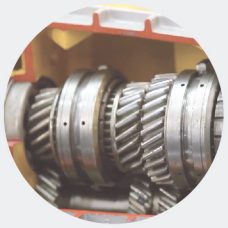
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO INDUSTRIAL






CLUSTERING OF RENEWABLE ENERGY ASSETS TO ENHANCE PERFORMANCE EVALUATION

SARA ISABEL GONÇALVES ABREU
Junho de 2024

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

CLUSTERING OF RENEWABLE ENERGY ASSETS TO ENHANCE PERFORMANCE EVALUATION

Sara Isabel Gonçalves Abreu

2024

Instituto Superior de Engenharia do Porto

Department of Mechanical Engineering

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

CLUSTERING OF RENEWABLE ENERGY ASSETS TO ENHANCE PERFORMANCE EVALUATION

Sara Isabel Gonçalves Abreu

1220525

Dissertation presented to the School of Engineering to fulfill the requirements necessary to obtain a Master's degree in Engineering and Industrial Management, carried out under the guidance of Doctor Maria de Fátima Coutinho Rodrigues.

2024

Instituto Superior de Engenharia do Porto

Department of Mechanical Engineering

AKNOWLEDGEMNTS

I want to start by expressing my deepest gratitude to all those who have supported me throughout the journey of completing this master's thesis.

First of all, I want to thank ISEP for all the knowledge it has given me over the last two years and for allowing me to further develop the skills I have acquired in a project in a business environment. I also want to show my deepest gratitude to my advisor, Doctor Fátima Rodrigues, for their unique and invaluable support, guidance, and insights that have significantly shaped this research. Their patience and encouragement have been instrumental in overcoming the challenges faced during this study. Also, thank Enlitia and, therefore, João Pereira for their insatiable support and for trusting in me to develop this project.

A special thanks to my family, whose love and support have been my anchor. To my parents, Fernanda and Luis, and my sister Petra, for their endless encouragement and for believing in me every step of the way. To my partner, David Freire, for their understanding and for providing me with much-needed breaks and motivation. And to the one that kept me company through the endless hours developing this project, Lua. Your presence in my life has been a constant source of strength and inspiration.

I would also like to acknowledge my colleagues and peers at ISEP for their camaraderie and for creating a stimulating academic environment.

Thank you all for making this journey a memorable and fulfilling one.

ABSTRACT

This study clusters solar inverters and wind turbines to aid Enlitia's clients in identifying assets similar to theirs based on historical power production, meteorological data, and power curve characteristics. This knowledge enables clients to optimize resource allocation and operational strategies, thereby avoiding unnecessary costs.

This project falls under the category of Data Mining and follows the CRISP-DM methodology. A crucial step in this approach is data cleaning, which involves treating null and duplicated values and reducing unnecessary features. During data cleaning, outlier values are identified and removed using various methods. For wind turbines, outliers are treated based on their power curve, which is defined by the power produced and the wind speed. For solar inverters, outliers are treated using the I-V curve, representing the DC power through the DC voltage and DC current.

Following data cleaning, the clustering phase begins. This project employs algorithms from three clustering categories: classical, ensemble, and time series clustering. Principal Component Analysis (PCA) is applied to the datasets to reduce computational costs while preserving at least 90% of the original variation in the data. If feature reduction results in less than the minimum variation, feature values are only normalized. The resultant datasets are used in classical and ensemble clustering.

In classical clustering, five hierarchical, two partitional, one soft, one model-based, and two density-based algorithms are applied. Five evaluation indexes, such as the silhouette score and the Davies-Bouldin index, assess the resulting segmentations. The top three classical algorithms proceed to ensemble clustering, where combinations of two and three algorithms are performed using major voting with weighted label assignment based on the best segmentations. Finally, two time series clustering algorithms are applied, with the data sets reduced to two components through the use of PCA.

The final step involves evaluating all obtained segmentations. The scores of each algorithm indicate that time significantly explains the variation in the data. For both solar and wind datasets, time series clustering produces the best segmentations.

KEYWORDS

Renewable Energy, Solar Inverters, Wind Turbines, Data Mining, CRISP-DM, Clustering, Time Series, PCA

RESUMO

Este estudo visa agrupar inversores solares e turbinas eólicas para ajudar os clientes da Enlitia a identificar ativos semelhantes aos seus com base na produção histórica de energia, dados meteorológicos e características das curvas de potência. Este conhecimento permite aos clientes otimizar a alocação de recursos e estratégias operacionais, evitando custos desnecessários.

Este projeto enquadra-se na categoria de *Data Mining* e segue a metodologia CRISP-DM. Um passo crucial nesta abordagem é a limpeza de dados, que envolve o tratamento de valores nulos e duplicados, e a redução de variáveis desnecessárias. Durante a limpeza de dados, os valores outliers são identificados e removidos utilizando vários métodos. Para as turbinas eólicas, os outliers são tratados com base na curva de potência, definida pela potência produzida e pela velocidade do vento. Para os inversores solares, os outliers são tratados utilizando a curva I-V, que representa a potência DC através da tensão DC e da corrente DC.

Após a limpeza de dados, inicia-se a fase de clustering. Este projeto utiliza algoritmos de três categorias de clustering: clustering clássico, ensemble e clustering de séries temporais. A Análise de Componentes Principais (PCA) é aplicada aos conjuntos de dados para reduzir os custos computacionais, preservando pelo menos 90% da variação original dos dados. Se a redução de características resultar em menos do que a variação mínima, os valores das características são apenas normalizados. Os conjuntos de dados resultantes deste procedimento são usados no clustering clássico e no ensemble clustering.

No clustering clássico, são aplicados cinco algoritmos hierárquicos, dois partitivos, um soft, um baseado em modelos e dois baseados em densidade. Cinco índices de avaliação, como o índice *silhouette* e o índice de Davies-Bouldin, avaliam as segmentações resultantes. Os três melhores algoritmos clássicos avançam para o ensemble clustering, onde combinações de dois e três algoritmos são realizadas usando votação majoritária com atribuição de pesos baseada nas melhores segmentações. Finalmente, dois algoritmos de clustering de séries temporais são aplicados, com os conjuntos de dados reduzidos a duas componentes, através do uso da PCA.

A etapa final consiste na avaliação de todas as segmentações obtidas. As pontuações de cada algoritmo indicam que o tempo tem grande importância na explicação da variação presente nos dados. Para ambos os conjuntos de dados, solar e eólico, o clustering de séries temporais produz as melhores segmentações.

PALAVRAS-CHAVE

Energia Renovável, Inversores Solares, Turbinas Eólicas, Data Mining, CRISP-DM, Clustering, Séries Temporais, PCA

INDEX

1 INTRODUCTION **1**
1.1 Problem of research, framework and relevance . . . 1
1.2 Research questions and objectives . . . 2
1.3 Methodological options . . . 3
1.4 Work Structure . . . 3

2 BIBLIOGRAFIC REVIEW **5**
2.1 Renewable Energy . . . 5
2.1.1 Wind Power . . . 6
2.1.2 Photovoltaic Power . . . 8
2.2 Data Mining . . . 10
2.2.1 CRISP-DM . . . 11
2.2.2 Clustering . . . 13
2.2.3 Clustering Algorithms . . . 14
2.2.4 Clustering Validation Metrics . . . 19
2.3 Related Work . . . 20

3 METHODOLOGY **23**
3.1 Business Understanding . . . 23
3.2 Data Understanding . . . 24
3.2.1 Wind Data . . . 24
3.2.2 Solar Data . . . 25
3.2.3 Satellite Data . . . 26
3.3 Data Preparation . . . 27
3.3.1 Wind Data . . . 27
3.3.2 Solar and Satellite Data . . . 31
3.4 Modelling . . . 33
3.4.1 Classic Clustering . . . 36
3.4.1.1 Hierarchical Clustering . . . 36
3.4.1.2 Partitional Clustering . . . 37
3.4.1.3 Soft Clustering . . . 38
3.4.1.4 Model-Based Clustering . . . 38
3.4.1.5 Density-Based Clustering . . . 39
3.4.1.6 Parameter Optimization . . . 39
3.4.1.7 Conclusions . . . 39
3.4.2 Ensemble Clustering . . . 40
3.4.3 Time Series Clustering . . . 40
3.5 Evaluation . . . 42
3.5.1 Clustering with Solar Data . . . 43
3.5.1.1 Classical Clustering . . . 43
3.5.1.2 Ensemble Clustering . . . 43

3.5.1.3 Time Series Clustering with all variables 44
3.5.1.4 Time Series Clustering with individual variables 45
3.5.1.5 Conclusions 46
3.5.2 Clustering with Wind Data 47
3.5.2.1 Classic Clustering 47
3.5.2.2 Ensemble Clustering 47
3.5.2.3 Time Series Clustering with all variables 48
3.5.2.4 Time Series Clustering with individual variables 48
3.5.2.5 Conclusions 49

4 CONCLUSION **51**
4.1 Final conclusions 51
4.2 Limitations and future work 52

A APPENDIX A **65**

B APPENDIX B **69**

C APPENDIX C **71**

D APPENDIX D **73**

E APPENDIX E **75**

F APPENDIX F **77**

G APPENDIX G **81**

H APPENDIX H **83**

I APPENDIX I **85**

J APPENDIX J **87**

A ANNEX A **89**

FIGURES INDEX

Figure 2.1 World Energy Capacity Over the Years (Data Source: (IRENA 2023)) 6
Figure 2.2 General Scheme of a WECS (Adapted from: (Hatziargyriou et al. 2000)) 7
Figure 2.3 Photovoltaic Effect (Penick & Louk 1998) 8
Figure 2.4 Electrical Circuit of a PV module (Adapted from: (Patel 1999)) 9
Figure 2.5 Grid-Connected PV System (Adapted from: (Mohamed & Sattar 2019)) 9
Figure 2.6 Phases of CRISP-DM (Adapted from: (Chapman 2000)) 12
Figure 2.7 Clustering Algorithms 15
Figure 2.8 Dendrogram of Agglomerative and Divisive clustering (Halkidi 2018) 16

Figure 3.1 Boxplots for Wind Data 25
Figure 3.2 Boxplots for Solar Data 26
Figure 3.3 Boxplots for Satellite Data 27
Figure 3.4 Pearson Correlation Matrix of Wind Data features 28
Figure 3.5 Power Curves with sigmoid restrictions 29
Figure 3.6 Power Curves with anomalous data behavior 30
Figure 3.7 Impact of each cleaning step on asset 49 30
Figure 3.8 Correlation Matrix of Solar and Satellite Data features 31
Figure 3.9 Initial I-V Curves 32
Figure 3.10 Results of Mean Square Method 33
Figure 3.11 Results of DBSCAN 33
Figure 3.12 Impact of each cleaning step on asset 12 34
Figure 3.13 Cumulative Explained Variance per number of components for each dataset 35
Figure 3.14 Dendrogram of Average Link (Clean Solar Data) 37
Figure 3.15 Elbow Plot 38
Figure 3.16 Visualization of Wind data points 38
Figure 3.17 Visualization of the Silhouette Scores for Clean Solar data 42
Figure 3.18 Clusters of K-Means with DTW distance metric with Clean Solar Data 45

Figure A.1 Histograms for Wind Data 65
Figure A.2 Histograms for Solar Data 66
Figure A.3 Histograms for Satellite Data 67

Figure B.1 MSM results for Wind data 69
Figure B.2 DBSCAN results for Wind data 69

Figure C.1 Power Curve Cleaning Results for assets 11, 12, 13, 14, 15, 16 71
Figure C.2 Power Curve Cleaning Results for assets 414, 415, 510, 511, 512 71

Figure D.1 I-V Curves after first clean 73

Figure E.1 I-V Curve Cleaning Results for assets 11, 12, 13, 14, 15, 16 75
Figure E.2 I-V Curve Cleaning Results for assets 37, 38, 39, 41, 42, 43 75

Figure F.1 Cluster Scatter Plots of Hierarchical algorithms with Clean Solar Data 77

Figure F.2 Cluster Scatter Plots of Partitional algorithms with Clean Solar Data 78
Figure F.3 Cluster Scatter Plot of Fuzzy C-Means (Soft Clustering) with Clean Solar Data 78
Figure F.4 Cluster Scatter Plot of Gaussian Mixture Model (Model-Based Clustering) with Clean Solar Data 78
Figure F.5 Cluster Scatter Plots of Density-Based algorithms with Clean Solar Data 79

Figure G.1 Cluster Scatter Plots of Ensemble Clustering with Clean Solar Data 81

Figure H.1 Clusters of SOM with Clean Solar Data 83

Figure J.1 Clusters of SOM with Non-Clean Solar Data (Time Series with *dc_power*) 87
Figure J.2 Clusters of K-Means with Euclidean distance metric, with Clean Wind Data (Time Series with *wind_direction*) 88

TABLES INDEX

Table 3.1 New number of rows in each dataset 34
Table 3.2 Percentage of variation of each feature present in the Principal Components (cleaned solar data) 36
Table 3.3 Best algorithms of Classic Clustering for each dataset 40
Table 3.4 Composition of time series of each dataset 41
Table 3.5 Metrics of the best three Classic Clustering algorithms for Solar data 43
Table 3.6 Ensemble Clustering Metrics Results for Clean Solar Data 43
Table 3.7 Ensemble Clustering Metrics Results for Non-Clean Solar Data 44
Table 3.8 Time Series Clustering Results for Solar Data with usual variables 44
Table 3.9 Time Series Clustering Results for Non-Clean Solar Data with Individual Variables . . . 46
Table 3.10 Classic Clustering Metrics Results for Wind Data 47
Table 3.11 Ensemble Clustering Metrics Results for Wind Data 48
Table 3.12 Time Series Clustering Results for Wind Data with usual variables 48
Table 3.13 Time Series Clustering Results for Clean Wind Data with individual variables 49

Table A.1 Estatistical Description of Wind Data 65
Table A.2 Estatistical Description of Solar Data 66
Table A.3 Estatistical Description of Satellite Data 67

Table I.1 Time Series Clustering Results for Clean Solar Data with individual variables 85
Table I.2 Time Series Clustering Results for Non-Clean Wind Data with individual variables . . . 85

LIST OF ABBREVIATIONS AND SYMBOLS

List of abbreviations

AC	Alternating Current
CH	Calinski-Harabasz
CLARA	Clustering for large applications
CRISP-DM	Cross-Industry Standard Process for Data Mining
DAS	Data Acquisition System
DM	Data Mining
DB	Davies-Bouldin
DC	Direct Current
EM	Expectation-Maximization
GMM	Gaussian Mixture Model
MI	Mutual Information
ML	Machine Learning
MSM	Mean Squared Method
NMI	Normalized Mutual Information
opt.	optimized
OPTICS	Ordering Points To Identify the Clustering Structure
PAM	Partitioning Around Medois
PCA	Principal Component Analysis
PC	Principal Component
PV	Photovoltaic
SS	Silhouette Score
SSE	Sum of Squared Errors
SVM	Support Vector Machines
WECS	Wind Energy Conversion System

1 INTRODUCTION

This chapter breaks down into three sections. The initial section outlines the research problem, explores its relevance in the current scenario, and delves into the framework within which it fits. The second section aims to describe the main research questions and the general and specific objectives. Finally, in the last section, the methodology options adopted by this study are presented.

1.1 Problem of research, framework and relevance

Nowadays, energy consumption originates from exhaustible sources such as fossil fuels or renewable sources like wind and solar. Unfortunately, energy consumption still highly relies on non-carbon-free sources, which implies considerable CO_2 emissions, the primary driver of global warming (Harrouz et al. 2020). The good news is that renewable energy production is growing yearly. In 2022, the industry noticed a 2% annual growth that balanced the significant fractional increases in solar power, accounting for 25%, and wind power, accounting for 14% of energy production (Haegel & Kurtz 2023).

Despite these positive trends in renewable energy, the efficient utilization and management of these resources pose unique challenges. To address these challenges, Data Mining (DM) is frequently used to allow companies and their customers to make more conscious decisions by providing them with essential knowledge about their assets in a simple and easy-to-interpret way, using statistics, mathematics, and Machine Learning (ML) techniques (Tougui et al. 2020).

Enlitia, a company dedicated to transforming the production, distribution, and consumption of renewable energy through the power of Artificial Intelligence, is value-oriented, human-centered, and forward-thinking. It also promotes simplicity in its work and a path to zero error to obtain excellence. The company originated as a spin-off from Smartwatt in 2023, already has more than 15 enterprise partners over three continents, and monitors renewable energy data from over 10,000 assets (*About - Enlitia* n.d.).

Therefore, Enlitia has put forth a study intending to develop an algorithm that provides valuable insights to clients to include in a product currently being developed by the company. The goal is to enable more informed decision-making while promoting sustainable power generation and enhancing renewable energy networks' overall performance and resilience.

Through applying analytical techniques and unsupervised ML algorithms, this study will offer valuable insights into optimizing resource allocation and operational strategies by grouping similar assets across multiple power plants based on historical power production, meteorological data, and power curve characteristics.

Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (Wirth & Hipp 2000), the solution for this problem involves implementing clustering algorithms using Python programming language to obtain groups of similar assets. These algorithms will use data relating to Enlitia's clients assets, which the company supplied.

1.2 Research questions and objectives

This study aims to answer two main research questions:

1. How can clustering techniques, supported by analytical methods and unsupervised ML algorithms, be effectively applied to group similar renewable energy assets based on historical power production, meteorological data, and power curve characteristics?
2. Furthermore, how does this clustering contribute to optimizing resource allocation and operational strategies, ultimately enhancing renewable energy networks' overall performance and resilience?

To be able to answer the questions thoroughly, two general objectives are defined:

- Create groups of similar assets according to different criteria to enhance the analysis of performance among different assets;
- Use analytical techniques and unsupervised ML algorithms.

To reach the general objectives with success, some specific objectives are set:

- Interpret the data;
- Analyze the data statistically;
- Apply different clustering algorithms;
- Evaluate the clustering algorithms through appropriate metrics;

1.3 Methodological options

The authors in (Feliciana et al. 2020) present a framework that organizes the methodological options into three topics: research approach, nature, and strategy.

In terms of the approach, one can classify the research as deductive, inductive, or hypothetical-deductive. When a researcher goes from established theories of the field of the study and tests them in the particular case in analysis, the research is deductive. The research is inductive if the researcher needs to take the opposite strategy and go from a particular case to a more general one due to the lack of an established body of knowledge (Feliciana et al. 2020, Sousa & Baptista 2014). Last, defended by (Walliman 2011), hypothetical-deductive research combines the first two through an interaction between experience and practice. Since this study uses existing knowledge about DM and clustering to implement a viable solution using quantitative data, the research adopts a deductive approach.

According to (Feliciana et al. 2020), the research nature can be defined as exploratory, explanatory, and descriptive, or, according to (Oliveira 2011), as descriptive, exploratory, analytical, synthetic, and active. This research has an analytical (or explanatory in the definition of (Feliciana et al. 2020)) nature since all the factors that can influence the problem will be carefully analyzed, so it is possible to evaluate their relative importance and how they are related (Feliciana et al. 2020, Oliveira 2011).

Finally, researchers can define the research strategy as experimental investigation, survey, case study, or investigation-action, among other methods. Since this study pretends to solve a real organizational problem, the adopted research strategy is investigation-action (Feliciana et al. 2020).

1.4 Work Structure

This dissertation comprises four chapters, incorporating this introductory chapter. The subsequent chapter, *Bibliographic Review* addresses the state of the art in renewable energies, emphasizing wind and solar power, which are the focal points of this study, along with Data Mining, with a specific focus on clustering. Following this, the *Methodology* chapter outlines the proposed case study for this project, adhering to the CRISP-DM approach. The *Conclusions* chapter presents final reflections, as well as faced limitations and suggestions for future work. The *Appendices* complement the presented work through graphs, cluster visualizations, and tables. Finally, *Annex A* presents the Integrity Declaration.

2 BIBLIOGRAFIC REVIEW

This chapter is composed of three sections. The first section presents an overview of the renewable energy industry and a more in-depth analysis of wind and photovoltaic power since they are the energy sources used for this study. The second section addresses the topic of data mining, the CRISP-DM methodology, an overview of clustering, some of its most used algorithms, and commonly used evaluation metrics. The last section analyzes studies related to this one.

2.1 Renewable Energy

Renewable energies, inherently natural, have been pivotal throughout history, serving diverse human needs. As humanity's primary energy source, they have been fueled by solar radiation, fuelwood, and draught animals. The Renaissance and Industrial Revolution ushered in advancements, leading to the dominance of coal. Fossil and nuclear energy gained prominence due to perceived cost advantages (Sørensen 1991), but the escalating energy demand, driven by population growth and recent industrial progress, has led more industrialized countries to actively pursue the development and utilization of renewable energies such as solar, hydropower, and wind power, among others (Olabi & Abdelkareem 2022, Saidi & Omri 2020).

However, some inconveniences come with this type of energy, such as the fact that the construction of renewable energy plants often requires the use of fossil fuels, the changes in water flows, and disturbance of land and ecosystems from hydroelectric reservoirs or the waste of ashes resulting of the combustion of biomass (Saidi & Omri 2020).

According to (Shahbaz et al. 2020), in 2013, only 19.1% of the final global energy consumption came from renewable energy sources. Since then, as presented in Figure 2.1, the installed capacity of non-fossil energy has been growing substantially. In 2022, the collective contribution of carbon-free energy sources, including hydro, nuclear, and renewables, represented 38% of the total electricity generation, with wind and solar power

being the major contributors (Haegel & Kurtz 2023). Some studies argue that it may be possible to transition completely to renewable energy by 2050 (Holechek et al. 2022, Ram et al. 2020).

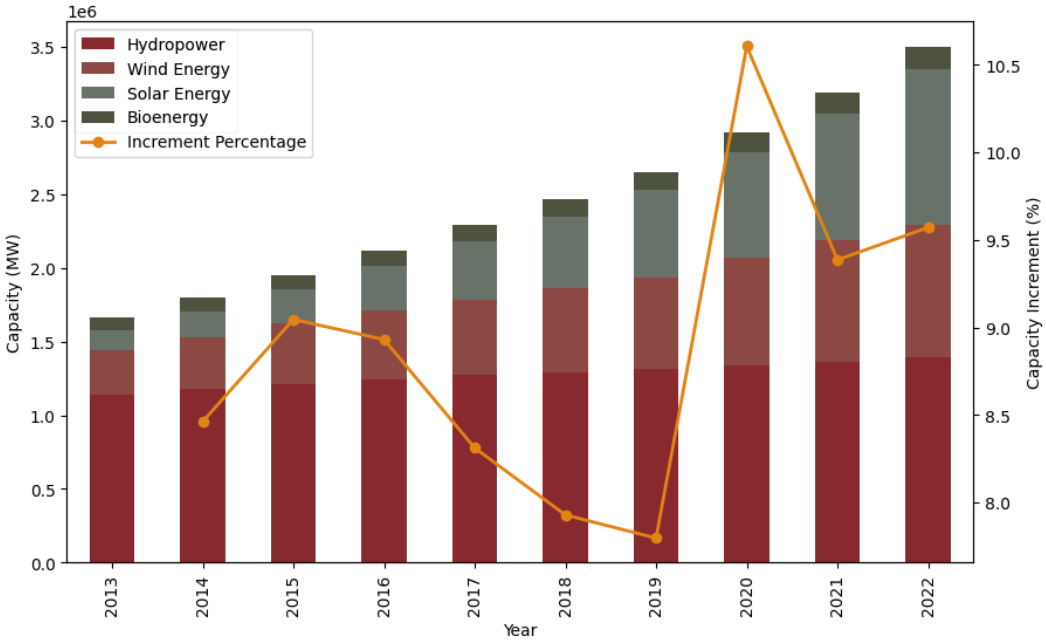


Figure 2.1 – World Energy Capacity Over the Years (Data Source: (IRENA 2023))

Presently, five primary forms of renewable energies are recognized: Biomass, Hydro, Geothermal, Solar, and Wind. Biomass energy involves converting organic materials, such as plants and animal waste, into heat or electricity through combustion, gasification, or anaerobic digestion. Hydropower is the utilization of falling or fast-flowing water to generate electricity or power machines; this involves converting a water source's gravitational potential or kinetic energy into usable power. Geothermal energy captures the Earth's heat, extending from the shallow ground to hot water and rocks beneath the surface (Alrikabi 2014, Egré & Milewski 2002). Solar and wind energy will be discussed in greater detail in the following sections, as they are the focus of this study.

2.1.1 Wind Power

Wind emerges as an alluring energy source, renowned for its renewable essence and economic prowess. Nevertheless, its intrinsic variability, dictated by spatial and temporal fluctuations in wind speed, introduces a nuanced complexity. Wind power production becomes subject to diverse variables, such as wind speed, direction, and temperature (Vargas et al. 2019).

The kinetic energy from the motion of the air, commonly called wind energy, is converted to electrical energy through the rotating blades of the wind turbines (Adnan et al. 2022, Genc & Ozden 2021). The mechanical power in the moving air (P, in W) can be translated in the equation 2.1, where ρ stands for air density (kg/m^3), A for the area swept by the rotor blades (m^3), and V the air velocity (m/s) (Patel 1999).

$$P = 1/2\rho AV \tag{2.1}$$

Nevertheless, the energy harnessed by the turbine blades constitutes only a portion of the upstream wind power, owing to losses and dissipation in the downstream wind. This portion can also be called the rotor efficiency (C_p, in %) and the power extracted by the rotor blades (P_0, in W) can be given by the equation 2.2 (Patel 1999).

$$P_0 = P \times C_p \tag{2.2}$$

Upon the extraction of mechanical energy by the turbine blades from the wind, the associated electrical generator, intricately connected to the rotor, converts this mechanical energy into electrical power, and subsequently, this electrical energy is transmitted to the grid. This transformation happens in the Wind Energy Conversion System (WECS), as represented in Figure 2.2, which can have different configurations explained in detail in (Hatziargyriou et al. 2000).

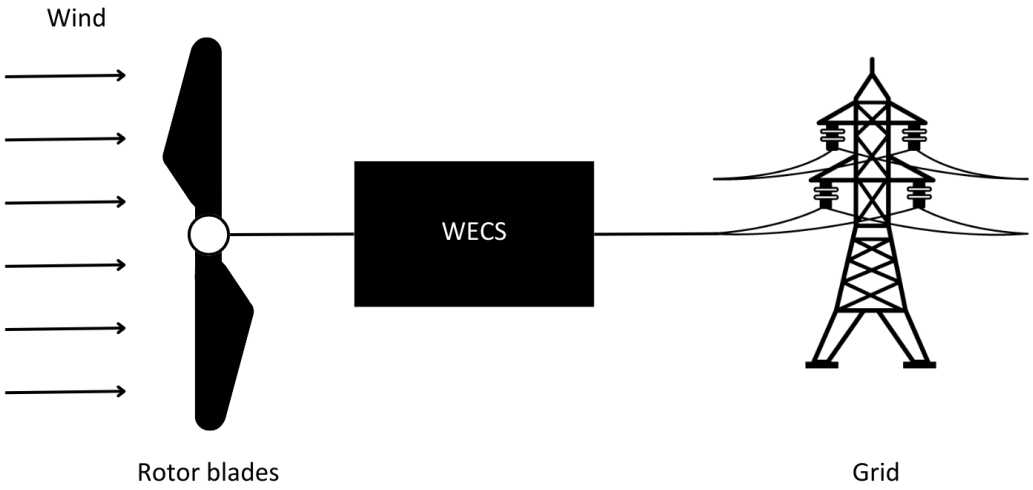


Figure 2.2 – General Scheme of a WECS (Adapted from: (Hatziargyriou et al. 2000))

To use the turbine data, the information from the sensors needs to get to the computers, and that process is called a Data Acquisition System (DAS). Data acquisition encompasses the automated collection of non-electricity and electricity signals from sensors and various equipment under test, transmitting them to a computer for subsequent analysis and processing (Qu & Ma 2020). A DAS serves as the instrumental device for this purpose, automatically gathering and recording data from sensors and electronic equipment, facilitating applications ranging from climate monitoring to relative humidity and temperature tracking. Its broad utilization extends across diverse technological systems in various electronic domains (Saleh et al. 2021).

The study of (Ma et al. 2023) explains how to select the best method to collect data from turbines, the most common being analog or digital data acquisition, communication acquisition, and computer simulation acquisition.

