for investment planning.

In the past multivariate statistical models have been widely used to model stochastic signals. For example, in [108] power injections from renewable sources of energy are modeled using Vector Autoregressive models to generate multivariate synthetic wind speed time series based on empirical observations. In [109], distribution system loads are represented through a Gaussian mixture model (GMM) whose parameters are estimated by the expectation-maximization (EM) algorithm.

In general, one of the most highly challenging tasks of a multivariate statistical framework, is its ability to model large-scale measurement data by accurately capturing the non-linear dependence structure of the data as well as their non-Gaussian marginal distributions. To this end, a vast number of applications have demonstrated that copulas are capable of accomplishing this task. In [110], a multivariate Gaussian copula is used to fit a model based on historical data and to generate synthetic wind power output from 15 sites in the Netherlands. Beyond that, a more advanced multivariate modeling framework based on vine copulas has been proposed in [111] to capture spatiotemporal interdependencies of wind power and to determine energy storage requirements. Although the increased computational burden has been resolved through a systematic truncation method, it is still impractical for the modelling of high-dimensional data.

In the present chapter, we propose a stochastic TNEP model considering multivariate dependent load and wind modeling based on a multivariate Gaussian copula. Load and wind generation measurements from French electricity system are mapped to the IEEE

118-bus test system, and modeled by the proposed multivariate Gaussian copula. Using the load and wind power samples generated from the constructed model, an appropriate transmission plan for a future year can be obtained by carrying out a Monte Carlo simulation. The proposed TNEP model is solved by using Benders decomposition. The solution quality of the proposed model is evaluated and compared to some conventional methods through constructing a confidence interval on the optimality gap. Results indicate that capturing spatial dependencies at the level of individual variables can have a substantial impact on transmission planning decisions. The proposed framework is shown to outperform existing approaches in this regard.

The outline of the remainder of this chapter is as follows: Chapter 6.2 introduces the procedure of stochastic dependence modeling using multivariate copulas. In Chapter 6.3, we present the mathematical formulation of the proposed TNEP model. Chapter 6.4 presents the simulation result of the proposed method. A comparison between the results of different methods is also performed here. Chapter 6.5 summarizes and concludes the chapter.

6.2 Stochastic Multivariate Dependence Modelling Using Copulas

Dependence modelling of stochastic variables in power systems faces two significant challenges; capturing non-linear dependence structures and non-Gaussian marginal distributions. In Figure 6-1. We show histograms and bivariate scatter plots of five stochastic variables taken from the French transmission system; variables L1, L2 and L3 are loads at different locations while W1 and W2 are wind injections from two wind farms.

It can be observed that stochastic variables have highly non-normal marginal distributions (diagonal histograms) and non-linear dependencies (scatter diagrams) [17]. The dependence between the different loads, between the outputs of the two wind farms as well as the dependence between individual loads and wind power injections are highly complex and non-standard. As such, it is of great potential benefit to move beyond the assumption of perfect correlation between loads at different locations etc. and investigate the dependence structure at the level of disaggregated variables in more detail. The first step towards this direction is being able to fit a model that can accurately

capture all these characteristics, which can then be subsequently used as an engine for the generation of a large number of stochastic scenarios.

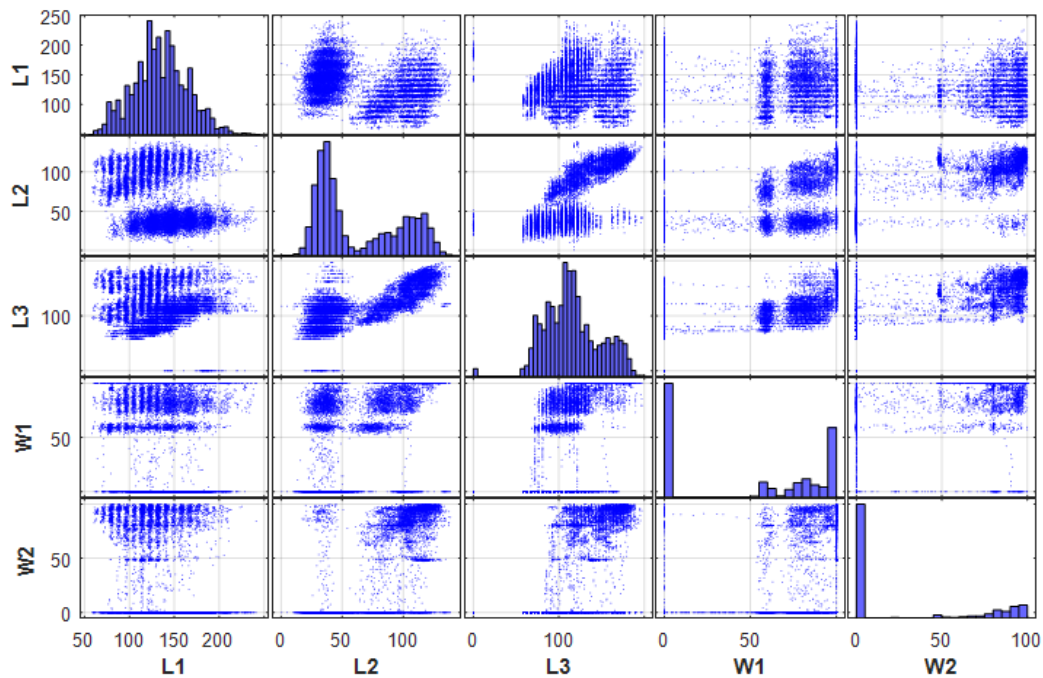


Figure 6-1. Histograms of marginals distributions and bivariate scatter plots of active loads (MW) and wind power output (MW) for randomly selected buses in France, measured between Jan 2012 and Feb 2012.

Generally speaking, there are three approaches to carrying out statistical modelling: 1) Nonparametric modelling; 2) Parametric modelling; 3) Semiparametric modelling. In general, nonparametric modelling techniques assume that the data do not have a characteristic underlying structure. As a result, such methods do not rely on any specific probability distributions, but the model is constructed directly from the data instead. On the other hand, a parametric model is a distribution that can be described using a finite number of parameters and relies on an analytic description of the joint probability distribution. Finally, semiparametric modelling involves constructing a model that has both parametric and non-parametric components. This is a suitable model for dealing with multi-dimensional data where a combination of different techniques may be essential to capture the important data characteristics. In the present research, the model of choice is semi-parametric due to its primary means of capturing the dependence structure of stochastic variables. The procedure of modelling load and wind based on copulas is presented in the following parts.

6.2.1 Probability Integral Transformation

In this chapter, probability density functions and cumulative distribution functions are denoted by f and F , respectively, and their copula versions by c and C . Consider m random variables $X = (X_1, X_2, \dots, X_m) \in \mathbb{R}^m$ with marginal density functions $f_i(x_i)$ and distribution functions $F_i(x_i)$, for $i = 1, \dots, m$. For each variable X_i , its corresponding variable U_i on $[0,1]$ can be obtained via the empirical cumulative distribution function (ECDF) transformation, mathematically formulated as:

$$U_i = F_i(X_i) \quad (6.1)$$

where U_i is uniformly distributed on $[0,1]$. Conversely, given U_i and the inverse ECDF, $F_i^{-1}(\cdot)$, the random variable X_i can be obtained by:

$$X_i = F_i^{-1}(U_i) \quad (6.2)$$

Using the equations (6.1) and (6.2), samples of stochastic variables with empirical marginal distributions of historical measurements can be generated for Monte Carlo simulation. Examples of this transformation are presented in Figure 6-2 based on the historical data shown in Figure 6-1. Also, the non-Gaussian margins of the examples show the benefits of using ECDFs instead of fitting parametric distributions.

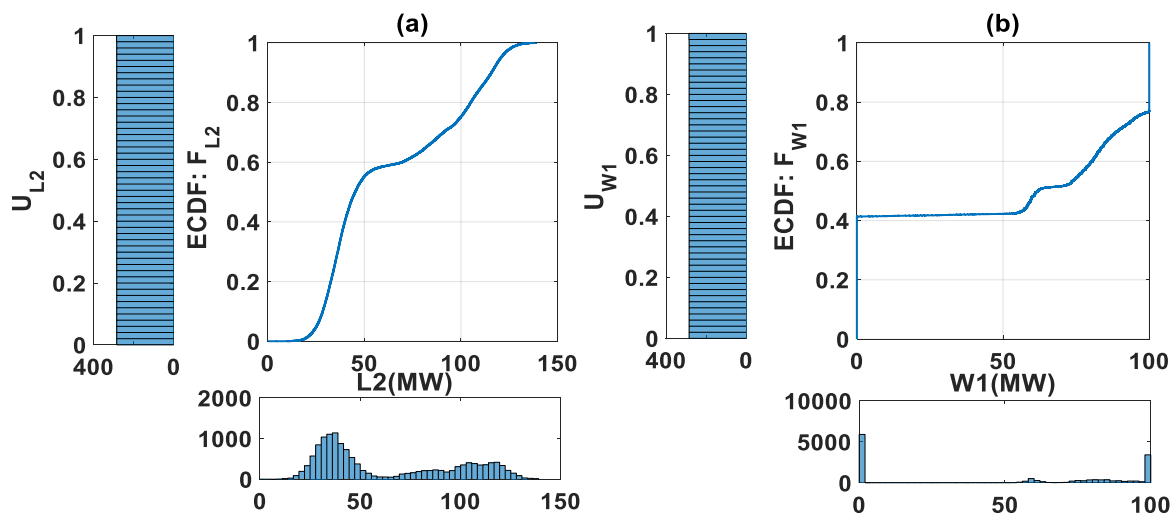


Figure 6-2. ECDF transformation of L2 (a) and W1 (b).

6.2.2 Multivariate Gaussian Copula

According to Sklar's theorem [112], the joint density function of m random variables $X = (X_1, X_2, \dots, X_m)$ is given by:

$$f(\mathbf{x}_1, \dots, \mathbf{x}_m) = \left(\prod_{i=1}^m f_i(\mathbf{x}_i) \right) c_{1\dots m}(F_1(\mathbf{x}_1), \dots, F_m(\mathbf{x}_m)) \quad (6.3)$$

where the function $c_{1\dots m}: [0,1]^m \rightarrow \mathbb{R}$ is an m -dimensional copula with uniform marginals $\{U_1, U_2, \dots, U_m\} = \{F_1(X_1), F_2(X_2), \dots, F_m(X_m)\}$. This way, even though a joint density function contains information on both the individual marginal distributions and on the relation between the variables, copulas provide a way for isolating the dependency structure. In general, a copula is an m -dimensional parametric function defined on the $[0,1]^m$ space and describing the dependency between m variables. Copulas have the convenient property that each of their marginal distributions are the uniform distribution on $[0,1]$. Moving beyond the above definition, there are many different types of copulas; empirical, parametric, time copulas etc. For the purposes of this research, we will focus on parametric copulas, which are copula functions that can be fully described by a family and corresponding parameters. The idea is to identify the family and parameters that give rise to a copula that best matches the dependency of historical variables.

In terms of the families of copulas, we use a multivariate Gaussian copula to model the stochastic variables. As demonstrated in [104], a Gaussian copula can efficiently capture the first-order notion of dependence. The Gaussian copula is also called normal copula and is usually signified by its defining distribution $N_m(0, R)$, where m is the number of dimensions ($m = 2$ for bivariate) and R is the correlation matrix. Gaussian copula is one kind of elliptical copula that only has mean dependence with no upper or lower tail dependencies. The range of pairwise correlations is $\rho \in [-1, 1]$. Mathematically, when using a multivariate Gaussian copula, equation (6.3) can be written as:

$$f(\mathbf{x}_1, \dots, \mathbf{x}_m) = \left(\prod_{i=1}^m f_i(\mathbf{x}_i) \right) |R|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} \mathbf{z}^T (R^{-1} - I) \mathbf{z} \right\} \quad (6.4)$$

where $z_i = \Phi^{-1}(F_i(X_i)) \sim \mathcal{N}_1(0, 1)$, $\mathbf{z} = (z_1, \dots, z_m)$, and Φ is the standard Gaussian cumulative distribution function. A visual illustration of the bivariate Gaussian copula is given in Figure 6-3.

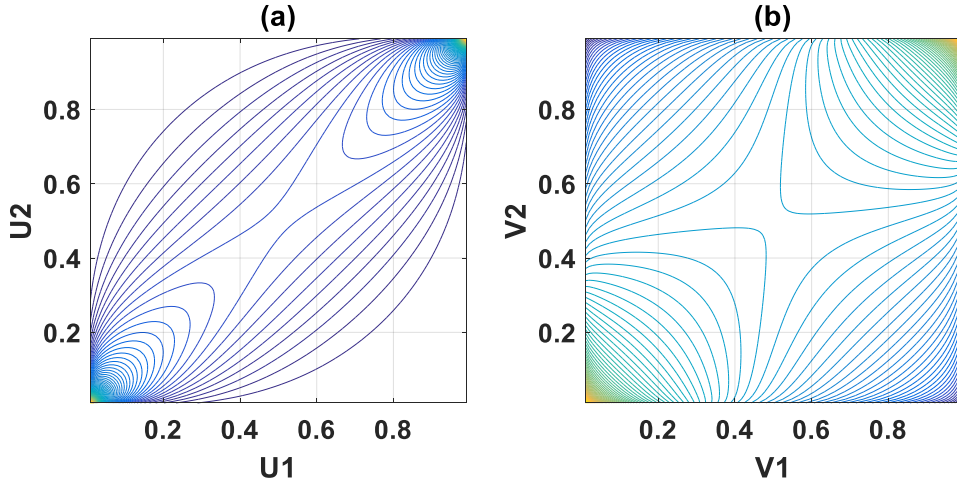


Figure 6-3. Examples of bivariate Gaussian copulas with parameter $\rho = 0.8$ (a) and with parameter $\rho = 0.2$ (b).

To generate dependent load and wind observations, the procedure for parameterizing and sampling from a multivariate Gaussian copula are as follows:

- 1) Using equation (6.1) to transform marginal distributions of historical variables from actual domain to the uniform domain in $[0,1]$ through the ECDFs.
- 2) Build the multivariate Gaussian copula model based on uniformly distributed variables. Parameters of the constructed model can be estimated by calculating the correlation matrix, as introduced in [110].
- 3) Generate sampled variables in $[0,1]$ and transform back to actual domain by using equation (6.2), the inverse ECDFs of historical data.

Note that the parameterized model is capable of providing as many samples as required in the Monte Carlo simulation for solving the TNEP problem.

6.3 Proposed TNEP Model

6.3.1 Mathematical Formulation

The TNEP problem aims to minimize the total system cost by determining the optimal transmission expansion decisions along with the optimal operational decisions. Taking into account inter-spatial correlations between load/wind farms across various locations, the proposed TNEP model can be formulated as follows:

$$\min \sum_{i,j \in \Lambda_{new}} C_{ij} n_{ij} + \sum_{t=1}^T \tau \cdot p^t \cdot \Xi(n_{ij}, D^t, P_W^{t,max}) \quad (6.5)$$

s.t.

$$n_{ij} \in \{0, 1\}, \quad \forall i, j \in \Lambda_{new} \quad (6.6)$$

where

$$\Xi(n_{ij}, D^t, P_W^{t, max}) = \min \sum_{g \in G} \pi_g \cdot P_g^t + \sum_{d \in Nd} \gamma \cdot U_d^t \quad (6.7)$$

s.t.

$$S(P_\Gamma^t + P_W^t + U^t - D^t) = F \quad (6.8)$$

$$0 \leq P_W^t \leq P_W^{t, max} \quad (6.9)$$

$$0 \leq P_\Gamma^t \leq P_\Gamma^{max} \quad (6.10)$$

$$0 \leq U^t \leq D^t \quad (6.11)$$

$$-f_{ij}^{max} \leq f_{ij}^t \leq f_{ij}^{max}, \quad \forall i, j \in \Lambda \quad (6.12)$$

$$f_{ij}^t = B_{ij}(\theta_i^t - \theta_j^t), \quad \forall i, j \in \Lambda \quad (6.13)$$

$$\begin{aligned} -n_{ij}f_{ij}^{max} \leq f_{ij}^t \leq n_{ij}f_{ij}^{max} \\ : \bar{\mu}_{ij}^t, \underline{\mu}_{ij}^t, \quad \forall i, j \in \Lambda_{new} \end{aligned} \quad (6.14)$$

$$\begin{aligned} -M(1 - n_{ij}) \leq f_{ij}^t - B_{ij}^{new}(\theta_i^t - \theta_j^t) \leq M(1 - n_{ij}) \\ : \bar{v}_{ij}^t, \underline{v}_{ij}^t, \quad \forall i, j \in \Lambda_{new} \end{aligned} \quad (6.15)$$

In the above model, transmission expansion decisions are determined in the first stage, represented by binary variables, n_{ij} . In the second stage, the linear objective function includes generation costs as well as the cost of load curtailment, as presented in (6.7). In (6.5), the expected operating cost, $\Xi(n_{ij}, D^t, P_W^{t, max})$, in the proposed model is approximated by the products of the operating hours (τ), the probability (p^t) of scenario t , and the optimal solution of (6.7), summed across all operating scenarios t . Note that in this case, $p^t = 1/T$, where T is the number of samples being considered. The branch-node incidence matrix $\mathbf{S}$ is built based on the configuration of the network. In addition, constraint (6.8) represents power balance including load curtailment. Maximum capacity limits of thermal and wind generators are given in (6.9) and (6.10). Load curtailment $\mathbf{U}$ is limited by the sampled load $\mathbf{D}$, as shown in (6.11). Finally, constraints (6.12)-(6.15) denote thermal capacity limits as well as DC power flow constraints for existing and candidate transmission lines. In particular, dual variables

$\bar{\mu}_{ij}^t, \underline{\mu}_{ij}^t, \bar{v}_{ij}^t, \underline{v}_{ij}^t$ are extracted from constraints (6.14) and (6.15). Also, M is a large value for the disjunctive constraint (6.15).

