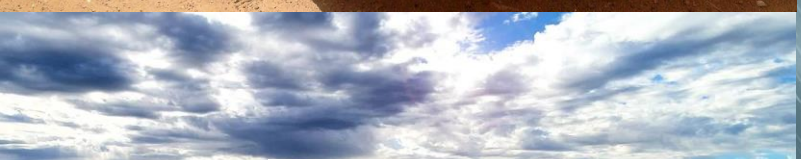
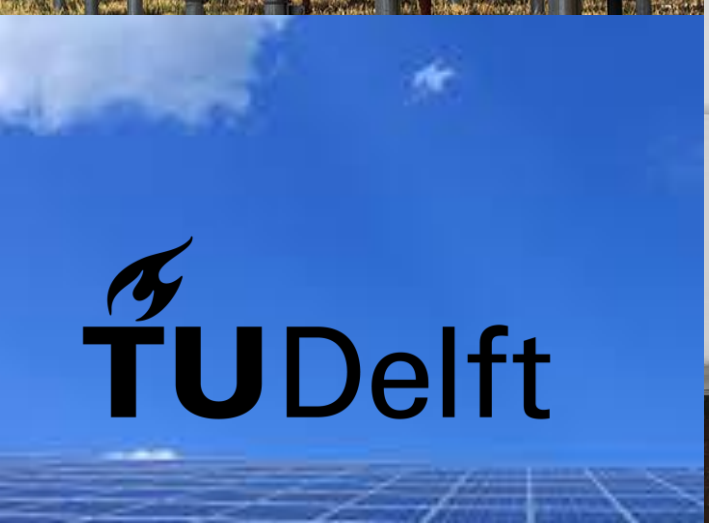



Machine Learning Driven PV-Climate Classification

F.J. Triana de las Heras






TU Delft

Machine Learning Driven PV-Climate Classification

Francisco Javier Triana de las Heras
Student Number: 5586976

Supervised by **Dr. Malte R. Vogt**

A thesis presented for the degree of
Master of Science in Sustainable Energy Technology

Photovoltaic Materials and Devices
EEMCS faculty
Delft University of Technology
Netherlands
13th of July 2023

Preface

"Machine Learning Driven PV-Climate Classification" is my MSc thesis as a student of Sustainable Energy Technology (SET). The project was carried out within the Photovoltaic Materials and Devices group (PVMD) at the Delft University of Technology. This report presents the motivation, background, methodology, reasoning, and conclusions of the research.

In short, a worldwide climate classification to guide Photovoltaic (PV) studies has been developed. For this purpose, Machine Learning techniques have been implemented. Even though being familiar with PV is advisable, this work does not deal with the complex theoretical aspects, and a general understanding and common sense are enough to understand the analysis. In contrast, a basic knowledge of Machine Learning seems essential to me. Readers uncomfortable with this point are referred to sections 1.4 and 2.2.

This topic appeared to me unexpectedly in a busy week where I had to change my thesis project in a few hours. In turn, the project has been an unexpectedly rewarding experience. I conclude the project very pleased, after having expanded my knowledge and enjoyed a great working atmosphere. I thank my supervisor Dr. M.R. Vogt for the opportunity, as well as his advice, dedication, and professionalism. I am also very thankful to A. Alcañiz Moya, whose recommendations facilitated my work and helped me achieve a more satisfactory result. Lastly, special greetings and thanks to the committee members, Dr. R.A.C.M.M. van Swaaij and Dr. P.P. Vergara Barrios, for their time and effort.

Delft, 25 June 2023
Francisco Javier Triana de las Heras

Abstract

Technological advances, cost reduction, depletion of fossil fuels, environmental concerns, and growing energy demand are expanding photovoltaic solar energy (PV) in more latitudes and locations. A simple and effective procedure to assess the PV potential of a particular region is to analyse its climatic conditions. In general, climatic studies use the Köppen-Geiger (KG) climate classification as a reference. However, KG is solely based on temperature and precipitation, resulting in an unsatisfactory scheme for analyses in the PV field, since the most important variable, solar irradiation, is not considered. Thus, in 2019 Ascencio-Vásquez et al. developed a new worldwide classification based on temperature, precipitation, and solar irradiation: the Köppen-Geiger-Photovoltaic (KGPV) climate classification. Even though KGPV is a good improvement, it just consists of a simplified version of the KG groups subdivided into four levels of irradiation: low, medium, high, and very high. Hence, the climate parameters are not considered in a combined manner in the sorting process.

In this project, a new worldwide climate classification directly applicable to PV has been developed. Machine Learning proved to be a convenient tool to achieve this objective. First, supervised learning served to identify and assess the climate variables more correlated to the specific energy yield. More specifically, a Linear Regression model was implemented. Subsequently, these variables were used to create the classification by applying k-means, a clustering algorithm. The classification was optimised following a comprehensive qualitative analysis, resulting in a scheme based on seven climate variables and 20 clusters. By contrast, KGPV considers five variables. Even though it contemplates 24 groups at first, half of them are neglected based on a land-surface ratio and population density criterion, resulting in a classification based on 12 clusters. Hence, the methodology proposed in this work enables identifying new relevant regions. Moreover, "Machine Learning driven PV-climate classification" presents a satisfactory correlation with the specific energy yield, except for very low values, where the correlation is minor.