2.1.2 Photovoltaic Power

Photovoltaic (PV) energy has rapidly grown in the last decade due to its natural availability and the decreasing cost of renewable energy equipment (Barka et al. 2020, Prajwal & Hegde 2022).

A typical PV cell is made with a semiconductor material, usually silicon, and includes a p-n junction. This junction refers to the interface between two types of semiconductor material: p-type, which is positively doped, and n-type, which is negatively doped (Al-Ezzi & Ansari 2022). The PV cell transforms solar energy through the Photovoltaic effect (Singh 2013) described in Figure 2.3. When sunlight strikes the cell, it raises the energy level of electrons, causing them to break free from their atomic shells. The p-n junction directs liberated electrons toward the n region while driving positive charges to the p region through the electric field. A metal grid on the cell's surface collects the electrons, while a metal backplate gathers the positive charges (Penick & Louk 1998).

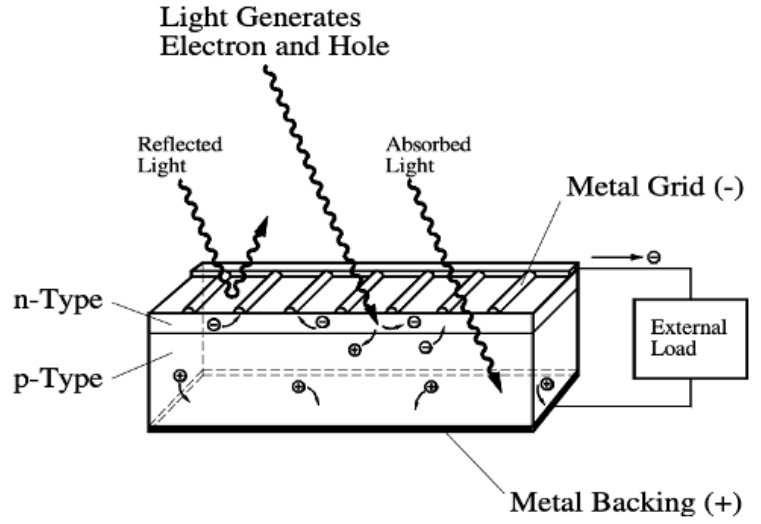


Figure 2.3 – Photovoltaic Effect (Penick & Louk 1998)

There are many available models to analyze the performance of solar modules (Araújo et al. 2020, Baig et al. 2020), but the most common is described by the mathematical equation 2.3. With this model is possible to notice that the load current (I, in A) results from the current produced by the solar cell (I_L) minus the current from the diode ($I_D \left(e^{\frac{QV_{OC}}{AKT}} - 1\right)$) and the current flowing through the shunt resistance ($\frac{V_{OC}}{R_{SH}}$):

$$I = I_L - I_D \left(e^{\frac{QV_{OC}}{AKT}} - 1\right) - \frac{V_{OC}}{R_{SH}} \tag{2.3}$$

I_L (A) is the light-generated current; I_D (A) is the diode saturation current; Q is the electron charge ($1.6 \times 10^{-19} C$); V_{OC} (V) is the open circuit voltage; A (C) is the curve fitting constant; K is the Boltzmann constant ($1.38 \times 10^{-23} J/°K$); T (°K) is the temperature on absolute scale; R_{SH} (Ω) is the shunt resistance.

Figure 2.4 describes the electric circuit of a PV module, where I_{SH} (A) represents

the shunt-leakage current, and R_S (Ω) the internal resistance of the current flow (Patel 1999).

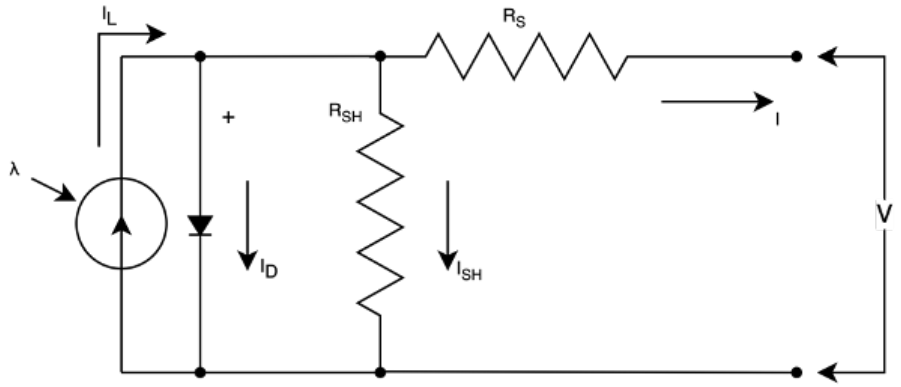


Figure 2.4 – Electrical Circuit of a PV module (Adapted from: (Patel 1999))

Solar panels produce variable DC voltage and current, influenced by sunlight intensity, and this regulated DC voltage undergoes conversion to AC voltage through a solar inverter (Shahria et al. 2023). In the study of (Parmar et al. 2019), several types of solar inverters are presented and explained, such as Off-Grid Inverters, On-Grid Inverters, and Hybrid Inverters. Regarding On-Grid Inverters, the type used in the context of this study, it is customary for the PV system to consist of PV arrays, the inverter, and the grid as in Figure 2.5. A collective assembly of PV solar cells interconnected electrically constitutes a PV module and, when replicated, gives rise to a PV array (Mohamed & Sattar 2019).

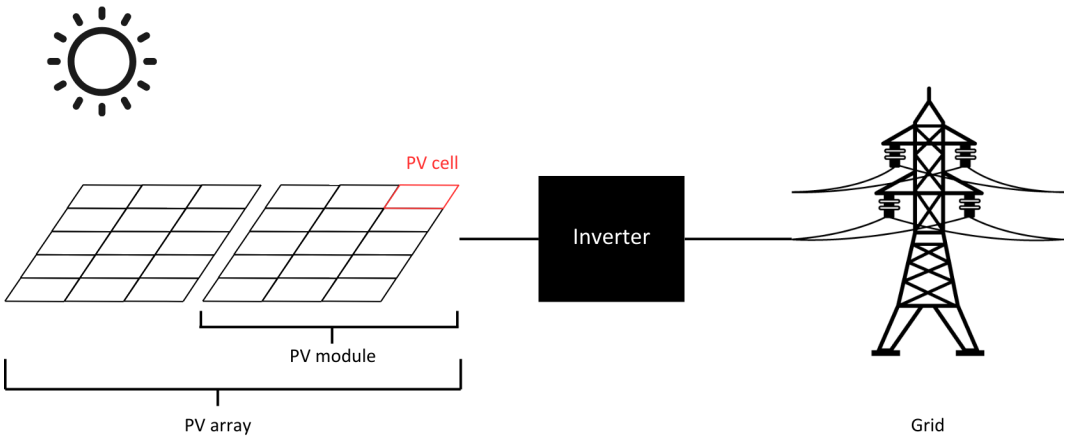


Figure 2.5 – Grid-Connected PV System (Adapted from: (Mohamed & Sattar 2019))

In PV, it is usual to use a DAS to get essential data from the solar panel or the inverter within the electrical domain, encompassing parameters such as voltage and current derived from either AC or DC. However, having sensors in each solar cell comes with a high economic cost, so it is usual to have the sensors in the solar inverter. Also, it extends its functionality beyond the electric domain to encompass non-electric data, including temperature, humidity, and light intensity (Barka et al. 2020, Prajwal & Hegde 2022).

In the study of (Barka et al. 2020), various DAS commonly used in this field are presented and explained.

2.2 Data Mining

DM is the process of discovering patterns, trends, correlations, or meaningful insights from large amounts of data. It includes techniques from statistics, machine learning algorithms, and databases (Fayyad et al. 1996). It analyzes data from various perspectives, summarizing it into valuable information from extensive volumes of structured or unstructured, inconsistent data that often changes. It proves instrumental in, for example, enhancing revenue, reducing costs, and making important decisions (Gul et al. 2021, Salem et al. 2022).

Firstly, data can be divided into structured, semi-structured, and unstructured and may require different DM methods. Structured data manifests when information assumes a predefined arrangement, wherein each data point exhibits a consistent set of attributes, encompassing fixed value ranges and imbued with specific semantic meaning. A typical example of this data type is data in database tables with established relations. Semi-structured data usually comes in graphs, for example, where the structure is not as defined as the first type described but has a semantic meaning. Unstructured data can usually be text, images, videos, or audio (Han et al. 2022).

Depending on the study's objective, researchers can divide DM into descriptive and predictive. Descriptive models frequently result in graphs or charts as their objective is to describe and summarize the dataset's properties. Predictive models use past information to project future outcomes through predictive analysis (Ramalingam & Ilakkiya 2021). Although it is possible to divide these two types of DM, it is more common that a study combines both by using descriptive DM as a way to learn the more important variables to be able to implement a better predictive model such as in the following studies (Anandi & Ramesh 2022, Dimić et al. 2019, Hanafiah et al. 2021, Irzavika & Supangkat 2018, Kaur et al. 2023).

ML uses historical examples to identify patterns in data and employs this knowledge to predict or classify events related to a specific problem, making these algorithms familiar in DM. ML algorithms can be supervised or unsupervised depending on the labeled data's existence in the training set. Usually, supervised learning is classified into classification and regression algorithms, while unsupervised learning is classified as clustering or association mining techniques (Alloghani et al. 2020).

According to (Alloghani et al. 2020), Decision Trees, Naïve Bayes, and Support Vector Machines (SVM) stand out as the most prevalent algorithms in supervised ML. Decision Trees function by copying a tree structure and organizing attributes into groups based on data values. In contrast, Naïve Bayes, rooted in the Bayesian probability theorem, can be considered a semisupervised method since it exhibits a versatile nature as it serves not only as a classification method but also as an algorithm suitable for clustering. It utilizes the Bayes theorem to assign variables to classes based on their probabilities. SVM takes a different approach by creating boundaries, represented as margins, to separate classes within the given dataset. At its core, the principle is to expand the distance between each class and its nearest margin, effectively minimizing classification errors.

The author of (Usama et al. 2019) divide unsupervised learning techniques into five:

- Hierarchical Learning involves extracting simple and complex features through a layered structure with multiple linear and nonlinear activations;
- Data Clustering involves organizing data into coherent and meaningful groups, relying on the similarity between various features as the determining factor for the grouping;
- Latent Variable Models are statistical models involving both observed (directly measured) and latent variables (not observed directly but inferred from the observed variables). It allows to describe complicated distributions by breaking them down into manageable joint distributions across a broader range of variables;
- Dimensionality Reduction search to represent the same data in fewer dimensions. It frequently finds use in data modeling and visualization;
- Outlier Detection, as the name indicates, aims to detect points distant from the other samples, commonly called an outlier.

In summary, DM provide powerful tools for extracting valuable insights from diverse datasets. DM, rooted in statistics, algorithms, and databases, facilitates the analysis of structured and unstructured data, offering the potential to enhance decision-making processes. ML, a key component of DM, employs historical examples to identify patterns, making it instrumental in predicting and classifying events. The continuous evolution of these techniques underscores their pivotal role in deriving meaningful knowledge from intricate datasets.

With the noticeable growth of DM came the need to have process models. Some of them are *KDD (Knowledge Discovery Databases)*, *SEMMA (Sample*, *Explore*, *Modify*, *Model*, *Assess)*, or the most considered as "de-facto" standard data mining methodology and adopted for this study, CRISP-DM (Shafique & Qaiser 2014).

2.2.1 CRISP-DM

The acronym *CRISP-DM* denotes the *CRoss-Industry Standard Process for Data Mining*. In 1996, the architects of this methodology started formulating a standardized process for DM projects and published the user guide in 2000. Their motivation was to spare aspiring entrants from the arduous trial-and-error journey into the data mining market and to cultivate a level of maturity in DM that would instill confidence within businesses, positioning it as an indispensable component of their operational processes (Chapman 2000).

Despite the emergence of novel frameworks designed to address certain limitations of CRISP-DM, such as challenges in cross-team collaboration (Li et al. 2023), this methodology continues to find widespread application in numerous DM projects across various

domains. It plays a pivotal role in diverse fields such as anomaly detection (Lima et al. 2023), the prediction of students' academic outcomes (Essayad & Abdella 2024, Santoso et al. 2023), and healthcare (Septiana et al. 2023). The industry standard status of CRISP-DM and the methodology's inherent simplicity and well-defined structure drive its adoption in contemporary research and practice (Schröer et al. 2021).

CRISP-DM has six phases, represented in Figure 2.6. These phases can assume the form of the life cycle of a DM project, but the sequence of the phases is not rigid as in this type of project, it is common to need to go back and forth (Chapman 2000), for example, to do more data preparation (Saltz 2021).

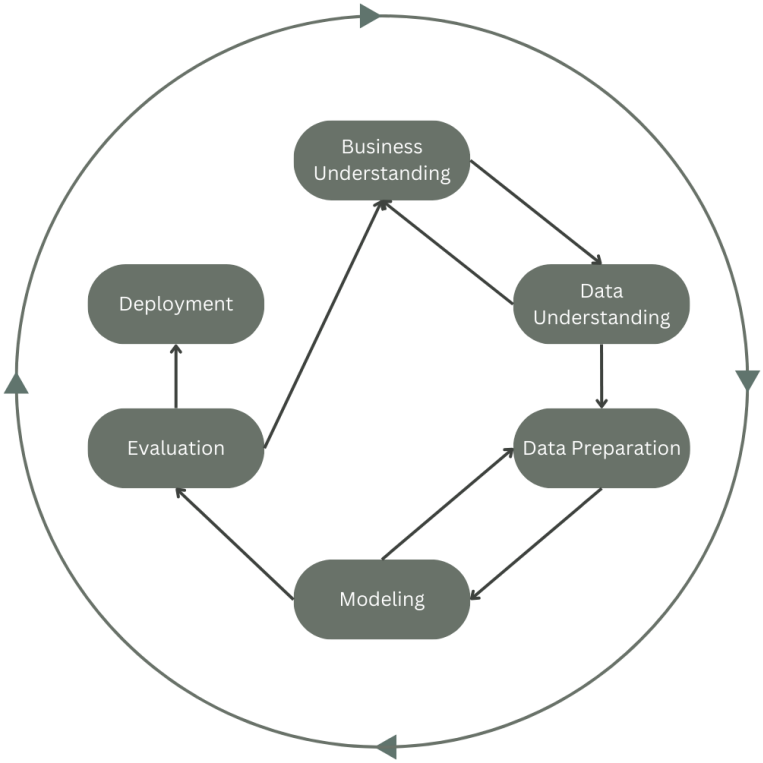


Figure 2.6 – Phases of CRISP-DM (Adapted from: (Chapman 2000))

The *Business Understanding* phase directs its focus toward comprehending the requisites from a business standpoint. It entails defining business objectives and delineating what the client intends to achieve. Subsequently, an in-depth assessment of the situation is imperative, involving a meticulous exploration of all pertinent factors crucial for defining the data analysis objective. This process entails understanding available resources, identifying necessary assumptions, recognizing project constraints, and discerning potential risks. Establishing DM goals within this framework and articulating technical objectives is essential. Also, delineating criteria, which will ultimately gauge the project's success, is critical. The culmination involves crafting a project plan that meticulously outlines the steps and their respective details, such as duration and dependencies, required to achieve business and DM goals (Chapman 2000).

The collection of data initiates in the *Data Understanding* phase. The data can come from various sources, so after collecting the data, integration either in this phase or the

subsequent one is needed. The following step involves a detailed description of the data, encompassing an understanding of data types, quantity, and adherence to stipulated requirements. Additionally, exploration and verification of data quality are imperative during this phase (Chapman 2000).

In the *Data Preparation* phase, the focus is the production of the dataset for modeling. Data selection needs to align with the project's goals and constraints. Addressing, for example, null values, a potential influence on subsequent phases that necessitates meticulous cleaning. Further, if warranted, the creation of derived attributes or new records is undertaken at this juncture. Finally, meticulous formatting is ensured, with attributes ordered appropriately for specific algorithms and records possessing uniform syntactic formats for a given attribute (Chapman 2000). Typically, 50-80% of the effort for the entire project is in this phase (Saltz 2021).

For the *Modelling* phase, the crux lies in defining the DM operation adequate to the problem, such as classification or clustering. In this phase, a typical approach involves formulating a preliminary procedure to assess the quality of the algorithms by training the model with the training set and applying metrics to the test set. Subsequently, the final model is crafted, often leveraging optimal algorithms or a combination (Chapman 2000).

Instead of basing the evaluation on metrics, the *Evaluation* phase focuses on determining the achievement of the initially defined business goals. If not, an exploration into the reasons behind any shortcomings is imperative. Consideration of remaining resources and budgets prompts the decision on whether to proceed to the deployment phase or contemplate revisiting the project, perhaps necessitating the initiation of a new DM endeavor (Chapman 2000).

Finally, the *Deployment* phase marks the integration of the DM project's outcomes into the business ecosystem. It mandates comprehensive planning for monitoring and maintenance to ensure the judicious utilization of DM results in daily business operations. Conducting a comprehensive project review entails outlining successes and identifying areas needing improvement (Chapman 2000).

2.2.2 Clustering

Stressing the importance of deriving actionable insights from data, in (Oyelade et al. 2019) the authors assert that data gains value when it yields information or knowledge for further reasoning. With that in mind, clustering is an unsupervised learning task that groups data objects or patterns based on similarity measures, where entities within a cluster exhibit more remarkable similarity among themselves than with entities in different clusters (Bhattacharjee & Mitra 2020). In essence, clustering aims to unveil distinct patterns, points, or objects inherent in the natural grouping of the data (Oyelade et al. 2019).

Throughout its history, clustering has been used for various purposes being the main

ones (Ghosal et al. 2020):

- Obtain useful knowledge from data;
- Identify the degree of similarity among data;
- Organize and summarize data through clusters.

Moreover, (Ghosal et al. 2020) and (Aggarwal & Reddy 2013) detail multiple typical clustering applications in the industry, such as customer segmentation, data summarization, image segmentation, anomaly detection, or as an intermediate step for other DM problems.

Different clustering techniques can be applied depending on the study's objective or field. It is common for certain algorithms to exhibit superior performance in specific contexts. The following section introduces various algorithms to illustrate this diversity.

2.2.3 Clustering Algorithms

Some researchers divide clustering algorithms into Partitional, Hierarchical, and Density-based methods (Bhattacharjee & Mitra 2020), while others add Grid-based, Soft clustering, Model-based and Ensemble clustering (Oyelade et al. 2019). Besides that, this section also discusses Time Series clustering given its importance to the present project. Figure 2.7 summarizes the algorithms and their division along this study.

Partitional clustering's primary objective is to uncover inherent groupings within the data by optimizing specific objective functions (Aggarwal & Reddy 2013). Partitional clustering is a non-hierarchical approach tailored for static data sets (Oyelade et al. 2019). These algorithms typically necessitate user-defined parameters, more precisely prototype points, to represent each cluster (Aggarwal & Reddy 2013). Due to its simplicity, efficiency, and ease of implementation, partitional clustering remains widely applied and popular in practice, consistently demonstrating empirical success (Oyelade et al. 2019).

One of the most famous and used partitional clustering algorithms is *K-Means*. This method requires the number of desired clusters (k) and a distance metric as input parameters. The algorithm initiates by randomly selecting k centroids. Subsequently, it assigns each data point to the cluster with the nearest centroid based on the distance metric, usually Euclidian's distance. After assigning each data point to the nearest centroid, the algorithm recalculates the centroids of the clusters by computing the mean of the data points within each cluster. This data point assignment and centroid update process iterate until convergence, characterized by no centroid changes between iterations, or, in cases of non-convergence, when it reaches a maximum of pre-defined iterations (Rodriguez et al. 2019). Commonly, the objective function that *K-Means* pretends to minimize is the Sum of Squared Errors (SSE) (Aggarwal & Reddy 2013). Noteworthy aspects of this algorithm include (Oyelade et al. 2019, Rodriguez et al. 2019)

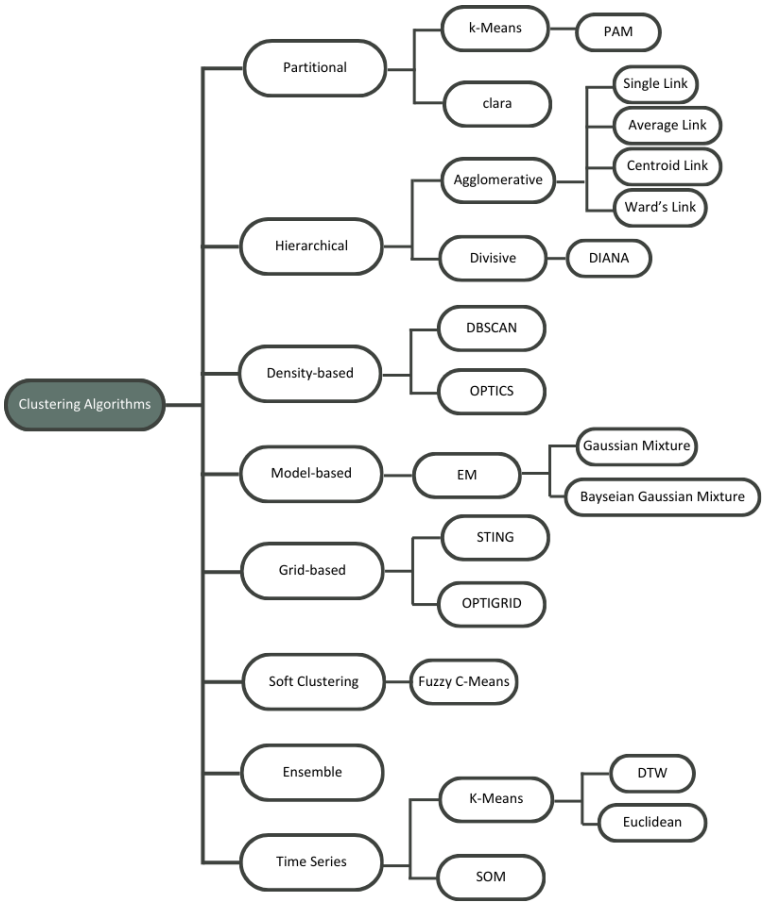


Figure 2.7 – Clustering Algorithms

- The algorithm's sensitivity to the initial centroid choice;
- Susceptibility to outliers;
- Computational efficiency for large datasets; and
- Users need to specify the desired number of clusters.

In order to overcome some of the drawbacks of *K-Means*, other algorithms have been proposed based on it, such as the *K-Medoid* approach. A medoid differs from a centroid as it is a point part of the dataset and not the mean of the points. The *Partitioning Around Medoid (PAM)* is one of the proposed algorithms for this approach. It works the same way as *K-Means*, but computing the distance between the data points in a cluster and its medoid and the objective function minimizes the sum of all the distances. This approach is more robust than *K-Means* in dealing with noise and outliers in the dataset (Mohammed & Abdulazeez 2017). The *CLARA (clustering large applications)* algorithm is one of the most used for large data sets since it does not explore all the data set, but selects random samples and applies *PAM* to obtain the optimal medoids (Janse van Vuuren & Vermeulen 2019, Rodriguez et al. 2019).

Hierarchical clustering groups similar data points into a hierarchical structure. Unlike partitioning methods, hierarchical clustering does not require specifying the number of

clusters beforehand (Oyelade et al. 2019). It organizes the data usually in a dendrogram structure, allowing for agglomerative (bottom-up) and divisive (top-down) approaches (Bhattacharjee & Mitra 2020) represented graphically in Figure 2.8. After construction, it is possible to cut the dendrogram to obtain the desired number of clusters (Aggarwal & Reddy 2013).

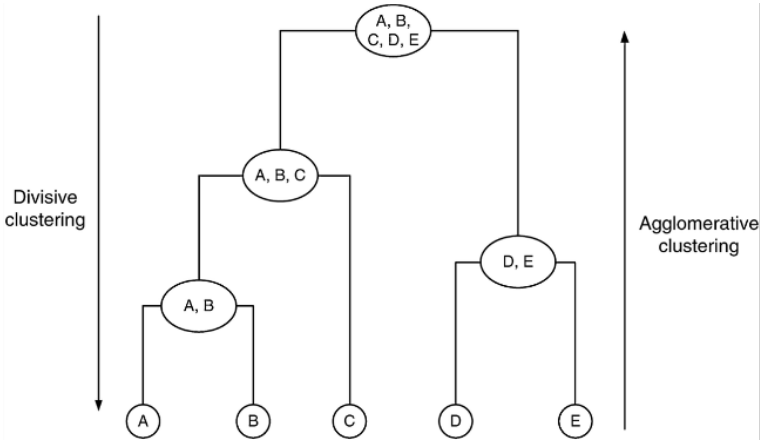


Figure 2.8 – Dendrogram of Agglomerative and Divisive clustering (Halkidi 2018)

In agglomerative clustering, the algorithm initially assigns each object to an individual cluster. Subsequently, the algorithm merges groups through successive iterations until it reaches specific stop conditions. This process entails placing each point in a cluster of its own and then identifying and combining the two points nearest to it. In this context, a point can refer to an individual object or a cluster of objects (Oyelade et al. 2019, Rodriguez et al. 2019). Linkage algorithms are typically used in hierarchical clustering and usually use Euclidean distance metric when computing the distance matrix between every data point. Some of the most famous and used linkages are, amongst others, Single, Average, Centroid, and Ward's Link. While single link merges two clusters where the distance between two points (one of each cluster) is the most minor, average link merges the clusters if the average distance between them is the smallest. Centroid link merges two clusters if their centroids have the smallest distance between each other compared to the centroids of other clusters. Ward's Link aims to minimize the increase in total within-cluster variance and merge clusters by combining the pair of clusters that results in the smallest growth in the sum of squared differences within all clusters (Sreedhar Kumar et al. 2019).

The top-down approach starts with all the data points in a single cluster and divides it through successive iterations into smaller clusters (Oyelade et al. 2019, Rodriguez et al. 2019). The *Divisive Analysis (DIANA)* algorithm uses divisive clustering by dividing the clusters based on the element that differs from the most on average (Elhassouny 2023). This algorithm is rigid and cannot identify non-spherical clusters (Oyelade et al. 2019).

Another well-known hierarchical algorithm, *Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)*, incrementally organizes data through a hierarchical method. BIRCH does not necessarily fit into agglomerative or divisive hierarchical clustering. Instead of a top-down or bottom-up approach, it employs its methodology by construct-

ing a hierarchical structure that groups data points into subclusters, representing a tree-based dataset (Nwadiugwu 2020). Its effective handling of large datasets and favorable time complexity make it stand out (Zhang et al. 1997).

Density-based clustering is a methodology that groups data points based on their density within the feature space. Unlike traditional clustering techniques that assume spherical or convex shapes for clusters, density-based clustering, exemplified by algorithms like *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)* and *OPTICS (Ordering Points To Identify the Clustering Structure)*, identifies clusters by locating regions of higher data point concentration (Bhattacharjee & Mitra 2020, Oyelade et al. 2019).

The *DBSCAN* algorithm begins with an arbitrary point and expands clusters by including all reachable points within a defined neighborhood. This approach makes it possible to accommodate clusters of arbitrary shapes and effectively handle noisy data, making it particularly suitable for datasets with irregular structures or varying densities (Oyelade et al. 2019). *OPTICS* is an algorithm founded on the principle of maximal density-reachability. The algorithm commences by selecting a data point and extending its neighborhood like the *DBSCAN* algorithm. However, *OPTICS* introduces a distinctive approach by initially expanding the neighborhood to points characterized by low core distance. This core distance, denoting the *m-th* smallest distance between the point and objects within its neighborhood, is governed by the algorithm's parameter '*m*' representing the minimum number of points required to constitute a cluster. A noteworthy strength of *OPTICS* is its ability to identify clusters exhibiting substantial density variations and irregular shapes, emphasizing its adaptability to diverse cluster structures within datasets (Rodriguez et al. 2019).