6.3.2 Benders' Decomposition

Benders' decomposition [113] is one of the most widely-used techniques to solve the mix integer problem. In this case, our proposed TNEP model can be decomposed into a master problem and a subproblem via Benders' decomposition. Note that a multi-cut formulation has been adopted, as in [114]. Let $\kappa \in K$ denote the index iterations, the master problem at the κ th iteration can be presented as:

$$\min \sum_{i,j \in \Lambda_{new}} c_{ij} n_{ij} + z \quad (6.16)$$

s.t.

$$n_{ij} \in \{0, 1\}, \quad \forall i, j \in \Lambda_{new} \quad (6.17)$$

$$z \geq \sum_{t=1}^T \tau \cdot p^t \cdot \{\hat{y}_{\kappa}^t + \sum_{i,j \in \Lambda_{new}} [(\bar{\mu}_{ij}^{t\kappa} + \underline{\mu}_{ij}^{t\kappa}) f_{ij}^{max} - M(\bar{v}_{ij}^{t\kappa} + \underline{v}_{ij}^{t\kappa})] \cdot (n_{ij} - \hat{n}_{ij}^{\kappa})\}, \quad \forall \kappa \in K \quad (6.18)$$

$$z \geq 0 \quad (6.19)$$

where $\hat{n}_{ij}^{\kappa}$ is the solution of the master problem at iteration κ , $\hat{y}_{\kappa}^t = \Xi(\hat{n}_{ij}^{\kappa}, D^t, P_W^{max})$ and $\bar{\mu}_{ij}^{t\kappa}, \underline{\mu}_{ij}^{t\kappa}, \bar{v}_{ij}^{t\kappa}, \underline{v}_{ij}^{t\kappa}$ is the optimal operation cost and dual variables of the subproblem at iteration κ under scenario t . z is an approximation of the second stage costs. For each scenario, the objective function of the subproblem is (6.7) with it constraints from (6.8) to (6.15). Note that the binary variables in (6.14) and (6.15) are also replaced by $\hat{n}_{ij}^{\kappa}$. The procedure of solving the proposed TNEP problem by using Benders' decomposition can be concluded as follows. Firstly, we initialize $\varepsilon = 0.02$, feasible integer solution $n_{ij} = 1, \quad \forall i, j \in \Lambda_{new}$, lower bound $LB = -\infty$, upper bound $UB = \infty$, and iteration $\kappa = 0$. After that the solutions of the subproblem with the above initial parameters are used as inputs to solve the master problem to get $\hat{n}_{ij}^{\kappa}$ and z^{κ} . Accordingly, the lower bound for the κ th iteration can be updated as $LB = \sum_{i,j \in \Lambda_{new}} c_{ij} \hat{n}_{ij}^{\kappa} + z^{\kappa}$. The next step is to solve the subproblem using $\hat{n}_{ij}^{\kappa}$ to get the optimal operation cost $\hat{y}_{\kappa}^t$ and dual variables $\bar{\mu}_{ij}^{t\kappa}, \underline{\mu}_{ij}^{t\kappa}, \bar{v}_{ij}^{t\kappa}, \underline{v}_{ij}^{t\kappa}$. Also, the upper bound is

updated by $UB = \min \{UB, \sum_{i,j \in \Lambda_{new}} C_{ij} n_{ij} + \sum_{t=1}^T \tau \cdot p^t \cdot \hat{y}_\kappa^t\}$. If the difference between LB and UB is lower than a determined value (e.g. $\varepsilon \cdot \min \{UB, LB\}$), the iteration should be terminated. Otherwise, another benders' cut (6.18) is built and appended to the master problem for the $(\kappa + 1)$ th iteration.

6.4 Simulation Study and Results Analysis

In this section, the performance of the proposed TNEP model along with the multivariate Gaussian copula modelling framework is demonstrated based on a modified version of the IEEE 118-bus system. As detailed in [104], a confidence interval (CI) on the optimality gap of the problem's solution is constructed to determine the solution quality of different planning models.

6.4.1 Test System and Multivariate Load and Wind Modelling

The original IEEE 118-bus system consisting of 54 generators and 186 lines has been modified to also include 10 wind farms of size 100MW as well as 10 more candidate transmission lines with capacity 500 MW. For the historical database of stochastic variables, we map 118 demand buses and 10 wind generators onto the test system from historical load and renewable injection measurements that were provided by RTE, the French Transmission System Operator. The original library includes 14,251 measurements at 15-minute intervals over the course of two months, from January to February 2012, and spans over 7,000 load buses and 200 wind plants. In this chapter we focus on a subset; a 128-dimensional dataset consisting of 118 demand buses and 10 wind generators chosen at random. The process of mapping the selected variables is to scale them by a constant that obtained by calculating the ratio between the maximum coincident peak demand of the original data and the sum of active power demand across all buses defined in the system. In this way, the complex dependence structure between the stochastic variables can be maintained. Also note that in this research, value of lost load Υ is set to \$4,000/MWh. The annualized investment cost for building one candidate transmission line is $C_{ij} = \$5,000,000/\text{year}$. Also, we use $\tau = 8,760$ and $M = 20,000,000$.

To illustrate the comparative advantage of sampling stochastic scenarios using a multivariate copula model that captures inter-spatial correlations, we generate the load and wind samples by the following methods for comparison;

- 1) Consider total load and total wind power as two independent variables (**BiInd**);
- 2) Consider total load and total wind power as two dependent variables and model them via a bivariate Gaussian copula (**BiGc**);
- 3) Consider load and wind power as multivariate dependent variables and model them via a multivariate Gaussian copula (**MGC**).

Note that in methods 1 and 2, load and wind at different locations are assumed to be perfectly correlated, respectively, which means static sharing factors between each bus/wind farm are assumed in these cases.

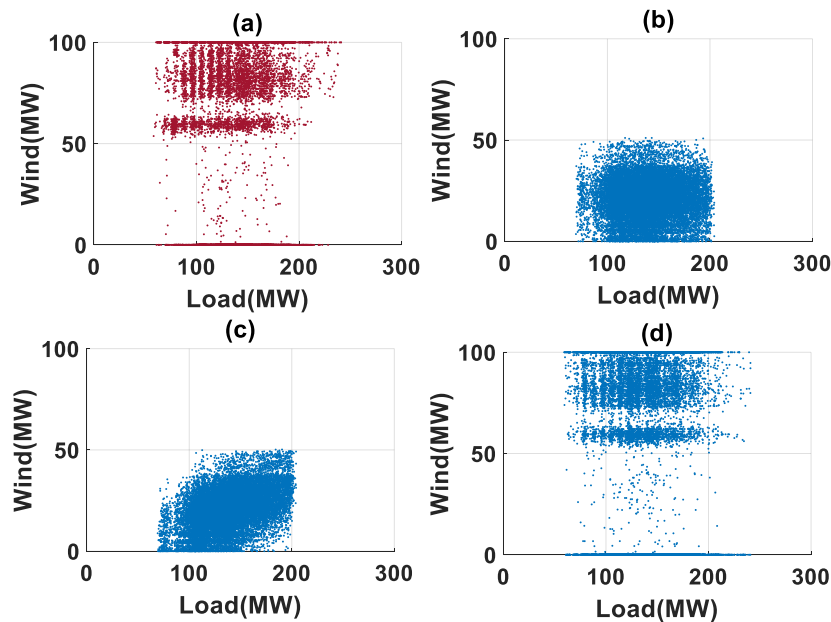


Figure 6-4. Scatter plots of historical data (a), scatter plots of generated data using BiInd method (d), BiGc method (c) and the proposed MGC method (d).

An example of comparing historical ($N_{hist} = 1,4251$) and sampled ($N_s = 2,0000$) data generated by the three different methods is shown in Fig. 4 via the scatter plot of a single load variable and a single wind infeed variable chosen at random. In Fig.4 (a), the point-cloud of historical data is shown to be composed of several sub-groups of highly non-linear dependence and complex data features. Moreover, the sampled data points in Figure 6-4 (b) and Figure 6-4 (c) are biased towards the bottom with almost zero and non-zero linear correlations, respectively, resulting from the assumption of the

static sharing factors. The superiority of the proposed MGC method (Figure 6-4 (d)) compared to the BiInd method (Figure 6-4 (b)) and the BiGc method (Figure 6-4 (c)), is clearly reflected in the reconstruction quality of the dependence structure. To this end, it is demonstrated that the disaggregated correlations between load/wind can be successfully retained in the sampled data by using the MGC method.

6.4.2 Solution Quality Evaluation

We proceed by assessing the impact that the proposed stochastic modelling methodology has on the TNEP problem. For this purpose, we make use of a Monte Carlo bounding technique originally proposed in [115] and implemented in [99] to calculate a confidence interval (CI) on the optimality gap of a candidate expansion plan. In this case, the optimality gap is denoted by the difference between a candidate solution $\hat{\psi} = \sum_{i,j \in \Lambda_{new}} C_{ij} \hat{n}_{ij} + \sum_{t=1}^T \tau \cdot p^t \cdot \mathbb{E}(\hat{n}_{ij}, \mathbf{D}^t, \mathbf{P}_W^{t, max})$ and an optimal solution ψ^* . In this way a 95% CI can be constructed by $P\{\hat{\psi} - \psi^* \leq \varpi\} \approx 0.95$, where ϖ is the CI width; a lower ϖ value indicates a better solution quality. For the proposed TNEP model with MGC modelling, the procedure of calculating the CI width ϖ is briefly summarized as follows.

Step 1. Lower bound $\underline{L}$ construction: Determine the number of samples $N_s = 500$ as well as the number of replications $N_L = 50$. For each replication n_L , we generate N_s samples for load $\mathbf{D}$ and wind $\mathbf{P}_W^{max}$ of size 500×118 and 500×10 , respectively, and then solve the TNEP problem to get an optimal solution $\hat{n}_{ij}^{n_L}$ and the corresponding optimal value $\hat{\psi}^{n_L}$. Then the lower bound can be calculated by $\underline{L} = \frac{1}{N_L} \sum_{n_L=1}^{N_L} \hat{\psi}^{n_L}$, with its sampling error $e_L = std(\hat{\psi}^{n_L}) \frac{t_{N_L-1, \alpha/2}}{\sqrt{N_L}}$, where $std(\cdot)$ is the standard deviation, and $t_{N_L-1, \alpha/2}$ is defined as the $(1 - \alpha/2)$ level quantile of a Student's t-distribution with $N_L - 1$ degrees of freedom.

Step 2. Candidate solution $\hat{n}_{ij}$ selection: Use the built MGC model to generate $N_c = 1000$ sampled load $\tilde{\mathbf{D}}$ and wind $\tilde{\mathbf{P}}_W^{max}$, and solve the TNEP problem with the solution $\hat{n}_{ij}^{n_L}$ of each replication in Step 1. The candidate solution can be obtained by $\hat{n}_{ij} \in \operatorname{argmin}_{\hat{n}_{ij}^1, \dots, \hat{n}_{ij}^{N_L}} \{\sum_{i,j \in \Lambda_{new}} C_{ij} \hat{n}_{ij}^{n_L} + \sum_{t=1}^T \tau \cdot p^t \cdot \mathbb{E}(\hat{n}_{ij}^{n_L}, \tilde{\mathbf{D}}^t, \tilde{\mathbf{P}}_W^{t, max})\}$.

Step 3. Upper bound $\bar{U}$ construction: Perform the simulation with $N_U = 20,000$ load and wind samples, $\bar{\mathbf{D}}$ and $\bar{\mathbf{P}}_w^{max}$. The upper bound can be constructed by $\bar{U} = \frac{1}{N_U} \sum_{n_U=1}^{N_U} \hat{\psi}^{n_U}$ with the sampling error $e_U = std(\hat{\psi}^{n_U}) \frac{t_{N_U-1, \alpha/2}}{\sqrt{N_U}}$. Note that we have $\hat{\psi}^{n_U} = \sum_{i,j \in \Lambda_{new}} C_{ij} \hat{n}_{ij} + \tau \cdot \Xi(\hat{n}_{ij}, \bar{\mathbf{D}}^t, \bar{\mathbf{P}}_w^{t, max})$.

Step 4. CI width ϖ calculation. The final CI width can be calculated by $\varpi = e_U + e_L + [\bar{U} - \underline{L}]^+$.

Detailed information of the above steps is illustrated in [99]. In addition, the CI widths of the TNEP solutions using BiInd and BiGc modelling algorithms are also calculated following the above procedures with the results shown in Table 6-1. Note that the modelling methods (i.e. BiInd, BiGc, and MGC) and the TNEP optimization problem were implemented in MATLAB and FICO Xpress, respectively, and run on an Intel Xeon E5-2690 PC with 8 cores. For comparison, the time consumption of these three methods for generating 500 samples are given in Table 6-2.

Table 6-1 Solution Quality Evaluation for Different Methods (500 Samples for Each Method)

	<i>Investment cost (\$ million/year)</i>	$\bar{U}$ (\$M/year)	$\underline{L}$ (\$M/year)	ϖ
BiInd	30	1,727.41	1,684.72	231.78
BiGc	30	1,625.93	1,649.83	57.83
MGC	15	<u>1,201.87</u>	<u>1,195.76</u>	<u>21.60</u>

Table 6-2 Computation Times for Different Methods (500 Samples for Each Method)

	<i>CPU Times (s)</i>
BiInd	2.24
BiGc	3.12
MGC	5.21

When using the proposed MGC method, only three lines are required to be built, whereas the other two methods both cost double (i.e. \$30 million/year) in the investment stage. Furthermore, regarding the upper bound $\bar{U}$ and the CI width ϖ , the superiority of the proposed MGC method can be well demonstrated by exhibiting much lower cost values than the other two methods. Compared to the BiInd method, there are approximately 6% and 75% expected cost reductions in $\bar{U}$ and ϖ , respectively, when taking into account the bivariate dependence between total load and total wind.

Most importantly, a substantial improvement on the solution variability (i.e. 30% lower $\overline{U}$ and 91% lower ϖ) is obtained by capturing the inter-spatial dependence between stochastic variables through the MGC. The other methods examined give rise to unrealistic and highly divergent scenarios which lead to inefficient investment decisions. The gradual improvement, from independent to bivariate to multivariate and from aggregated to disaggregated variables that pertain to the specific network locations highlights the importance of capturing stochastic dependence at the level of individual variables when performing transmission expansion planning. The amount of operational variability and importance of capturing spatial correlation is bound to increase in the future due to higher penetration of renewables and the looming electrification of the transport and heating sectors [116].

6.5 Conclusions

The present chapter proposes a two-stage TNEP model integrated with a multivariate Gaussian copula modelling framework. A modified version of IEEE 118-bus system is used as the test network in this research. Based on real historical stochastic variables from the French transmission system, multivariate dependent load and wind samples of 128 dimensions are generated by the proposed MGC algorithm, and compared with the results of BiInd and BiGc methods. A visual comparison between these three methods shows that the inter-spatial dependence can be re-constructed much more accurately when employing a multivariate Gaussian copula model. Furthermore, to evaluate the solution quality, a Monte Carlo bounding technique is implemented to build a 95% CI on the optimality gap of the solution of each method. The results of the comparison suggest that the CI width and the expected total cost of the TNEP problem can be substantially reduced when considering the multivariate dependence of load and wind.

Further research will focus on applying some more advanced high-dimensional modelling techniques in the TNEP problem to deal with more complex dependence structure between stochastic variables. In the future, contingency constraints could also be appended to the TNEP model. Another application of interest is to solve a multistage TNEP problem considering load and wind farm uncertainties.

7. A Target-Based Scenarios Selection Approach for Transmission Network Expansion Planning

7.1 Introduction

Decarbonization of electricity system is expected to be achieved by the high penetration of renewable energy sources (RES). Large scale transmission network expansion is required to facilitate the integration of the remotely located RES. Transmission network expansion planning (TNEP) aims to determine the optimal time, locations and the number of new transmission lines that are necessary to be built in an electrical network while minimizing the total investment and operation costs [117]. TNEP is traditionally solved by worst-case based deterministic approaches [118]. However, due to the significant variability and uncertainty associated with RES [119], it is not sufficient anymore to carry out the planning only considering the worst case operating profile. In recent years, a series of probabilistic TNEP models, such as two-point estimation method [120], fuzzy method [121], and Monte Carlo simulation [122], have been proposed to deal with the uncertainties stemming from various sources such as forced outage rates of lines [123], wind farms [124], and liberalized electricity markets [125].

In general, uncertainties in TNEP problem can be classified as non-random and random [126]. In details, statistics of random uncertainties can be derived from the historical observations such as load variability, whereas non-random uncertainties such as generation expansion/closures cannot be characterized by probability distribution functions (PDFs). In this thesis, the operational uncertainty of the TNEP problem belongs to the random uncertainties, including the demand variability and wind output variability. To handle the variability and uncertainty, a vast number of scenarios are required to represent the various operating conditions, which may lead to computational burdens in the nonlinear and mixed-integer TNEP. A series of methods have been proposed to improve the computational efficiency in the simulation. As stated in [127], important sampling and the distributed computing methods have been implemented to reduce the number of scenarios and to accelerate the simulation, respectively. Nevertheless, it can not be guaranteed that the selected samples can represent the other discarded data in terms of the TNEP problem. In addition, a few scenarios are usually

manually selected as representative snapshots by experts that can describe the system critical states according to the variations of the load and renewable generation outputs [128]. However, choosing the considered scenarios by this way can result in imprecise system operations and investment, especially in the context of highly integrated intermittent generation. Therefore, it is imperative to propose a novel scenario selection method specifically for the TNEP problem.