Model-based clustering assumes that the underlying structure of the data follows a probabilistic model. Rather than relying solely on the density or distance between data points, model-based clustering seeks to fit a statistical model to the data and identify clusters based on the parameters of this model. Practitioners commonly employ the *Expectation-Maximization (EM)* algorithm in this type of clustering (Rodriguez et al. 2019). The EM algorithm is a powerful tool used to estimate the parameters of a probabilistic model. The algorithm operates iteratively, starting with an initial guess of the model parameters. The E-step computes the probability of each data point belonging to each cluster based on the current parameter estimates. By updating the parameters, the M-step aims to maximize the likelihood of the observed data given the computed probabilities. This iterative process continues until convergence, with the algorithm refining its estimates of the model parameters and the cluster assignments of data points. The EM algorithm is valuable when data distribution is not readily apparent and assumes a latent structure governed by a probabilistic model (Rodriguez et al. 2019, Sammaknejad et al. 2019). A typical probabilistic model used along EM is the *Gaussian Mixture Model (GMM)* where it is assumed that the variation of the dataset can be described as a mixture of different Gaussian distributions (He et al. 2011).

Grid-based clustering is a technique that organizes the data space into a grid structure and assigns data points to grid cells based on their spatial proximity. Unlike other clus-

tering methods, grid-based clustering does not require the specification of the number of clusters beforehand. The algorithm divides the data space into a predefined number of cells or a grid and considers each cell a cluster. Grid-based clustering is particularly effective in handling datasets with varying densities and irregular shapes. Grouping data points into cells simplifies the process of identifying clusters and can efficiently detect clusters with different shapes and sizes. This approach is beneficial in scenarios where other clustering methods may struggle, providing a straightforward and computationally efficient solution for certain types of spatial data analysis (Aggarwal & Reddy 2013, Oyelade et al. 2019). Two of the most known algorithms for this type of clustering are the *Statistical Information Grid (STING)* algorithm and the *Optimal Grid-Clustering (OptiGrid)*. STING organizes the data into a grid, using statistical measures like mean and standard deviation to identify clusters. Employing a recursive subdivision approach, *STING* dynamically adapts the grid structure to detect homogeneous regions (Aggarwal & Reddy 2013). This algorithm has a low computational cost, but the quality of the clusters highly depends on the density input parameter chosen by the user (Oyelade et al. 2019). OptiGrid, strategically partitions the dataset into regions of low density, ensuring that the cutting plane effectively distinguishes clusters. This method prioritizes unbiased cluster size and shape representation while adeptly handling high-dimensional data (Aggarwal & Reddy 2013, Oyelade et al. 2019).

Soft clustering is a technique that allows data points to belong to multiple clusters with varying degrees of membership, unlike hard clustering, where each point can only be part of a single cluster. In soft clustering, points are assigned based on probabilities or membership values, expressing the likelihood of a point belonging to each cluster. One prominent algorithm for soft clustering is *Fuzzy C-Means*. The algorithm iteratively updates cluster centroids and membership values, minimizing an objective function considering the distances between data points and cluster centroids. However, the algorithm tends to assign indistinguishable membership to outliers across clusters, which can result in less desirable outcomes (Ezugwu et al. 2022, Oyelade et al. 2019).

Ensemble clustering combines multiple clustering methods on a dataset to achieve a more robust consensus clustering. It uses a consensus function to aggregate results from various techniques into a single clustering outcome. This approach avoids the need for a priori input of the number of clusters by employing cluster validation indices to determine the optimum cluster numbers for each dataset (Oyelade et al. 2019).

Time series clustering pertains to classifying temporal sequences into clusters according to shared patterns or behaviors. A time series constitutes a dataset wherein the constituent features evolve over time. This clustering methodology facilitates categorizing successive data points into homogeneous groups driven by their temporal characteristics and intrinsic similarities. Practitioners commonly apply algorithms like *K-Means* in which Euclidean distance is employed, but *Dynamic Time Warping (DTW)* emerges as a frequently used technique for aligning points from distinct time series. DTW is a distance metric capable of yielding superior results in time series clustering (Javed et al. 2020). The *Self-Organizing Maps (SOM)* algorithm represents another commonly applied method. SOM, a variant of neural network architecture, arranges data within a one-

or two-dimensional grid and comprises competitive and interconnected neurons. These neurons employ a neighborhood function to adjust their weights, enabling SOM to associate input data with the nearest neuron, thereby facilitating the formation of clusters (Cherif et al. 2011).

2.2.4 Clustering Validation Metrics

In unsupervised learning, particularly in clustering, verifying partition quality becomes imperative. With a means to assess the efficacy of clustering outcomes, the utility of various results would be easier to discern. Evaluating the goodness of clustering outcomes can be described as validating the clustering. Different validation measures broadly fall into two categories: external and internal validation. The key distinction lies in the incorporation of external information for validation. External validation metrics, such as the *Normalized Mutual Information (NMI)*, assess the agreement between obtained clusters and accurate partitions when external information like class labels is available. Conversely, internal validation measures become the sole recourse for evaluating clustering effectiveness in scenarios where such external information is lacking (Aggarwal & Reddy 2013).

External validation metrics assess the degree to which the clustering structure identified by a clustering algorithm aligns with an external structure, utilizing information not present in the data. These metrics prove valuable when the true cluster number is known in advance, aiding in selecting an optimal clustering algorithm for a specific dataset (Aggarwal & Reddy 2013).

One such metric is the *Consistency Index (CI)*, which quantifies the proportion of shared objects between matching clusters of a given partition and the actual data partition obtained from ground-truth information (Duarte et al. 2013).

Entropy is another external validation measure that gauges the purity of cluster class labels. The entropy is zero if all clusters comprise objects with a single class label. However, as the class labels within a cluster become more diverse, the entropy increases, reflecting a decrease in purity (Rendón et al. 2011).

Another famous external metric is the *NMI*. This metric is the normalized version of the *Mutual Information (MI)* metric that computes the common information in the identified clusters and in the actual partitions. The *NMI* is normalized using the entropy of the true class labels and the clustering partition. The *NMI* metric operates on a scale from zero to one, with values closer to one indicating a more accurate clustering performance, *i.e.*, a high *NMI* value, nearing one, signifies a strong agreement between the identified cluster labels and the actual class labels (Kachouie & Shutaywi 2020).

Internal validation metrics assess the quality of a clustering structure independently of external information. These metrics serve the dual purpose of selecting the optimal clustering algorithm and determining the ideal cluster number without relying on additional information (Aggarwal & Reddy 2013). Some of the most famous metrics of this

type are the *Modified Hubert statistic*, *Dunn's index*, *Silhouette score (SS)*, *Davies-Bouldin (DB) index*, and the *Calinski-Harabasz (CH) index*.

The *Modified Hubert statistic* is a metric that quantifies differences between clusters by counting disagreements among pairs of data objects in two partitions. *Dunn's index* further contributes to internal validation, utilizing the minimum pairwise distance between objects in different clusters as the intercluster separation and the maximum diameter among all clusters as the intracluster compactness. A higher Dunn's index means better segmentation. Another metric, the *Silhouette score*, evaluates clustering performance based on the pairwise differences between within- and between-cluster distances. Additionally, it aids in determining the optimal cluster number by maximizing its value (Aggarwal & Reddy 2013).

The *Davies-Bouldin index* aims to identify sets of clusters that are both compact and well-separated (Rendón et al. 2011). Each cluster is assigned the highest value as its similarity, and the DB index is obtained by averaging all cluster similarities. A smaller *DB* index indicates a better clustering result, with minimized index values suggesting that clusters are more distinct, achieving an optimal partition (Aggarwal & Reddy 2013).

Finally, the *CH index* assesses the quality of clustering by examining the dispersion between clusters and within clusters, where a higher value indicates a greater separation between clusters and tighter cohesion within clusters (Wang & Xu 2019).

Evaluating clustering outcomes is crucial for assessing the quality of results in unsupervised learning. The distinction between external and internal validation metrics provides a comprehensive approach to gauge clustering effectiveness. External validation metrics leverage external information when available, while internal validation metrics offer insights when external information is absent. This dual approach equips researchers with a versatile toolkit for making informed decisions and advancements in the dynamic field of unsupervised learning.

2.3 Related Work

Many studies apply clustering to the renewable energy field. However, only a few studies seek data comprehension as their final objective. With that in mind, this section presents work done in recent years with clustering for renewable energies, more precisely, for wind and photovoltaic.

(AZIZI et al. 2019) highlighted various applications of wind speed analysis and emphasized its essential role in wind farms. Applications included selecting suitable sites for wind turbine installation, predicting optimal turbine sizes for specific locations, and optimizing operating costs. Regardless, the study goal was to achieve meaningful wind power generation scenarios. The study employed the *Linkage-Ward* clustering method and compared it with *K-Means*, noting that while *Linkage-Ward* offered better accuracy, it came with a higher computation cost. After clustering, the authors computed the probability of occurrence for each cluster, identifying the one with the highest probability for

wind farm planning. The study concluded that these results were crucial for precise planning, significantly reducing fossil fuel consumption for energy production.

Using the *CLARA* algorithm, the researchers of (Janse van Vuuren & Vermeulen 2019) clustered temporal wind speed profiles in South African Renewable Energy Development Zones. The study significantly reduced the computational cost of capacity allocation optimization studies, providing practical insights for optimizing the geographical allocation of wind generation capacity.

To simplify complex data structures for better decision-making, (Vankov et al. 2020) propose a clustering method for reducing spatial and temporal data size in nodal time series data of renewable power networks. The results indicated that the proposed temporal and spatial clustering algorithm outperformed spatial-only clustering methods, showcasing its efficiency in addressing high-dimensional energy problems and improving the representation of renewable energy networks.

(Sabitha & Punhani 2019) focused on identifying wind speed patterns in potential locations for wind energy generation. The authors employed *K-Means* and *Rapidminer's X-Means* clustering algorithms to analyze wind speed data from a district in south India. The results suggested that the approach could aid in site selection for wind turbine installation. (Mehrjoo et al. 2021) also employed *K-Means* clustering to classify turbines within a wind farm into homogeneous groups based on performance features. Results show significant improvements in accuracy and a more than 90% reduction in computing costs compared to individual turbine power curve modeling. This study effectively addressed the trade-off between complexity and accuracy, providing a solution for large wind farms with diverse turbine types or control parameters.

Unfortunately, in some regions land resources are limited, so (Uti et al. 2023) explored the untapped potential of ocean renewable energy development in Malaysia. They leveraged advancements in space technology, employing altimetry data and spatial-temporal clustering through the *K-Means* technique to identify promising locations for harnessing wind and wave energy. Seasonal analyses and spatial-temporal clustering, aided by silhouette analysis, contributed to identifying optimal energy resource clusters. The identified clusters provided valuable information for developers interested in creating energy devices suitable for such regions.

Many of the studies regarding clustering analysis and PV are about fault detection (Gunda et al. 2020, Rahman et al. 2021, Zhu et al. 2018) or as an intermediate step for forecasting models (Gu et al. 2021, Jinpeng et al. 2022), so only a few studies with the final objective of getting representative groups are presented.

The study of (Holloway et al. 2023) focuses on optimizing the selection of locations for a distributed hybrid renewable energy system in rural Western Australia using *K-Means* and *K-Medoids* clustering. While *K-Means* show superior clustering, *K-Medoids* identified specific locations with higher solar and wind energy potential. Despite inconclusive results, the study emphasizes the significance of national energy planning.

(Mabuggwe & Morsi 2020) identify 17 representative profiles from 123 homes with

locally distributed energy resources, plug-in electric vehicles, and PV systems. The study demonstrates computational efficiency by leveraging ML techniques such as *PCA* and *K-Means* clustering, compressing the annual PV dataset with minimal energy information loss. These representative profiles offer valuable insights for accelerating distribution system time-series analysis studies.

(Schütz et al. 2018) compare six clustering algorithms in their ability to identify typical demand days for energy system optimization. In the same line, (Zatti et al. 2019) introduce a novel *Mixed Integer Linear Program* clustering model, termed *k-MILP*, for selecting representative days in the design optimization of multi-energy systems. The model simultaneously identifies typical and extreme periods, controlling the features of the selected days with a maximum deviation tolerance on the load duration curves. The authors compare *k-MILP* with *K-Means* and *K-Medoids*. Results demonstrate that *k-MILP* provides a superior representation of typical and extreme operating conditions, improving accuracy in reproducing load duration curves and total yearly attribute values compared to the clustering methods used for comparison. The approach is particularly advantageous for optimizing designs regarding cost and reliability, showcasing its effectiveness in capturing extreme operating periods critical for energy systems.

Most of the presented studies used validation metrics famous for clustering models. (Munshi 2020) details eight of the most popular metrics; among them, the *Dunn Index* was the most occurrent in the studies described in this section (Holloway et al. 2023, Janse van Vuuren & Vermeulen 2019) followed by the *Silhouette Index* (Janse van Vuuren & Vermeulen 2019).

In conclusion, clustering methodologies emerge as a pivotal thread across diverse studies within the renewable energy domain, specifically in wind and PV energy contexts. Researchers widely employ clustering techniques to enhance understanding, optimize decision-making, and address critical challenges in renewable energy systems. Whether applied to wind speed analysis for wind farm optimization, classification of wind turbine performance, or identification of optimal locations for distributed hybrid renewable energy systems, *K-Means* consistently features as a preferred clustering algorithm. This ubiquity underscores its effectiveness in grouping similar entities and revealing patterns within complex renewable energy datasets. However, the present work differs from the discussed papers in this section as it aims to search among multiple solar inverters and wind turbines, understand their behavior using various characteristics of each, and segment the assets in clusters to get groups of the most similar assets among each other. Nevertheless, identical to the presented papers, this work tries multiple algorithms to understand which ones work the best in clustering solar and wind data.

3 METHODOLOGY

The development of this work relied on the CRISP-DM standard process. Each section of this chapter corresponds to a phase of CRISP-DM methodology.

The initial section provides an overview of the business problem, delineating specific objectives and success criteria. Subsequently, the second section elucidates the data utilized for solar and wind clustering, followed by a segment addressing all requisite treatments and cleaning procedures necessary to ensure a robust dataset for segmentation. The fourth section encompasses the clustering algorithms employed. Finally, the fifth section presents a detailed discussion of the obtained evaluation metrics.

3.1 Business Understanding

The primary business goal is for Enlitia's clients to render more informed decisions regarding their solar or wind assets by leveraging comparative performance insights derived from analogous assets using diverse performance characteristics of the assets.

The Data Mining objectives for this project are:

- Understand and interpret the variables;
- Clean and process data;
- Understand what variables are more indicated and important to do the clustering;
- Understand what algorithms work the best with each type of data through the use of objective evaluation internal metrics;
- Understand how data cleaning affects the allocation of assets in clusters and the quality of segmentation.

Clustering results can be challenging to evaluate and define subjective success criteria, given the lack of ground truth. For that reason, the success criteria for this project are the scores from the internal metrics: *Dunn Index*, *Xie Beni Index*, *Silhouette Score*, *Davies-Bouldin Index*, and *Calinski-Harabasz Index*.

3.2 Data Understanding

Enlitia has provided the present project's solar, satellite, and wind data. The solar inverter and satellite data result from 5 solar farms, while the wind turbine data comes from 5 wind farms.

Each solar farm has one satellite but hosts multiple solar inverters:

- *Farm 1:* 12 solar inverters;
- *Farms 2, 3, and 4:* 17 solar inverters each;
- *Farm 5:* 14 solar inverters.

The wind farms have the following composition:

- *Farm 1 and 3:* 17 wind turbines each;
- *Farm 2:* 10 wind turbines;
- *Farm 4:* 15 wind turbines;
- *Farm 5:* 12 wind turbines.

The ensuing sections outline the analysis conducted on wind, solar, and satellite data for clustering purposes. Appendix A provides a comprehensive statistical description of the features, complemented by histograms.

3.2.1 Wind Data

8 367 759 data entries and 71 assets compose the wind dataset. The nine features comprised the wind data are:

- *asset_id:* unique identifier of the wind turbine;
- *read_at:* timestamp variable with 10 minute interval, from January 1st of 2021 to March 30th of 2023 in the format *yyyy-mm-dd hh:mm:ss*;
- *wind_speed:* speed of the wind in m/s^2;
- *wind_direction:* direction of the wind in $degrees$;

- *power_average:* average power generated by the wind turbine in kW;
- *rotation_average:* average rotation per minute of the wind blades;
- *exterior_temperature:* air temperature in $^{\circ}C$;
- *nacelle_temperature:* temperature inside the nacelle in $^{\circ}C$;
- *farm:* wind farm within which the wind turbine belongs to.

These features have the distributions represented by the boxplots in Figure 3.1. Except for the *rotation_average*, all features exhibit outliers. It is also noticeable that the *wind_speed* presents negative values, which is not conventional. The boxplot relative to *wind_direction* indicates that the variable assumes unusual values since the direction of the wind typically has a value between -360° and 360°. These values go under treatment in the Data Preparation phase.

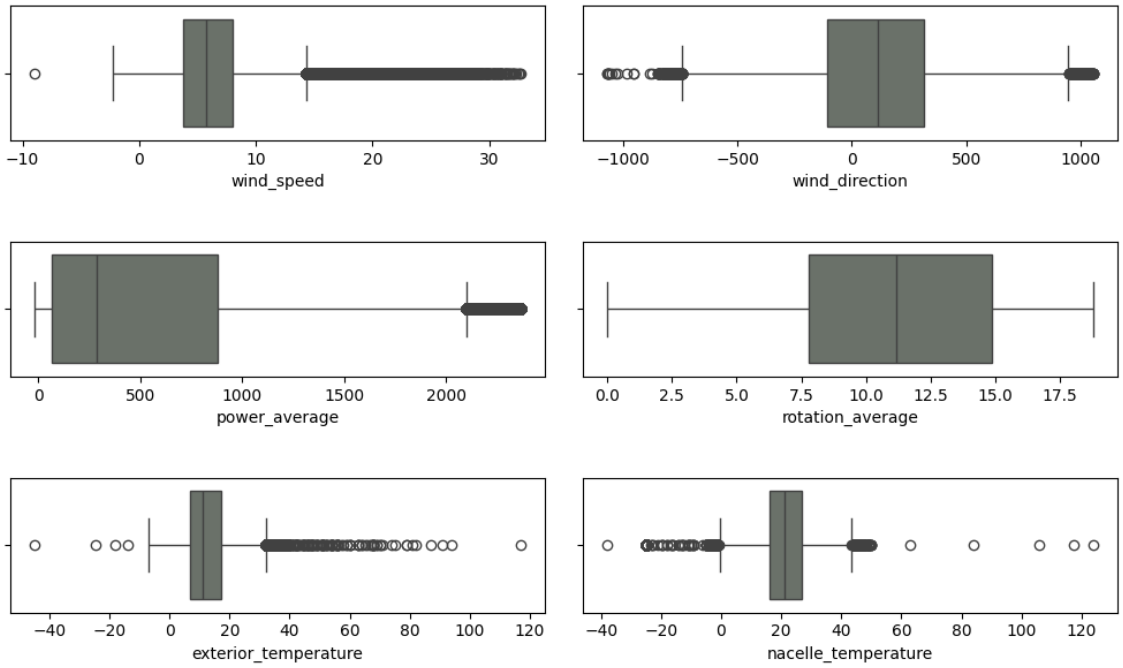


Figure 3.1 – Boxplots for Wind Data

3.2.2 Solar Data

Solar data comprises nine features, 4 081 774 data entries, and 77 assets, also known as solar inverters. The features of this data set are:

- *asset_id:* unique identifier of the solar inverter;
- *read_at:* timestamp variable with 10 minute interval, from January 1st of 2022 to January 9th of 2024 in the format *yyyy-mm-dd hh:mm:ss*. Nighttime hours not included;

- *ac_power:* AC power in kW;
- *ac_voltage:* AC voltage in V;
- *ac_current:* AC current in A;
- *dc_power:* DC power in kW;
- *dc_voltage:* DC voltage in V;
- *dc_current:* DC current in A;
- *farm:* solar farm within which the solar inverter belongs to.

Through Figure 3.2, it is possible to understand that the distribution of the variables composing the data set presents a right-skewed distribution, which indicates a tendency to lower values. It is also possible to see that some variables have outliers that need treatment in the data preparation phase.

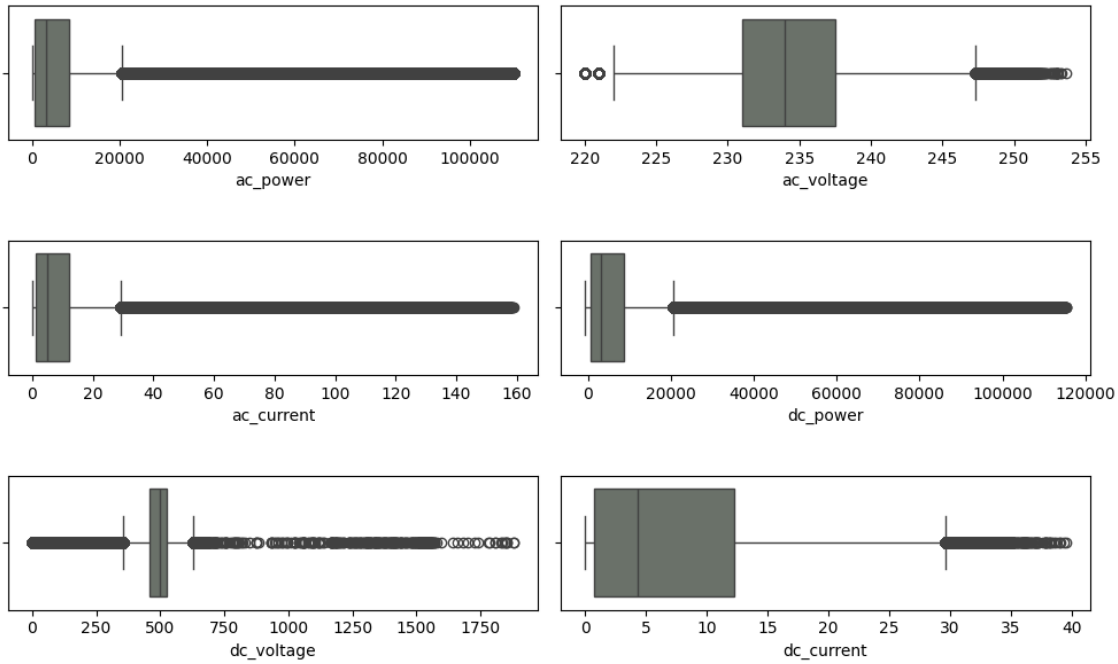


Figure 3.2 – Boxplots for Solar Data

3.2.3 Satellite Data

The satellite data consists of five satellites, each corresponding to a farm. Notably, all inverters within a particular farm share the same satellite data.

This data set is composed of seven variables and 476 012 data entries. The variables are:

- *satellite_id:* unique identifier of the satellite that corresponds to the solar farm they belong to;

- *read_at:* timestamp variable with 10 minute interval, from January 1st of 2022 to January 9th of 2024 in the format *yyyy-mm-dd hh:mm:ss*;
- *global_tilted_irradiance:* total irradiance directly incident (90°) in a surface at a certain tilt and azimuth in W/m^2;
- *global_horizontal_irradiance:* total irradiance incident at an horizontal surface W/m^2;
- *temperature:* air temperature in °C;
- *cloud_cover:* relative sunlight attenuation due to clouds in %.

In the *boxplots* referent to satellite data (Figure 3.3), it is possible to note a similar skewness behavior as the solar data as well as some outlier values.

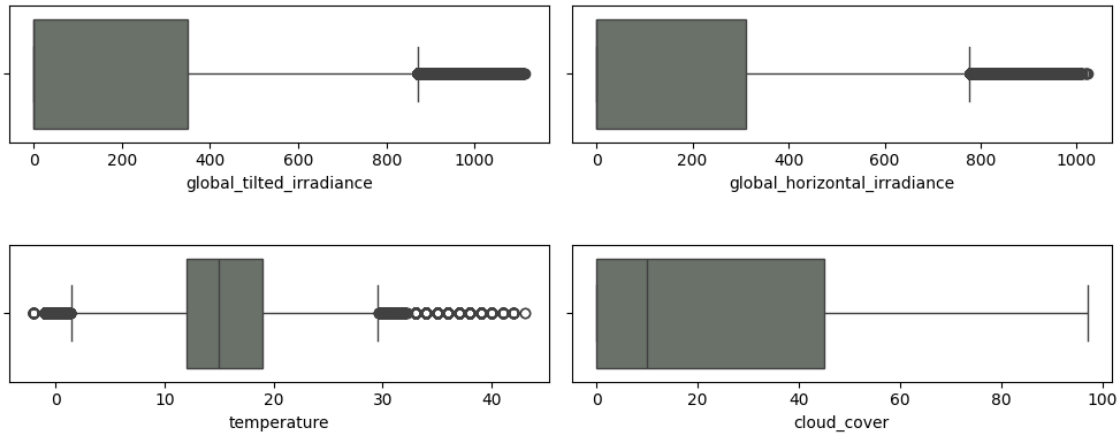


Figure 3.3 – Boxplots for Satellite Data

3.3 Data Preparation

In this project phase, the data undergoes treatment, so the datasets are adequate for clustering. This section delineates all processing of the data flaws, such as null or duplicated values, feature redundancy, and outliers.

In the correlation analysis, used with the objective of feature reduction, features exhibiting correlations exceeding 80% with others are considered candidates for exclusion from the final dataset.

When treating outliers, unless stated otherwise, each asset is treated individually instead of the entire dataset at once.

3.3.1 Wind Data

Correlation Analysis

A correlation analysis is imperative to ascertain the absence of multicollinearity among data variables. Highly correlated variables may signify redundancy in information. This

analysis facilitates the identification of redundant features for exclusion from the final dataset for the modeling phase.

Through Figure 3.4, it is possible to observe that some variables present a high correlation with others. Starting top to bottom, *wind_speed* is highly correlated with *power_average* (93%) and with *rotation_average* (88%). The existing correlation is expected as wind speed is the source of the energy. For that reason, *wind_speed* is going to be left out of the final dataset. However, since *power_average* also presents an 80% correlation with *rotation_average*, the latter variable will also be excluded as *power_average* is more relevant to this analysis since it is a more direct measure of wind assets performance and is typically more intuitive for stakeholders and *rotation_average* is significant but an intermediary measure that ultimately contributes to power production. Finally, *exterior_temperature* and *nacelle_temperature* have a strong correlation of 88% and since *nacelle_temperature* is more negatively correlated to other variables than *exterior_temperature*, it is the one to exclude.

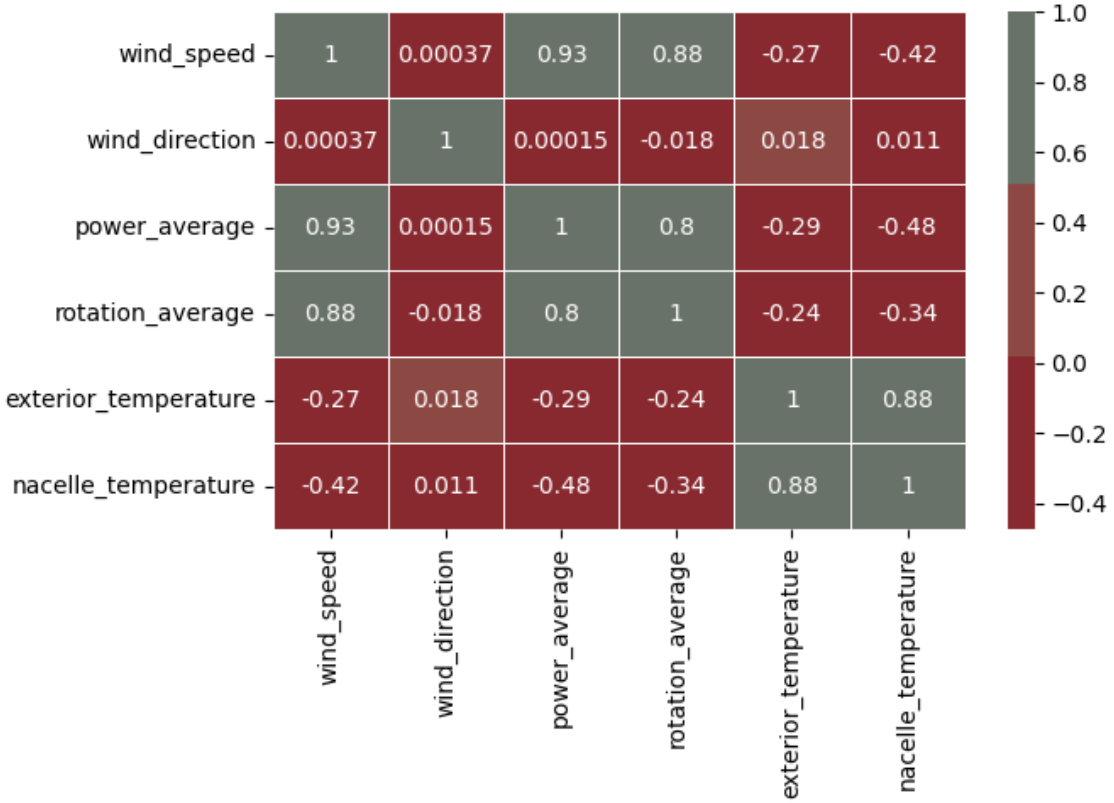


Figure 3.4 – Pearson Correlation Matrix of Wind Data features

The dataset to use for the segmentation phase is composed of the features: *wind_direction*, *power_average*, and *exterior_temperature*.