Data mining techniques (e.g. unsupervised and supervised clustering methods) are widely employed to significantly decrease the vast number of scenarios by grouping the clustering variables with similar characteristics (e.g. profile patterns of demand-wind data). In general, data clustering algorithms can be categorized as connectivity-based (e.g. hierarchical clustering), centroid-based (e.g. k-means clustering), model-based (e.g. Gaussian mixture model), and density-based clustering (e.g. DBSCAN, peak-density method). Specifically, connectivity-based clustering constructs a hierarchy of groups based on their distance, whereas centroid-based clustering aims to find a determined number of centers and assign the data points to the nearest one. Different from the above distance based algorithms, model-based and density-based clustering methods are performed according to the distributions and the density of the input data, respectively. Comprehensive comparisons have been performed among these clustering techniques in terms of their statistical performance [129], [130]. Various applications of clustering methods in power systems have been proceeded such as modeling correlation among demand, climate and office building features [131], time-of-use tariff design [132], domestic electricity load profile characterization [133], and energy consumption prediction [134].

Although varieties of clustering techniques have been proposed in the literature, only a few of them have been applied in the scenario selection procedure for TNEP. For example, in [128], k-means algorithm was used to cluster the operation features based on price differences and non-controllable injections. In [135], a tri-level reliability-constrained robust power system expansion planning framework was proposed to model the uncertainty of load, wind, and availability of generators and lines. To characterize the load and wind generation uncertainty, representative scenarios were selected according to the centroid, α -quantile, and $(1 - \alpha)$ -quantile of each cluster's demand-

wind profiles, grouped by the k-means technique. Instead of k-means, k-medoids clustering was employed in [136] to select the actual median present in the historical data rather than the mean values. In addition, a self-adaptive clustering technique integrated with graph partitioning method and rough fuzzy clustering was proposed in [137] to deal with the uncertainties in the TNEP problem based on the historical wind power data set.

All the above methods select representative scenarios based on clustering input information (e.g. energy price, demand, renewable generation). However, the target outputs of TNEP are the operational cost and the investment decisions of the candidate lines, which are nonlinearly related to the input random variables. Therefore, directly performing clustering on these uncertain parameters is not efficient as the effects of the selected scenarios on the TNEP problem can not be identified properly. To this end, in [138], operational states are characterized by their optimal power flow (OPF) patterns and clustered via a moments technique as transmission investments are losses and congestions driven. According to the results of numerical examples, the proposed OPF based scenario selection algorithm indeed has an effective reduction in the required number of scenarios for achieving an accurate TENP solution when comparing with the input-based clustering. In addition, an operational state aggregation technique has been proposed in [139] to select the representative scenarios based on line benefits. However, this algorithm needs to solve a relaxed TEP problem with all the operating points, resulting in huge computational cost before the clustering procedure.

Furthermore, although the performance of different clustering technologies to represent the statistical information of the data set have been widely compared, their impacts on the particular applications has not been fully investigated. The scenario reduction techniques based on k-means method, peak-density method, and typicality method have been compared when applied in stochastic unit commitment [140], it is also important to choose the most appropriate clustering method for the TNEP problem, which is one of the major contributions in this thesis.

In this thesis, a target-based scenarios selection framework was proposed by performing bi-level clustering based on the OPF patterns of the existing lines and the investment costs of candidate lines for each scenario, respectively, to further improve

the computational efficiency of the TNEP problem. Within the framework, different clustering methods (i.e. k-means, k-medoids, hierarchical, and Gaussian mixture model (GMM)) are compared and evaluated according to their TNEP solutions across a range of number of clusters. Furthermore, the proposed framework is evaluated and compared with other tested scenarios selection methods based on different clustering variables. The superior performance of the proposed scenarios selection framework is demonstrated through a numerical case study on the modified IEEE 118-bus system.

The outline of the remainder of this chapter is as follows: Chapter 7.2 gives an introduction to various types of clustering variables and different clustering techniques for the TNEP problem. Chapter 7.3 explicitly introduces the structure of the proposed target-based scenarios selection framework. Chapter 7.4 presents the simulation results among different clustering methods and clustering variables. Chapter 7.5 contains the concluding remarks.

7.2 Clustering variables and clustering techniques

In the context of large-scale systems, transmission network expansion planning is encountered with various challenges especially when taking into account the uncertainties in demand and renewable power generation. In particular, it is essential to capture the temporal correlations and inter-spatial dependencies between loads in various locations and intermittent generation units' output. This can be achieved by representing the uncertainties via a set of historical operating points, each of them indicates a demand-wind pattern with an estimated probability that can be exploited in the proposed TNEP model. Nevertheless, the vast number of historical data renders it indispensable to select the representative scenarios as less as possible on condition that an accurate planning result is ensured. As illustrate in Section 2, clustering the scenarios based on their similarity is one of the most broadly used methods to compress the size of input data for the transmission expansion problem. To this end, three major directions in terms of the clustering based scenarios selection method can be concluded as:

- 1) For the TNEP problem, the selection of clustering variables is the main topic of interest to be investigated;

- 2) Given a type of clustering variables, among the varieties of clustering methods, it is hard but imperative to choose the most appropriate one for this specific problem;
- 3) After obtaining the constructed clusters of data, the selection of representative profile in each cluster also has great influence on the final solution of the target problem.

In this thesis, we make efforts to mainly explore the first two issues of clustering variables and clustering methods, and then preliminarily analyze the effects of selecting different scenarios (e.g. mean value or actual median point) in each cluster.

7.2.1 Fundamental Clustering variables

As already mentioned above, one of the most important issues of the scenarios selection problem is to choose the appropriate clustering variables for the TNEP problem. Nodal operational information is usually employed as clustering variables to select the representative scenarios to obtain an accurate solution efficiently. For example, demand-generation patterns multiplied by the nodal price were used as clustering variables to select snapshots from historical data in [128]. Also, in [135], a polyhedral uncertainty set was built by clustering demand and wind power generation data to characterize uncertainties for power system expansion planning. The wind generation was also selected as clustering variables in [137] to generate scenarios based on the proposed self-adaptive clustering technique. However, it is imperative to claim that such a selection method based on these input stochastic variables is not efficient for TNEP problem as some significantly different scenarios can also lead to the identical investment decisions, as illustrated in [138].

To this end, it is more reasonable to perform the clustering procedure directly on the target outputs of the TNEP problem for each historical operating points. This is because of the fact that the solutions of the transmission expansion problem are not linearly related to the input demand-generation patterns due to the complexity of the network structure and the optimization problem. Power flow patterns in network expansion planning are of interest to be investigated while selecting the scenarios as they are highly related to transmission congestions and losses, resulting in potential investments.

Furthermore, the most relevant outputs of the transmission expansion problem are the investment costs of candidate lines for each scenario.

In this thesis, we propose three types of clustering variables based on the OPFs and the investment costs of the simulation for each operating point. In the proposed TNEP model, transmission lines are categorized into existing lines and candidate lines for investments. After solving the investment problem for each of all the scenarios, the first proposed clustering variables are the OPFs of existing transmission lines $\hat{F}_E$ whose element f_{ij} satisfies $\forall i, j \in \Lambda$. In the context of no investments case, it is effective to select the representative scenarios by clustering the operating points based on $\hat{F}_E$ as the final total cost only contains the operational cost which is highly related to the power flow patterns. However, the target problem with transmission expansions leads to the requirements of obtaining accurate investment decisions when using the selected scenarios. Consequently, another type of clustering variable taken use of selecting useful scenarios are the OPFs of candidate lines $\hat{F}_c$ whose element f_{ij} satisfies $\forall i, j \in \Lambda_{new}$. Although $\hat{F}_c$ can indirectly signify the investment solutions, different OPFs' values of candidate lines may also cause inaccurate solution as the scenarios with different investment decisions can be grouped together due the fact that most of clustering procedures are performed based on the Euclidean distance. To this end, the proposed framework is constructed based on the direct target outputs, investment costs of candidate lines $\hat{I}_c$, another type of clustering variable whose element $\hat{I}_{ij} = C_{ij}n_{ij}$ satisfies $\forall i, j \in \Lambda_{new}$.

Furthermore, all the proposed clustering variables can be further categorized as investment scenarios and non-investment scenarios on the basis of the simulation result for each scenario. Accordingly, the scenarios selection process should consider these two sets of operating points separately via a bi-level clustering framework. Note that all the fundamental clustering variables will also be processed in different combinations to evaluate their performances.

7.2.2 Clustering methods

For large-scale systems, it is imperative to accurately represent the operational uncertainty in an efficient way for solving the TNEP problem in the context of high

penetration levels of renewable resources. In general, operational uncertainty is represented by a vast number of operational states that each of them (also called a scenario), including a demand-generation pattern, can result in an OPF pattern in the network [138]. In order to reduce the number of scenarios required to solve the TNEP problem, clustering methods are widely used to classify the multi-dimensional stochastic input into different groups according to their similarity; each one has an estimated probability. When solving the TNEP problem, the input scenarios are not linearly related to the target outputs are the operational cost and the investment decisions of the candidate transmission lines. Therefore, it is important to assess the performance of different clustering methods for this specific problem because the clustering method with good statistical performance may not guarantee an accurate solution. The clustering methods considered here can be classified into three categories: centroid-based clustering (k-means, k-medoids clustering), connectivity-based clustering (hierarchical clustering), and distribution-based clustering (Gaussian mixture model clustering (GMM)). Note that details of the abovementioned clustering technique are given in Chapter 2.

7.3 Proposed target-based scenarios selection framework

Conventionally, the clustering based scenarios selection method can be described in Figure 7-1. Note that the clustering procedure in the workflow can be directly performed on the input data when regarding demand-generation patterns as clustering variables. Otherwise, the other types of clustering variables should be constructed by solving the TNEP problem for each operating point. Based on this framework, the tested clustering variables are:

V1) $\mathbf{X}$: demand-wind patterns;

V2) $\hat{\mathbf{F}} = [\hat{\mathbf{F}}_E, \hat{\mathbf{F}}_C]$: OPFs patterns of all transmission lines including both existing and candidate lines;

The purpose this comparison is to illustrate the benefits of using the effects (i.e. OPF patterns) instead of the historical data itself.

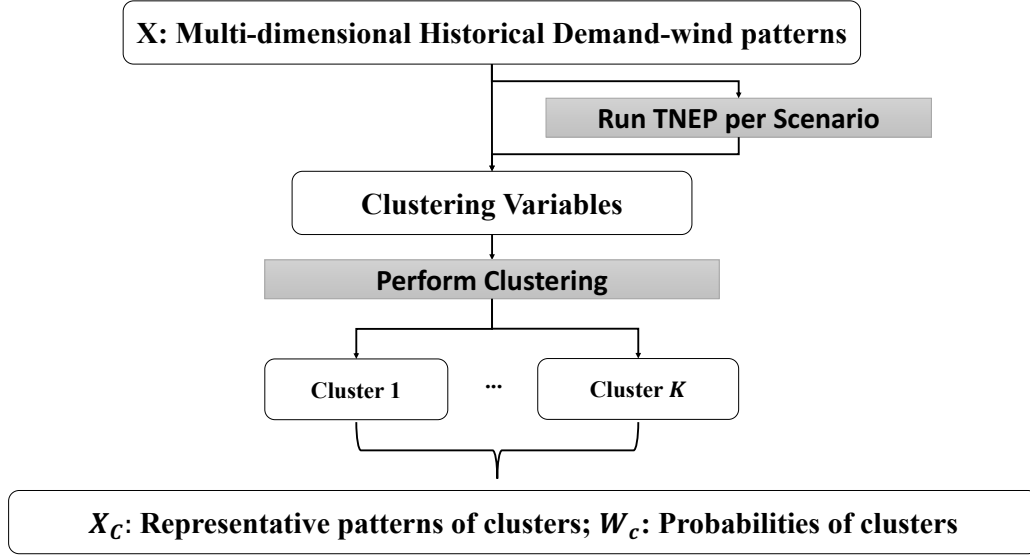


Figure 7-1. The structure of conventional scenarios selection framework

The performance of the conventional scenarios selection framework is limited by considering the existing lines and candidate lines as well as the investment scenarios and no investment scenarios together, which may result in inaccurate TNEP solutions. To this end, we propose a bi-level clustering scenarios selection framework based on both of the clustering variables $\hat{F}_E$ and $\hat{F}_C$ as illustrated in Figure 7-2. In general, the whole workflow, termed V3, can be categorized into three stages:

- 1) **Pre-solving Stage:** Given the input multi-dimensional historical demand-wind data, the proposed TNEP problem is primarily solved for each scenario to obtain the clustering variables, optimal power flows (OPFs) for the existing lines $\hat{F}_E$ and candidate transmission lines $\hat{F}_C$.
- 2) **Clustering Stage:** According to the power flow patterns $\hat{F}_C$ of the candidate lines, the indices idx_{C0} of scenarios $\hat{F}_{C0}$ that result in all zeros power flow values (no investments) are extracted from $\hat{F}_C$, whereas the rest of scenarios with continuous power flow patterns construct matrix $\hat{F}_{Cinv}$. For the purpose of getting an accurate investments decisions, clustering method is performed in the firstly level based on $\hat{F}_{Cinv}$ with a determined number of clusters K_1 . Indices idx_{Cinv} is obtained in this level. Subsequently, the scenario set $\hat{F}_{E0}$ is filtered from scenarios $\hat{F}_E$ indicated by idx_{C0} . For the TNEP problem, operational costs are directly related

to $\hat{F}_{E0}$ as there no transmission extensions considered under these scenarios. Therefore, the second level clustering is carried out on the basis of $\hat{F}_{E0}$ with another number of clusters K_2 to get the corresponding indices idx_{E0} . By this way, all the scenarios can be clustered according to the combined indices IDX_{final} . Perform clustering techniques in different variables to get the indices of each scenario with a certain probability.

- 3) **Scenarios Selection Stage:** In each constructed cluster, choose one appropriate profile to represent this group for the further employment of the full TNEP problem.

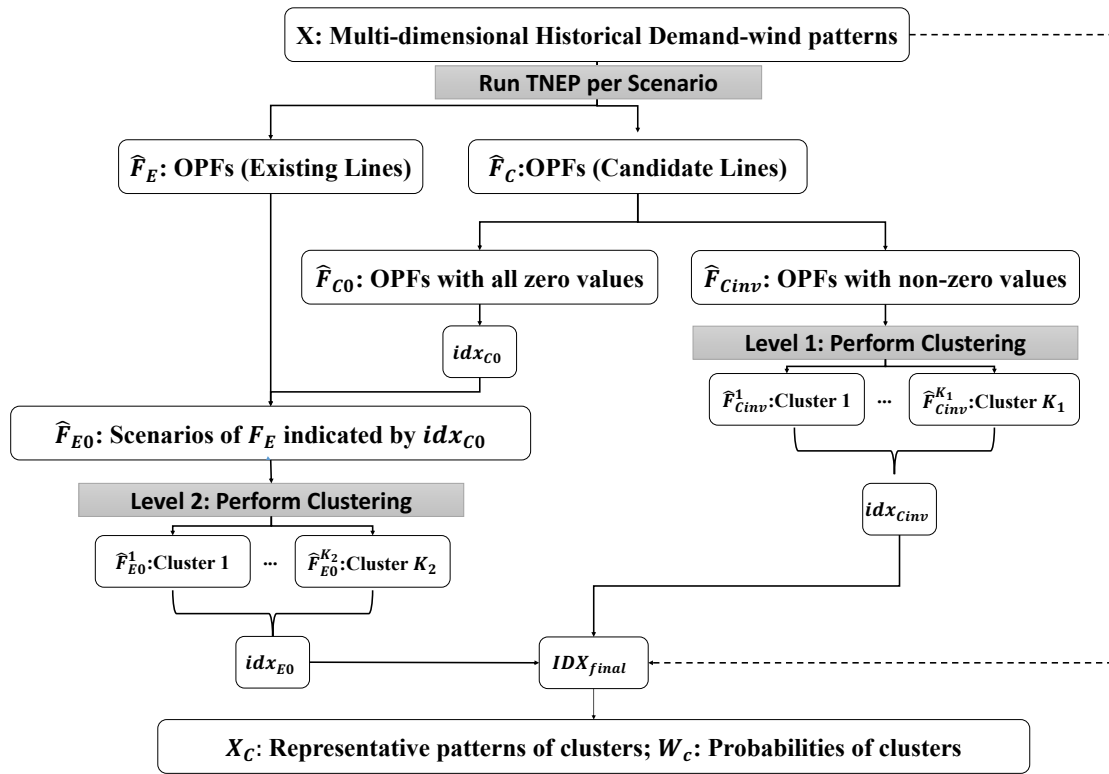


Figure 7-2. The structure of the proposed OPF-based scenarios selection framework

Although the framework V3 has taken into account the differences between types of lines and conditions of scenarios, the target-based scenarios selection framework, termed V4, is proposed to obtain further improvements by employing the direct target outputs, $\hat{I}_c$ instead of $\hat{F}_c$ in the Clustering Stage. As shown in Figure 7-3, the indices idx_{c0} of scenarios $\hat{I}_{c0}$, indicating no investments, are extracted from $\hat{I}_c$, whereas the rest of scenarios with continuous investment costs construct matrix $\hat{I}_{cinv}$. Same as V3, the no investment scenarios are also extracted using idx_{c0} and clustered based on the OPFs

of existing lines in the second level. It is important to highlight that another advantage of V4 is the determined maximum number of clusters that required for the TNEP problem. For example, given the number of candidate lines $n_c = 5$, there are $2^5 = 32$ possible combinations of investment decisions/ investment costs that can be used as the maximum number of clusters. Practically, this number should be significantly less than 2^{n_c} with the increasing number of candidate lines because some investment combinations will never happen within a limited number of operating points.