Null and Duplicated Values Treatment

This dataset has only one data entry containing null values. These values happen in the features *exterior_temperature* and *nacelle_temperature*. Instead of deleting this data entry, the median of the data entries from the asset containing the null values that occurred on the same day and time of other years replaced it.

Regarding the presence of duplicated values, this dataset does not possess any.

Outliers Treatment

As mentioned before, negative values when talking about wind speed are not common, and due to their existence in this dataset, all data entries containing values below 0 get deleted. Also, when *wind_speed* assumes the value 0, the turbine was not working and, therefore, does not interest this analysis. Similarly, the final dataset does not include entries with *wind_direction* values below -360° or above 360°.

Researchers like, for example, (Xie et al. 2023, Shen et al. 2019, Wang et al. 2023) commonly use the turbine's power curve to treat outlier values in wind data, and this study uses the same approach.

The power curves follow a sigmoid curve, and therefore, the first approach to clean outliers is to find the sigmoid curve that characterizes the power curve of each asset and apply an upper and lower limit from where the points outside the boundaries are considered outliers. The derivative of the sigmoid curve determines the boundaries. The derivative is added for the upper limit while subtracted for the lower limit of a fraction of the maximum value of the feature *power_average*. Figure 3.5 shows the results of this approach for the first two solar assets where the *inliers* corresponds to the data points that are going to be maintained and the *outliers* the values to discard.

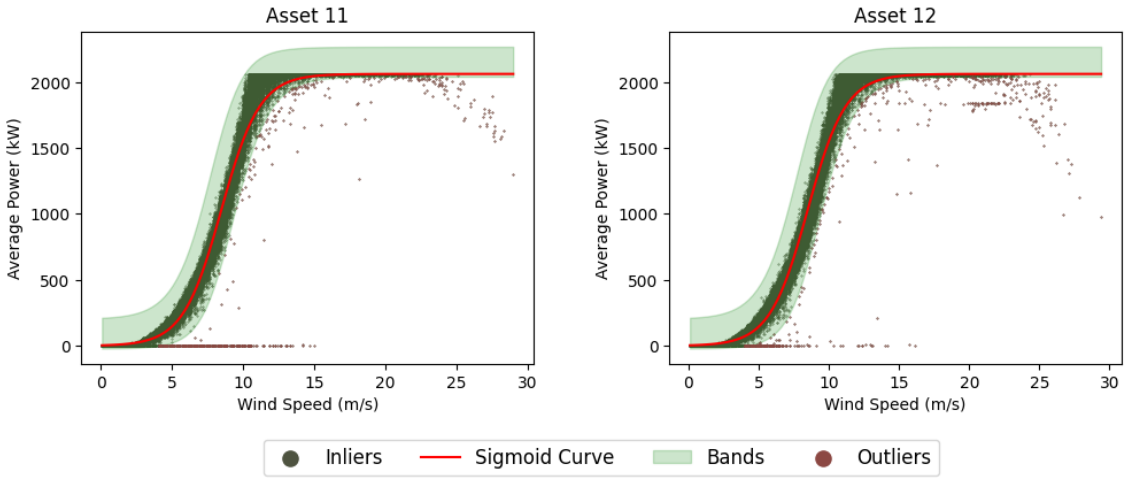


Figure 3.5 – Power Curves with sigmoid restrictions

This approach sufficed for nearly all assets. However, certain assets, exemplified by those depicted in Figure 3.6, exhibit distinctive characteristics. Assets 46 and 49 display anomalous points marked in red alongside outlier values denoted in blue.

One of the methods used to clean the abnormal behavior of the power curves is the *Mean Square Method (MSM)* proposed by (Xie et al. 2023) as an efficient method to clean unusual power curve data points with Figure B.1 in Appendix B depicting its result. MSM can effectively remove the values marked in red in Figure 3.6 as long as some other outlier values, but not the ones marked in blue. So, the DBSCAN algorithm is applied to remove the blue-marked data points. Being a density-based algorithm, it can recognize further away points as outliers, as shown in Appendix B, Figure B.2.

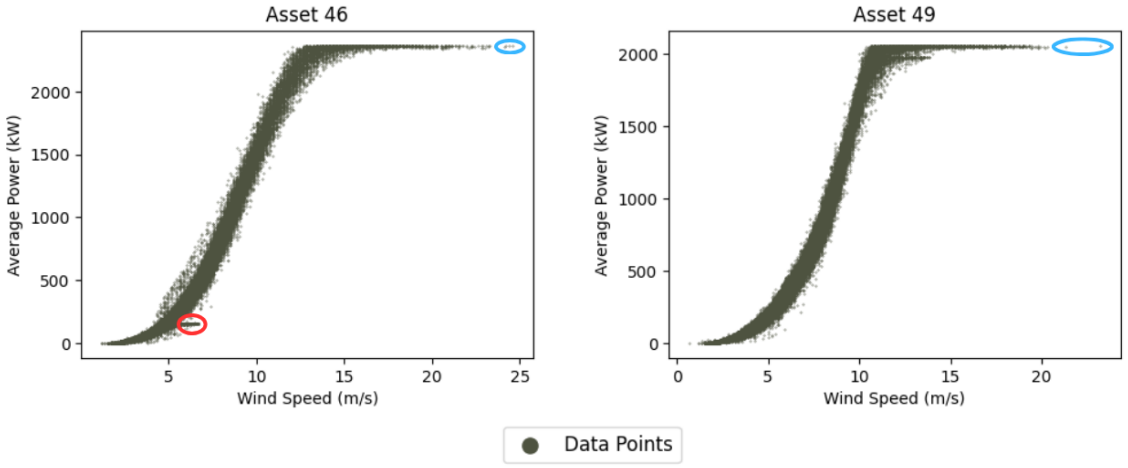


Figure 3.6 – Power Curves with anomalous data behavior

To fully understand the impact of every cleaning step, Figure 3.7 shows the results of each cleaning step on asset 49, starting with the data after cleaning the wind speed values equal or under 0 and wind direction outside the expected range, until the final result that passed by the sigmoid bands' procedure, MSM, and DBSCAN.

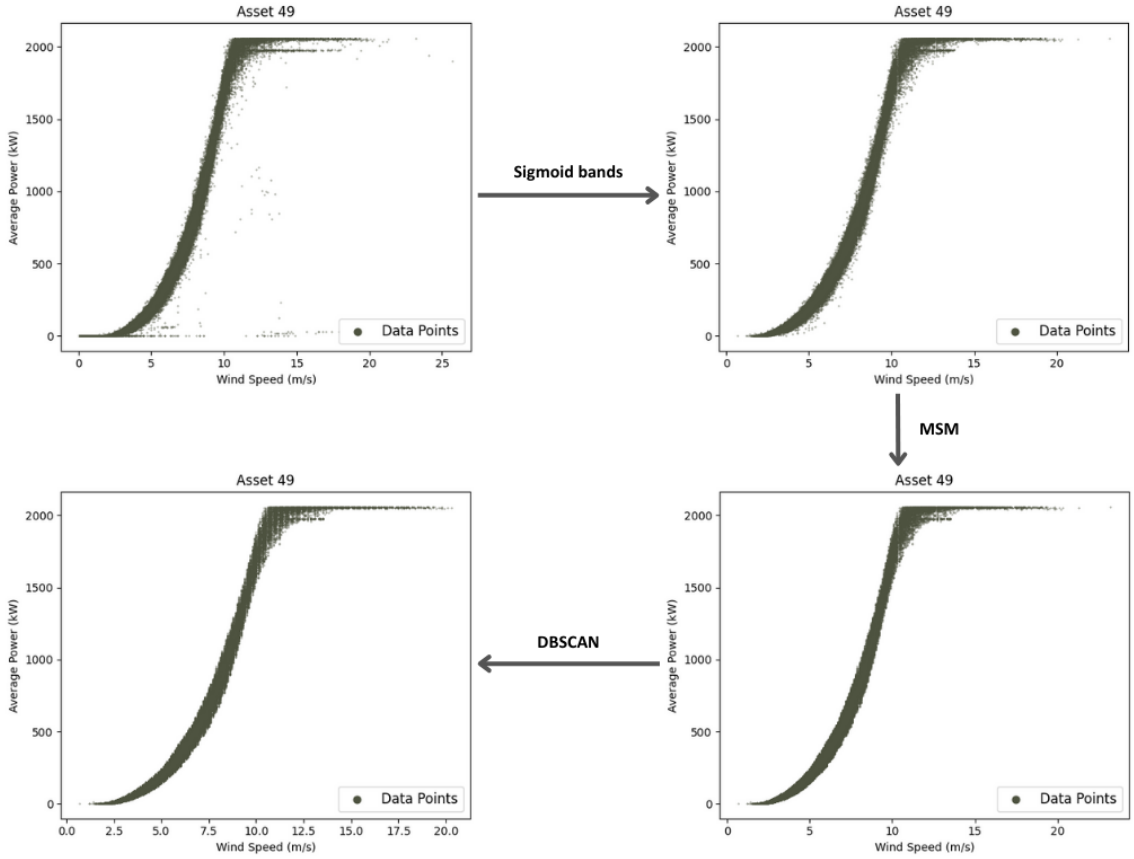


Figure 3.7 – Impact of each cleaning step on asset 49

Appendix C shows the cleaning results of the data of the first six and last five assets as an example.

After the cleaning, the concluding dataset encompasses 5 384 535 data entries, indi-

cating that 2 983 224 data points (35.65% of the original dataset) were deemed outliers.

3.3.2 Solar and Satellite Data

Solar clustering involves the utilization of solar and satellite data. So, the cleaning occurs with solar and satellite data combined. After this section, these two datasets will be referred to as solar data.

Correlation Analysis

Figure 3.8 presents a solar and satellite features correlation matrix. Excluding the principal diagonal, it is possible to observe pronounced correlations between distinct variables, such as *ac_power* with *ac_current* (99%) and *dc_power* (100%), with the latter two also exhibiting substantial correlation (99%) between themselves. The variable *ac_voltage* displays a moderate correlation with all variables except *dc_voltage*.

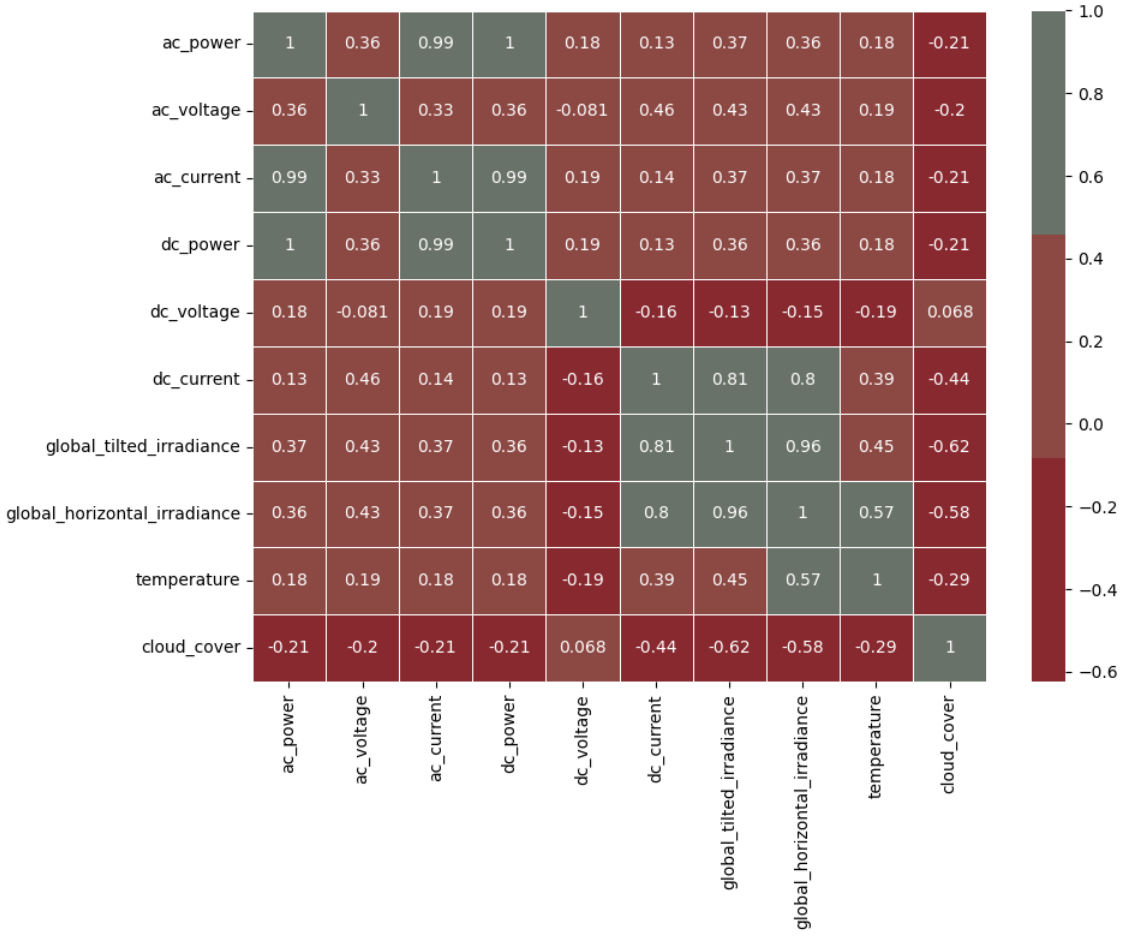


Figure 3.8 – Correlation Matrix of Solar and Satellite Data features

As previously indicated, the DC variables represent the solar panel's output, while the AC variables denote the output after conversion by the inverter. Consequently, the DC variables hold greater significance for this analysis, as the AC variables primarily convey information regarding the grid's nature rather than the inverter's characteristics. Considering the above statement, the dataset for the Modelling phase excludes AC variables.

The dataset will also exclude the features *global_tilted_irradiance* and *global_horizontal_irradiance* due to their notable correlations with *dc_current* (81% and 80% respectively).

Despite irradiance being the primary energy source, it is not as highly correlated with DC current as wind speed is with the produced wind power in the wind dataset. That can be explained by seasonality and by the fact that irradiance is not measured in the inverters, as wind speed is measured in the wind turbine, but by satellites, which results in a considerable associated error.

To clarify, the dataset for the clustering phase will include the following features: *dc_power*, *dc_voltage*, *dc_current*, *temperature*, and *cloud_cover*.

Null and Duplicated Values Treatment

In satellite data, there is no presence of null values, but in the solar dataset, 99 722 rows contain null values in the *ac_voltage* and *ac_current* features. However, since these variables are not present in the dataset used for modeling purposes, it is not essential to address these null values.

Neither satellite nor solar data present any duplicated values.

Outliers Treatment

Observing and treating outlier values in solar data use DC current and voltage variables since their multiplication represents the inverter's DC power, depicting it as an I-V curve. Figure 3.9 shows the initial I-V curves of the first two solar assets.

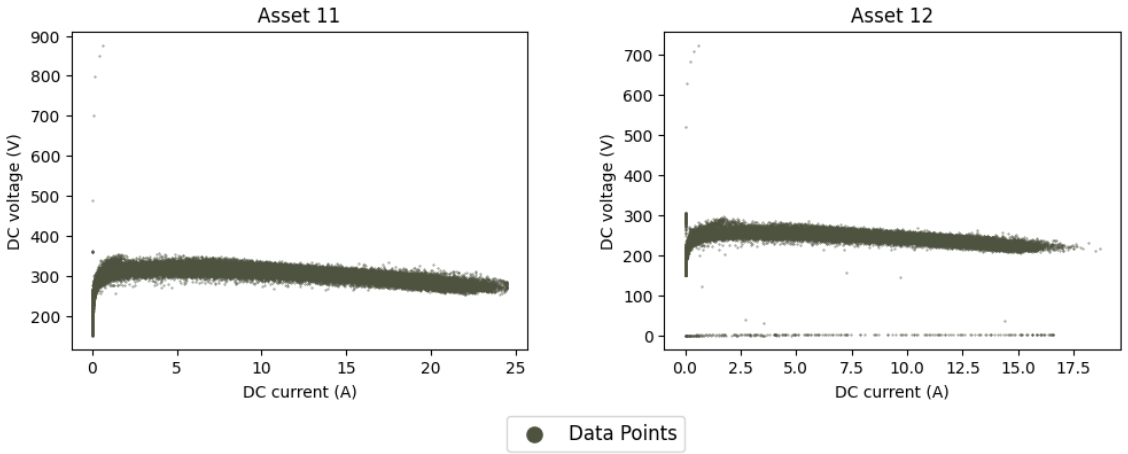


Figure 3.9 – Initial I-V Curves

Excluding all instances where DC power equals zero is necessary as these likely indicate faults in the inverter since the dataset does not contain any nighttime data to justify this value. This exclusion represents 9.43% of the data entries.

The data from the beginning and end of the day presents values with a lot of variability in voltage and current, which leads to it not representing the typical behavior of the rest of the I-V curve. For that reason, it is necessary to disregard the first 0.8% of the lowest values of each asset for DC power. In Figure 3.9 evident atypical values can be

noticed. Therefore, values above 17.5% and below 20% of the median value observed for an asset's DC voltage get eliminated. Appendix D shows the results from this first cleaning process, where 10.17% of the data entries were considered outliers.

Two techniques are tested to remove the not-so-apparent outliers: the Mean Squared Method, given its success in cleaning the wind data, and DBSCAN. Through Figures 3.10 and 3.11, it is possible to understand that the MSM can clean the outliers more efficiently than the DBSCAN, and for that reason, this is the technique applied to the dataset.

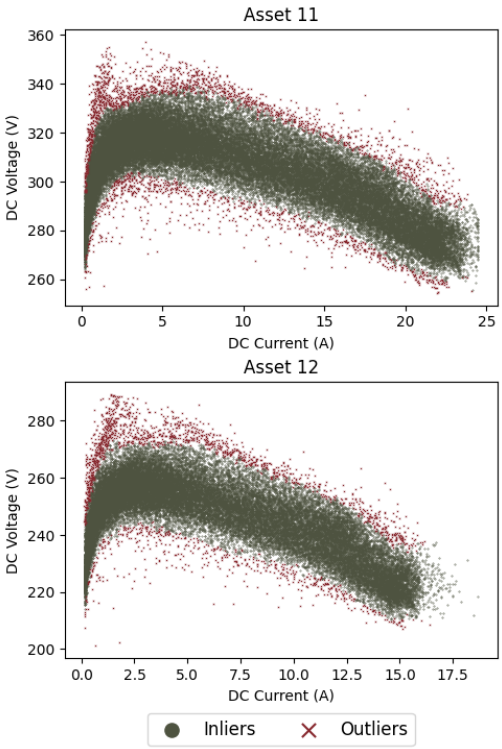


Figure 3.10 – Results of Mean Square Method

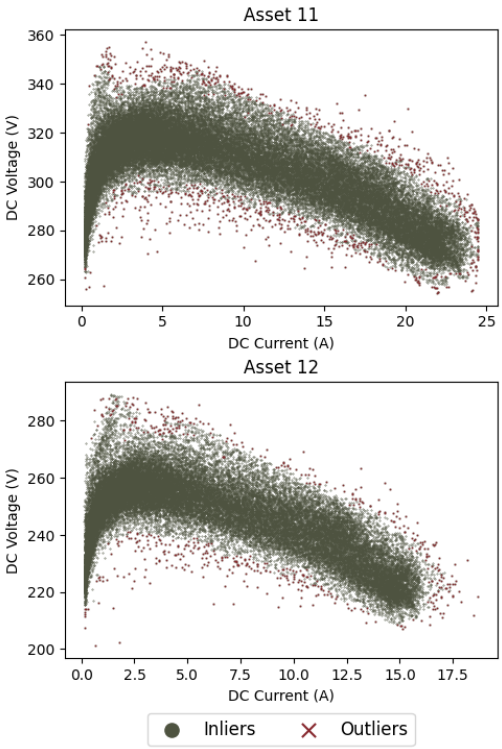


Figure 3.11 – Results of DBSCAN

The final dataset contains 3 160 060 entries, meaning that 921 714 data entries (22.58% of the original dataset) were considered outliers.

Figure 3.12 shows the results of each cleaning step for asset 12. The scatter plot on the left depicts the result of disregarding the lowest 0.8% of DC power and the DC voltage values above or below the defined limits, and the one on the right, the final result after the data passed through the MSM.

Appendix E provides an illustrative example, showcasing the final I-V curves of the first six assets representing typical I-V curve behavior in Figure E.1, alongside an additional six assets in Figure E.2 that depict I-V curves with an intriguing shape.

3.4 Modelling

This phase consist on the implementation of three types of clustering: Classic Clustering, Ensemble Clustering, and Time Series Clustering.

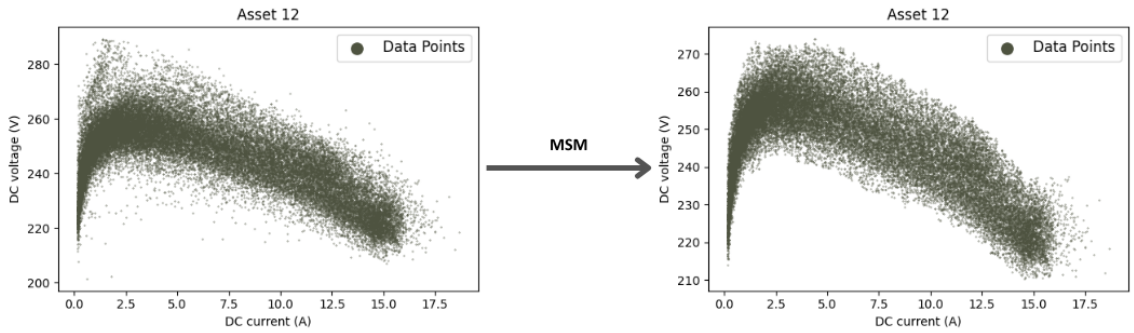


Figure 3.12 – Impact of each cleaning step on asset 12

The solar and wind data underwent the same algorithmic procedures, with all algorithms tested on the cleaned data from the preceding section and the original, non-cleaned data to understand what impact cleaning has on the asset assignment and what type of dataset produces better results.

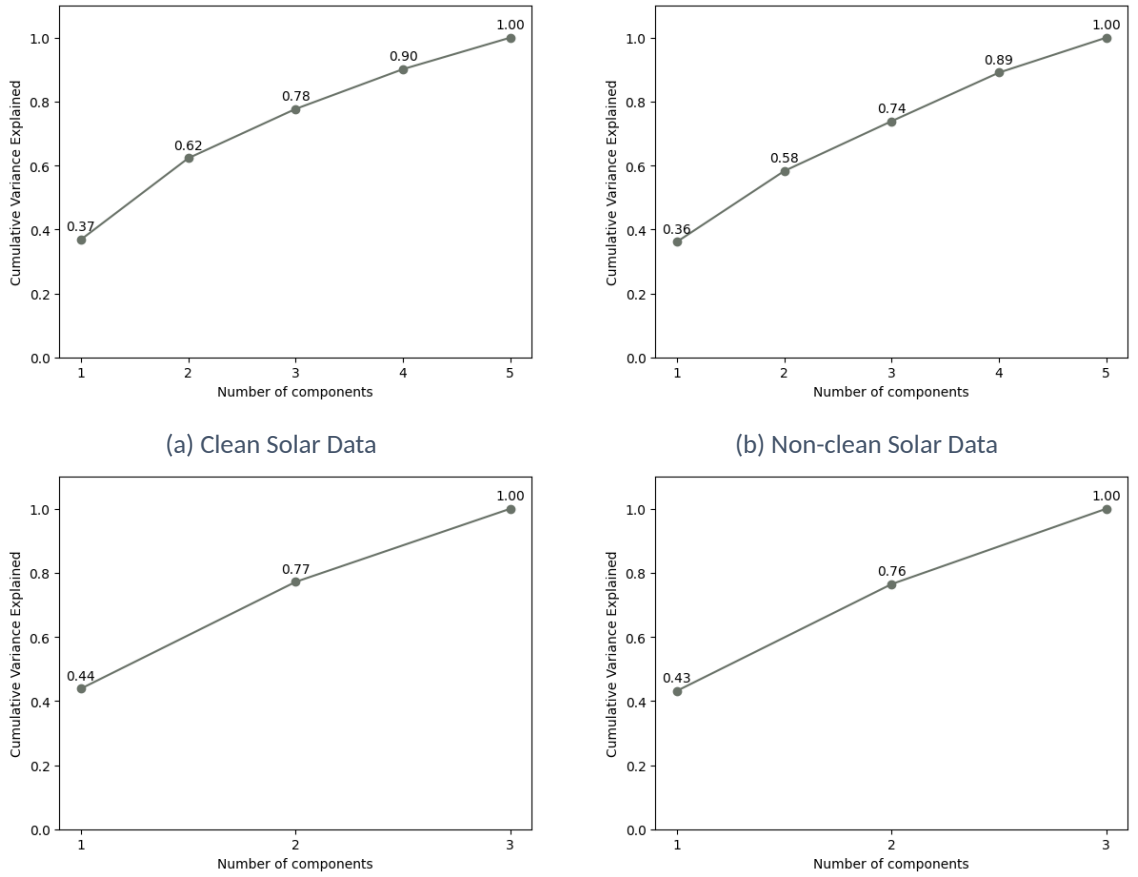
First and foremost, it is imperative to note that the substantial computational expenses associated with utilizing the entire datasets for segmentation, owing to their dimensions in rows and columns, prompted the adoption of two techniques to mitigate this cost. Establishing a predetermined quantity of rows per asset alleviates the volume of rows. This approach not only diminishes the number of rows but also ensures uniformity across assets by maintaining an identical number of rows per asset, thereby preserving parity in information content. Table 3.1 specifies each dataset's allocated number of rows per asset and the total number of rows in the dataset. Determining the number of rows to utilize relies on a fraction of the existing rows of the first asset in each dataset. The process selects the rows to include in the dataset randomly.

Data	Cleaned	Rows per asset	Total rows in dataset
Solar	Yes	664	51128
	No	521	40117
Wind	Yes	631	44801
	No	707	50197

Table 3.1 – New number of rows in each dataset

To further reduce the expenses associated with the datasets, PCA diminishes their dimensions by identifying the principal components that represent linear combinations of the original variables. This process preserves most of the variance inherent in the dataset while facilitating dimensionality reduction. To define the number of components that the data must be reduced to, is common to establish a value, that explains, usually, at least 70% of the total variance present in the original data (Jolliffe & Cadima 2016). In the case of this study, the components resulting from PCA should explain at least 90% of the original data variance. However, as Figure 3.13 illustrates, this goal is not achievable for most datasets, with clean solar data being the exception (Figure 3.13a). Applying PCA is impractical for the other datasets, as maintaining at least 90% of the variance would require preserving all original dimensions since reducing the dimensions of any wind dataset would imply reducing the explained variance from 100% to around 76%

and to 89% for the non-cleaned solar dataset. For this reason, PCA is only applied to the clean solar dataset, reducing its dimensions from five to four, and this new dataset is then used for segmentation. The non-cleaned solar data and all wind data retain their original dimensions and undergo value normalization to ensure that differences in magnitude do not disproportionately influence clustering results.