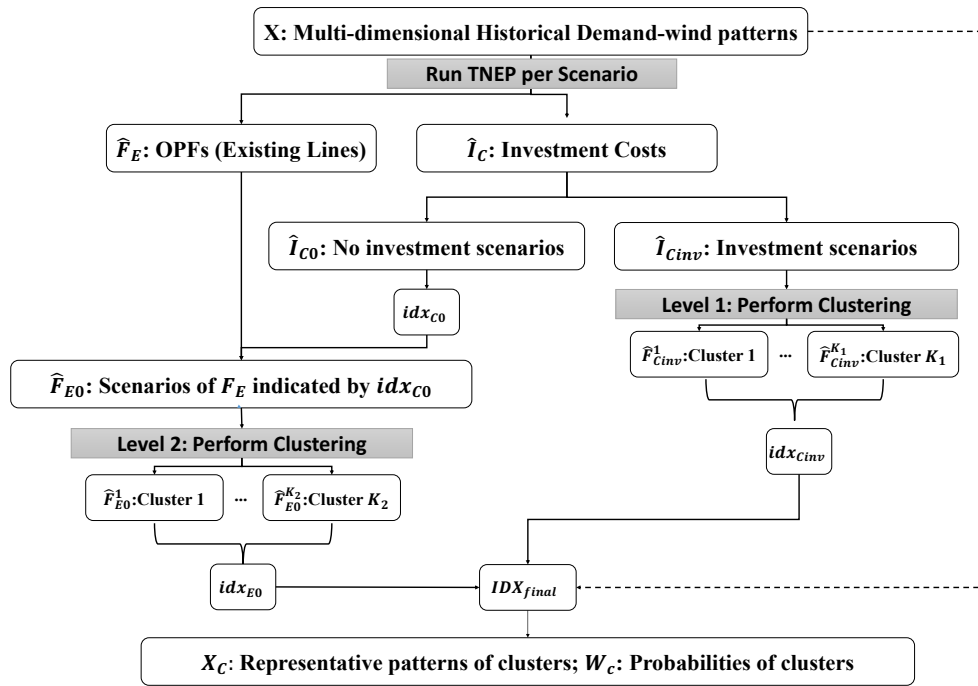


Figure 7-3. The structure of the proposed target-based scenarios selection framework

7.4 Simulation study and results analysis

7.4.1 Historical data and test system

To demonstrate the performance of the proposed framework, the original IEEE 118-bus system consisting of 54 generators and 186 existing transmission lines has been modified by including 10 wind farms of size 100 MW as well as 10 more candidate transmission lines with capacity 1000 MW. For the historical database of stochastic variables, a large library of historical measurements of load and wind power injections was provided by RTE, the French Transmission System Operator (TSO). The library consists of 14,251 measurements at 15-minute time intervals over the course of two months, from January to February 2012, and spans over 7,000 load buses and 200 wind

plants [16]. In this work, we focus on a subset; a 128-dimensional dataset consisting of 118 demand buses and 10 wind generators chosen at random. To map the selected variables onto the test system, the historical dataset is scaled by calculating the ratio between the maximum coincident peak demand of the original data and the sum of active power demand across all buses defined in the system. By this way, the temporal correlations and inter-spatial dependencies between the stochastic variables can be retained. An example is given in Figure 7-4 that shows the histograms and bivariate scatter plots of five stochastic variables taken from the French transmission system; variables L1, L2 and L3 are loads at different locations while W1 and W2 are wind injections from two wind farms. It can be observed that stochastic variables have highly non-normal marginal distributions (diagonal histograms) and non-linear dependencies (scatter diagrams). The dependence between the different loads, between the outputs of the two wind farms as well as the dependence between individual loads and wind power injections are highly complex and non-standard. As such, for TNEP problem, it is of great potential benefit to move beyond the assumption of perfect correlation between loads at different locations etc. and investigate the dependence structure at the level of disaggregated variables in more detail. To this end, the proposed clustering framework is designed to select representative scenarios in a multi-dimensional space.

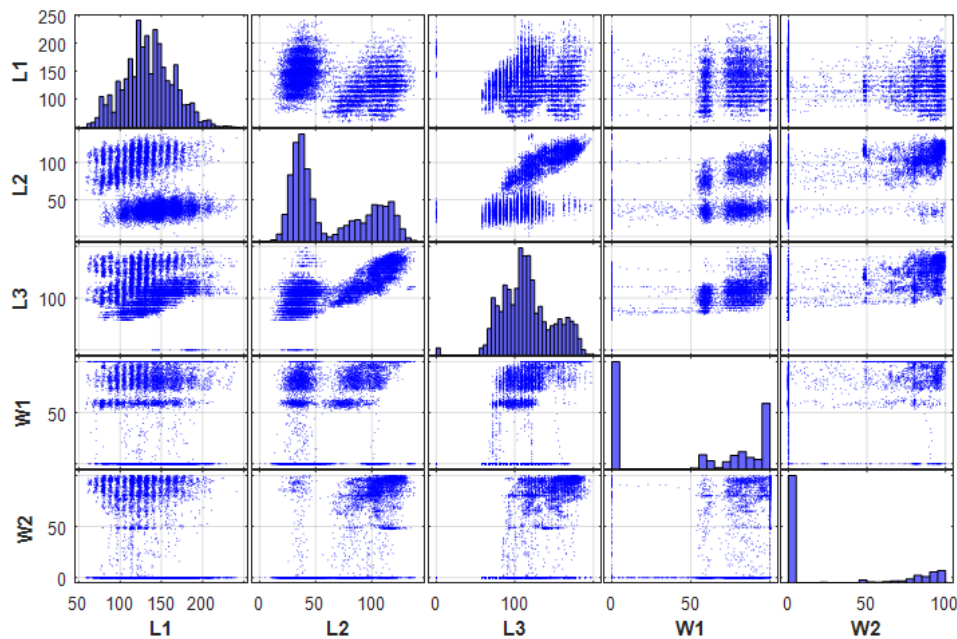


Figure 7-4. Histograms of marginals distributions and bivariate scatter plots of active loads (MW) and wind power output (MW) for randomly selected buses in France, measured between Jan 2012 and Feb 2012.

Table 7-1 shows the 20 candidate lines and their annualized investment costs that estimated by the product of cost multipliers, cost per mile and line length [151]. Regarding the parameters of the proposed TNEP model, value of lost load γ is set to \$400,000/MWh. In addition, we use $\tau = 8,760$ hours and $M = 10,000,000$ MW.

Table 7-1. Candidate Lines for the IEEE 118-bus System

<i>No.</i>	<i>Lines</i>	<i>Cost (\$ million/year)</i>	<i>No.</i>	<i>Lines</i>	<i>Cost (\$ million/year)</i>
1	(8 - 9)	5.49	11	(89 - 90)	17.95
2	(9 - 10)	5.80	12	(83 - 84)	19.80
3	(26 - 30)	15.48	13	(69 - 75)	18.30
4	(64 - 65)	5.44	14	(17 - 113)	5.42
5	(68 - 116)	0.91	15	(77 - 78)	2.79
6	(49 - 51)	20.55	16	(76 - 118)	9.79
7	(70 - 74)	19.85	17	(105 - 106)	9.85
8	(44 - 45)	16.22	18	(103 - 110)	27.19
9	(80 - 96)	27.30	19	(114 - 115)	2.34
10	(92 - 100)	44.25	20	(12 - 117)	21.00

7.4.2 Performance Evaluation

To illustrate the comparative advantage of the proposed framework, the tested scenario selection methods considered for comparison are:

M1: direct clustering operating points based on clustering variable X , demand-wind patterns;

M2: clustering operating points based on clustering variable $\hat{F} = [\hat{F}_E, \hat{F}_C]$, OPFs patterns of all transmission lines;

M3: bi-level clustering based on clustering variable $\hat{F}_E$ and $\hat{F}_C$, respectively;

M4: target-based bi-level clustering based on clustering variable $\hat{F}_E$ and $\hat{I}_C$, where $\hat{I}_C$ consists of the investment cost for each candidate line per scenario;

Furthermore, the clustering methods performed here are k-means, k-medoids clustering, hierarchical clustering with average linkage (H(Average)), with Ward linkage (H(Ward)), with complete linkage (H(Complete)) and GMM. The performances of all the 24 tested methods are evaluated according to their estimated investment costs, operation costs and total costs. Note that, given investment decisions, operation costs are computed by solving the economic dispatch problem considering all the scenarios. Therefore, same investment decisions directly result in the same operation costs and total costs of the TNEP problem. In this chapter, we firstly compare the investment decisions for each clustering method from Chapter 7.4.2.1 to Chapter 7.4.2.4. Then

Chapter 7.4.2.5 performs the overall comparison considering the total cost of the best perform clustering method for each scenarios selection framework (i.e. M1, M2, M3, and M4).

7.4.2.1 TNEP results of M1

In this case, the representative scenarios are selected by directly clustering the historical stochastic variables according to their statistical similarities via the above-mentioned clustering techniques. Note that, for this method, it is not required to solve the TNEP problem per operating points before carrying out the clustering process. From the view of statistical performance, clustering validity indicators have been proposed and broadly used to quantitatively analyze the quality of clustering schemes based on their similarities. Consider that K partitioned scenarios clusters with computed centroids $C = \{c_1, \dots, c_K\}$, are obtained by a clustering technique. Let X_k denote the scenarios that belongs to cluster k , for $k = 1, \dots, K$, the two indicators considered in this thesis are computed as follows.

one of the commonly used indicators applied in this thesis, the mean index adequacy (MIA) [75], indicates the average of the distances between the centroid of a cluster and each scenario in this cluster:

$$MIA(X, K) = K^{-1} \sum_{k=1}^K d^2(X_k, c_k) \quad (7.1)$$

The clustering dispersion indicator (CDI), defined as the ratio of the mean intraset distance between the scenarios in the same cluster and the intraset distance between the centroids of all the K clusters [46]:

$$CDI(X, K) = \hat{d}(C)^{-1} \sqrt{K^{-1} \sum_{k=1}^K \hat{d}^2(X_k)} \quad (7.2)$$

where $d(.,.)$, $\hat{d}(.)$ represent cluster-to-cluster distance and intraset distance (Euclidean distance), respectively. Note that both indicators have the same characteristic that a lower criterion value indicates better clustering performance. Before running the TNEP problem, the original data set of demand-wind patterns are clustered via k-means, k-medoids clustering, hierarchical clustering with average linkage (H(Average)), Ward linkage (H(Ward)), complete linkage (H(Complete)) and GMM. Figure 7-5 shows the

results of adequacy comparisons among the tested clustering methods by using the CDI and MIA validity indicators for clusters counts K ranging from 2 to 100.

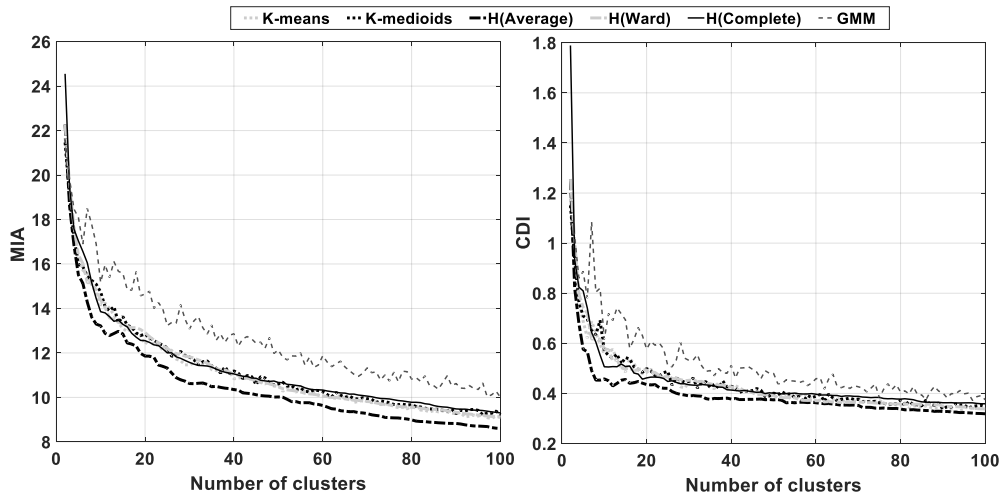


Figure 7-5. Comparisons among the clustering techniques by using (left) the MIA indicator and (right) the CDI indicator for $K = 2$ to 100.

The results illustrated in Figure 7-5 shows that H(Average) is the best-performed method among these tested techniques, for the historical load and wind dataset, indicated by its lowest indicator values for both indicators. In contrast, the limitations of GMM is indicated by the large indicator values across all the number of clusters. In addition, it is constructive to note that K-means, K-medoids, H(Complete) and H(Ward) have comparable performances in terms of the statistical similarities.

The next step is to assess the TNEP solution quality of the above clustering methods when using demand-wind patterns as clustering variables. To obtain the benchmark, the TNEP problem based on the modified IEEE-118 system is firstly solved with all the historical dataset of 14,251 operating points. The overall total cost consisting of optimal operational cost and investment cost are given in Table 7-2.

Table 7-2. TNEP Solution Benchmark With All The Scenarios

	Operational Cost (\$million/year)	Investment cost (\$ million/year)	Total cost (\$ million/year)
All scenarios solution	2667.14	69.98	2737.12

Based on the constructed clusters for each method, most of the representative scenarios are selected by using the mean value of each group with corresponding probabilities except the K-medoids method whose final scenarios are presented by the actual median. Figure 7-6 show the TNEP solutions of investment cost obtained by

performing the tested clustering methods on the historical demand-wind patterns X of size $14,251 \times 128$, for the number of clusters K from 2 to 300.

Regarding the optimal investment costs, $H(\text{Ward})$ and $H(\text{Complete})$ exhibit superior performance compared to the other techniques indicated by their ability to approach the benchmark solution at $K = 300$, whereas the other clustering methods cannot achieve in the tested range of K . Specifically, most of clustering methods result in an under-estimation of the investment cost, whereas $H(\text{Average})$ method makes over-investment decisions as more outliers (e.g. worst-case patterns) are detected by this method with higher weights. In addition, model-based clustering method, GMM, has relatively stable performance after about $K=20$. For the centroid-based clustering methods, solutions of K -means are slightly closer to the target value than K -medoids which indicates that taking the mean value of each cluster has better performance than using the actual median in this case. It is imperative to note that, for each clustering method, the inconsistency of performance orders between the validation indicator values and the TNEP solutions can be observed by comparing Figure 7-6 with Figure 7-5. This demonstrates the reasonability and the importance of investigating various types of clustering methods beyond their statistical performances for such a complex problem.

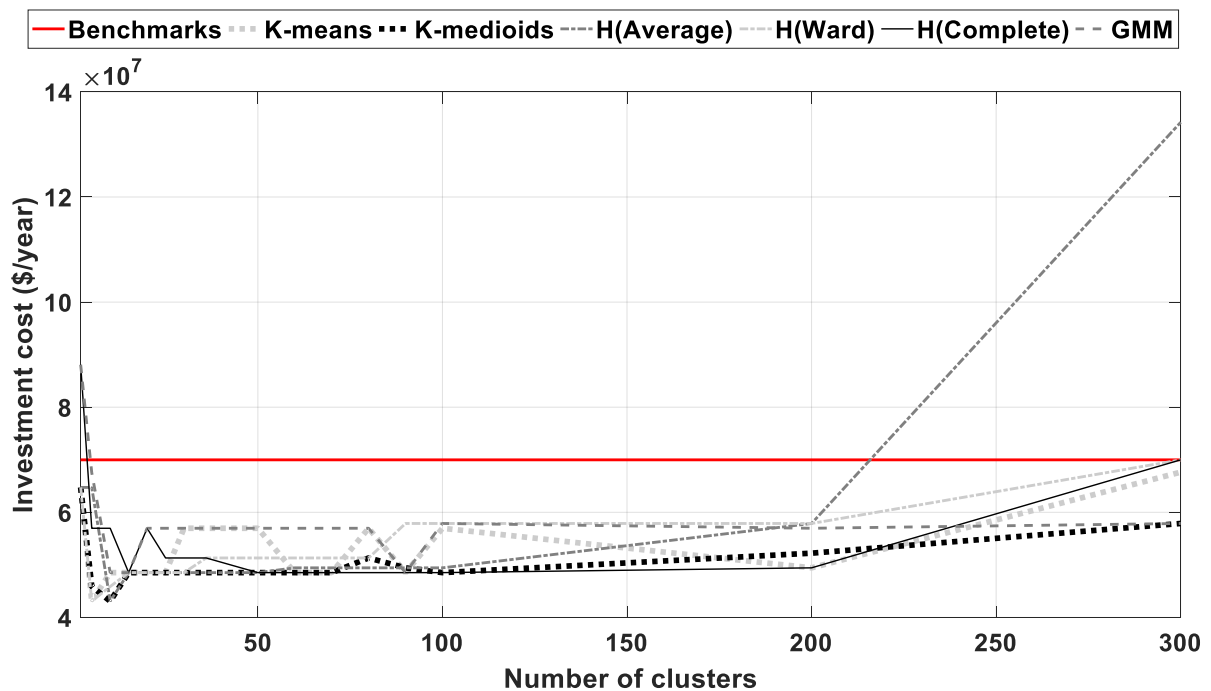


Figure 7-6. Comparisons among clustering methods based on the TNEP solutions of investment costs. The input scenarios are selected via M1.

7.4.2.2 TNEP results of M2

As illustrate in Chapter 7.3, optimal power flow patterns of transmissions line are also of interest to be employed as clustering variables to select the representative scenarios by different clustering methods following the conventional steps introduced in Figure 7-1. After solving the TNEP problem for each scenario, clustering variables $\hat{F}$ are formed of size $14,251 \times 206$ that consists of $\hat{F}_E$ ($14,251 \times 186$) and $\hat{F}_C$ ($14,251 \times 10$). An example of the OPF patterns of five existing lines and five candidate lines is given in Figure 7-7. It can be observed that there are indeed some obvious groups existing in the clustering variables $\hat{F}_E$ and $\hat{F}_C$ which renders the clustering procedure to be reasonable.