(a) Clean Solar Data

(b) Non-clean Solar Data

(c) Clean Wind Data

(d) Non-clean Wind Data

Figure 3.13 – Cumulative Explained Variance per number of components for each dataset

The dataset in use for the segmentation of clean solar data, composed of Principal Components (PC) 1, 2, 3, and 4, explains 90.14% of the variance of the original data. The percentage of explained variance is divided between the PCs as follows: 36.80%, 25.48%, 15.41%, and 12.45%, respectively. This means that the first two PCs will have more impact on clustering as they explain more of the original data variance.

Table 3.2 presents the percentage of variation of each original variable explained in each PC. PC1 accounts for more than half of the variation in *dc_current*, *temperature*, and *cloud_cover*. PC2 explains almost 70% of the variation in *dc_power* and *dc_voltage*. PC3 contains 60.23% of the variation in *temperature* and 50.88% of the variation in *cloud_cover*, while PC4 primarily captures 59.20% of the variation in *dc_voltage*. The last row of the table shows the total variation of a feature explained in the PCA version of the data. With that in mind, it is possible to state that *dc_power* and *dc_voltage* are the features that will have more impact on clustering and *dc_current* the feature with less impact.

Principal Component	dc_power	dc_voltage	dc_current	temperature	cloud_cover
1	0.2234	0.2229	0.5699	0.5302	0.5427
2	0.6978	0.6805	0.1604	0.0050	0.1556
3	0.3450	0.3459	0.3452	0.6023	0.5088
4	0.4173	0.5920	0.3343	0.4621	0.3875
Total	1.6835	1.8413	1.4098	1.5996	1.5946

Table 3.2 – Percentage of variation of each feature present in the Principal Components (cleaned solar data)

The described datasets serve the purpose of segmentation for classic and ensemble algorithms. The datasets undergo a distinct process for time series clustering to incorporate the time variable. Section 3.4.3 explains the method applied to the time series datasets in more detail.

The following subsections detail the implementation of each algorithm tested for each type of clustering.

3.4.1 Classic Clustering

This type of clustering involves typical algorithms that cannot capture the data variability caused by time. The classical algorithms for this project fall under the following categories: Hierarchical Clustering, Partitional Clustering, Soft Clustering, Model-Based Clustering, and Density-Based Clustering.

The classic clustering categories divide this subsection. Appendix F presents the resulting cluster visualizations for all classic clustering algorithms for clean solar data as an example.

3.4.1.1 Hierarchical Clustering

This clustering method applies four agglomerative algorithms: Single Link, Average Link, Centroid Link, Ward's Link; and the divisive algorithm BIRCH. The agglomerative algorithms utilize a linkage distance threshold of 0.01, wherein clusters solely merge if the linkage distance between them is less than or equal to 0.01 (*sklearn.cluster.AgglomerativeClustering — scikit-learn 1.4.2 documentation* n.d.).

Average Link applied to clean solar data serves as an example of the process undergone by each agglomerative algorithm. Figure 3.14 represents the dendrogram with five hierarchical levels resulting from the algorithm. The *y-axis* of a dendrogram portrays the dissimilarity between clusters. Therefore, when determining the number of clusters, selecting a value with a notably high dissimilarity is customary, but also where there is a considerable difference in dissimilarity between hierarchical levels. This approach facilitates the creation of distinct clusters.

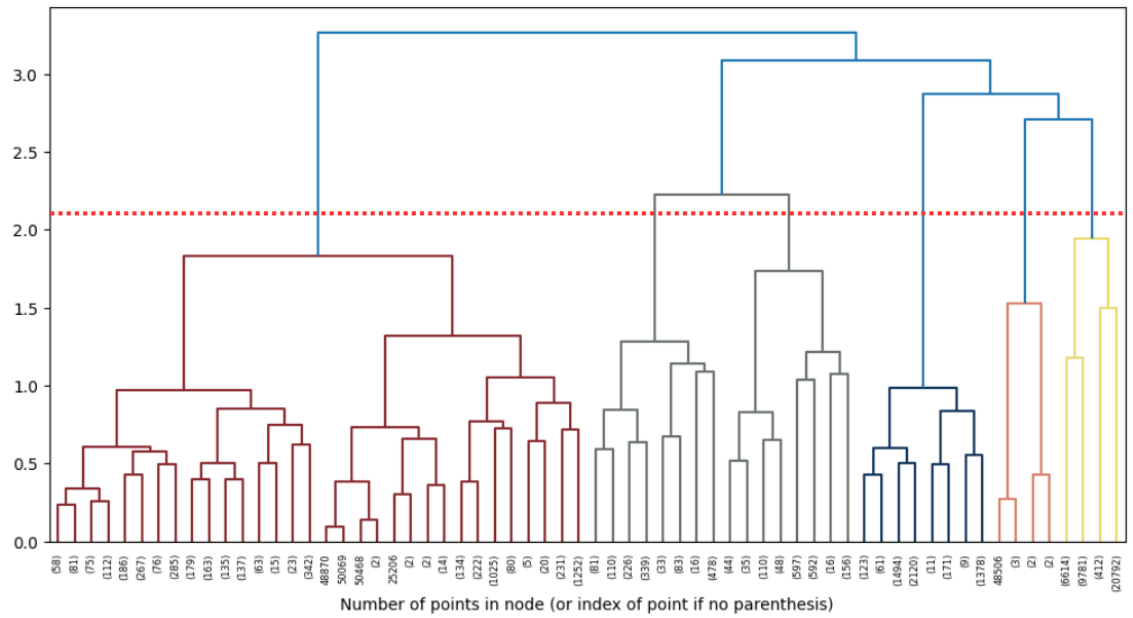


Figure 3.14 – Dendrogram of Average Link (Clean Solar Data)

In this algorithm, the designated number of clusters is six, as indicated by the red dotted line in Figure 3.14. However, the dataset containing multiple rows per asset results in the algorithm assigning one asset in more than one cluster, and since the objective is to identify the most similar assets, it is more logical that one asset belongs to just one cluster. We can assume that the cluster most often assigned to an asset is the one to which the asset belongs. With that assumption, in this example case, all the assets are divided into three clusters instead of the initial six. The procedure applies to the remaining hierarchical algorithms and partitional, model-based, and density-based clustering.

Cutting the dendrogram is impossible for BIRCH, as the used *Python* function does not present the necessary attributes to compute it. So, to avoid a large number of cluster labels, the distance threshold in use is 1.5, which results in three clusters in the case of clean solar data.

3.4.1.2 Partitional Clustering

The algorithms applied for this type of clustering are K-Means, since it is one of the most famous partitional clustering algorithms, and CLARA, as an adaptation of K-Medoids since this project uses large datasets.

One of the primary challenges in partitional clustering revolves around determining the appropriate number of clusters for data division. The elbow method provides a solution to this quandary by identifying the optimal number of clusters, marked by the point where inertia (defined on the y-axis of the plot) begins to decrease at a decelerated rate. For instance, Figure 3.15 depicts the plot of applying the elbow method to K-Means on clean solar data, where the selected number of clusters is five.

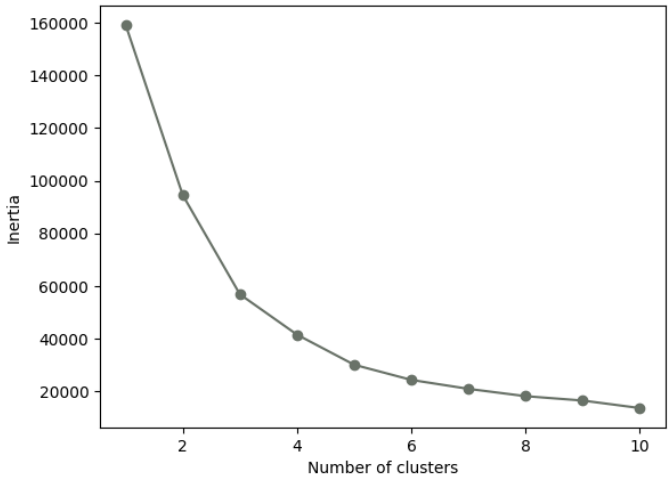
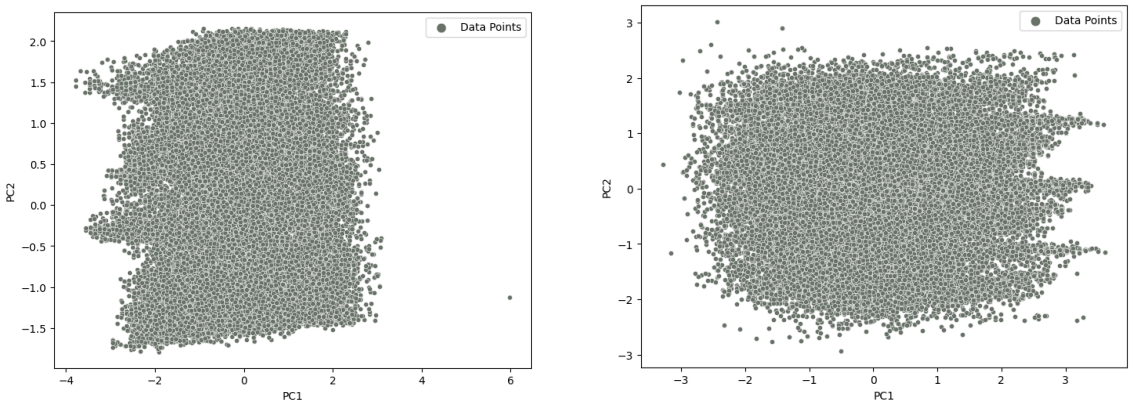


Figure 3.15 – Elbow Plot

3.4.1.3 Soft Clustering

Due to the absence of distinct cluster boundaries in the datasets, observed in solar data, as Appendix F depicts, but essentially in wind data that exhibits an even greater degree of cluster indistinctiveness, as illustrated in Figure 3.16, soft clustering methods may offer a practical approach for analyzing this data type.



(a) Clean Data (b) Non-clean Data

Figure 3.16 – Visualization of Wind data points

The utilized algorithm is Fuzzy C-Means, configured with an identical number of predefined clusters as determined by the elbow method for partitional clustering since Fuzzy C-Means can be categorized as a form of partitional clustering and employs a soft clustering approach as opposed to the rigid nature of algorithms such as K-Means and CLARA. The applied degree of fuzziness is two and is equal for all datasets.

3.4.1.4 Model-Based Clustering

The algorithm that constitutes the model-based clustering in this project is the GMM.

Like partitional, model-based clustering requires specifying the number of clusters

as input. However, in this scenario, the Python function utilized for GMM does not furnish an inertia attribute akin to K-Means. Consequently, computing the elbow method is unfeasible. Therefore, the number of clusters that result in the segmentation with the highest silhouette score determines the number of clusters for data segmentation.

3.4.1.5 Density-Based Clustering

Given the heightened sensitivity of density-based algorithms to their input parameters, selecting parameters for DBSCAN and OPTICS entails employing a Grid Search methodology. This approach involves utilizing a custom scoring function that integrates both the silhouette and the Davies Bouldin scores.

3.4.1.6 Parameter Optimization

Until now, the selected number of clusters relies on scores, inertia, or, in the case of hierarchical clustering, on subjective determination, as the dendrogram is truncated based on the researcher's judgment. Therefore, parameter optimization ensures that every algorithm receives an equitable opportunity by finding the optimal parameters for each algorithm.

This phase utilizes a frequently employed function in the *R* programming language for algorithms available within it, including Single, Average, Centroid, and Ward's Link, as well as K-Means and K-Medoids (utilized for CLARA), named *NbClust*. This function applies twenty-six metrics to evaluate the best partition for every algorithm (*NbClust function - RDocumentation* n.d.).

A Grid Search is applied with a custom function for the remaining algorithms *NbClust* does not provide. The function computes the Silhouette and DB scores, allows the selection of weights for each index, and employs Cross-Validation.

After obtaining the optimized parameters, the procedures described for the Classic Clustering algorithms applied in this project are repeated with the optimized parameters.

3.4.1.7 Conclusions

Section 3.5 conducts a more in-depth analysis of the best segmentations obtained for classic clustering and its corresponding indexes scores. However, given the ensemble clustering process developed in the subsequent stage of this project, it becomes imperative to showcase the three algorithms that exhibit the best metrics. Table 3.3 presents the best algorithms obtained for this type of clustering.

Data	Cleaned	Best	Second Best	Third Best
Solar	**Yes**	K-Means	CLARA	Ward's Link
	No	K-Means optimized	Ward's Link optimized	CLARA
Wind	**Yes**	K-Means optimized	CLARA	Fuzzy C-Means
	No	K-Means optimized	BIRCH	Fuzzy C-Means

Table 3.3 – Best algorithms of Classic Clustering for each dataset

3.4.2 Ensemble Clustering

Ensemble clustering is the integration of multiple classic clustering algorithms. However, instead of testing all potential combinations of these algorithms, the ensemble is constructed using the three most effective algorithms identified in Table 3.3.

For each dataset, the combination of the three algorithms undergoes testing. Cluster label assignment employs Major Voting, which amalgamates the results derived from each algorithm. This combination involves tallying one vote for each algorithm's cluster label assignment, with the final selected label receiving the most votes. A weight is assigned to each algorithm based on its performance in classical clustering, giving greater weight to higher-ranking algorithms.

The ensemble technique is employed to enhance the metric results derived from classic clustering methods. This technique typically facilitates the combination of various algorithms, wherein one may mitigate the limitations of the other, thereby attenuating the weaknesses inherent in each algorithm.

As an example, Appendix G shows the visualizations of the segmentations obtained for the Clean Solar dataset.

3.4.3 Time Series Clustering

Conducting this form of clustering analysis is imperative, considering the data's inherent nature and temporal variability. However, the absence of temporal consideration thus far necessitates preparatory measures, particularly in incorporating the time variable. Ensuring uniformity in the dataset regarding the quantity of data across the time series is also indispensable.

This clustering approach involves the utilization of both K-Means and SOM. K-Means employs two distinct distance metrics: Euclidean and DTW. To thoroughly examine the variables that have the most significant impact on clustering, besides the datasets containing all variables used until now plus the time variable, the algorithms utilize all possible datasets consisting solely of the time variable and one additional variable from those previously employed.

Firstly, the data contained within a single asset defines a time series. Since not every asset comprises an equal number of data entries, ensuring uniformity of data entry counts in each time series is crucial. The uniformization is accomplished by obtaining

the asset with the fewest data entries and ensuring that every asset possesses that same quantity of data. The choice of data entries to include is made randomly. This process results in time series with the composition presented in Table 3.4.

Data	Cleaned	Time Series	Entries
Wind	Yes	71	60 633
	No		117 854
Solar	Yes	77	5 980
	No		12 889

Table 3.4 – Composition of time series of each dataset

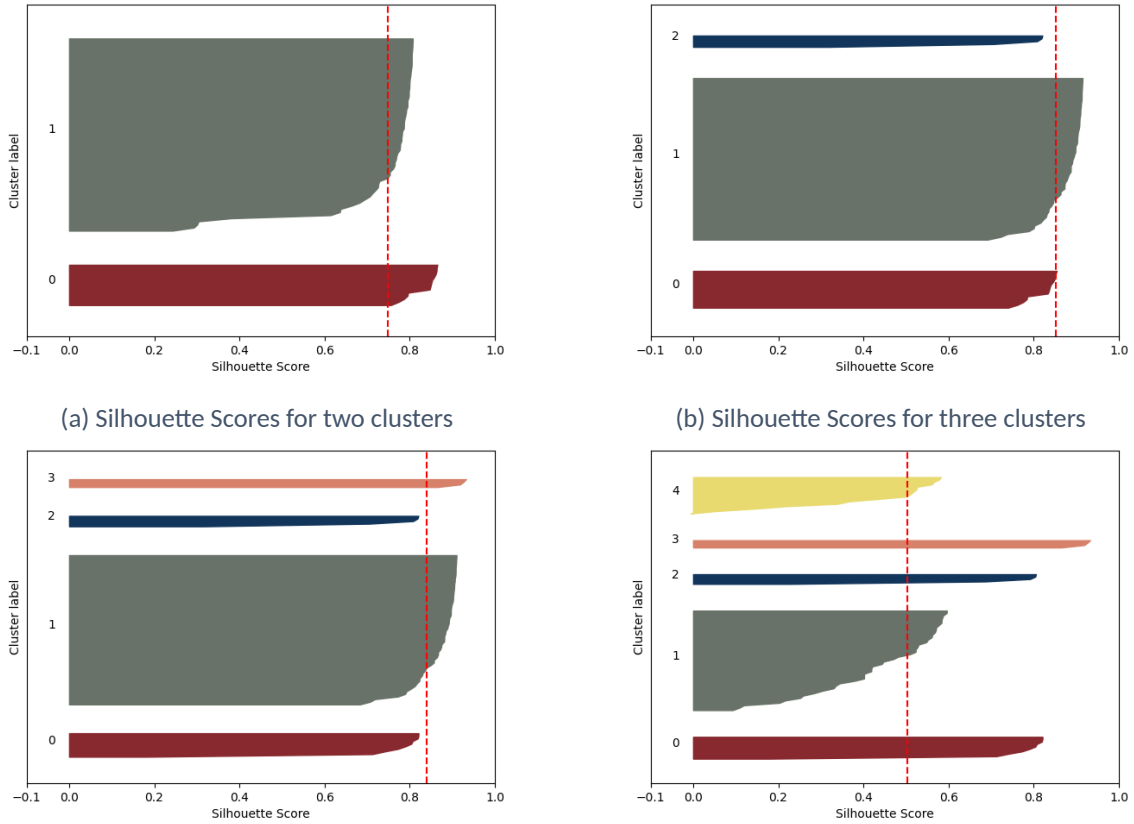
The time variable is transformed into Unix timestamps, representing the date in float format. Given the difference in numerical magnitude among the dataset variables, the data undergoes transformation to ensure uniform scaling. This scaling prevents any single variable from exerting a disproportionate influence on cluster label assignment due to its larger magnitude than others with smaller magnitudes. PCA is applied to reduce the feature dimensionality of the data, resulting in datasets composed of two components.

Once again, defining the optimal number of clusters poses a challenge. Therefore, for K-Means, the input quantity of clusters is determined by the partition that achieves the highest average silhouette score while maintaining balanced silhouette scores across individual clusters. The average SS is obtained by fitting the clustering algorithm to the entire dataset and predicting the cluster labels for each time series. The silhouette score for each individual time series is then calculated using these predicted labels. The SS for each time series considers how similar it is to other time series within the same cluster compared to those in the nearest different cluster. By averaging these individual silhouette scores, we get the overall average SS for the entire dataset. The individual cluster silhouette scores and the average SS can be visualized in Figure 3.17 for each tested partition with clean solar data.

Observing the difference in the values on the y-axis within each cluster label makes it apparent how many samples correspond to each silhouette score value. For instance, in Figure 3.17a, it is evident that there are more samples with a silhouette score of around 0.3 compared to those with a score of 0.8.

In addition to considering the average and individual silhouette scores, another criterion involves assessing whether the clusters are of similar size. Clean solar data does not achieve this balance in any tested partition. Figure 3.17 reveals that the partition with three clusters attains the highest average score, represented by the dotted red line. In this scenario, the individual silhouette scores are also relatively balanced. Nevertheless, even if the highest silhouette score is associated with the partition featuring three clusters, the decision could sway towards a partition with four clusters if it demonstrated significantly more balanced individual clusters than the three-cluster partition.

Instead of specifying the exact number of clusters, SOM requires the dimensions of the grid onto which it will map the time series. For instance, if the chosen grid dimension is 2×2, the maximum number of clusters will be 4. Determining the number of clusters



(a) Silhouette Scores for two clusters

(b) Silhouette Scores for three clusters

(c) Silhouette Scores for four clusters

(d) Silhouette Scores for five clusters

Figure 3.17 – Visualization of the Silhouette Scores for Clean Solar data

for SOM relies on each partition's average silhouette score. Appendix H shows the visualization of the clusters obtained for clean solar data with SOM.

3.5 Evaluation

This section discusses the achievement of the success criterion for solar and wind segmentations.

To determine the most effective models for each dataset type, a custom function applies five metrics to the segmentation results: Silhouette Score, Davies-Bouldin, Calinski-Harabasz, and Dunn Indexes.

Upon obtaining the scores for each segmentation, ordering ensues to discern which algorithms exhibit the most favorable combination of scores. As seen in Section 2.2.4, higher SS, Dunn Index, and CH Index values indicate superior segmentation, while the opposite happens for DB index. Subsequently, the ideal combination of metrics, as well as an anti-ideal combination, is computed. Following this, the Euclidean distance between the indexes results for each segmentation and the ideal combination is calculated, along with the distance between the scores results for each segmentation and the anti-ideal combination. Finally, a relative proximity to the ideal solution is determined, ordering the algorithms from the closest approximation to the ideal solution to the farthest.

3.5.1 Clustering with Solar Data

This subsection presents the results obtained for all clustering algorithms tested with the solar datasets, categorized by the types of clustering utilized. The final product of this project will include the best algorithm and dataset for clustering the solar inverters.

3.5.1.1 Classical Clustering

Table 3.5 presents the three best algorithms ordered in descending order for each dataset and the indexes results for those algorithms.

Cleaned	Algorithms	Indexes			
		SS	DB	CH	Dunn
Yes	K-Means	0.3734	0.8725	32078.0102	0.0008
	CLARA	0.3503	0.8829	29460.1706	0.0011
	Ward's Link	0.3305	0.9194	28015.5203	0.0037
No	K-Means	0.3287	1.0048	17873.5275	0.0013
	Ward's Link	0.2573	1.0994	13809.9036	0.0042
	CLARA	0.2358	1.2362	13367.3665	0.0008

Table 3.5 – Metrics of the best three Classic Clustering algorithms for Solar data

As noticeable, K-Means is the best algorithm for solar data as it is the 1st placed in both clean and non-clean versions. However, it is possible to note that the scores are better for the segmentations with clean data. Starting by comparing the best algorithm in both cases, every index, except Dunn, presents a more favorable value in the clean data segmentation. Although the index values are not optimal, the K-Means clustering indexes for clean solar data suggest relatively compact and separated clusters. The SS of 37.48% suggests clusters with moderate separation, indicating that, on average, data points in the same cluster are closer than those in other clusters, but some overlap might happen. The CH value of 32078.0102 reinforces K-Means' superiority over the alternative algorithms.

3.5.1.2 Ensemble Clustering

Table 3.6 delineates the results for ensemble clustering with clean solar data.

Ensembled Algorithms	Indexes			
	SS	DB	CH	Dunn
K-Means + CLARA + Ward's Link	0.0693	1.8108	8365.5599	0.0009

Table 3.6 – Ensemble Clustering Metrics Results for Clean Solar Data

The index scores of the ensemble with the three algorithms suggest that they do not tend to agree on the label assignment, producing worse results than the individual algorithms. The clustering results indicate poor performance, as evidenced by a very low SS, high DB Index, relatively low CH Index, and extremely low Dunn Index. These values suggest that the clusters are overlapping, poorly defined, and not well-separated.

Table 3.7 presents the obtained scores of the ensembling with the non-cleaned data.

Ensembled Algorithms	Indexes			
	SS	DB	CH	Dunn
K-Means + Ward's Link + CLARA	0.2621	1.5023	12091.4507	0.0010

Table 3.7 – Ensemble Clustering Metrics Results for Non-Clean Solar Data

Despite yielding poorer results than the segmentations with the classic algorithms for non-cleaned data, the ensemble segmentation performed better with this version of the data than with the cleaned data. The Dunn and DB indexes indicate that the segmentation for the uncleaned dataset is only slightly better separated and compact. Still, the remaining indexes demonstrate an improvement over the segmentation obtained from the cleaned data.

3.5.1.3 Time Series Clustering with all variables

For the segmentations generated using the datasets utilized thus far for solar data with the addition of the time variable, K-Means consistently yields identical partitions regardless of the distance metric employed. Consequently, when discussing these datasets, focusing only on the results obtained from K-Means utilizing DTW is appropriate.

Table 3.8 presents the results achieved for the case in analysis. Previously, all computed index values used the Euclidean distance metric, which aligned with the algorithms' utilized distance metric. Currently, the calculation of the indexes for each algorithm incorporates the specific distance metric employed by that algorithm. For instance, in the case of K-Means, DTW is used to compute the scores instead of the Euclidean distance. This methodology applies to all time series segmentations.

Cleaned	Algorithm	Distance metric	Indexes			
			SS	DB	CH	Dunn
Yes	K-Means	DTW	0.8511	0.2242	521.9335	0.6327
	SOM	Euclidean	0.8262	0.2397	442.5247	0.8187
No	K-Means	DTW	0.8484	0.2641	462.8346	0.4023
	SOM	Euclidean	0.8272	0.2227	370.7730	0.5608

Table 3.8 – Time Series Clustering Results for Solar Data with usual variables

The presented scores unmistakably indicate the suitability of time series clustering for these datasets. In this instance, the cleaned dataset version yielded more noteworthy index values.

Time series clustering produces promising results compared to the other types of clustering. It is characterized by a high SS and a low DB score, suggesting substantial intra-cluster similarity and inter-cluster dissimilarity. Nevertheless, the low CH score hints at potential overlap or reduced cluster cohesion. Moreover, the Dunn implies a separation between clusters above the average.

Figure 3.18 illustrates the cluster visualization of the segmentation derived from clean solar data to elucidate the obtained scores. Each black line represents a time series, with

the red line depicting the cluster's centroid. The plot on the left displays the entire time series, while the one on the right provides a zoomed-in view, facilitating a comprehensive understanding of the behavior of the clustered time series and their centroids.

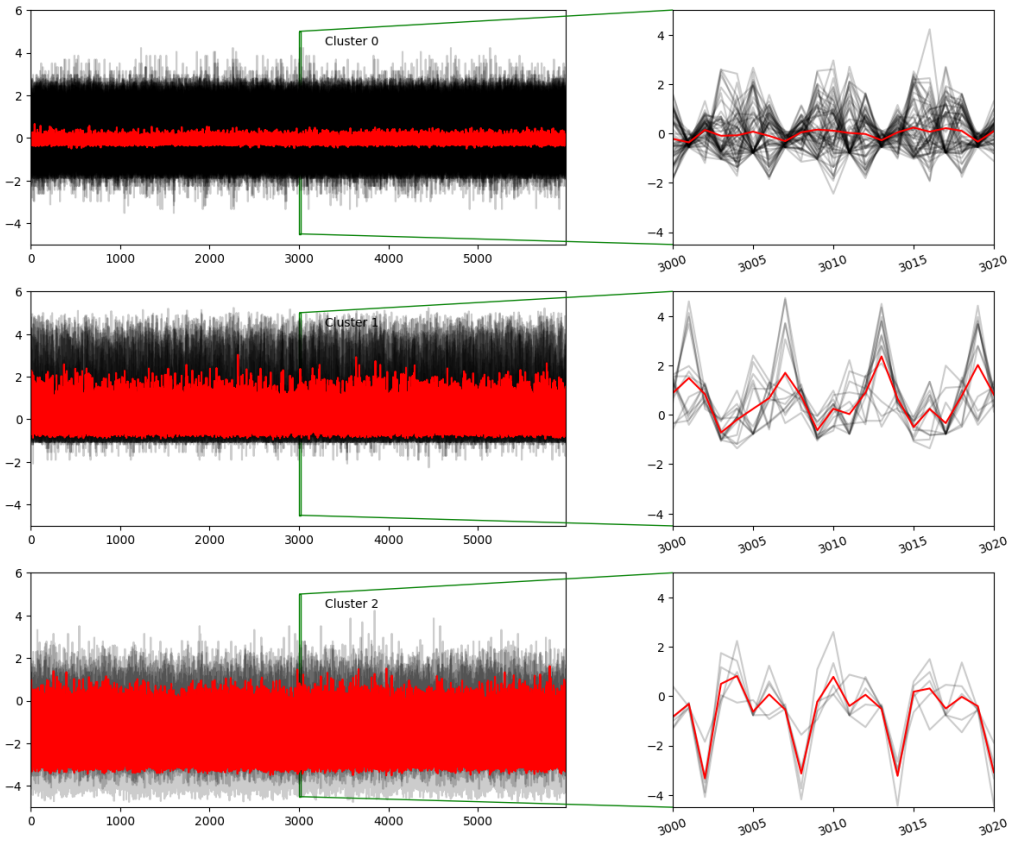


Figure 3.18 – Clusters of K-Means with DTW distance metric with Clean Solar Data

Representing all time series in one plot would cause them to overlap, which could justify the values obtained for the CH index. Additionally, it is observable that the time series within one cluster exhibit similarity while being distinct from those in different clusters supporting the SS, DB, and Dunn index values.