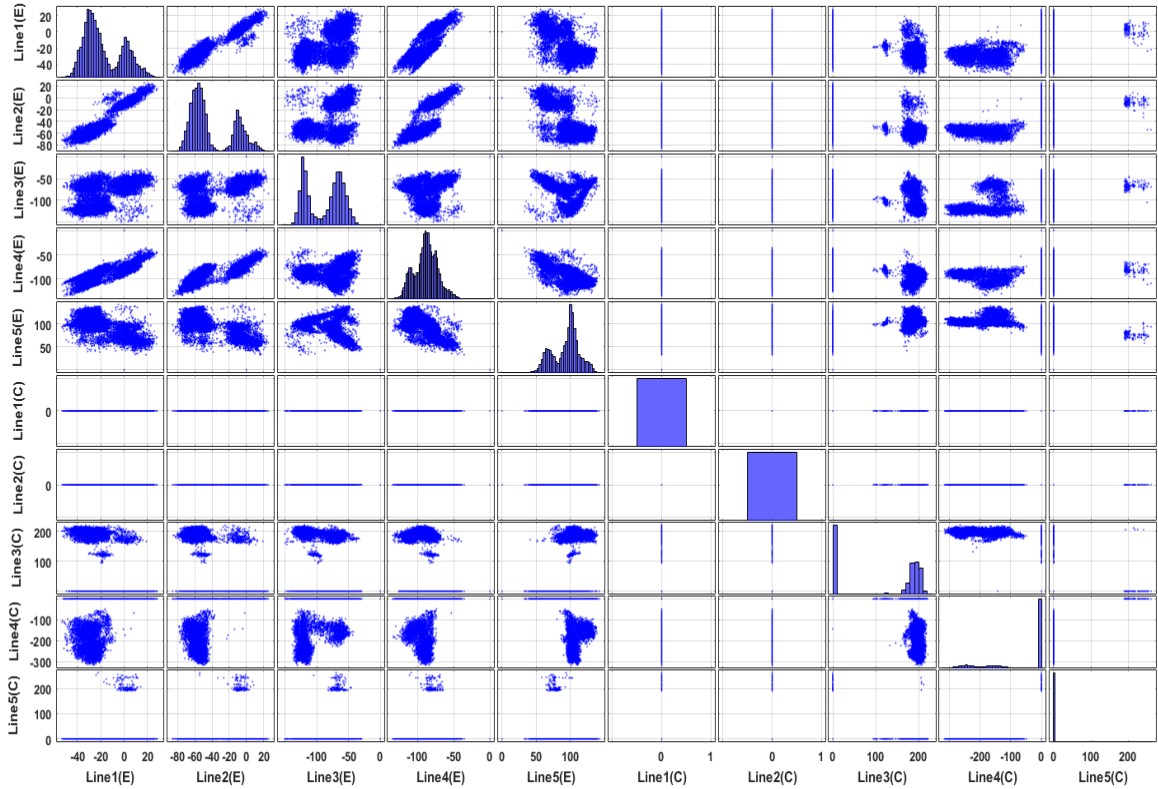


Figure 7-7. Histograms of marginals distributions and bivariate scatter plots of power flow patterns of five existing transmission lines (No.1 to No.5 defined in the test system) and five candidate lines (No.1 to No.5 defined in Table 7-1).

Applying the tested clustering methods on $\hat{F}$, the results of indices and corresponding probabilities are assigned to the historical stochastic variables X for the number of clusters K from 2 to 300. According to the constructed groups for each K , representative scenarios are finally used as inputs for the proposed TNEP problem with a certain

probability for each operating point. The TNEP solutions of investment cost for M2 are shown in Figure 7-8.

In this case, different clustering methods show large discrepancy in their performance orders comparing with the results obtained from Figure 7-6. Concretely, K-medoids performs the best in this case that the solutions achieve a relatively accurate level from $K=80$ and converge to the exact benchmark investment cost at $K=200$. Then GMM and H(Ward) methods also exhibit high efficient to approach approximate but not exact optimal solutions that may also result in huge different in terms of the total cost. On the contrary, H(Average), H(Complete) and K-means methods cannot achieve the benchmark value in the tested range of number of clusters, and even increase to a overestimated value from about $K=100$, $K=230$, and $K=240$, respectively.

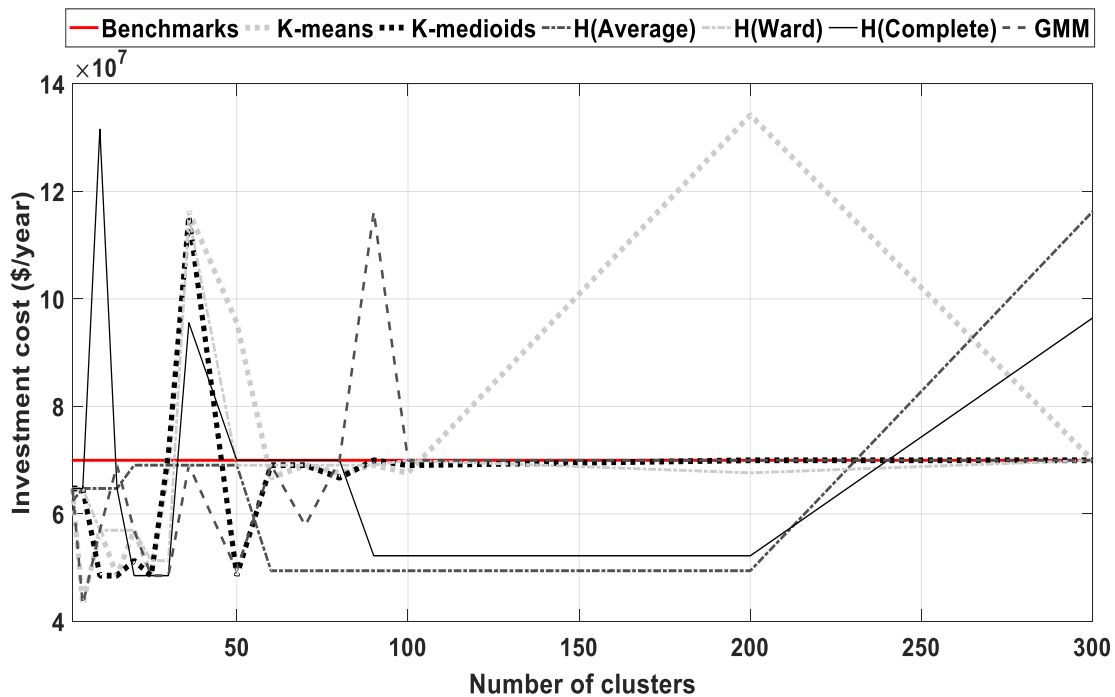


Figure 7-8. Comparisons among clustering methods based on the TNEP solutions of investment costs. The input scenarios are selected via M2.

7.4.2.3 TNEP results of M3

For the proposed bi-level clustering framework M3, both of the constructed $\hat{F}_E$ and $\hat{F}_C$ are further classified into a series of useful dataset for clustering according to the power flow patterns of candidate lines. In particular, 3,594 scenarios are extracted from $\hat{F}_C$ to form the first stage clustering variables, $\hat{F}_{Cinv}$, which only containing the non-zero

OPF patterns. Afterwards, the second level clustering variables $\hat{F}_{E0}$ of size $10,657 \times 10$ are indicated and constructed based on the scenarios $\hat{F}_{C0}$ that result in all zeros power flow values (no investments) in $\hat{F}_C$. Note that although nearly 75% percent of scenarios are included in $\hat{F}_{E0}$, the major scenarios that lead to investment are identified by $\hat{F}_{Cinv}$. To this end, we determine the number of clusters $K_1 = 2, \dots, 300$ for the first stage and then set $K_2 = 5$ for the second level clustering for convenience.

Using the selected scenarios obtained from M3, the results of investment for the tested clustering methods are shown in Figure 7-9. All the methods are exposed to achieve the benchmark after in the tested range of K. In details, K-means presents the most inefficient performance that obtains the steady optimal solution from roughly $K = 90$; Afterward, GMM and K-medoids converge to the target investment cost at $K = 70$ and $K = 80$, respectively. Comparing to the centroid-based and the distribution-based clustering techniques, the tested connectivity-based methods exhibit better solution qualities with the increasing number of K, especially for H(Ward), the fastest approach to reach the exact optimal investment decisions from $K = 36$.

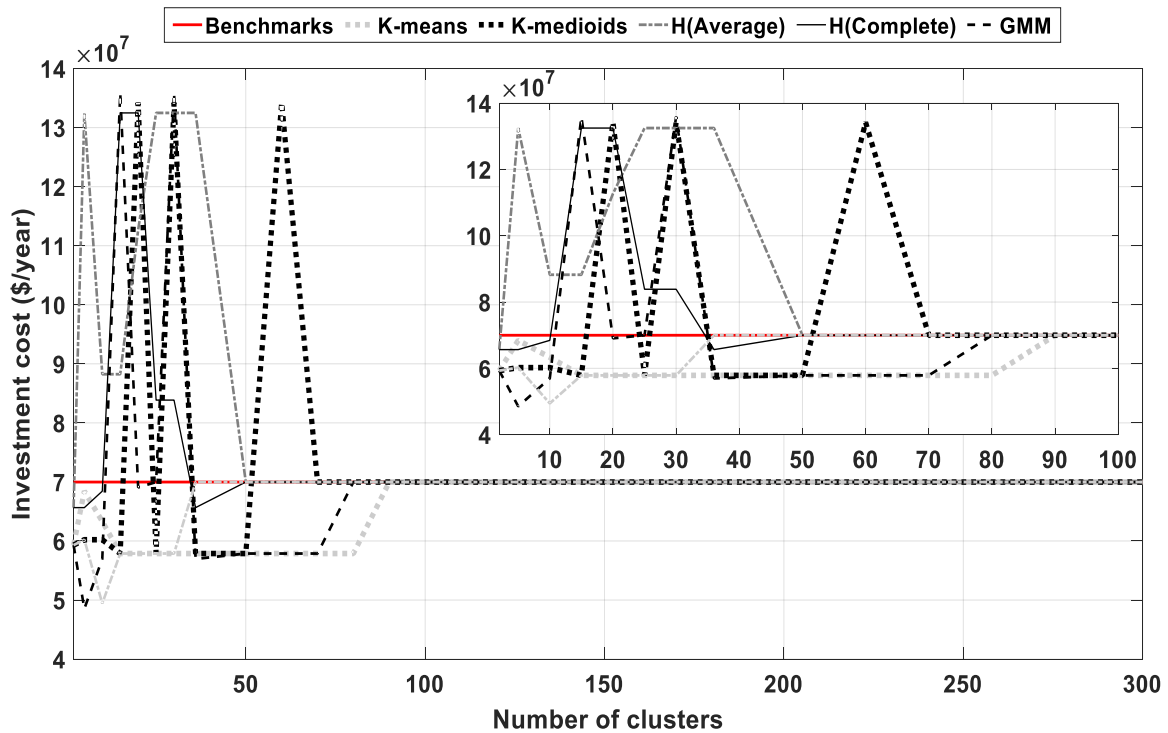


Figure 7-9 Comparisons among clustering methods based on the TNEP solutions of investment costs. The input scenarios are selected via M3.

7.4.2.4 TNEP results of M4

Finally, the proposed target-based scenarios selection framework is implemented based on the constructed OPF patterns of the existing lines $\hat{F}_E$ and the estimated investment cost of the candidate lines $\hat{I}_C$. In the first stage clustering, 3,594 scenarios $\hat{I}_{C0}$ with non-zero investment costs are picked from $\hat{I}_C$. Then the clustering variable $\hat{F}_{E0}$ of the second-level is extracted from $\hat{F}_E$ indicated by idx_{C0} obtained from the no investment scenarios $\hat{I}_{C0}$. It is imperative to note that one of the important benefits of performing the first stage clustering on $\hat{I}_C$ is the determined maximum number of clusters that requires for this problem. Mathematically, there should be $2^{20} = 1,048,576$ possible combinations of investment cost as it is a constant value for each candidate line. However, only 56 sets of expansion plans actually occur in the investment solutions when solving each scenario independently. Therefore, $\bar{K} = 56$ is defined as the maximum essential number of clusters because all the scenarios should get perfectly clustered (i.e exactly identical investment decision in each cluster) at this number, and it is meaningless to further group them into more clusters. Nevertheless, for the purposed of comparing with other methods, the number of clusters are also defined as $K_1 = 2, \dots, 300$ and $K_2 = 5$ in the first and second stage, respectively.

The TNEP solutions of the proposed target-based framework (M4) are shown in Figure 7-10 regarding its ability to approach the benchmark investment cost within a significantly reduced number of representative scenarios. Among the tested clustering techniques, GMM requires the most number of scenarios to obtain the expected result ($K = 60$). In addition, it is evident that the scenarios selected via M4 exhibit superior performance when using the centroid-based and connectivity based clustering techniques. When applying K-means, the most efficient method for M4, only 15 selected scenarios are required to approaches the target investment cost. Meanwhile, both of H(Ward) and H(Complete) achieve the accurate investment cost from $K = 20$. Then K-medoids and H(Average) converge to the benchmark solution at $K = 25$ and $K = 30$, respectively. This results highlights the importance of performing the first level clustering on the basis of investment costs.

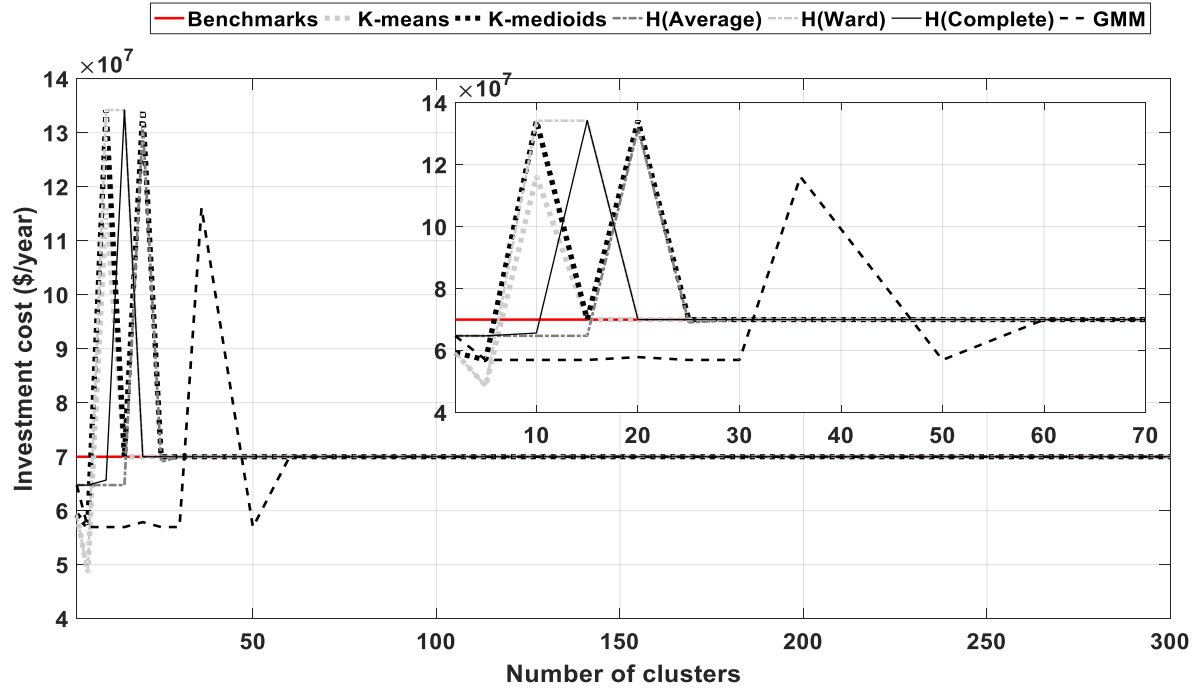


Figure 7-10. Comparisons among clustering methods based on the TNEP solutions of investment costs. The input scenarios are selected via M4.

7.4.2.5 Overall comparison

In all the preceding analysis we have analyzed the performances of different types of clustering techniques for each scenarios selection method considering their efficiency to get the accurate TNEP solution. In this part, a comprehensive comparison is given among the overall performance of different scenarios selection methods to highlight the contributions of the proposed framework.

Firstly, selecting the representative scenarios based on conventional clustering variable, demand-wind patterns (M1), have the benefits of simple implementation and readily comprehensible clustering procedure. However, it suffers from the issues of inaccurate and underestimated investment costs with the increasing number of clusters even for the best-performed method, H(Ward), whose estimated solution approaches the optimal value at $K = 300$. To this end, the optimal power flow patterns of all transmission lines $\hat{F}$ are proposed to be used as clustering variables in the conventional framework (M2) to decrease the required number of scenarios for the optimal TNEP solution. Comparing Figure 7-8 with Figure 7-6, it is constructive to note that the estimated investment costs of all the clustering technique indeed get improved by using the OPF patterns as clustering variables, especially for K-medoids method that achieves

the objective investment cost steadily after $K = 200$ clusters. Nevertheless, as shown in Figure 7-9, an obvious improvement has been observed when considering the OPF patterns of existing lines and candidate lines as well as investment scenarios and no investment scenarios separately (M3). Regarding the network expansion plan, the proposed OPF based bi-level clustering framework with H(Ward) exhibits the best estimation that achieves the optimal decisions from $K = 36$, resulting in approximately 80% scenarios reduction. This enhancement is contributed to the fact that the scenarios with investment requirements in the full TNEP problem (i.e. 14,251 scenarios in this case) should also lead to investment when solving the problem per operating points. However, it is important to claim that the contrary of the above statement is not always true because it is a probabilistic problem when solving the TNEP problem considering all the scenarios together. For example, the 9th candidate line is planned to be built in some scenarios when solving the problem individually, however, it is not worth to invest this line eventually according to the optimal investment decision obtained by solving the full problem. Although such great achievements have been attained by M3, the main contribution of this thesis is to propose the idea that it is the most efficient way to select the representative scenarios via clustering directly on the target of problem, which is the expansion plan for TNEP problem. Consequently, the target-based scenarios selection framework (M4) presents its superior performance than all the other methods that it is capable of reaching the optimal solution at $K=15$, which is about 50% further reduction of M3, when using the most efficient clustering technique, K-means, in this case. Overall, the best-performed clustering technique for each method can be concluded as follows: H(Ward) is the most appropriate clustering technique for M1 and M3; Centroid-based clustering techniques, K-medoids and K-means, can be regarded as the suitable clustering techniques for M2 and M4, respectively. Figure 7-11 shows the optimal total costs associated with the transmission expansion plan obtained by solving the TNEP problem using the tested methods M1, M2, M3, and M4 with their most appropriate clustering techniques. It is evident that the performance regarding the total costs is highly consistent with the results concluded from the view of investment costs.

Note that the tested scenarios selection methods and the TNEP optimization problem were implemented in MATLAB and FICO Xpress, respectively, and run on an Intel Xeon E5-2690 PC with 8 cores. For comparison, time consumptions of solving the full TNEP problem as well as the TNEP problem with the minimum required number of scenarios for each type of selection method with its best-performed clustering technique are given in Table 7-3. Comparing to the computation times of the full problem, it is constructive to highlight that the computational cost has been successfully reduced from hours to seconds when employing the clustering-based scenarios selection methods. In details, clustering based on OPF patterns (M2) can reduce about 47% computational cost comparing to M1, while the OPF based bi-level clustering method (M3) further improve M2 by saving roughly 70% computing times. Finally, another 73% reduction in the computational time of M2 is accomplished via the proposed target-based framework (M4).

Table 7-3. TNEP Solutions and Computation Times(Seconds)

Benchmarks	Optimal clustering technique	Minimum required K	CPU Times(s)
M1	H(Ward)	14,251	1.47×10^5
M2	K-medoids	300	198.28
M3	H(Ward)	200	103.78
M4	K-means	36	30.13
	K-means	15	8.07

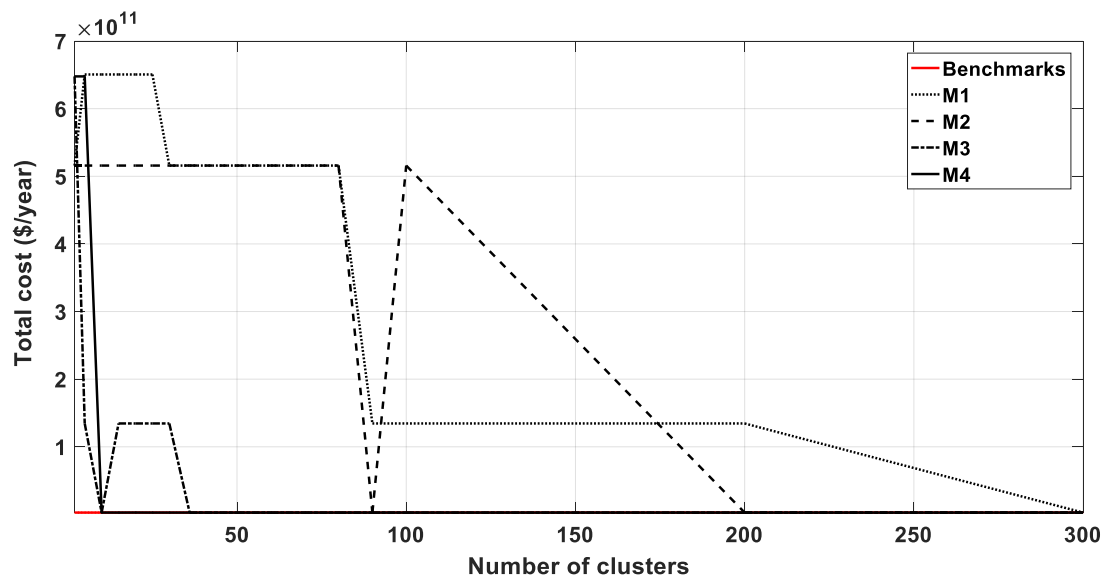


Figure 7-11: Comparisons among the TNEP solutions of different selection algorithms according to total cost. Note that the result of the best-performed clustering method is chosen for each algorithm (i.e. M1:H(Ward), M2: K-medoids, M3: H(Ward), M4(K-means)).

7.5 Conclusions

This thesis has proposed a novel target-based scenarios selection framework for choosing the representative operating points for the TNEP problem with large penetration of renewable energy resources. Among varieties of clustering methods, we choose the unsupervised clustering techniques: centroid-based clustering (K-means, K-medoids), connectivity-based clustering (H(Average), H(Complete), and H(Ward)), and distribution-based clustering (GMM). Also, beyond the conventional clustering variable, demand-generation patterns, we propose three new clustering variables $\hat{F}_E$, $\hat{F}_C$, and $\hat{I}_C$, allowing the historical scenarios to be classified directly based on their effects. In addition, conventional clustering based framework and the proposed bi-level clustering framework has been introduced and integrated with the abovementioned clustering variables to construct four types of tested scenarios selection methods: clustering based on demand-wind patterns (M1), clustering based on all OPF patterns (M2), bi-level clustering based on OPF patterns (M3), target-based bi-level clustering (M4). To evaluate the performance of the proposed method and the other combinations, the selected scenarios are used to solve the TNEP problem based on a modified IEEE-118 system with real historical data in France.

One major conclusion stemming from the analysis is that the proposed target-based scenarios selection framework (M4) exhibits superior performance when compared to the other methods, indicated by its fastest convergence speed to approach the optimal TNEP solutions. In addition, another key advantage of the proposed method is that we can define the maximum number of clusters that required for the TNEP problem by counting the actual number of existing combinations of expansion plans which is far less than the theoretical number 2^{n_c} in practice. Furthermore, the main contributions of this work is to propose three major directions in terms of the clustering based scenarios selection method: (1) clustering variables selection; (2) clustering technique selection; (3) representative scenario selection in each constructed cluster. According to the results, we conclude that clustering based on the effects of the input data indeed improves the performance of the selected scenarios obtained by clustering the input dataset itself. Regarding the OPF patterns, constructed by pre-solving the TNEP problem per scenarios, it is of interest to consider the existing lines and candidate lines as well as

investment scenarios and no investment scenarios, separately, realized by the proposed bi-level clustering framework. Although a great achievement has been obtained via implementing the OPF based bi-level clustering framework, directly based on the target outputs, investment costs, exhibit the highest efficiency to reach the optimal solution. For the second direction, various clustering techniques have been demonstrated to have significantly different performances for each scenario selection method. Also, it shows that the conventional clustering validation indicators cannot be used as criteria to select the most appropriate clustering technique for this specific problem.

Finally, a primary conclusion of choosing the scenario in each cluster is that taking the actual median in each cluster is more suitable for OPF patterns based methods (M2) and (M3), whereas using the mean value can result in a good performance for the target-based method (M4).

Future research should be mainly devoted to further improve the efficiency of the proposed framework, and implementing it into a larger system with more candidate transmission lines. With the increasing number of candidate lines in the TNEP problem, traditional clustering methods may suffer from the high-dimensionality problem. Therefore, dimensionality reduction techniques (e.g. PCA) can be integrated into the framework. Otherwise, more advanced clustering techniques could be considered to handle this issue and to enhance the clustering quality. Moreover, the idea of target-based clustering can also be extended to the other applications such as generation planning and energy storage planning.

8. Conclusion and Future Works

This thesis proposes novel data analytic methods for dealing with the challenges arising from the influx of extremely large datasets, termed big data, in power system with the large-scale integration of intermittent energy sources. This chapter highlights the key contributions and conclusions of this thesis and outlines the most promising avenues for further research.

8.1 Evaluating Composite Approaches to Modelling High-Dimensional Stochastic Variables in Power Systems

To handle the challenge Q1 described in Chapter 1.2, a novel composite modelling approach has been proposed for modelling high-dimensional stochastic variables in power systems (Chapter 3). The main contribution is an in-depth evaluation of potential synergies between different components with the aim to identify well-performing composite modelling approaches. Among varieties of data mining techniques and vine copulas types, we choose the unsupervised clustering techniques: k-means, hierarchical clustering and GMM, the dimensionality reduction techniques: PCA and LPP as well as C-vine, D-vine copulas. Before assessing combinations of the above techniques, clustering validation indicators CDI and MIA are used to determine the optimal number of clusters and the optimal clustering method. On the basis of a 100-dimensional measurement dataset from the French transmission system, the calculation of these indicators for each clustering method shows that the most appropriate clustering technique is the hierarchical clustering with average linkage. To evaluate the performance of the proposed composite modelling schemes with different combinations of the aforementioned techniques, visual comparisons, two-sample K-S tests, and multivariate two-sample tests are performed to test the reconstructed marginal distributions and dependency patterns of the sampled measurements. The results of the comparisons suggest the use of a modelling scheme that combines hierarchical clustering (average linkage), LPP, and D-vine copulas. Also, the results show that the performance of the proposed modelling schemes mostly depends on the choice of dimensionality reduction techniques. Overall, the composite modelling approach with LPP performs better than with PCA; in addition, D-vine copulas are more suitable for modelling the high-dimensional dependence encountered in the examined dataset.

Furthermore, case study applications on the calculation of a system's PNS profile and security assessment based on DTs demonstrate the benefits of integrating the developed statistical technique within a Monte Carlo simulation framework in the context of system security. The modeling approach presented here is set to become increasingly relevant with the advent of the smart grid and increasing penetration of intermittent generation. On one hand, many TSOs are bound to reduce controllability of a large part of their generation fleet and are already relying more and more on statistical techniques to forecast available resources and ensure robustness. On the other hand, measurement devices are being deployed across the network, gathering data that can be useful in characterizing in detail the statistical properties of stochastic parameters. All the above entail a fundamental shift from deterministic analysis to Monte Carlo simulation frameworks used to explore the uncertainty state-space. The present research, with its suitability for high-dimensional applications, enables operators to move beyond the selection bias of methods that rely exclusively on past observations.

8.2 C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data

In Chapter 4, a novel copula-based mixture model clustering algorithm has been proposed based on C-vine representations for grouping electrical customers according to consumption data for dealing with Q2 in Chapter 1.2. C-vine copulas, which can capture complex dependency structures by employing cascades of bivariate distribution, have been, for the first time, integrated into a mixture model-based clustering method. In order to address the high computational burden of the proposed scheme, model fitting takes place in a lower-dimension space where we exploit the synergy between PCA and the C-Vine structure's hierarchy. To evaluate the performance of the proposed CVMM algorithm, a set of well-established clustering validation indicators (CDI, MDI, DBI and MIA) have been used.

One major conclusion stemming from the analysis performed on a large dataset containing half-hourly smart meter recordings is that the CVMM method exhibits superior performance when compared to existing methods, indicated by the lower validation indicators' values. We also highlight the method's ability to detect outliers,

in contrast to conventional techniques. The classification result of different techniques was further assessed according to a decision-tree based classification module. The result of the DTs' performance is highly consistent with the outcome of the clustering validation assessment. This result highlights the superiority of the proposed approach in a practical context when assigning new consumers to existing classes. In addition, the proposed method is used to analyze load partitioning behavior under different loading conditions such as different calendar seasons and days. This increases the dimensionality of the clustering problem but could uncover particular customer classes that have markedly daily and seasonal changes in their behavior, opening up the possibility for more tailored tariff structures. The superior performance of the proposed method is also well demonstrated under different conditions in terms of the indicator values and the visual comparisons.

8.3 Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data

For Q3, a comprehensive analysis of the dataset, high-resolution residential demand measurements across 2,639 customers collected via the LCL smart meter, has been carried out in Chapter 5 that enables us to gain valuable insights towards the evolution of residential electricity consumption patterns in the UK and evaluate how they are affected by consumer demographic characteristics such as household occupancy and wealth level. In addition, peak and diversified demand metrics have been analysed and quantified, aiming to inform future distribution planning processes. The thesis first introduces distribution planning practices currently applied in the UK, based on the concept of ADMD, highlighting the practical significance of accurate diversified demand metrics. Subsequently, the LCL dataset is presented and analysed. Major conclusions stemming from this analysis are that the ADMD value calculated across all customers suggests that currently adopted values may be overestimating residential demand while also overestimating the degree with which ADMD converges as a function of connected customers. In addition, household occupancy is shown to be a primary driver of peak demand of individual households. Finally, both household

occupancy and wealth level are shown to drive diversified demand, suggesting that there may be merit in novel planning methods that consider customers' demographic makeup.

8.4 Transmission Network Expansion Planning with Stochastic Multivariate Load and Wind Modelling

For the purpose of generating sufficient scenarios for simulating the TNEP problem (Q4), this thesis proposes a two-stage TNEP model integrated with a multivariate Gaussian copula modelling framework. A modified version of IEEE 118-bus system is used as the test network in this research. Based on real historical stochastic variables from the French transmission system, multivariate dependent load and wind samples of 128 dimensions are generated by the proposed MGC algorithm, and compared with the results of BiInd and BiGc methods. A visual comparison between these three methods shows that the inter-spatial dependence can be re-constructed much more accurately when employing a multivariate Gaussian copula model. Furthermore, to evaluate the solution quality, a Monte Carlo bounding technique is implemented to build a 95% CI on the optimality gap of the solution of each method. The results of the comparison suggest that the CI width and the expected total cost of the TNEP problem can be substantially reduced when considering the multivariate dependence of load and wind.

8.5 A Target-Based Scenarios Selection Approach for Transmission Network Expansion Planning

Finally, a novel target-based scenarios selection framework for choosing the representative operating points (Q5) has been proposed for the TNEP problem with large penetration of renewable energy resources. Among varieties of clustering methods, we choose the unsupervised clustering techniques: centroid-based clustering (K-means, K-medoids), connectivity-based clustering (H(Average), H(Complete), and H(Ward)), and distribution-based clustering (GMM). Also, beyond the conventional clustering variable, demand-generation patterns, we propose three new clustering variables $\hat{F}_E$, $\hat{F}_C$, and $\hat{I}_C$, allowing the historical scenarios to be classified directly based on their effects. In addition, conventional clustering based framework and the proposed bi-level clustering framework has been introduced and integrated with the abovementioned clustering variables to construct four types of tested scenarios selection methods: clustering based on demand-wind patterns (M1), clustering based on all OPF patterns

(M2), bi-level clustering based on OPF patterns (M3), target-based bi-level clustering (M4). To evaluate the performance of the proposed method and the other combinations, the selected scenarios are used to solve the TNEP problem based on a modified IEEE-118 system with real historical data in France.

One major conclusion stemming from the analysis is that the proposed target-based scenarios selection framework (M4) exhibits superior performance when compared to the other methods, indicated by its fastest convergence speed to approach the optimal TNEP solutions. In addition, another key advantage of the proposed method is that we can define the maximum number of clusters that required for the TNEP problem by counting the actual number of existing combinations of expansion plans which is far less than the theoretical number 2^{n_c} in practice. Furthermore, the main contributions of this work is to propose three major directions in terms of the clustering based scenarios selection method: (1) clustering variables selection; (2) clustering technique selection; (3) representative scenario selection in each constructed cluster. According to the results, we conclude that clustering based on the effects of the input data indeed improves the performance of the selected scenarios obtained by clustering the input dataset itself. Regarding the OPF patterns, constructed by pre-solving the TNEP problem per scenarios, it is of interest to consider the existing lines and candidate lines as well as investment scenarios and no investment scenarios, separately, realized by the proposed bi-level clustering framework. Although a great achievement has been obtained via implementing the OPF based bi-level clustering framework, directly based on the target outputs, investment costs, exhibit the highest efficiency to reach the optimal solution. For the second direction, various clustering techniques have been demonstrated to have significantly different performances for each scenario selection method. Also, it shows that the conventional clustering validation indicators cannot be used as criteria to select the most appropriate clustering technique for this specific problem. Finally, a primary conclusion of choosing the scenario in each cluster is that taking the actual median in each cluster is more suitable for OPF patterns based methods (M2) and (M3), whereas using the mean value can result in a good performance for the target-based method (M4).

8.6 Future Work

According to the conclusions of this thesis, developing the proposed methods and employing them into more research areas are of interest to be investigated in future work.

For the proposed data modelling framework, further research will focus on applying the proposed modelling approach to a broad range of problems that involve decision-making under high-dimensional uncertainty and require the generation of large populations of system states. Such problems include stability analysis around a population of anticipated operating points as well as the application of machine learning techniques to train stability predictors. There is also benefit in carrying out further tests on larger datasets spanning thousands of variables, as in large-area power systems. In terms of the model evaluation, additional metrics (e.g. [47]) could be considered to further assess the quality of the dependency modelling.

In addition, the proposed CVMM method can be further improved by employing more possible types of vine constructions (e.g. R-vine), and exploring alternative dimension reduction techniques to further relieve the method's computational burden. Another topic of interest is the extension of the proposed model to mixed data sets so as to accommodate additional information such as load conditions; this is important to render the algorithm suitable for real-time pricing applications.

In terms of the demand diversity analysis, further research should propose novel methods for using the information extracted from the smart-meter data to design and operate distribution network in the smart grid era. Potential applications include the assessment of new connections for different mixes of consumers based on the detailed data. In addition, further temporal analysis of the dataset can provide useful insights towards electricity consumption patterns across different calendar seasons, days and hours. The outputs of such an analysis can be used for the optimization of outage management processes to minimize service disruptions.

For the TNEP problem, scenarios generation and scenarios selection are two important research areas to accommodate the computational issues arising from the significantly increasing size of data in power systems.

Regarding the proposed scenarios generation method, some more advanced high-dimensional modelling techniques should be implemented to deal with more complex dependence structure between stochastic variables. Also, contingency constraints could also be appended to the TNEP model. Another application of interest is to solve a multistage TNEP problem considering load and wind farm uncertainties.