3.5.1.4 Time Series Clustering with individual variables

Because of their highly similar behavior, Appendix I, Table I.1 provides the metrics derived from the cleaned version of the solar data for time series clustering, while Table 3.9 displays the results for the non-cleaned version, which yielded superior segmentations. The variable used in the clustering, along with the time variable, divides the table, followed by the index values obtained for each segmentation produced using both variables.

Similar to the time series clustering utilizing all typical variables, in this instance, overall, the CH index exhibits inferior values compared to those obtained by other clustering types. However, the scores derived from segmenting using the variable *dc_power* are notable, indicating that the resulting segmentations are the most compact and well-separated of all the segmentations obtained with time series clustering and demonstrate the highest similarity among assets within the same cluster observed thus far. K-Means

Algorithms	Distance metric	Indexes			
		SS	DB	CH	Dunn
cloud_cover					
K-Means	DTW	0.7025	0.5537	128.0480	0.2811
SOM	Euclidean	0.6983	0.2734	134.5243	0.3337
dc_current					
K-Means	DTW	0.6076	0.5775	145.8639	0.0276
SOM	Euclidean	0.6208	0.4058	160.1544	0.0449
dc_power					
K-Means	DTW	0.9383	0.1173	1556.0316	1.2260
SOM	Euclidean	0.8995	0.1012	2597.4490	1.4085
dc_voltage					
K-Means	DTW	0.7249	0.4458	139.2027	0.3269
SOM	Euclidean	0.7148	0.2590	38.9054	0.3386
temperature					
K-Means	DTW	0.7054	0.2588	551.8100	0.1456
SOM	Euclidean	0.6791	0.2685	1021.6430	0.2140

Table 3.9 – Time Series Clustering Results for Non-Clean Solar Data with Individual Variables

achieved a 93.83% SS compared to almost 90% obtained by SOM. Nonetheless, the remaining indexes suggest a superior segmentation with SOM, with a CH exceeding 2500, while other variables obtain values between 100 and 200 and a Dunn index surpassing 1.

3.5.1.5 Conclusions

In conclusion, the analysis illuminates several critical insights into clustering techniques applied to solar data.

Classic and ensemble clustering proved incompatible with the dataset type used in this project. Classic and ensemble clustering, regardless of whether the data was cleaned or not, produced poorly separated and compact segmentations. However, ensembling the three algorithms simultaneously produced better segmentation with the non-cleaned version of the data.

The time series results indicate that time highly influences the variation in the solar dataset. Despite the segmentations obtained using all typical variables along with the time variable producing outstanding results, the focus remains on the segmentation produced by the non-cleaned dataset composed of the variables related to DC power and time.

A discernible pattern emerges regarding the impact of data cleanliness on clustering performance. Despite the data preprocessing enhancing the clustering outcomes of classic algorithms and time series with a dataset with more than two variables, in the remaining clustering methodologies, segmentations derived from non-cleaned data consistently exhibit superior score values compared to their cleaned counterparts. The optimal segmentation arises from non-clean data, implying that while data cleaning significantly influences clustering outcomes in some contexts, the inclusion of data points considered

outliers during the data processing phase aids algorithms in producing improved time series segmentations when utilizing only two variables.

The algorithm chosen for inclusion in the final product of this project for clustering solar inverters is SOM, utilizing the non-cleaned dataset containing only the time and *dc_power* variables (Figure J.1).

3.5.2 Clustering with Wind Data

This subsection outlines the segmentations obtained by each algorithm using cleaned and uncleaned wind data. The objective is to determine which algorithm and dataset to incorporate into the final product for clustering wind turbines.

3.5.2.1 Classic Clustering

Table 3.10 shows the results produced by the three best classic algorithms using wind data.

Cleaned	Algorithms	Indexes			
		SS	DB	CH	Dunn
Yes	K-Means opt.	0.3276	0.9612	24421.6546	0.0008
	CLARA	0.3074	1.0334	22714.9429	0.0011
	Fuzzy C-Means	0.3074	1.0334	22714.9429	0.0011
No	K-Means opt.	0.3186	0.9818	26172.5058	0.0007
	BIRCH	0.3005	1.0957	23267.3057	0.0011
	Fuzzy C-Means	0.2321	1.2961	19851.2167	0.0008

Table 3.10 – Classic Clustering Metrics Results for Wind Data

Once more, K-Means has produced the best segmentations for both clean and non-clean data. However, in the two cases, the metrics present little difference in the first two ranked algorithms. This trend does not hold for Fuzzy C-Means, which ranked in third place in both scenarios, where the clustering metrics indicate that the cleaned dataset significantly outperforms the uncleaned dataset in terms of clustering quality. For the cleaned dataset, the Fuzzy C-Means algorithm achieved a higher SS, CH Index, and Dunn Index, as well as a lower DB Index. These results suggest that the cleaned dataset has better-defined clusters with improved separation and compactness compared to the uncleaned dataset for the algorithm ranked in the third position.

3.5.2.2 Ensemble Clustering

Table 3.11 presents the metric results for the segmentations achieved through ensemble clustering for wind data.

Unexpectedly, the ensemble with the non-cleaned wind dataset presents a better segmentation than the cleaned version. The DB and Dunn indexes are better in the

Cleaned	Ensembled Algorithms	Indexes			
		SS	DB	CH	Dunn
Yes	K-Means opt.+ CLARA + + Fuzzy C-Means	0.1002	1.1782	5472.0219	0.0008
No	K-Means + BIRCH + + Fuzzy C-Means	0.24488	1.8291	13358.8489	0.0006

Table 3.11 – Ensemble Clustering Metrics Results for Wind Data

cleaned dataset, but the most significant difference lies in the SS, which achieved 24.49% in the non-cleaned version versus 10% in the cleaned version, and the CH Index, where the value is around 13359 versus only 5472.

Nonetheless, as with solar data, this type of clustering does not improve the segmentations obtained with the individual algorithms.

3.5.2.3 Time Series Clustering with all variables

When applying time series clustering algorithms to the wind dataset containing all variables used for the other clustering types plus the time variable, the tested distance metrics, DTW and Euclidean, yielded identical segmentations. Consequently, only the scores obtained for K-Means using DTW are presented for that algorithm in Table 3.12, alongside the values obtained for SOM.

Cleaned	Algorithms	Distance metric	Indexes			
			SS	DB	CH	Dunn
Yes	K-Means	DTW	0.5705	0.5547	145.0963	0.0867
	SOM	Euclidean	0.5021	0.5076	103.8943	0.1034
No	K-Means	DTW	0.5830	0.5520	139.2560	0.2141
	SOM	Euclidean	0.4817	0.7379	103.1113	0.0638

Table 3.12 – Time Series Clustering Results for Wind Data with usual variables

The attained scores confirm that the data variation is primarily justified by time, as the values are the most favorable observed for wind data thus far. However, only the SS, DB, and Dunn indexes exhibit better yet mediocre values. The CH index displays less amusing values, which can be attributed to the same reasons as those presented for time series clustering with solar data in the section 3.5.1.3.

The scores do not present a significant difference between the achieved segmentations with cleaned or non-cleaned data. K-Means with the non-cleaned data version possess the best SS and Dunn index, but CH achieves the higher value in K-Means with clean data, and the lowest value for DB also happens in the clean dataset, but with SOM.

3.5.2.4 Time Series Clustering with individual variables

K-Means yield distinct segmentations for the first time when employing Euclidean and DTW in two cases: one when the dataset contains the variable *power_average* and time,

and the other when the dataset contains *wind_direction* and time. Therefore, the scores obtained for K-Means using both metrics for the mentioned cases are presented in Table 3.13, alongside the scores obtained for SOM and the segmentations produced with the dataset containing *exterior_temperature*. Due to the identical behavior observed between the non-cleaned version of wind data and the optimal results of the clean version, Appendix I, Table I.2 presents the attained scores of the segmentations derived from the non-cleaned wind data.

Algorithms	Distance metric	Indexes			
		SS	DB	CH	Dunn
power_average					
K-Means	Euclidean	0.4398	0.7751	55.9600	0.0220
	DTW	0.4153	0.8163	56.6061	0.0331
SOM	Euclidean	0.4232	0.6287	43.3935	0.0179
wind_direction					
K-Means	Euclidean	0.7371	0.3831	135.5360	0.4958
	DTW	0.4991	0.9333	106.2956	0.0155
SOM	Euclidean	0.6043	0.4703	98.9106	0.2538
exterior_temperature					
K-Means	DTW	0.4609	0.7836	72.7874	0.0293
SOM	Euclidean	0.4552	0.5339	138.9356	0.2462

Table 3.13 – Time Series Clustering Results for Clean Wind Data with individual variables

The segmentation produced with K-Means using the Euclidean distance metric for *wind_direction* in the cleaned dataset obtains the best scores among the other variables and datasets. The achieved SS of 73.71% and the remaining indexes, except for CH, are better than any other segmentations with any clustering type for wind data.

3.5.2.5 Conclusions

Concluding, wind data clustering generally yielded poorer results than solar data clustering, as detailed throughout this subsection, across all types of clustering. However, wind and solar data clustering achieved optimal segmentations with a dataset containing only the time variable and another corresponding variable in time series clustering.

For wind data, data preprocessing has little influence on classical and time series clustering with all typical variables. Nonetheless, in classic clustering, despite the scores falling short of ideal segmentation, the use of clean data resulted in indexes showing less unfavorable values than with non-clean data. However, the non-clean dataset is able to produce better results in ensemble clustering.

Furthermore, the best segmentation attained for wind data, and the one to include in the final product for clustering wind turbines, employed data that underwent cleaning processes. The algorithm that will cluster the wind turbines in the final product is K-Means for time series clustering employing the Euclidean distance metric and using the cleaned dataset composed by the time variable and *wind_direction* (Figure J.2).

4 CONCLUSION

This final chapter consists of two sections. The first section presents the conclusions of this project, addressing the initial proposed questions. Remembering the mentioned questions:

1. How can clustering techniques, supported by analytical methods and unsupervised ML algorithms, be effectively applied to group similar renewable energy assets based on historical power production, meteorological data, and power curve characteristics?

2. Furthermore, how does this clustering contribute to optimizing resource allocation and operational strategies, ultimately enhancing renewable energy networks' overall performance and resilience?

The second section discusses the limitations of this project and suggests future work.

4.1 Final conclusions

Effectively applying clustering techniques to group similar renewable energy assets based on historical power production, meteorological data, and power curve characteristics required several critical steps. One of the essential phases of this project was data cleaning, which involved identifying and excluding outliers from the dataset using various approaches. Following the data cleaning, several algorithms from classic, ensemble, and time series clustering categories were tested on both cleaned and non-cleaned data. Although the cleaned data often resulted in better segmentations for classic and ensemble algorithms, the cleaning had a minimal impact on segmentations obtained from these types of clustering. Conversely, the data cleaning proved beneficial for wind data in the context of time series clustering, yielding the best segmentation of wind turbines in this scenario.

Clustering solar inverters and wind turbines optimizes resource allocation and operational strategies, enhancing the overall performance and resilience of renewable energy networks. For clients, understanding which inverters or wind turbines are similar offers multiple advantages. This knowledge facilitates predictive maintenance by leveraging data from similar equipment, enabling more accurate maintenance schedules and reducing downtime. Additionally, performance benchmarking against similar equipment allows clients to identify and address underperforming assets, leading to improved efficiency. Optimized resource allocation becomes possible as clients can tailor maintenance strategies and allocate resources based on the specific needs of clustered equipment, thereby reducing waste and improving operational efficiency. Furthermore, enhanced operational strategies can be developed by understanding how similar equipment operates under various conditions, ultimately leading to reduced operational costs and improved performance.

Understanding the performance and reliability of different types of inverters and turbines facilitates informed investment decisions, guiding clients toward more strategic investments. Additionally, access to richer data analytics and insights derived from clustered equipment enables more informed decision-making and strategic planning. Overall, clustering provides clients with actionable insights that drive operational efficiency, cost savings, and strategic advantages, thereby enhancing the performance and resilience of their renewable energy assets.

4.2 Limitations and future work

This section addresses the limitations encountered during the study and proposes directions for future research.

The main limitations of this project included development time and computational costs. Due to limited computational resources, certain choices were made that could impact the clustering process, such as reducing the number of data entries in the datasets and utilizing PCA to decrease the number of variables included in the segmentation process. Bearing this in mind, various approaches can be implemented to enhance the results obtained.

Firstly, as mentioned in Section 4.1, the cleaning of the datasets did not result in a significant difference in the scores obtained by the segmentations produced. Future work should include testing new cleaning techniques for both solar and wind data. Given that time series clustering algorithms achieved the best segmentations, including the time variable during data pre-processing and using methods like a moving average could enhance the cleaning of outliers for this specific case. Additionally, the use of better computing resources would enable the utilization of all data entries for classic and ensemble clustering.

This project did not use geographical information about the assets for confidentiality reasons. However, the inclusion of this information in the datasets could improve the

segmentations obtained, especially in classic and ensemble clustering.

Regarding ensemble clustering, more algorithms should be considered for ensembling. For instance, rather than limiting the ensemble to the three best algorithms for each type of dataset, incorporating combinations of the best four or even the best five algorithms can enhance the resulting segmentations.

Being data mining an experimental field and knowing *a priori* what type of clustering and correspondent algorithms will work the best for the data is almost impossible, different algorithms must be tested, especially algorithms that can capture the time variation present in the time series, that as seen throughout this project, is critical when clustering the data in use.

Finally, using the algorithms that obtained the better scores for wind and solar data, the development of a software that identifies the N most similar assets, given a dataset with a pre-determined group of assets, could add value to Enlitia's portfolio.

REFERENCES

About - Enlitia (n.d.).
URL: *https://www.enlitia.com/about*

Adnan, W. M. A. B. M., Aziz, A. B. A. & Raya, L. B. (2022), 'Feasibility study of wind power generation system using small scale wind turbines', *2022 IEEE 10th Conference on Systems, Process and Control, ICSPC 2022 - Proceedings* pp. 166–169.
URL: *https://doi.org/10.1109/ICSPC55597.2022.10001785*

Aggarwal, C. & Reddy, C. (2013), *DATA CLUSTERING Algorithms and Applications*, Taylor & Francis Group.

Al-Ezzi, A. S. & Ansari, M. N. M. (2022), 'Photovoltaic solar cells: A review', *Applied System Innovation* **5**.
URL: *https://doi.org/10.3390/asi5040067*

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. & Aljaaf, A. J. (2020), *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*, Springer International Publishing, Cham, pp. 3–21.
URL: *https://doi.org/10.1007/978-3-030-22475-2_1*

Alrikabi, N. (2014), 'Renewable energy types', *Journal of Clean Energy Technologies* **2**, 61–64.
URL: *https://doi.org/10.7763/JOCET.2014.V2.92*

Anandi, V. & Ramesh, M. (2022), Descriptive and predictive analytics on electronic health records using machine learning, *in* '2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)', pp. 1–6.
URL: *https://doi.org/10.1109/ICAECT54875.2022.9808019*

Araújo, N., Sousa, F. & Costa, F. (2020), 'Equivalent models for photovoltaic cell - a review', *Revista de Engenharia Térmica* **19**, 77–98.

AZIZI, E., KHARRATI-SHISHAVAN, H., MOHAMMADI-IVATLOO, B. & SHOTORBANI, A. M. (2019), 'Wind speed clustering using linkage-ward method: A case study of khaaf, iran',

Gazi University Journal of Science **32**, 945–954.
URL: *https://doi.org/10.35378/gujs.459840*

Baig, M. Q., Khan, H. A. & Ahsan, S. M. (2020), 'Evaluation of solar module equivalent models under real operating conditions—a review', *Journal of Renewable and Sustainable Energy* **12**.
URL: *https://doi.org/10.1063/1.5099557*

Barka, C., Messaoudi-Abid, H., Setthom, H. B. A., Abdelghani, A. B.-B., Slama-Belkhodja, I. & Sammoud, H. (2020), 'A real time, wireless and low cost data acquisition system for residential pv modules', *6th IEEE International Energy Conference, ENERGYCon 2020* pp. 417–422.
URL: *https://doi.org/10.1109/ENERGYCON48941.2020.9236592*

Bhattacharjee, P. & Mitra, P. (2020), 'A survey of density based clustering algorithms', *Frontiers of Computer Science* **15**, 151308.
URL: *https://doi.org/10.1007/s11704-019-9059-3*

Chapman, P. (2000), Crisp-dm 1.0: Step-by-step data mining guide.
URL: *https://api.semanticscholar.org/CorpusID:59777418*

Cherif, A., Cardot, H. & Boné, R. (2011), 'Som time series clustering and prediction with recurrent neural networks', *Neurocomputing* **74**, 1936–1944. Adaptive Incremental Learning in Neural Networks Learning Algorithm and Mathematic Modelling Selected papers from the International Conference on Neural Information Processing 2009 (ICONIP 2009).
URL: *https://doi.org/10.1016/j.neucom.2010.11.026*

Dimić, G., Rančić, D., Rančić, O. P. & Spalević, P. (2019), Descriptive statistical analysis in the process of educational data mining, *in* '2019 14th International Conference on Advanced Technologies, Systems and Services in Telecommunications (TELSIKS)', pp. 388–391.
URL: *https://doi.org/10.1109/TELSIKS46999.2019.9002177*

Duarte, J. M., Fred, A. L. & Duarte, F. J. F. (2013), Data clustering validation using constraints., *in* 'KDIR/KMIS', pp. 17–27.
URL: *https://doi.org/10.5220/0004543800170027*

Egré, D. & Milewski, J. C. (2002), 'The diversity of hydropower projects', *Energy Policy* **30**, 1225–1230.
URL: *https://doi.org/10.1016/S0301-4215(02)00083-6*

Elhassouny, A. (2023), 'Neutrosophic logic-based diana clustering algorithm', *Neutrosophic Sets and Systems* **55**, 30.

Essayad, A. & Abdella, K. M. (2024), 'Predicting baccalaureate student result to prevent failure: a hybrid model approach', *IAES International Journal of Artificial Intelligence*

(IJ-AI) .
URL: *https://api.semanticscholar.org/CorpusID:266588105*

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I. & Akinyelu, A. A. (2022), 'A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects', *Engineering Applications of Artificial Intelligence* **110**, 104743.
URL: *https://doi.org/10.1016/j.engappai.2022.104743*

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996), 'From data mining to knowledge discovery in databases', *AI Magazine* **17**, 37.
URL: *https://doi.org/10.1609/aimag.v17i3.1230*

Feliciana, C., José, A. L. F. & Machado (2020), *Methodology Used for Determination of Critical Success Factors in Adopting the New General Data Protection Regulation in Higher Education Institutions*, Springer International Publishing, pp. 71–109.
URL: *https://doi.org/10.1007/978-3-030-40896-1_4*

Genc, M. S. & Ozden, K. S. (2021), 'Flow physics analysis of a vertical axis wind turbine using floefd', *7th Iran Wind Energy Conference, IWEC 2021* .
URL: *https://doi.org/10.1109/IWEC52400.2021.9467011*

Ghosal, A., Nandy, A., Das, A. K., Goswami, S. & Panday, M. (2020), A short review on different clustering techniques and their applications, *in* J. K. Mandal & D. Bhattacharya, eds, 'Emerging Technology in Modelling and Graphics', Springer Singapore, Singapore, pp. 69–83.
URL: *https://doi.org/10.1007/978-981-13-7403-6_9*

Gu, B., Shen, H., Lei, X., Hu, H. & Liu, X. (2021), 'Forecasting and uncertainty analysis of day-ahead photovoltaic power using a novel forecasting method', *Applied Energy* **299**, 117291.
URL: *https://doi.org/10.1016/j.apenergy.2021.117291*

Gul, S., Bano, S. & Shah, T. (2021), 'Exploring data mining: facets and emerging trends', *Digital Library Perspectives* **37**, 429–448.
URL: *https://doi.org/10.1108/DLP-08-2020-0078/FULL/PDF*

Gunda, T., Hackett, S., Kraus, L., Downs, C., Jones, R., McNalley, C., Bolen, M. & Walker, A. (2020), 'A machine learning evaluation of maintenance records for common failure modes in pv inverters', *IEEE Access* **8**, 211610–211620.
URL: *https://doi.org/10.1109/ACCESS.2020.3039182*

Haegel, N. M. & Kurtz, S. R. (2023), 'Global progress toward renewable electricity: Tracking the role of solar (version 3)', *IEEE Journal of Photovoltaics* **13**, 768–776.
URL: *https://doi.org/10.1109/JPHOTOV.2021.3104149*

Halkidi, M. (2018), *Hierarchical Clustering*, Springer New York, pp. 1684–1689.
URL: *https://doi.org/10.1007/978-1-4614-8265-9_604*

Han, J., Pei, J. & Tong, H. (2022), *Data mining: concepts and techniques*, 4 edn, Morgan kaufmann.

Hanafiah, A. M., Barakbah, A. R., Karlita, T. & Muliawati, T. H. (2021), Data analytics for medical record data of covid-19 patient with descriptive & predictive mining, *in* '2021 International Electronics Symposium (IES)', pp. 304–311.
URL: *https://doi.org/10.1109/IES53407.2021.9593982*

Harrouz, A., Belatrache, D., Boulal, K., Colak, I. & Kayisli, K. (2020), 'Social acceptance of renewable energy dedicated to electric production', *9th International Conference on Renewable Energy Research and Applications, ICRERA 2020* pp. 283–288.
URL: *https://doi.org/10.1109/ICRERA49962.2020.9242904*

Hatziargyriou, N., Donnelly, M., Papathanassiou, S., Lopes, J. A. P., Takasaki, M., Chao, H., Usaola, J., Lasseter, R., Efthymiadis, A., Karoui, K. & Arabi, S. (2000), 'Cigre tf38.01.10 modeling new forms of generation and storage'.
URL: *https://fglongatt.org/OLD/Archivos/Archivos/SistGD/CIGRE-TF-380110.pdf*

He, X., Cai, D., Shao, Y., Bao, H. & Han, J. (2011), 'Laplacian regularized gaussian mixture model for data clustering', *IEEE Transactions on Knowledge and Data Engineering* **23**(9), 1406–1418.
URL: *https://doi.org/10.1109/TKDE.2010.259*

Holechek, J. L., Geli, H. M. E., Sawalhah, M. N. & Valdez, R. (2022), 'A global assessment: Can renewable energy replace fossil fuels by 2050?', *Sustainability* **14**.
URL: *https://doi.org/10.3390/su14084792*

Holloway, R., Ho, D., Delotavo, C., Xie, W. Y., Rahimi, I., Nikoo, M. R. & Gandomi, A. H. (2023), 'Optimal location selection for a distributed hybrid renewable energy system in rural western australia: A data mining approach', *Energy Strategy Reviews* **50**, 101205.
URL: *https://doi.org/10.1016/j.esr.2023.101205*

IRENA (2023), 'Renewable capacity statistics 2023'.
URL: *https://rb.gy/1ks3z2*

Irzavika, N. & Supangkat, S. H. (2018), Descriptive analytics using visualization for local government income in indonesia, *in* '2018 International Conference on ICT for Smart Society (ICISS)', pp. 1–4.
URL: *https://doi.org/10.1109/ICTSS.2018.8550006*

Janse van Vuuren, C. Y. & Vermeulen, H. J. (2019), Clustered wind resource domains for the south african renewable energy development zones, *in* '2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)', pp. 616–623.
URL: *https://doi.org/10.1109/RoboMech.2019.8704832*

Javed, A., Lee, B. S. & Rizzo, D. M. (2020), 'A benchmark study on time series clustering', *Machine Learning with Applications* **1**, 100001.
URL: *https://doi.org/10.1016/j.mlwa.2020.100001*

Jinpeng, W., Yang, Z., Xin, G., Jeremy-Gillbanks & Xin, Z. (2022), 'A hybrid predicting model for the daily photovoltaic output based on fuzzy clustering of meteorological data and joint algorithm of gaps and rbf neural network', *IEEE Access* **10**, 30005–30017.
URL: *https://doi.org/10.1109/ACCESS.2022.3159655*

Jolliffe, I. T. & Cadima, J. (2016), 'Principal component analysis: a review and recent developments', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202.
URL: *https://doi.org/10.1098/rsta.2015.0202*

Kachouie, N. N. & Shutaywi, M. (2020), 'Weighted mutual information for aggregated kernel clustering', *Entropy* **22**, 351.
URL: *https://doi.org/10.3390/e22030351*

Kaur, B., Gupta, A. & Singla, R. K. (2023), Descriptive statistical analysis and discretization of academic data for machine learning techniques, *in* '2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)', pp. 1494–1499.

Li, K., Griffin, M. A., Barker, T., Prickett, Z., Hodkiewicz, M. R., Kozman, J. & Chirgwin, P. (2023), 'Embedding data science innovations in organizations: a new workflow approach', *Data-Centric Engineering* **4**, e26.
URL: *https://doi.org/10.1017/dce.2023.22*

Lima, P. G. D. S., Maciel, A. M. A., Resnick, N. E., Neto, A. T. & Leite, D. (2023), 'Machine learning models to identify anomalies in the production of flat glass', *Revista de Engenharia e Pesquisa Aplicada* .
URL: *https://api.semanticscholar.org/CorpusID:266893557*

Ma, T., Li, F., Ma, J., Wang, Y., Ma, H. & Li, Y. (2023), Simulation model and data acquisition method for wind turbine group, *in* '2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)', pp. 753–758.
URL: *https://doi.org/10.1109/CIPAE60493.2023.00146*

Mabuggwe, D. J. & Morsi, W. G. (2020), Representative profiling of prosumers with local distributed energy resources and electric vehicles using unsupervised machine learning, *in* '2020 IEEE Electric Power and Energy Conference (EPEC)', pp. 1–7.
URL: *https://doi.org/10.1109/EPEC48502.2020.9320051*

Mehrjoo, M., Jozani, M. J., Pawlak, M. & Bagen, B. (2021), 'A multilevel modeling approach towards wind farm aggregated power curve', *IEEE Transactions on Sustainable Energy* **12**, 2230–2237.
URL: *https://doi.org/10.1109/TSTE.2021.3087018*

Mohamed, S. A. & Sattar, M. A. E. (2019), 'A comparative study of p&o and inc maximum power point tracking techniques for grid-connected pv systems', *SN Applied Sciences* **1**, 174.
URL: *https://doi.org/10.1007/s42452-018-0134-4*

Mohammed, N. N. & Abdulazeez, A. M. (2017), Evaluation of partitioning around medoids algorithm with various distances on microarray data, *in* '2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)', pp. 1011–1016.
URL: *https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.155*

Munshi, A. A. (2020), 'Clustering of wind power patterns based on partitional and swarm algorithms', *IEEE Access* **8**, 111913–111930.
URL: *https://doi.org/10.1109/ACCESS.2020.3001437*

NbClust function - RDocumentation (n.d.).
URL: *https://www.rdocumentation.org/packages/NbClust/versions/3.0.1/topics/NbClust*

Nwadiugwu, M. C. (2020), 'Gene-based clustering algorithms: Comparison between denclue, fuzzy-c, and birch', *Bioinformatics and Biology Insights* **14**, 1177932220909851. PMID: 32284672.
URL: *https://doi.org/10.1177/1177932220909851*

Olabi, A. G. & Abdelkareem, M. A. (2022), 'Renewable energy and climate change', *Renewable and Sustainable Energy Reviews* **158**, 112111.
URL: *https://doi.org/10.1016/j.rser.2022.112111*

Oliveira, L. A. (2011), *Dissertação e Tese em Ciências e Tecnologia segundo Bolonha*, Lidel.

Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghiren, E., Olaniyan, D. & Olawole, O. (2019), Data clustering: Algorithms and its applications, *in* '2019 19th International Conference on Computational Science and Its Applications (ICCSA)', pp. 71–81.
URL: *https://doi.org/10.1109/ICCSA.2019.000-1*

Parmar, R., Tripathi, A. K., Kumar, S., Banerjee, C., Yadav, K. & Kumar, M. (2019), 'Solar photovoltaic power converters: Technologies and their testing protocols for indian inevitabilities', pp. 1–6.
URL: *https://doi.org/10.1109/ICPECA47973.2019.8975463*

Patel, M. R. (1999), *Wind and Solar Power Systems*, CRC Press LLC.
URL: *https://library.uniteddiversity.coop/Energy/Wind/Wind_and_Solar_Power_Systems.pdf*

Penick, T. & Louk, B. (1998), 'Photovoltaic power generation', *Final report presented to Gale Greenleaf on December* **4**.
URL: *https://api.semanticscholar.org/CorpusID:18300234*

Prajwal, S. & Hegde, V. (2022), 'Data acquisition systems for monitoring real time parameters of rooftop solar panels', *MysuruCon 2022 - 2022 IEEE 2nd Mysore Sub Section*

International Conference .
URL: *https://doi.org/10.1109/MYSURUCON55714.2022.9972631*

Qu, M. F. & Ma, D. B. (2020), 'Research on integrated data acquisition method of wind power generation based on deep learning', *2020 IEEE International Conference on Industrial Application of Artificial Intelligence, IAAI 2020* pp. 481–485.
URL: *https://doi.org/10.1109/IAAI51705.2020.9332835*

Rahman, M. M., Khan, I. & Alameh, K. (2021), 'Potential measurement techniques for photovoltaic module failure diagnosis: A review', *Renewable and Sustainable Energy Reviews* **151**, 111532.
URL: *https://doi.org/10.1016/j.rser.2021.111532*

Ram, M., Aghahosseini, A. & Breyer, C. (2020), 'Job creation during the global energy transition towards 100% renewable power system by 2050', *Technological Forecasting and Social Change* **151**, 119682.
URL: *https://doi.org/10.1016/j.techfore.2019.06.008*

Ramalingam, M. & Ilakkiya, R. (2021), 'Data mining algorithms(knn & dt) based predictive analysis on selected candidates in academic performance', *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* pp. 332–337.
URL: *https://doi.org/10.1109/Confluence51648.2021.9377203*

Rendón, E., Abundez, I., Arizmendi, A. & Quiroz, E. M. (2011), 'Internal versus external cluster validation indexes', *International Journal of computers and communications* **5**, 27–34.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., da F. Costa, L. & Rodrigues, F. A. (2019), 'Clustering algorithms: A comparative approach', *PLOS ONE* **14**, e0210236.
URL: *https://doi.org/10.1371/journal.pone.0210236*

Sabitha, A. S. & Punhani, R. (2019), Identification of potential regions for wind power development using data mining, *in* '2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)', pp. 306–313.
URL: *https://doi.org/10.1109/COMITCon.2019.8862192*

Saidi, K. & Omri, A. (2020), 'The impact of renewable energy on carbon emissions and economic growth in 15 major renewable energy-consuming countries', *Environmental Research* **186**, 109567.
URL: *https://doi.org/10.1016/j.envres.2020.109567*

Saleh, U. A., Jumaat, S. A., Johar, M. A. & Jamaludin, W. A. (2021), 'Photovoltaic-thermoelectric generator monitoring system using arduino based data acquisition system technique', *3rd IEEE International Conference on Artificial Intelligence in Engineering and Technology, IICAIET 2021* .
URL: *https://doi.org/10.1109/IICAIET51634.2021.9573815*

Salem, I. E., Mijwil, M., Abdulqader, A. W., Ismaeel, M. M., Alkhazraji, A. & Alaabdin, A. M. Z. (2022), 'Introduction to the data mining techniques in cybersecurity', *Mesopotamian Journal of CyberSecurity* **2022**, 28–37.
URL: *https://doi.org/10.58496/MJCS/2022/004*

Saltz, J. S. (2021), Crisp-dm for data science: Strengths, weaknesses and potential next steps, *in* '2021 IEEE International Conference on Big Data (Big Data)', pp. 2337–2344.
URL: *https://doi.org/10.1109/BigData52589.2021.9671634*

Sammaknejad, N., Zhao, Y. & Huang, B. (2019), 'A review of the expectation maximization algorithm in data-driven process identification', *Journal of Process Control* **73**, 123–136.
URL: *https://doi.org/10.1016/j.jprocont.2018.12.010*

Santoso, P. H., Santosa, H. S., Istiyono, E., Haryanto, H. & Retnawati, H. (2023), 'Predicting physics students' achievement using in-class assessment data: A comparison of two machine learning models', *Physics Education Research Journal* .
URL: *https://api.semanticscholar.org/CorpusID:266382393*

Schröer, C., Kruse, F. & Gómez, J. M. (2021), 'A systematic literature review on applying crisp-dm process model', *Procedia Computer Science* **181**, 526–534. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.
URL: *https://doi.org/10.1016/j.procs.2021.01.199*

Schütz, T., Schraven, M. H., Fuchs, M., Remmen, P. & Müller, D. (2018), 'Comparison of clustering algorithms for the selection of typical demand days for energy system synthesis', *Renewable Energy* **129**, 570–582.
URL: *https://doi.org/10.1016/j.renene.2018.06.028*

Septiana, Y., Agustin, Y. H., Mudzakir, M. N., Mulyani, A., Fatimah, D. D. S. & Julianto, I. T. (2023), 'Implementation of classification algorithm c4.5 in determining the emergency patient in the maternity hospital queue system', *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)* pp. 790–794.
URL: *https://api.semanticscholar.org/CorpusID:258869702*

Shafique, U. & Qaiser, H. (2014), 'A comparative study of data mining process models (kdd, crisp-dm and semma)', *International Journal of Innovation and Scientific Research* **12**, 217–222.

Shahbaz, M., Raghutla, C., Chittedi, K. R., Jiao, Z. & Vo, X. V. (2020), 'The effect of renewable energy consumption on economic growth: Evidence from the renewable energy country attractive index', *Energy* **207**, 118162.
URL: *https://doi.org/10.1016/j.energy.2020.118162*

Shahria, M. N., Anik, M. R. M., Shufian, A., Kabir, S., Islam, M. A. & Kumar, S. U. (2023), 'Solar pv panel automatic shading analysis using boost regulator and inverter system',

2023 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2023 - Proceedings pp. 285–289.
URL: *https://doi.org/10.1109/ICICT4SD59951.2023.10303418*

Shen, X., Fu, X. & Zhou, C. (2019), 'A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm', *IEEE Transactions on Sustainable Energy* **10**, 46–54.
URL: *https://doi.org/10.1109/TSTE.2018.2822682*

Singh, G. K. (2013), 'Solar power generation by pv (photovoltaic) technology: A review', *Energy* **53**, 1–13.
URL: *https://doi.org/10.1016/j.energy.2013.02.057*

sklearn.cluster.AgglomerativeClustering — scikit-learn 1.4.2 documentation (n.d.).
URL: *https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html*

Sousa, M. J. & Baptista, C. S. (2014), *Como Fazer Investigação, Dissertações, Teses e Relatórios: Segundo Bolonha*, 2011 edn, Pactor.

Sreedhar Kumar, S., Madheswaran, M., Vinutha, B., Manjunatha Singh, H. & Charan, K. (2019), 'A brief survey of unsupervised agglomerative hierarchical clustering schemes', *Int J Eng Technol* **8**(1), 29–37.
URL: *https://doi.org/10.14419/ijet.v8i1.13971*

Sørensen, B. (1991), 'A history of renewable energy technology', *Energy Policy* **19**, 8–12.
URL: *https://doi.org/10.1016/0301-4215(91)90072-V*

Tougui, I., Jilbab, A. & Mhamdi, J. E. (2020), 'Heart disease classification using data mining tools and machine learning techniques', *Health and Technology* **10**, 1137–1144.
URL: *https://doi.org/10.1007/s12553-020-00438-1*

Usama, M., Qadir, J., Raza, A., Arif, H., lim Alvin Yau, K., Elkhatib, Y., Hussain, A. & Al-Fuqaha, A. (2019), 'Unsupervised machine learning for networking: Techniques, applications and research challenges', *IEEE Access* **7**, 65579–65615.
URL: *https://doi.org/10.1109/ACCESS.2019.2916648*

Uti, M. N., Din, A. H. M., Yusof, N. & Yaakob, O. (2023), 'A spatial-temporal clustering for low ocean renewable energy resources using k-means clustering', *Renewable Energy* **219**, 119549.
URL: *https://doi.org/10.1016/j.renene.2023.119549*

Vankov, D., Zorin, I. & Pozo, D. (2020), Clustering time series over electrical networks, *in* '2020 International Conference on Smart Energy Systems and Technologies (SEST)', pp. 1–6.
URL: *https://doi.org/10.1109/SEST48500.2020.9203491*

Vargas, S. A., Esteves, G. R. T., Maçaira, P. M., Bastos, B. Q., Oliveira, F. L. C. & Souza, R. C. (2019), 'Wind power generation: A review and a research agenda', *Journal of Cleaner Production* **218**, 850–870.
URL: *https://doi.org/10.1016/J.JCLEPRO.2019.02.015*

Walliman, N. (2011), *Research Methods: The Basics*, Routledge.
URL: *https://doi.org/10.4324/9781003141693*

Wang, X. & Xu, Y. (2019), 'An improved index for clustering validation based on silhouette index and calinski-harabasz index', *IOP Conference Series: Materials Science and Engineering* **569**, 52024.
URL: *https://dx.doi.org/10.1088/1757-899X/569/5/052024*

Wang, Y., Zhang, M., Ren, X., Meng, X., Yu, J., Wang, E., Wang, J. & Ge, Y. (2023), Comparison of artificial intelligence-based power curve cleaning algorithms for wind farms, *in* '2023 7th International Conference on Power and Energy Engineering (ICPEE)', pp. 329–336.
URL: *https://doi.org/10.1109/ICPEE60001.2023.10453807*

Wirth, R. & Hipp, J. (2000), 'Crisp-dm: Towards a standard process model for data mining', *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* .

Xie, H., Zhou, Q., Zhong, S., Xi, Z. & Wang, H. (2023), 'Data cleaning and modeling of wind power curves', *2023 10th International Conference on Power and Energy Systems Engineering (CPESE)* pp. 188–193.
URL: *https://doi.org/10.1109/CPESE59653.2023.10303189*

Zatti, M., Gabba, M., Freschini, M., Rossi, M., Gambarotta, A., Morini, M. & Martelli, E. (2019), 'k-milp: A novel clustering approach to select typical and extreme days for multi-energy systems design optimization', *Energy* **181**, 1051–1063.
URL: *https://doi.org/10.1016/j.energy.2019.05.044*

Zhang, T., Ramakrishnan, R. & Livny, M. (1997), 'Birch: A new data clustering algorithm and its applications', *Data Mining and Knowledge Discovery* **1**, 141–182.
URL: *https://doi.org/10.1023/A:1009783824328*

Zhu, H., Lu, L., Yao, J., Dai, S. & Hu, Y. (2018), 'Fault diagnosis approach for photovoltaic arrays based on unsupervised sample clustering and probabilistic neural network model', *Solar Energy* **176**, 395–405.
URL: *https://doi.org/10.1016/j.solener.2018.10.054*

APPENDIX A

	Features					
	wind_ _speed	**wind_ _direction**	**power_ _average**	**rotation_ _average**	**exterior_ _temperature**	**nacelle_ _temperature**
count	8367759	8367759	8367759	8367759	8367759	8367759
mean	6.15e	100.89	574.39	11.01	12.43	21.41
std	3.27	302.08	657.03	4.68	7.09	8.012
min	-8.97	-1070	-18.18	0	-44.99	-37.89
25%	3.80	-109	67	7.79	7.00	16
50%	5.70	115	289	11.18	11.02	21
75%	8.00	314	880	14.87	17	27
max	8.00	314	880	14.87	17	124

Table A.1 – Estatistical Description of Wind Data

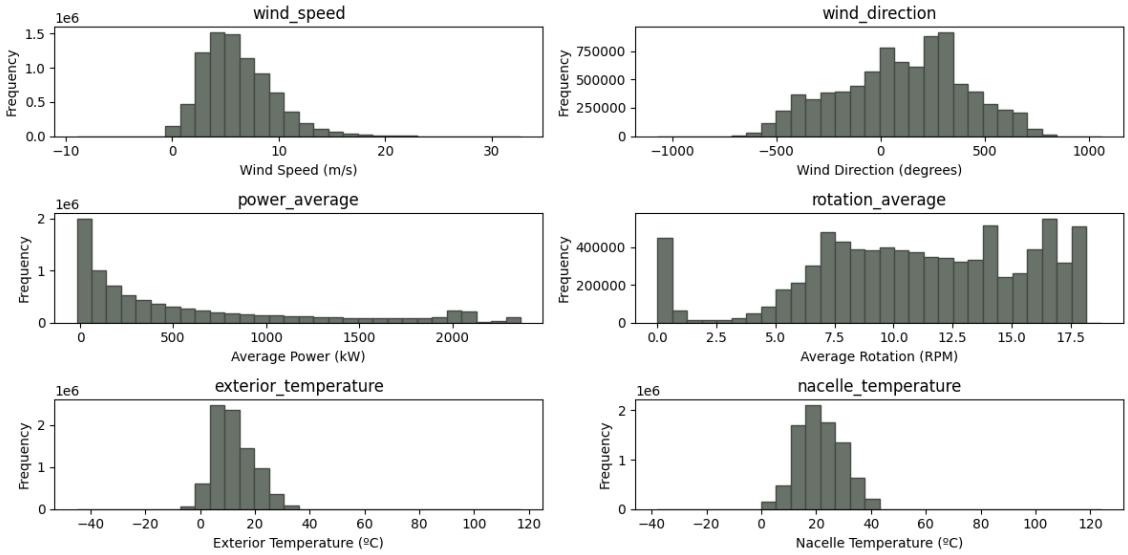


Figure A.1 – Histograms for Wind Data

	Features					
	ac_power	**ac_voltage**	**ac_current**	**dc_power**	**dc_voltage**	**dc_current**
count	4081774	3982052	3982052	4081774	4081774	4081774
mean	8250.52	223.40	11.76	8343.59	468.99	7.56
std	16363.67	50.01	22.97	16762.18	104.93	8.30
min	0.00	0.00	0.00	-855.92	0.00	-0.06
25%	3273.00	234.00	4.68	3173.52	497.87	4.35
50%	3273.00	234.00	4.68	3173.52	497.87	4.35
75%	8505.00	238.00	12.04	8544.12	524.77	12.30
max	110000.00	253.63	158.93	115313.00	1878.70	39.52

Table A.2 – Estatistical Description of Solar Data

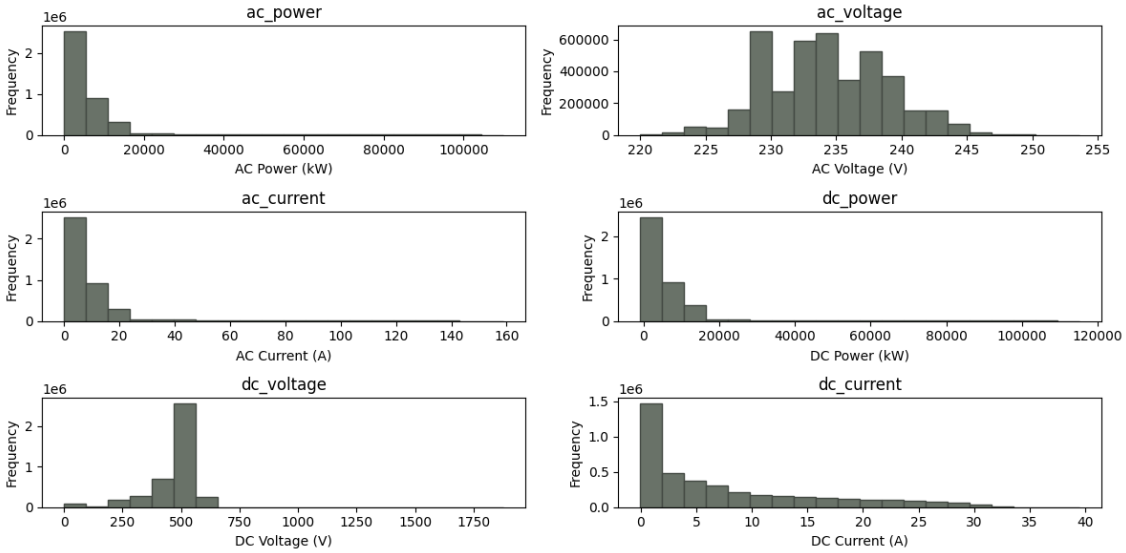


Figure A.2 – Histograms for Solar Data

	Features			
	global_tilted_ _irradiance	**global_horizontal_ _irradiance**	**temperature**	**cloud_cover**
count	476012.00	476012.00	476012.00	476012.00
mean	209.56	183.42	15.58	24.08
std	312.79	28.41	5.67	28.41
min	0.00	0.00	-2.00	0.00
25%	0.00	0.00	12.00	0.00
50%	0.00	0.00	15.00	10.00
75%	348.00	311.00	19.00	45.00
max	1113.00	1022.00	43.00	97.00