On the other hand, future research of scenarios selection method should be mainly devoted to further improve the efficiency of the proposed framework, and implementing it into a larger system with more candidate transmission lines. With the increasing number of candidate lines in the TNEP problem, traditional clustering methods may suffer from the high-dimensionality problem. Therefore, dimensionality reduction techniques (e.g. PCA) can be integrated into the framework. Otherwise, more advanced clustering techniques could be considered to handle this issue and to enhance the clustering quality.

References

- [1] IBM. (Jul. 1, 2015). *The Four V's of Big Data*. [Online]. Available: <http://www.ibmbigdatahub.com/infographic/four-vs-big-data>.
- [2] M. Kezunovic, L. Xie and S. Grijalva, "The role of big data in improving power system operation and protection," *Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid (IREP), 2013 IREP Symposium*, Rethymno, 2013, pp. 1-9.
- [3] P. Panciatici, G. Bareux and L. Wehenkel, "Operating in the fog: security management under uncertainty", *IEEE Power & Energy Magazine*, vol. 10, pp. 40-49, Sep. 2012.
- [4] M. Misiti, Y. Misiti, G. Oppenheim, and J.-M. Poggi, "Optimized clusters for disaggregated electricity load forecasting," *REVSTAT Statist. J.*, vol. 8, no. 2, pp. 105–124, Nov. 2010.
- [5] M. Chaouch, "Clustering-Based Improvement of Nonparametric Functional Time Series Forecasting: Application to Intra-Day Household-Level Load Curves," *IEEE Trans. Smart Grid*, vol. 5, pp. 411-419, 2014.
- [6] J. Nazarko, A. Jurczuk, and W. Zalewski, "ARIMA models in load modelling with clustering approach," in *Proc. IEEE Russia Power Tech*, St. Petersburg, Russia, Jun. 27–30, 2005, pp. 1–6.
- [7] R. F. Chang and C. N. Lu, "Load profiling and its applications in power market," in *Proc. IEEE Power Eng. Soc. General Meeting*, Jul. 13–17, 2003, vol. 2.
- [8] G. Chicco, "Overview and performance assessment of clustering methods for electrical load pattern grouping," *Energy*, vol. 42, no. 1, pp. 68–80, Jun. 2012.
- [9] P. Jokar, N. Arianpoo and V. C. M. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns," in *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216-226, Jan. 2016.
- [10] R. Granell, C. J. Axon and D. C. H. Wallom, "Impacts of Raw Data Temporal Resolution Using Selected Clustering Methods on Residential Electricity Load Profiles," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3217-3224, Nov. 2015.
- [11] H. Yu, C. Y. Chung, K. P. Wong and J. H. Zhang, "A Chance Constrained Transmission Network Expansion Planning Method With Consideration of Load and Wind Farm Uncertainties," in *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1568-1576, Aug. 2009.
- [12] H. Park and R. Baldick, "Transmission Planning Under Uncertainties of Wind and Load: Sequential Approximation Approach," in *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2395-2402, Aug. 2013.
- [13] H. Park, R. Baldick and D. P. Morton, "A Stochastic Transmission Planning Model With Dependent Load and Wind Forecasts," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3003-3011, Nov. 2015.
- [14] Fitiwi, Desta Z., F. de Cuadra, L. Olmos, and M. Rivier. "A new approach of clustering operational states for power network expansion planning problems dealing with RES (renewable energy source) generation operational variability and uncertainty." *Energy* 90 (2015): 1360-1376.
- [15] I. Konstantelos, M. Sun, S. Tindemans, G. Strbac, S. Issad, P. Panciatici "Modelling of High-Dimensional Stochastic Variables Using Vine Copulas," *IEEE Trans. Power Syst*, Submitted.
- [16] M. Sun, I. Konstantelos, S. Tindemans and G. Strbac, "Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems," *PSCC '16*, Genoa, pp. 1-8, 2016.
- [17] M. Sun; I. Konstantelos; G. Strbac, "C-Vine Copula Mixture Model for Clustering of Residential Electrical Load Pattern Data," in *IEEE Transactions on Power Systems*, vol. PP, no.99, pp.1-1 doi: 10.1109/TPWRS.2016.2614366.
- [18] M. Sun, I. Konstantelos and G. Strbac, "Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data," *IEEE PES GM 2016*, Boston, Jul. 2016.
- [19] M. Sun, I. Konstantelos and G. Strbac, "Transmission Network Expansion Planning with Stochastic Multivariate Load and Wind," *PMAPS 2016*, Beijing, Oct. 2016.
- [20] M. Sun, F. Teng, I. Konstantelos and G. Strbac, "A Target-Based Scenario Selection Approach for Transmission Network Expansion Planning," *Applied Energy*, Submitted.
- [21] J. MacQueen, "Some methods for classification and analysis of multivariate observations." in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Univ. California Press, Berkeley, CA, Vol. 1, pp. 281-297, 1967.
- [22] E.M. Mirkes, K-means and K-medoids applet. University of Leicester, 2011.
- [23] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, 32(3), pp. 241–254, Sep. 1967.
- [24] A. D. Gordon, "A review of hierarchical classification," *J Roy Stat Soc A Sta*, vol. 150, pp. 119–137, 1987.
- [25] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.
- [26] G.J. McLachlan and K.E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

- [27] A Dempster, N. Laird, and D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39 (Series B): 1-38, 1977.
- [28] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1989.
- [29] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.
- [30] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems* 16.. Cambridge, MA: MIT Press, 2003.
- [31] X. He, D. Cai, S. Yan, and H. Zhang, "Neighborhood preserving embedding," in *Proc. IEEE Int. Conf. Computer Vision*, vol. 2, pp. 1208–1213, 2005.
- [32] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
- [33] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems* 14, Vancouver, British Columbia, Canada, 2002.
- [34] H. Joe, "Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters", *Distribution with Fixed Marginals and Related Topics*, 1996.
- [35] J. Mai and M. Scherer, *Simulating Copulas: Stochastic Models, Sampling Algorithms and Applications*, World Scientific, 2012.
- [36] T. Bedford and R.M. Cooke, "Probability density decomposition for conditionally dependent random variables modeled by vines", *Annals of Mathematics and Artificial Intelligence*, vol. 32, pp. 245-268, 2001.
- [37] K. Aas, C. Czado, A. Frigessi and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 182-198, 2009.
- [38] D. Kurowicka, R.M. Cooke, *Uncertainty Analysis with High Dimension Dependence Modeling*, John & Wiley, 2012, Chichester, 2012.
- [39] T. Bedford and R.M. Cooke, "Vines – a new graphical model for dependent random variables", *Annals of Statistics*, vol. 30, no. 4, pp. 1031-1068, 2002.
- [40] M. Rosenblatt, Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3), pp.470-472. 1952.
- [41] G. Papaefthymiou and D. Kurowicka, "Using copulas for modelling stochastic dependence in power system uncertainty analysis," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 40–49, Feb. 2009.
- [42] H. Valizadeh Haghi, M. Tavakoli Bina, and M. A. Golkar, "Nonlinear modelling of temporal wind power variations," *IEEE Trans. Sustain. Energy*, vol. 4, no. 4, pp. 838–848, Oct. 2013.
- [43] FP7 iTesla project, <http://www.itesla-project.eu/>
- [44] F. J. Massey, "The Kolmogorov–Smirnov test for goodness of fit," *J. Amer. Statist. Assoc.*, vol. 46, no. 253, pp. 68–78, Mar. 1951. [Online]. Available: <http://www.jstor.org/stable/2280095>
- [45] B. Aslan and G. Zech, "New test for the multivariate twosample problem based on the concept of minimum energy," *Journal of Statistical Computation and Simulation*, vol. 75, no. 2, pp. 109–119, 2005.
- [46] G. Chicco, R. Napoli, and F. Piglion, "Comparisons among clustering techniques for electricity customer classification," *IEEE Trans. Power Syst.*, vol. 21, no. 2, pp. 933–940, May 2006.
- [47] P. Pinson, R. Girard, "Evaluating the quality of scenarios of short-term wind power generation," *Applied Energy*, vol.96, pp. 12-20, 2012.
- [48] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang and R. L. Szabados, "Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, March 2015.
- [49] I. Benítez, A. Quijano, J.-L. Díez, and I. Delgado, "Dynamic clustering segmentation applied to load profiles of energy consumption from Spanish customers," *International Journal of Electrical Power & Energy Systems*, vol. 55, pp. 437-448, 2014.
- [50] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1120–1128, Aug. 2007.
- [51] F. McLoughlin, A. Duffy, M. Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, *Applied Energy*, Vol. 141, pp. 190-199, Mar. 2015.
- [52] G. Chicco, O.-M. Ionel, and R. Porumb, "Electrical load pattern grouping based on centroid model with ant colony clustering," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1706–1715, May 2013.
- [53] S.V. Verdu, M.O. Garcia, C. Senabre, A.G. Martin and F.J.G. Franco, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," *IEEE Transactions in Power Systems*, vol. 21, no. 4, pp. 1672-1682, Nov. 2006.
- [54] J. Nazarko and Z. A. Styczynski, "Application of statistical and neural approaches to the daily load profile modelling in power distribution systems," *Proc. IEEE Transm. Dist. Conf.*, New Orleans, LA, Apr. 11–16, 1999, vol. 1, pp. 320–325.
- [55] Y. Wang; Q. Chen, C. Kang; M. Zhang; K. Wang and Y. Zhao, "Load profiling and its application to demand response: A review," *Tsinghua Science and Technology*, vol.20, no.2, pp.117-129, Apr. 2015.

- [56] C. Fraley and A.E. Raftery, "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578-588, 1998.
- [57] R. Li, Z. Wang, C. Gu, F. Li, H. Wu, "A novel time-of-use tariff design based on Gaussian Mixture Model," *Applied Energy*, vol. 162, pp. 1530-1536, Jan. 2016.
- [58] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt and P. Järventausta, "Enhanced Load Profiling for Residential Network Customers," in *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 88-96, Feb. 2014.
- [59] W. Labeeuw, G. Deconinck, "Residential electrical load model based on mixture model clustering and Markov models," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1561-1569, Aug. 2013.
- [60] R. Granell, C. J. Axon, D. Wallom, "Clustering disaggregated load profiles using a Dirichlet process mixture model," *Energy Conversion and Management*, pp. 507 – 516, Mar. 2015.
- [61] A. Tewari, M. J. Giering, and A. Raghunathan, "Parametric characterization of multimodal distributions with non-gaussian modes," *Proc. IEEE International Conference on Data Mining Workshops*, pp. 286-292, 2011.
- [62] M. Rey, V. Roth, "Copula mixture model for dependency-seeking clustering," in: *Proceedings of the 29th International Conference on Machine Learning, ICML*, pp. 927-934, 2012.
- [63] I. Kosmidis and D. Karlis, "Model-based clustering using copulas with applications", arXiv:1404.4077, 2014.
- [64] K. Aas, C. Czado, A. Frigessi and H. Bakken, "Pair-copula constructions of multiple dependence," *Insurance: Mathematics and Economics*, vol. 44, no. 2, pp. 182-198, 2009.
- [65] C.J. Stone, "Optimal rates of convergence for nonparametric estimators," *The Annals of Statistics*, vol. 8, pp. 1348-1360, 1980.
- [66] T. Nagler and C. Czado, "Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas," arXiv:1503.03305v3, May 2016.
- [67] A. Albert and R. Rajagopal, "Smart Meter Driven Segmentation: What Your Consumption Says About You," *IEEE Trans. Power Systems*, vol. 28, pp. 4019-4030, 2013.
- [68] Y. Wang, Q. Chen, C. Kang and Q. Xia, "Clustering of Electricity Consumption Behavior Dynamics toward Big Data Applications," *IEEE Trans. Power Syst.*, vol. PP, no. 99, pp.1-11, Apr. 2016.
- [69] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 334, pp. 1492-1496, 2014.
- [70] E. D. Varga, S. F. Beretka, C. Noce and G. Sapienza, "Robust Real-Time Load Profile Encoding and Classification Framework for Efficient Power Systems Operation," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1897-1904, July 2015.
- [71] S. Haben, C. Singleton and P. Grindrod, "Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 136-144, Jan. 2016.
- [72] R. Al-Otaibi, N. Jin, T. Wilcox and P. Flach, "Feature Construction and Calibration for Clustering Daily Load Curves from Smart-Meter Data," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp. 645-654, April 2016.
- [73] M. Sun, I. Konstantelos, S. Tindemans and G. Strbac, "Evaluating composite approaches to modelling high-dimensional stochastic variables in power systems," *PSCC '16*, Genoa, pp. 1-8, 2016.
- [74] F. Dellaert "The expectation maximization algorithm," *College of Comput., GATECH*, Atlanta, GA, Tech. Rep. GIT-GVU-02-20, 2002.
- [75] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on datamining techniques," *IEEE Transactions on Power Systems*, vol. 20, no. 2, pp. 596-602, May 2005.
- [76] M. Sun, I. Konstantelos and G. Strbac, "Analysis of Diversified Residential Demand in London using Smart Meter and Demographic Data," *IEEE PES GM 2016*, Boston, Jul. 2016.
- [77] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery and W.L. Ruzz, "Model-Based Clustering and Data Transformations for Gene Expression Data," *Bioinformatics*, vol. 17, pp. 977-987, 2001.
- [78] Elexon, Definitions of seasons and day types: v.1.01, 2014.
- [79] J. Boggis, "Diversity, bias and balance," *Distrib. Elect.*, pp. 357-362, July 1953.
- [80] C. Barteczko-Hibbert, After Diversity Maximum Demand (ADMD) Report. Report for the 'Customer-Led Network Revolution' project: Durham University, Feb, 2015.
- [81] D. H. O. McQueen, P. R. Hyland, and S. J. Watson, "Monte Carlo simulation of residential electricity demand for forecasting maximum demand on distribution networks," *IEEE Trans. Power Syst.*, vol. 19, no. 3, pp. 1685-1689, Aug. 2004.
- [82] I. Richardson, M. Thomson, D. Infield, and C. Clifford, "Domestic electricity use: A high-resolution energy demand model," *Energy Buildings*, vol. 42, no. 10, pp. 1878-1886, 2010.
- [83] A. Seneviratne, Information – Electrical Design for Distribution Networks : After Diversity Maximum Demand, HPC-3DC-07-0001-2012, 1-12.

- [84] K. C. Schneider and R. F. Hoad, "Initial transformer sizing for single-phase residential load," *IEEE Trans. Power Del.*, vol. 7, no. 4, pp. 2074–2081, Oct. 1992.
- [85] C. Bary, "Coincidence factor relationships of electric service load characteristics", *Trans AIEE*, vol. 64, September 1945.
- [86] C. Ziser, ZY Dong, "Probabilistic Modelling of Demand Diversity and its Relationship with Electricity Market Outcomes", *Proc. IEE PES General Meeting* 2007.
- [87] J. Nazarko, R. P. Broadwater, and N. I. Tawalbeh, "Identification of statistical properties of diversity and conversion factors from load research data," in *Proc. 9th Mediterranean Electrotechnical Conf. (MELECON 98)*, vol. 1, 1998, pp. 217–220.
- [88] Nabeel I. Tawalbeh, A.M. Al-Fahed Nuseirat " Peak and daily profile estimation for substations in rural networks" *Electrical Power and Energy Systems* 30 (2008) 202-206.
- [89] R. P. Broadwater, A. Sargent, A. Yarali, H. E. Shaalan, and J. Nazarko, "Estimating substation peaks from load research data," *IEEE Trans. Power Delivery*, vol. 12, pp. 451–456, Jan. 1997.
- [90] V. P. Chatlani, D. J. Tylavsky, D. C. Montgomery, and M. Dyer, "Statistical properties of diversity factors for probabilistic loading of distribution transformers," in *Proc. 39th North Amer. Power Symp.*, 2007, pp. 555–561.
- [91] E.ON, "Network Design Manual," 2006.
- [92] UK Power Networks, Low Carbon London Project. Available at: [http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Low-Carbon-London-\(LCL\)](http://innovation.ukpowernetworks.co.uk/innovation/en/Projects/tier-2-projects/Low-Carbon-London-(LCL))
- [93] UK Power Networks, "Use of smart meter information for network planning and operation," Report C1 for the "Low Carbon London" LCNF project, 2014.
- [94] H. Park and R. Baldick, "Transmission Planning Under Uncertainties of Wind and Load: Sequential Approximation Approach," in *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2395-2402, Aug. 2013.
- [95] R. Hemmati, R. A. Hooshmand and A. Khodabakhshian, "Comprehensive review of generation and transmission expansion planning," in *IET Generation, Transmission & Distribution*, vol. 7, no. 9, pp. 955-964, Sept. 2013.
- [96] L. L. Garver, "Transmission network estimation using linear programming," *IEEE Trans. Power App. Syst.*, vol. PAS-89, no. 7, pp. 1688–1697, Sep./Oct. 1970.
- [97] R. Romero, A. Monticelli, A. Garcia, and S. Haffner, "Test systems and mathematical models for transmission network expansion planning," *Proc. Inst. Elect. Eng., Gen., Transm., Distrib.*, vol. 149, no. 1, pp. 27–36, Jan. 2002.
- [98] A.S. Sousa and E.N. Asada, "Combined heuristic with fuzzy systems to transmission system expansion planning" *Elect. Power Syst. Res*, vol. 81, no. 1, pp 123-128, Jan. 2011.
- [99] H. Park, R. Baldick and D. P. Morton, "A Stochastic Transmission Planning Model With Dependent Load and Wind Forecasts," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3003-3011, Nov. 2015.
- [100] A. Hajebrahimi; A. Abdollahi; M. Rashidinejad, "Probabilistic Multiobjective Transmission Expansion Planning Incorporating Demand Response Resources and Large-Scale Distant Wind Farms," in *IEEE Systems Journal*, vol. PP, no.99, pp.1-1, 2015.
- [101] J. Choi, T. Tran, A. A. El-Keib, R. Thomas, H. Oh, and R. Billinton, "A method for transmission system expansion planning considering probabilistic reliability criteria," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1606–1615, Aug. 2005.
- [102] P. Sanchez-Martin, A. Ramos, and J. F. Alonso, "Probabilistic midterm transmission planning in a liberalized market," *IEEE Trans. Power Syst.*, vol. 20, no. 4, pp. 2135–2142, Nov. 2005.
- [103] M. Moeini-Aghtaie, A. Abbaspour and M. Fotuhi-Firuzabad, "Incorporating Large-Scale Distant Wind Farms in Probabilistic Transmission Expansion Planning—Part II: Case Studies," in *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1594-1601, Aug. 2012.
- [104] G. A. Orfanos, P. S. Georgilakis and N. D. Hatziaargyriou, "Transmission Expansion Planning of Systems With Increasing Wind Power Integration," in *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1355-1362, May 2013.
- [105] H. Yu, C. Y. Chung, K. P. Wong and J. H. Zhang, "A Chance Constrained Transmission Network Expansion Planning Method With Consideration of Load and Wind Farm Uncertainties," in *IEEE Transactions on Power Systems*, vol. 24, no. 3, pp. 1568-1576, Aug. 2009.
- [106] H. Park and R. Baldick, "Transmission Planning Under Uncertainties of Wind and Load: Sequential Approximation Approach," in *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 2395-2402, Aug. 2013.
- [107] C. Li, Z. Dong, G. Chen, F. Luo and J. Liu, "Flexible transmission expansion planning associated with large-scale wind farms integration considering demand response," in *IET Generation, Transmission & Distribution*, vol. 9, no. 15, pp. 2276-2283, 11 19 2015.