Table A.3 – Estatistical Description of Satellite Data

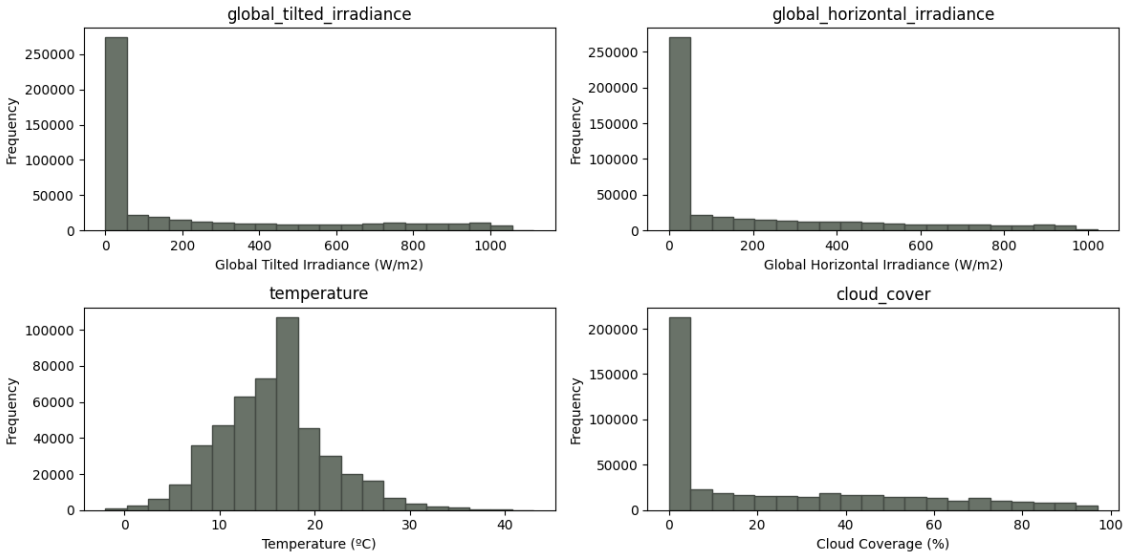


Figure A.3 – Histograms for Satellite Data

APPENDIX B

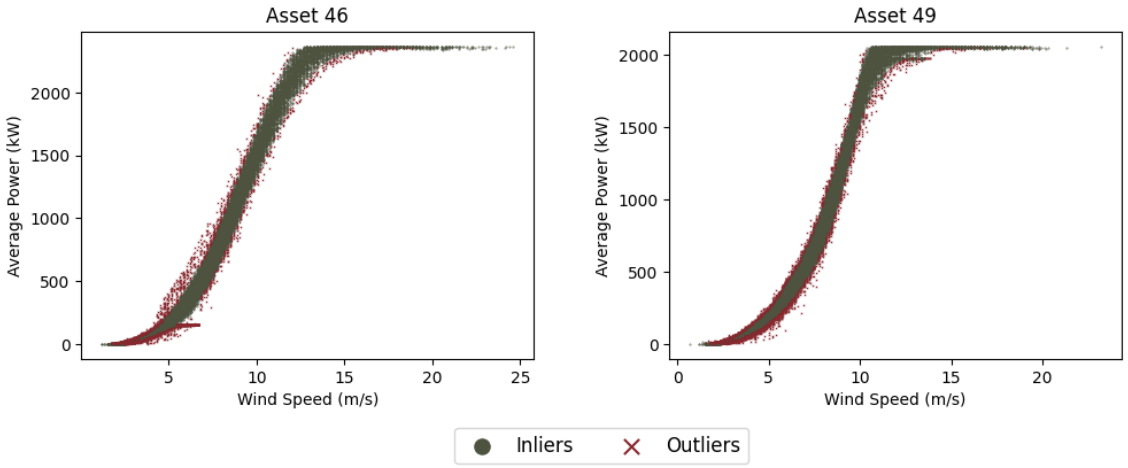


Figure B.1 – MSM results for Wind data

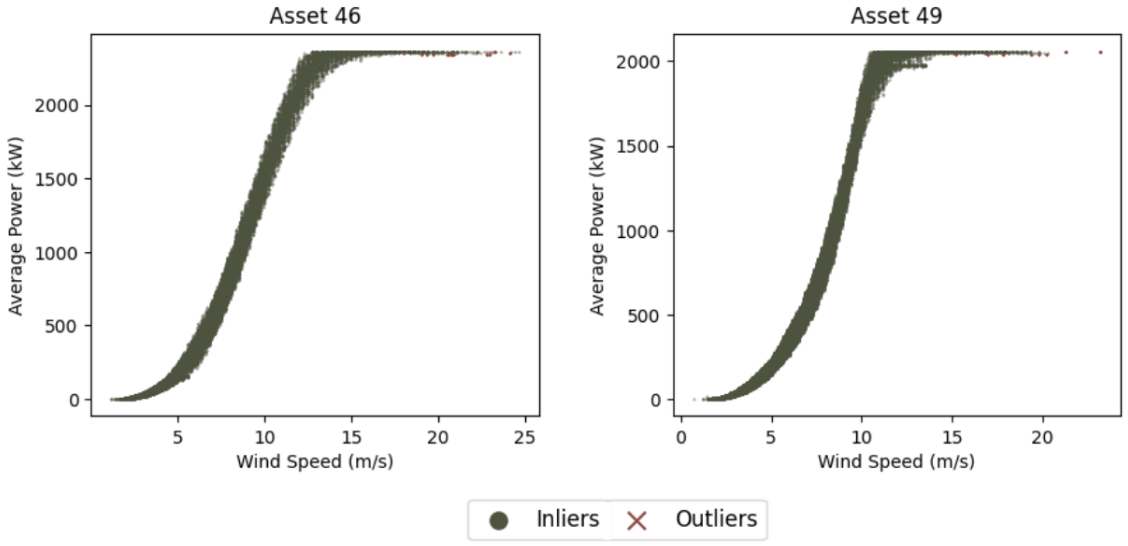


Figure B.2 – DBSCAN results for Wind data

APPENDIX C

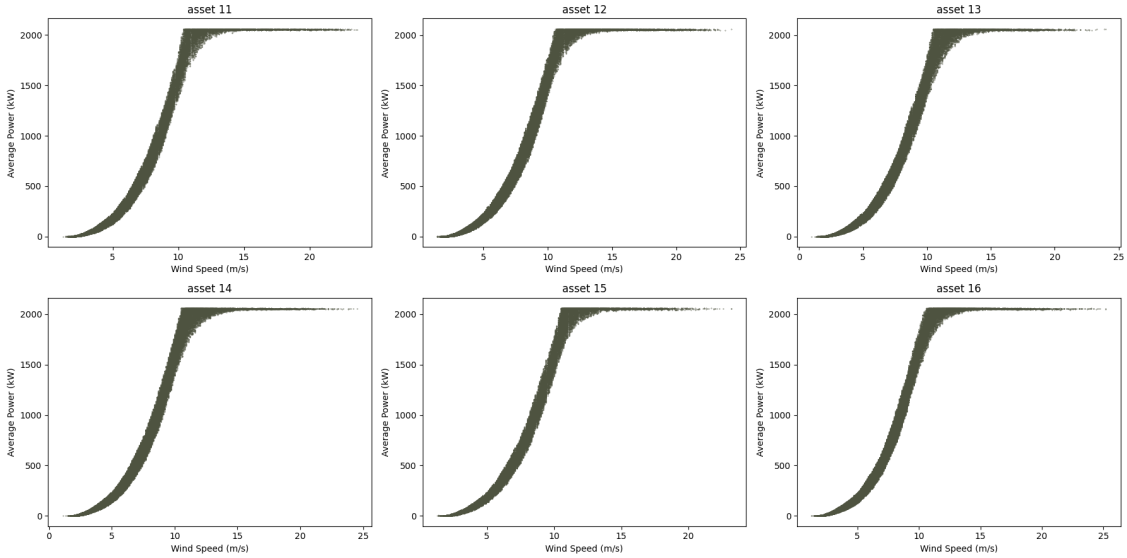


Figure C.1 – Power Curve Cleaning Results for assets 11, 12, 13, 14, 15, 16

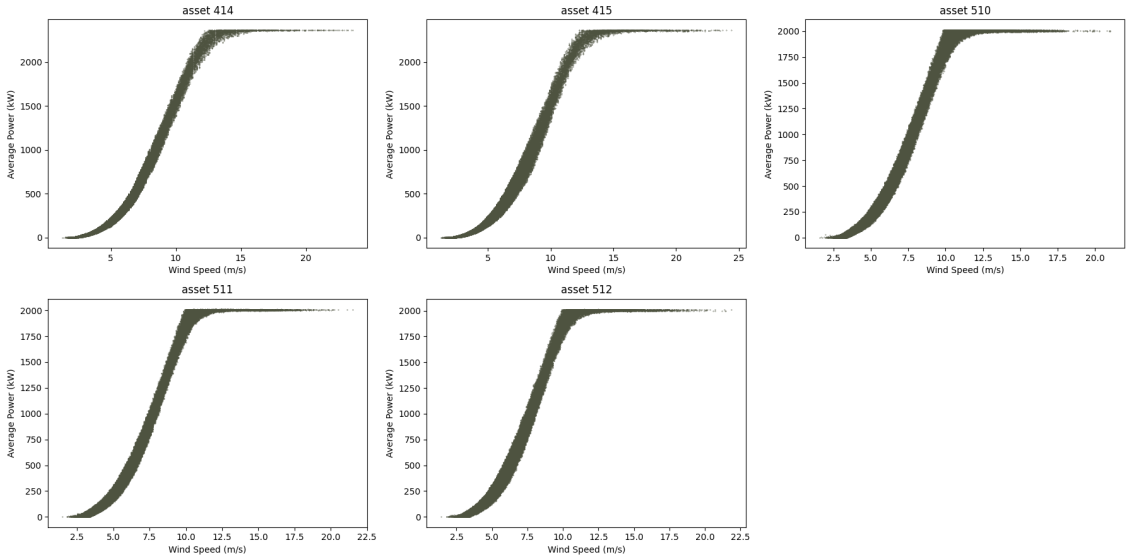


Figure C.2 – Power Curve Cleaning Results for assets 414, 415, 510, 511, 512

APPENDIX D

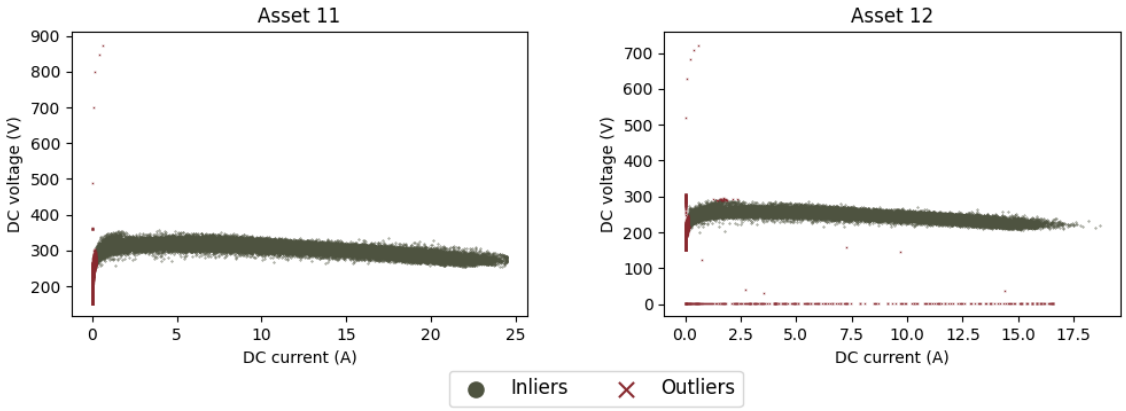


Figure D.1 – I-V Curves after first clean

APPENDIX E

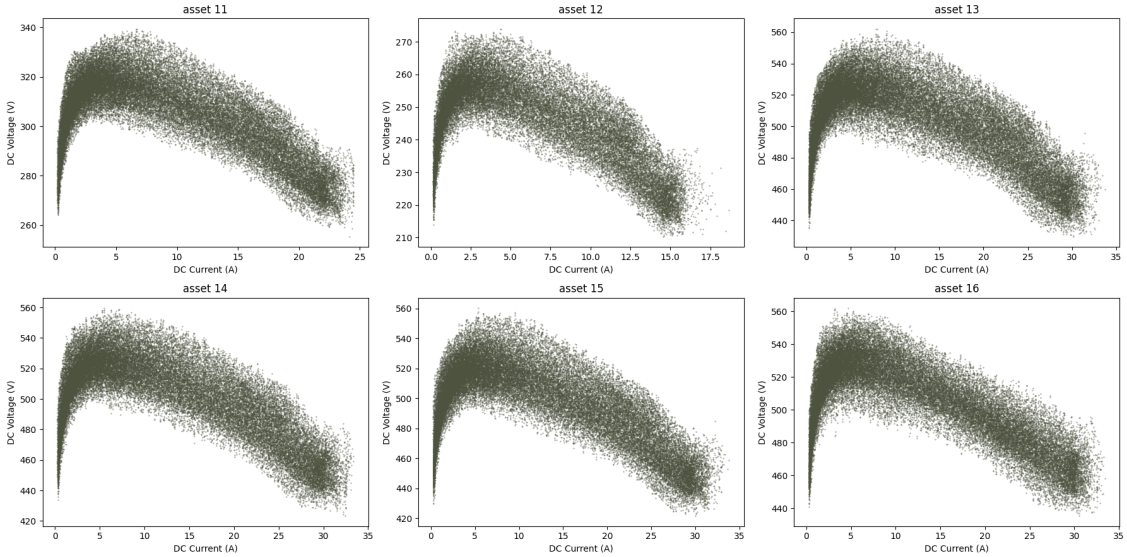


Figure E.1 – I-V Curve Cleaning Results for assets 11, 12, 13, 14, 15, 16

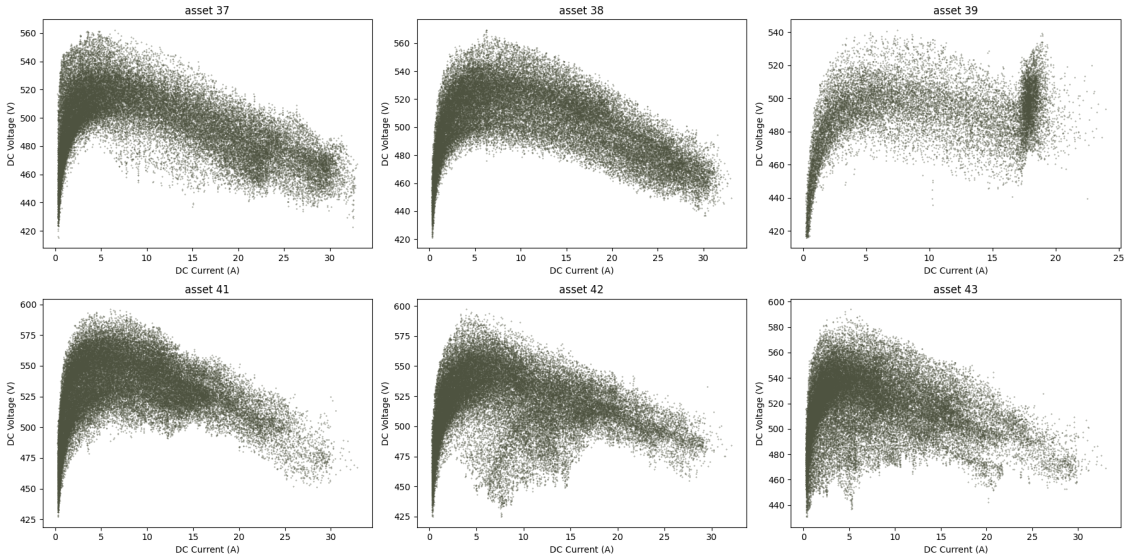
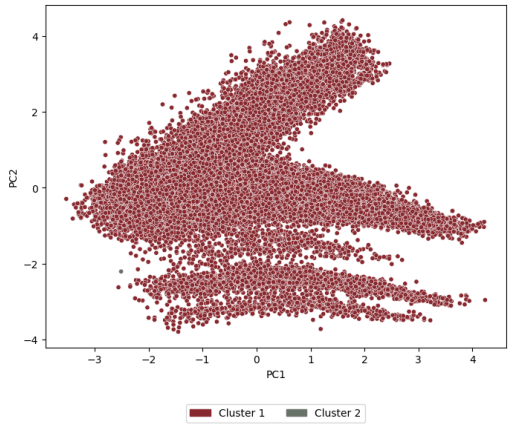


Figure E.2 – I-V Curve Cleaning Results for assets 37, 38, 39, 41, 42, 43

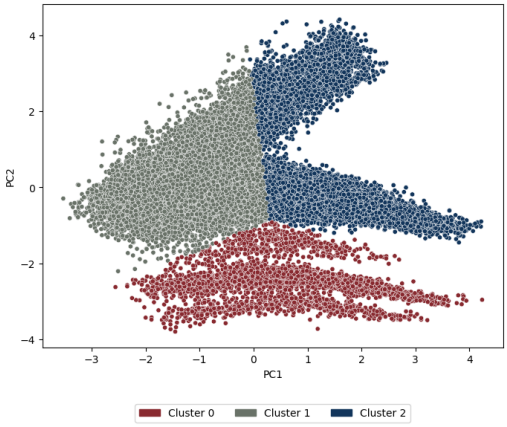
APPENDIX F



(a) Clusters of Single Link

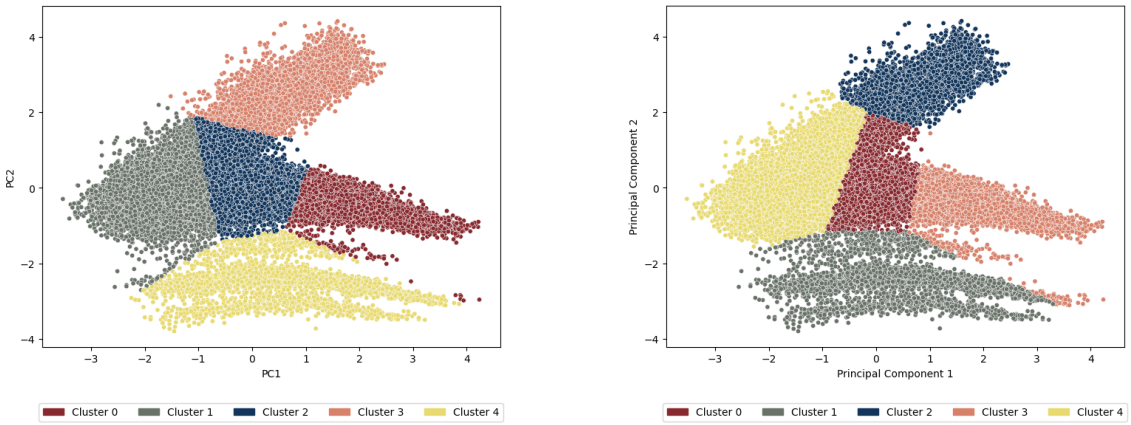
(b) Clusters of Ward's Link

(c) Clusters of Centroid Link



(d) Clusters of BIRCH

Figure F.1 – Cluster Scatter Plots of Hierarchical algorithms with Clean Solar Data



(a) Clusters of K-Means (b) Clusters of CLARA

Figure F.2 – Cluster Scatter Plots of Partitional algorithms with Clean Solar Data

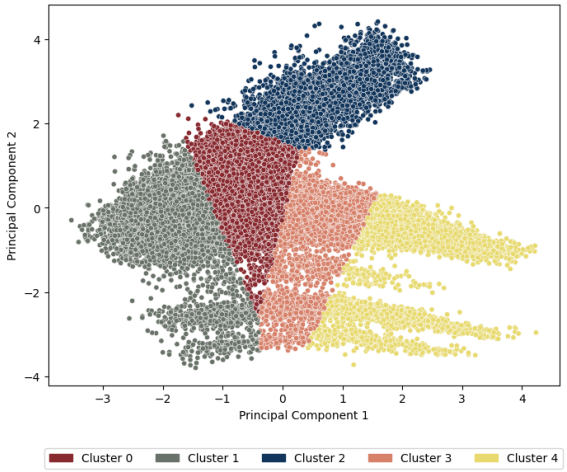


Figure F.3 – Cluster Scatter Plot of Fuzzy C-Means (Soft Clustering) with Clean Solar Data

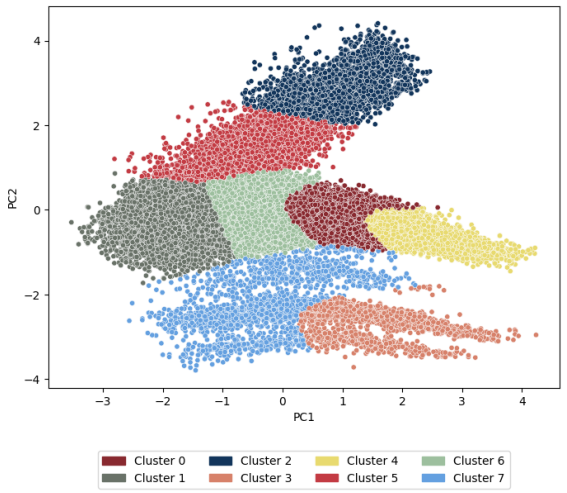
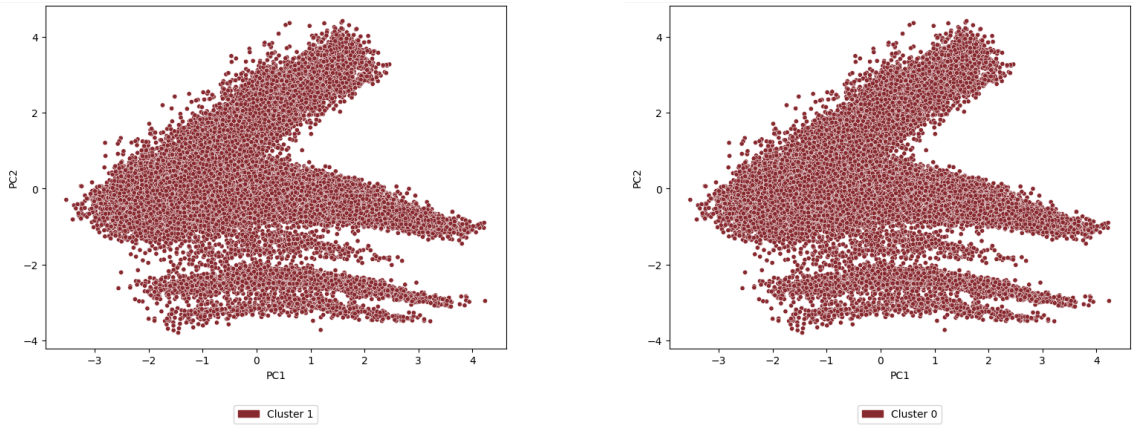


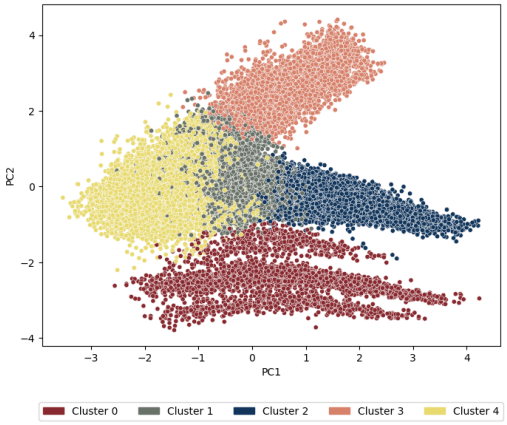
Figure F.4 – Cluster Scatter Plot of Gaussian Mixture Model (Model-Based Clustering) with Clean Solar Data



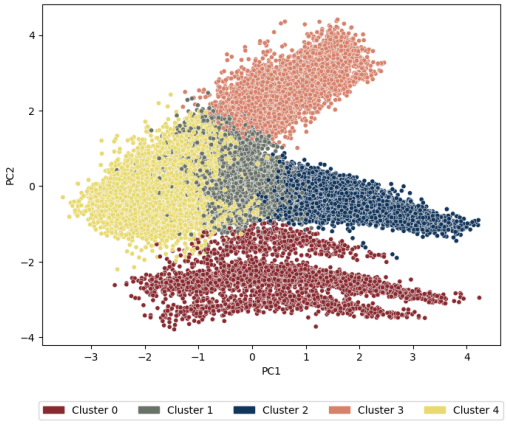
(a) Clusters of DBSCAN (b) Clusters of OPTICS

Figure F.5 – Cluster Scatter Plots of Density-Based algorithms with Clean Solar Data

APPENDIX G

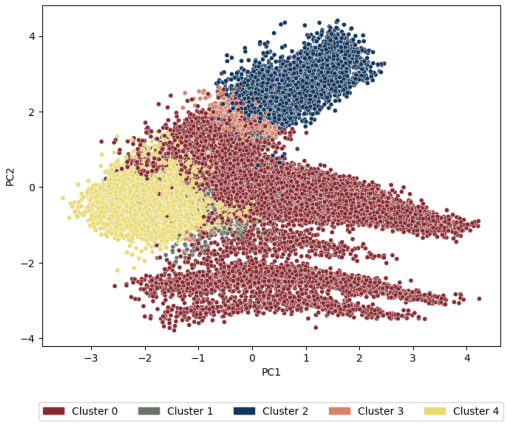


(a) Clusters of K-Means and CLARA ensembled



(b) Clusters of K-Means and Ward's Link ensembled

(c) Clusters of CLARA and Ward's Link ensembled



(d) Clusters of K-Means, CLARA, and Ward's Link ensembled

Figure G.1 – Cluster Scatter Plots of Ensemble Clustering with Clean Solar Data

APPENDIX H

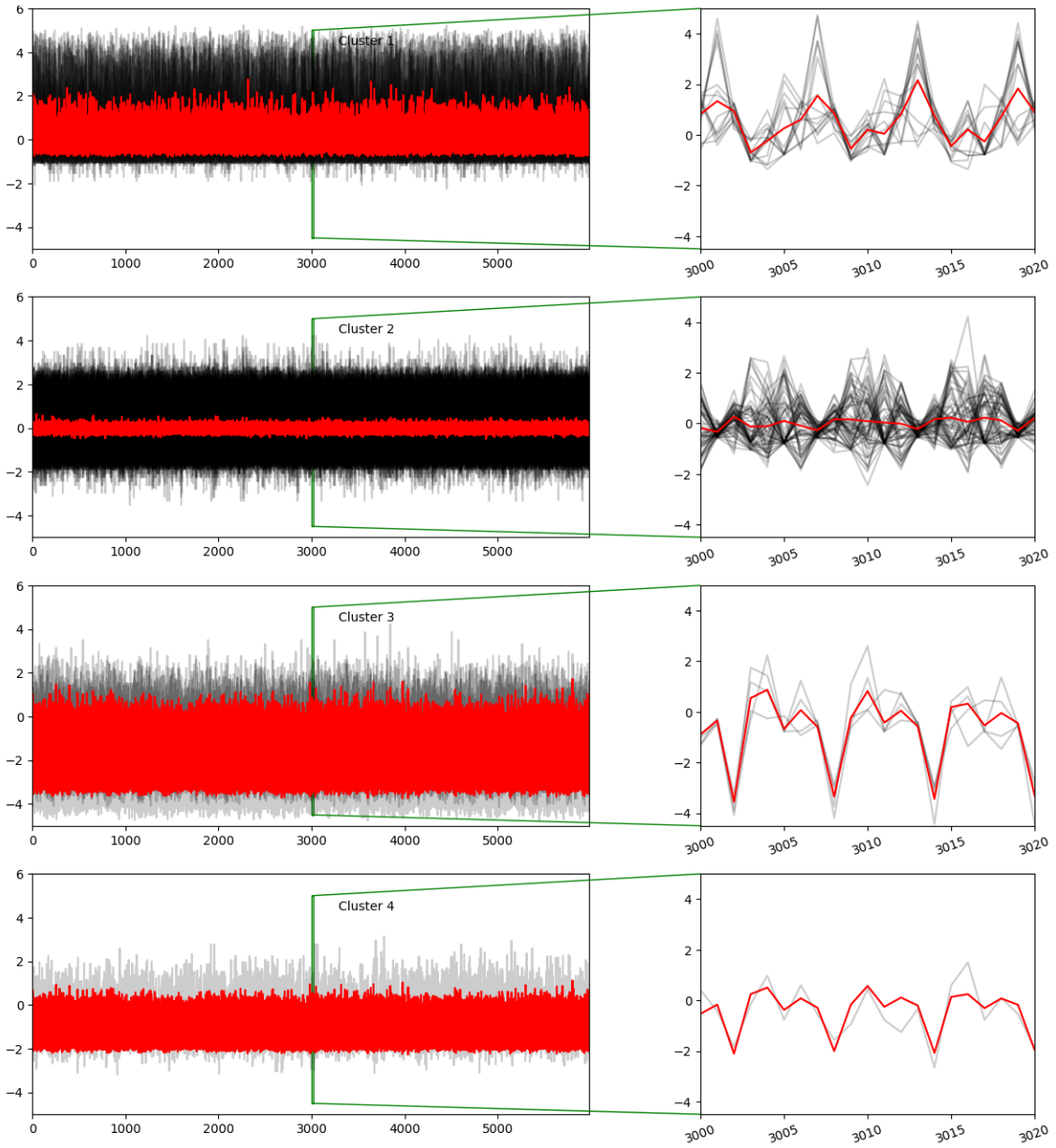


Figure H.1 – Clusters of SOM with Clean Solar Data

APPENDIX I

Algorithms	Distance metric	Indexes			
		SS	DB	CH	Dunn
cloud_cover					
K-Means	Euclidean	0.4582	0.9107	71.5153	0.0369
	DTW	0.6741	0.5078	116.5587	0.4552
SOM	Euclidean	0.5159	0.5443	129.3921	0.0740
dc_current					
K-Means	Euclidean	0.5276	0.7041	146.1390	0.0429
	DTW	0.6927	0.3347	203.9580	0.1133
SOM	Euclidean	0.6462	0.4498	171.9371	0.1133
dc_power					
K-Means	Euclidean	0.8856	0.1175	419.9386	0.4842
	DTW	0.8856	0.1175	419.9386	0.4842
SOM	Euclidean	0.7751	0.2929	271.3257	0.3565
dc_voltage					
K-Means	Euclidean	0.6554	0.4124	161.1631	0.2221
	DTW	0.6767	0.3737	163.3957	0.2221
SOM	Euclidean	0.6527	0.4258	165.6105	0.2698
temperature					
K-Means	Euclidean	0.5760	0.5337	121.0774	0.0631
	DTW	0.5760	0.5337	121.0774	0.0631
SOM	Euclidean	0.5673	0.3569	73.7426	0.0527

Table I.1 – Time Series Clustering Results for Clean Solar Data with individual variables

Algorithms	Distance metric	Indexes			
		SS	DB	CH	Dunn
power_average					
K-Means	DTW	0.5031	0.6288	54.8177	0.0418
SOM	Euclidean	0.5528	0.4289	73.7969	0.0860
wind_direction					
K-Means	DTW	0.4579	0.9037	44.8531	0.0200
SOM	Euclidean	0.4719	0.5558	70.2833	0.1695
exterior_temperature					
K-Means	Euclidean	0.4970	0.6928	112.6338	0.1442
	DTW	0.5175	0.6472	114.6432	0.1930
SOM	Euclidean	0.4591	0.6887	98.2891	0.1433

Table I.2 – Time Series Clustering Results for Non-Clean Wind Data with individual variables

APPENDIX J

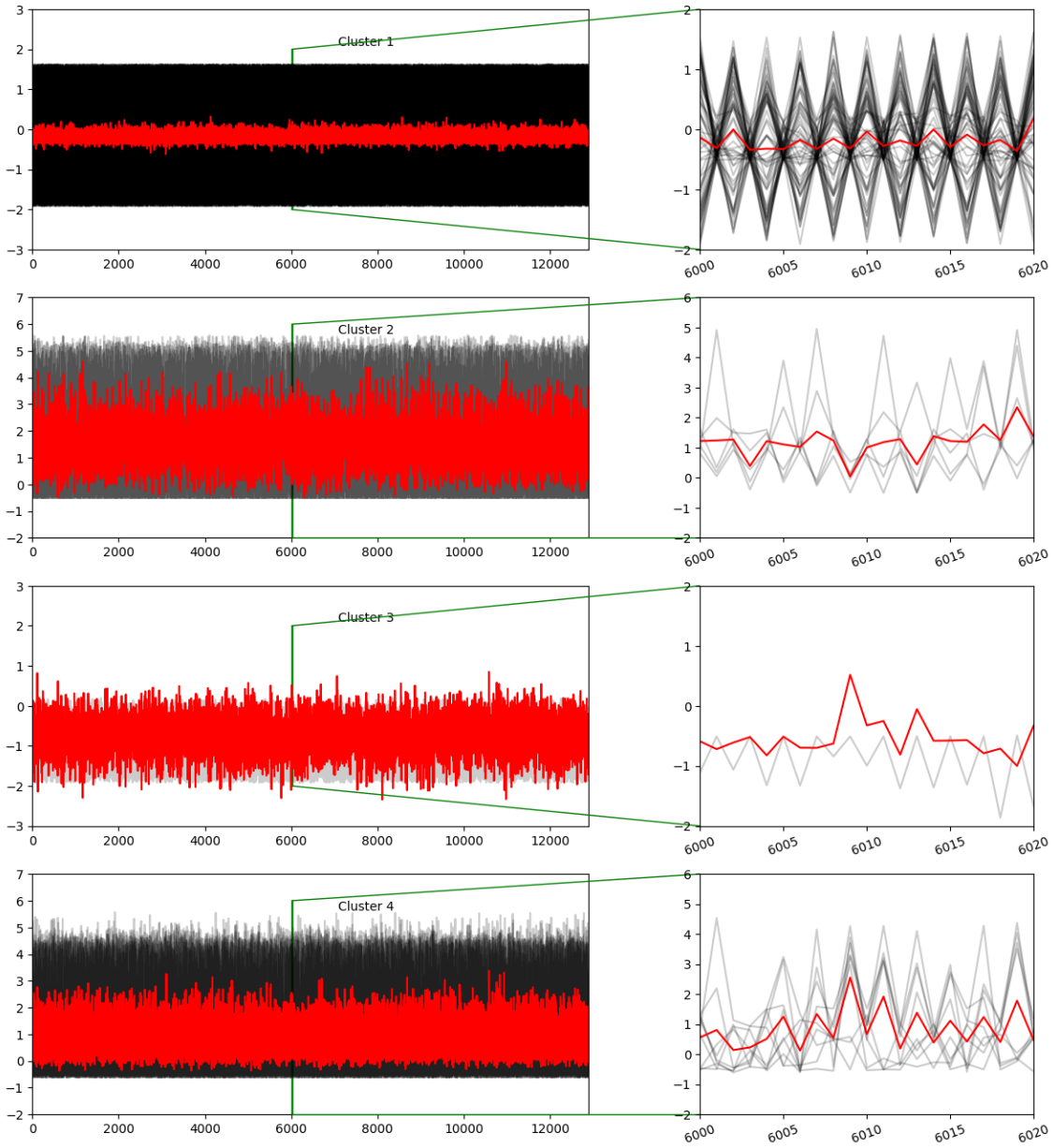


Figure J.1 – Clusters of SOM with Non-Clean Solar Data (Time Series with *dc_power*)

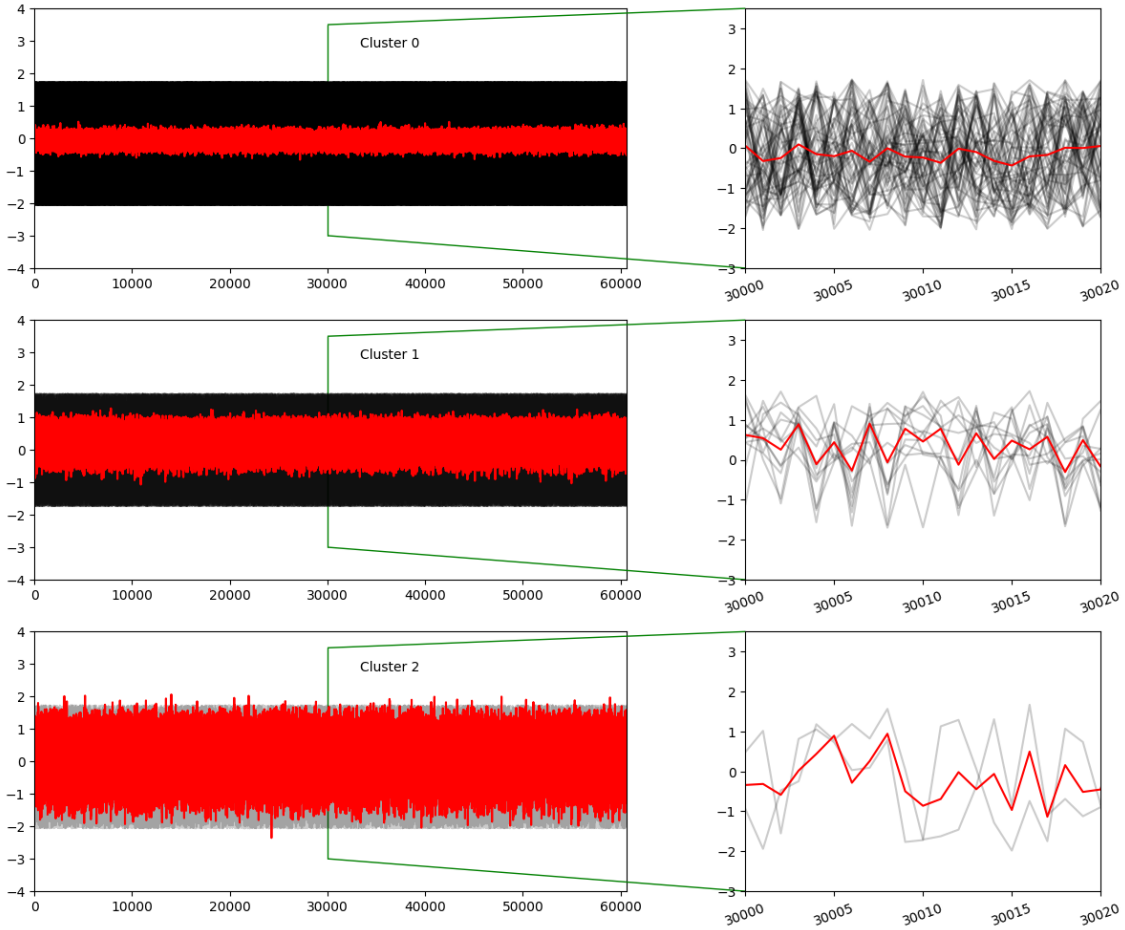


Figure J.2 – Clusters of K-Means with Euclidean distance metric, with Clean Wind Data (Time Series with *wind_direction*)

ANNEX A

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

P.PORTO

DECLARAÇÃO DE INTEGRIDADE

DECLARAÇÃO DE INTEGRIDADE

Declaro ter conduzido este trabalho académico com integridade. Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Declaro que o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

Sara Isabel Gonçalves Abreu

ISEP, Porto, 17 de junho de 2024