- [108] B. Klöckl and G. Papaefthymiou, "Multivariate time series models for studies on stochastic generators in power systems", *Electric Power Systems Research*, vol. 80, no. 3, pp. 265-276, 2010.
- [109] R. Singh, B. C. Pal and R. A. Jabr, "Statistical Representation of Distribution System Loads Using Gaussian Mixture Model," in *IEEE Transactions on Power Systems*, vol. 25, no. 1, pp. 29-37, Feb. 2010.
- [110] G. Papaefthymiou and D. Kurowicka, "Using Copulas for Modelling Stochastic Dependence in Power System Uncertainty Analysis", *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 40-49, 2009.
- [111] H. Valizadeh Haghi and S. Lotfifard, "Spatiotemporal Modelling of Wind Generation for Optimal Energy Storage Sizing," in *IEEE Transactions on Sustainable Energy*, vol. 6, no. 1, pp. 113-121, Jan. 2015.
- [112] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges" *Publ. Inst. Statist. Univ. Paris* **8**: 229-231, 1959.
- [113] J. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numer. Math.* vol. 4, no. 1, pp. 238-252, 1962 [Online]. Available: <http://dx.doi.org/10.1007/BF01386316>.
- [114] I. Konstantelos and G. Strbac, "Valuation of Flexible Transmission Investment Options Under Uncertainty," in *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 1047-1055, March 2015.
- [115] W. Mak, D. P. Morton, and K. Wood, "Monte Carlo bounding techniques for determining solution quality in stochastic programs," *Oper. Res. Lett.*, vol. 24, pp. 47-56, 1999.
- [116] G. Strbac, I. Konstantelos, M. Pollitt and R. Green, Delivering future-proof energy infrastructure, report for the National Infrastructure Commission, Feb. 2016.
- [117] R. Hemmati, R. A. Hooshmand and A. Khodabakhshian, "Comprehensive review of generation and transmission expansion planning," in *IET Generation, Transmission & Distribution*, vol. 7, no. 9, pp. 955-964, Sept. 2013.
- [118] L. L. Garver, "Transmission network estimation using linear programming," *IEEE Trans. Power App. Syst.*, vol. PAS-89, no. 7, pp. 1688-1697, Sep./Oct. 1970.
- [119] P. Panciatici, G. Bareux, and L. Wehenkel, "Operating in the fog: Security management under uncertainty," *IEEE Power Energy Mag.*, vol. 10, no. 5, pp. 40-49, Sep. 2012.
- [120] J. Choi, T. Tran, A. A. El-Keib, R. Thomas, H. Oh, and R. Billinton, "A method for transmission system expansion planning considering probabilistic reliability criteria," *IEEE Trans. Power Syst.*, vol. 20, no. 3, pp. 1606-1615, Aug. 2005.
- [121] H. Park, R. Baldick and D. P. Morton, "A Stochastic Transmission Planning Model With Dependent Load and Wind Forecasts," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3003-3011, Nov. 2015.
- [122] A. Hajebrahimi; A. Abdollahi; M. Rashidinejad, "Probabilistic Multiobjective Transmission Expansion Planning Incorporating Demand Response Resources and Large-Scale Distant Wind Farms," in *IEEE Systems Journal*, vol. PP, no.99, pp.1-1, 2015.
- [123] P. Sanchez-Martin, A. Ramos, and J. F. Alonso, "Probabilistic midterm transmission planning in a liberalized market," *IEEE Trans. Power Syst.*, vol. 20, no. 4, pp. 2135-2142, Nov. 2005.
- [124] G. A. Orfanos, P. S. Georgilakis and N. D. Hatziargyriou, "Transmission Expansion Planning of Systems With Increasing Wind Power Integration," in *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1355-1362, May 2013.
- [125] M. Moeini-Aghtaie, A. Abbaspour and M. Fotuhi-Firuzabad, "Incorporating Large-Scale Distant Wind Farms in Probabilistic Transmission Expansion Planning—Part II: Case Studies," in *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1594-1601, Aug. 2012.
- [126] Buygi MO, Balzer G, Shanechi HM, Shahidehpour M. Market-based transmission expansion planning. *IEEE Trans Power Syst* 2004; 19:2060e7. <http://dx.doi.org/10.1109/TPWRS.2004.836252>.
- [127] Zhao JH, Foster J, Dong ZY, Wong KP. Flexible transmission network planning considering distributed generation impacts. *IEEE Trans Power Syst* 2011;26: 1434e43. <http://dx.doi.org/10.1109/TPWRS.2010.2089994>.
- [128] S. Agapoff, C. Pache, P. Panciatici, L. Warland and S. Lumbreras, "Snapshot selection based on statistical clustering for Transmission Expansion Planning," *PowerTech, 2015 IEEE Eindhoven*, Eindhoven, 2015, pp. 1-6.
- [129] O. A. Abbas. "Comparisons between data clustering algorithms." *International Arab Journal of Information Technology (IAJIT)*, vol. 5. no. 3.2008.
- [130] S. Wagner and D. Wagner, "Comparing clusterings—An overview," Univ. Karlsruhe (TH), Karlsruhe, Germany, Tech. Rep. 2006-04, 2007.
- [131] G.Ciulla, V.L. Brano, A. D'Amico, Modelling relationship among energy demand, climate and office building features: A cluster analysis at European level, *Applied Energy*, Volume 183, 1 December 2016, Pages 1021-1034, ISSN 0306-2619, <http://dx.doi.org/10.1016/j.apenergy.2016.09.046>.
- [132] R. Li, F. Wang, C. Gu, F. Li, H. Wu, "A novel time-of-use tariff design based on Gaussian Mixture Model," *Applied Energy*, vol. 162, pp. 1530-1536, Jan. 2016.

- [133] Fintan McLoughlin, Aidan Duffy, Michael Conlon, A clustering approach to domestic electricity load profile characterisation using smart metering data, *Applied Energy*, Volume 141, 1 March 2015, Pages 190-199, ISSN 0306-2619, <http://dx.doi.org/10.1016/j.apenergy.2014.12.039>.
- [134] David Hsu, Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data, *Applied Energy*, Volume 160, 15 December 2015, Pages 153-163, ISSN 0306-2619, <http://dx.doi.org/10.1016/j.apenergy.2015.08.126>.
- [135] Dehghan, Shahab, Nima Amjady, and Antonio J. Conejo. "Reliability-constrained robust power system expansion planning." *IEEE Transactions on Power Systems* 31, no. 3 (2016): 2383-2392.
- [136] Torbaghan, Shahab Shariat, Madeleine Gibescu, Barry G. Rawn, and Mart Van der Meijden. "A market-based transmission planning for HVDC grid—case study of the North Sea." *IEEE Transactions on Power Systems* 30, no. 2 (2015): 784-794.
- [137] Li, Yunhao, and Jianxue Wang. "Flexible transmission network expansion planning under uncertainty based on a self-adaptive clustering technique." In *International Conference on Renewable Power Generation (RPG 2015)*, pp. 1-6. IET, 2015.
- [138] Fitiwi, Desta Z., F. de Cuadra, L. Olmos, and M. Rivier. "A new approach of clustering operational states for power network expansion planning problems dealing with RES (renewable energy source) generation operational variability and uncertainty." *Energy* 90 (2015): 1360-1376.
- [139] Q. Ploussard; L. Olmos; A. Ramos, "An operational state aggregation technique for transmission expansion planning based on line benefits," in *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1-1 doi: 10.1109/TPWRS.2016.2614368.
- [140] Du, Ershun, Ning Zhang, Chongqing Kang, Jianhua Bai, Lu Cheng, and Yi Ding. "Impact of Wind Power Scenario Reduction Techniques on Stochastic Unit Commitment." In *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)*, pp. 202-210. IEEE, 2016.
- [141] H. Park, R. Baldick and D. P. Morton, "A Stochastic Transmission Planning Model With Dependent Load and Wind Forecasts," in *IEEE Transactions on Power Systems*, vol. 30, no. 6, pp. 3003-3011, Nov. 2015.
- [142] J. Benders, "Partitioning procedures for solving mixed-variables programming problems," *Numer. Math.* vol. 4, no. 1, pp. 238–252, 1962 [Online]. Available: <http://dx.doi.org/10.1007/BF01386316>.
- [143] I. Konstantelos and G. Strbac, "Valuation of Flexible Transmission Investment Options Under Uncertainty," in *IEEE Transactions on Power Systems*, vol. 30, no. 2, pp. 1047-1055, March 2015.
- [144] J. MacQueen, "Some methods for classification and analysis of multivariate observations." in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Univ. California Press, Berkeley, CA, Vol. 1, pp. 281-297, 1967.
- [145] E.M. Mirkes, K-means and K-medoids applet. University of Leicester, 2011.
- [146] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, 32(3), pp. 241–254, Sep. 1967.
- [147] A. D. Gordon, "A review of hierarchical classification," *J Roy Stat Soc A Sta*, vol. 150, pp. 119–137, 1987.
- [148] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.
- [149] G.J. McLachlan and K.E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.
- [150] A Dempster, N. Laird, and D. Rubin. "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, 39 (Series B): 1-38, 1977.
- [151] H. Zhang, G. T. Heydt, V. Vittal, and J. Quintero, "An improved network model for transmission expansion planning considering reactive power and network losses," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3471–3479, Aug. 2013.

Appendix A

In this chapter, we give analytical expressions for the distribution functions of the five copula families used in the continuous sampling module. They have been compiled from and verified across different sources such as [34], [35], [37].

A1. The bivariate Gaussian copula

The bivariate Gaussian copula is given by:

$$C(x_1, x_2; \rho) = \Phi_\rho \left(\Phi^{-1}(x_1), \Phi^{-1}(x_2) \right), \quad (\text{A. 1})$$

where Φ is the standard normal distribution $N(0,1)$ and Φ^{-1} is its inverse. Φ_ρ is the standard normal distribution with correlation ρ .

The density of the bivariate Gaussian copula is given by:

$$c(x_1, x_2; \rho) = \frac{1}{\sqrt{1-\rho^2}} \exp \left(-\frac{\rho^2(x_1^2 + x_2^2) - 2\rho x_1 x_2}{2(1-\rho^2)} \right), \quad (\text{A. 2})$$

and the h and h^{-1} functions are:

$$h(x_1, x_2; \rho) = \Phi \left(\frac{\Phi^{-1}(x_1) - \rho \Phi^{-1}(x_2)}{\sqrt{1-\rho^2}} \right), \quad (\text{A. 3})$$

$$h^{-1}(x_1, x_2; \rho) = \Phi \left(\Phi^{-1}(x_1) \sqrt{1-\rho^2} + \rho \Phi^{-1}(x_2) \right). \quad (\text{A. 4})$$

A2. The bivariate Student-t copula

The bivariate Student's t-copula is given by:

$$C(x_1, x_2) = t_{v,\rho} \left(t_v^{-1}(x_1), t_v^{-1}(x_2) \right), \quad (\text{A. 5})$$

where $t_{v,\rho}$ is the cumulative distribution function of a bivariate $t_{v,\rho}$ distribution with correlation ρ .

The density of the bivariate Student's t-copula is given by:

$$c(x_1, x_2) = \frac{\Gamma(\frac{v+2}{2})/\Gamma(\frac{v}{2})}{v \pi dt(y_1, v) dt(y_2, v) \sqrt{1-\rho^2}} \left(1 + \frac{y_1^2 + y_2^2 - 2\rho y_1 y_2}{v(1-\rho^2)}\right)^{-\frac{v+1}{2}}, \quad (\text{A. 6})$$

where v and ρ are the parameters of the copula, the standard univariate student-t distribution with v degrees of freedom, expectation 0 and variance $\frac{v}{v-2}$. In addition, $y_1 = t_v^{-1}(x_1)$, $y_2 = t_v^{-1}(x_2)$, $t_v^{-1}(\cdot)$ and $dt(\cdot, v)$ are the quantile function and probability density.

The h and h^{-1} functions of the bivariate student's t-copula are:

$$h(x_1, x_2, \rho, v) = t_{v+1} \left(\frac{t_v^{-1}(x_1) - \rho t_v^{-1}(x_2)}{\sqrt{\frac{(v + (t_v^{-1}(x_2))^2)(1-\rho^2)}{v+1}}} \right), \quad (\text{A. 7})$$

$$h^{-1}(x_1, x_2, \rho, v) = t_v \left[t_{v+1}^{-1}(x_1) \sqrt{\frac{(v + (t_v^{-1}(x_2))^2)(1-\rho^2)}{v+1}} + \rho t_v^{-1}(x_2) \right]. \quad (\text{A. 8})$$

A3. The bivariate Clayton copula

The bivariate Clayton copula is given for $0 < \delta < \infty$ by:

$$C(x_1, x_2; \delta) = (x_1^{-\delta} + x_2^{-\delta} - 1)^{-\frac{1}{\delta}}, \quad (\text{A. 9})$$

The Clayton copula density is given by:

$$c(x_1, x_2; \delta) = (1 + \delta)(x_1 x_2)^{-\delta-1} (x_1^{-\delta} + x_2^{-\delta} - 1)^{-2-\frac{1}{\delta}}, \quad (\text{A. 10})$$

The h and h^{-1} functions of the bivariate Clayton copula are:

$$h(x_1, x_2; \delta) = x_2^{-\delta-1} (x_1^{-\delta} + x_2^{-\delta} - 1)^{-1-\frac{1}{\delta}}, \quad (\text{A. 11})$$

$$h^{-1}(x_1, x_2; \delta) = [(x_1 x_2^{\delta+1})^{-\frac{\delta}{\delta+1}} + 1 - x_2^{-\delta}]^{-\frac{1}{\delta}}. \quad (\text{A. 12})$$

A4. The bivariate Frank copula

The bivariate Frank copula is given for $0 \leq \delta < \infty$ and its distribution function is given by:

$$C(x_1, x_2; \delta) = -\delta^{-1} \log \left(\frac{[y - (1 - e^{\delta x_1})(1 - e^{\delta x_2})]}{y} \right), \quad (\text{A. 13})$$

where $y = 1 - e^{-\delta}$.

The copula density is given by:

$$c(x_1, x_2; \delta) = \frac{\delta y e^{-\delta(x_1+x_2)}}{[y - (1 - e^{\delta x_1})(1 - e^{\delta x_2})]^2} \quad (\text{A. 14})$$

The h and h^{-1} functions of the bivariate Frank copula are:

$$h(x_1, x_2; \delta) = \frac{e^{-\delta x_2}}{\frac{1 - e^{-\delta}}{1 - e^{-\delta x_1}} + e^{-\delta x_2} - 1} \quad (\text{A. 15})$$

$$h^{-1}(x_1, x_2; \delta) = -\log \frac{\left\{ 1 - \frac{1 - e^{-\delta}}{(x_1^{-1} - 1)e^{-\delta x_2} + 1} \right\}}{\delta} \quad (\text{A. 16})$$

A5. The bivariate Gumbel copula

The bivariate Gumbel copula is given for $1 \leq \delta < \infty$ by:

$$C(x_1, x_2; \delta) = \exp \left\{ -[(-\log x_1)^\delta + (-\log x_2)^\delta]^{\frac{1}{\delta}} \right\}, \quad (\text{A. 17})$$

The copula density is given by:

$$c(x_1, x_2; \delta) = C(x_1, x_2; \delta)(x_1 x_2)^{-1} \frac{(\log x_1 \log x_2)^{\delta-1}}{((- \log x_1)^\delta + (- \log x_2)^\delta)^{2-\frac{1}{\delta}}} \times \left[((- \log x_1)^\delta + (- \log x_2)^\delta)^{\frac{1}{\delta}} + \delta - 1 \right], \quad (\text{A. 18})$$

The h function of the bivariate Gumbel copula is:

$$h(x_1, x_2; \delta) = x_2^{-1} \exp \left(- [(- \log x_1)^\delta + (- \log x_2)^\delta]^{\frac{1}{\delta}} \right) \times \left(1 + \left(\frac{- \log x_1}{- \log x_2} \right)^\delta \right)^{-1+\frac{1}{\delta}} \quad (\text{A. 19})$$

As $h^{-1}(x_1, x_2; \delta)$ cannot be written in closed form, a numerical routine has to be used to invert the h function. The bisection method, explained below, was chosen for this task.