Lastly, the relationship between climate and degradation rate was explored. The complexity and non-linear behaviour of degradation demand an alternative approach. Random Forests was proposed, but it showed poor performance. It is necessary to be able to predict non-linearities and, at the same time, keep a logical mathematical relation between the supervised and clustering algorithms. In this regard, Multivariate Adaptive Regression Spline (MARS) might be a promising option.

Contents

Introduction

Photovoltaic solar energy (PV) is strongly affected by climatic conditions. This fundamental relationship between climate and PV performance can be used to make accurate predictions on energy yield, reliability, or service lifetime to facilitate financial analysis and decision-making processes. For climate studies, the Köppen-Geiger (KG) climate classification has been widely used [30]. However, KG is solely based on temperature and precipitation, so its applicability to the PV field is limited. For this reason, in 2019, Ascencio-Vásquez et al. developed the Köppen-Geiger-Photovoltaic (KGPV) climate classification, based on temperature, precipitation, and solar irradiation [6]. Even though this was a significant enhancement, KGPV has plenty of room for improvement.

The objective of this report is to develop a new PV-climate classification. The focus will be on finding a relationship between climate and the specific energy yield. This task entails two questions. First, which climate variables must the classification consider? And secondly, how to classify them? To address these challenges, a method based on Machine Learning is proposed. More specifically, Linear Regression (supervised learning) is used to select the climate variables, while k-means (unsupervised learning) is applied to create the classification. Furthermore, the report includes an exploratory analysis of the method's suitability for studying degradation, which provides further insights.

The report is structured as follows. Chapter 1 provides background on the motivation, KG and KGPV climate classifications, and Machine Learning. It concludes with a detailed description of the methodology. The data collection and algorithms used in the project are explained in Chapter 2. Chapter 3 is devoted to selecting the climate variables. Then, the classification is created and evaluated in Chapter 4. The analysis of the degradation is illustrated in Chapter 5. Finally, Chapter 6 lays out the key takeaways and conclusions.

Chapter 1

Foundations

This first chapter covers the vision and motivation of the project, as well as the tools and methodology implemented. It comprises five sections. First, in "PV and climate", the motivation, and background of the project are presented. The relationship between climate and PV performance, and the necessity of a climate classification are introduced. Furthermore, previous works are reviewed. In particular, two schemes stand out: the KG and KGPV climate classifications. These are explained in detail in sections two and three, respectively. KG, based only on temperature and precipitation variables, is unsuitable for PV studies. KGPV includes the global horizontal irradiation to achieve a significant improvement. However, it might oversimplify the classification.

Section four, "Machine Learning", presents the tools used in this work to develop a new classification, and provides the background required to understand the methodology. Supervised and unsupervised learning are explained. Finally, the remainder of the chapter is devoted to presenting this methodology and introducing the rest of the chapters.

1.1 PV and climate

Global solar photovoltaic (PV) energy capacity surpassed the magic number of 1 TW in 2022 [2]. PV growth during the last decade has been impressive, being the current capacity more than 21 times greater than in 2010. This has been accompanied by an 88 percent reduction in the global weighted average Levelized Cost of Energy (LCOE) of utility-scale photovoltaics [3]. Solar PV has shown the highest learning rates of all renewable energy technologies, becoming the lowest-cost option for new electricity generation in most of the world. Even though it already contributes to more than 3.6 percent of global electricity generation [22], worldwide deployment of PV seems to have just begun. Technological advances, cost reduction, depletion of fossil fuels, environmental concerns, and growing energy demand are expanding PV in more latitudes and locations. Certainly, PV is becoming a worldwide energy source.

With this global trend, several questions arise. How does solar PV perform in different locations? What issues should be expected? Which are the most suitable technologies? The accurate prediction and evaluation of energy yield, reliability, or service lifetime are essential for any financial analysis and decision-making process. Furthermore, it can help to identify specific engineering improvement points and fields of research, optimising PV applications and boosting its implementation [33].

PV performance depends on numerous variables. As Ascencio-Vásquez et al. [6] indicate, it is possible to divide them into solar irradiation, weather (temperature, wind, humidity), local conditions (shading, albedo), and technical specifications of the technology and PV module itself (efficiency, temperature coefficient). Clearly, climate plays a fundamental role in PV performance. Therefore, a simple and effective procedure for assessing PV potential for a particular location is to analyse its climatic conditions. Broad research has been made in this direction. Dash et al. [13] divided India into 6 climatic zones and studied their most efficient PV technologies. Skandalos et al. [42] analysed the effect of local climatic conditions on photovoltaic building integration for some global locations, concluding that the optimised design depends on the climate zone. Karin et al. [26] developed a climate classification to identify which types of degradation may be expected in different geographic areas in the USA. Micheli et al. [35] studied the impact of performance losses due to soiling. The International Energy Agency (IEA) [1] presented guidance for customized O&M service in seven different climate zones.

In general, climatic studies, including those for PV applications, have been based on the Köppen-Geiger (KG) climate classification ([23], [35], [41]). However, as many of these works pointed out ([23], [26], [42]) the KG classification is solely based on temperature and precipitation, resulting in an unsatisfactory scheme for conducting a comprehensive analysis of PV performance. Therefore, several authors aimed to supplement KG including other relevant parameters. For instance, Skandalos et al. [42] extended it for some particular locations by including the annual global horizontal irradiation, while Karin et al. [26] focused on PV module degradation stressors. Consequently, the original KG classification has been modified differently by several authors, resulting in a variety of classifications without any standardisation or global acceptance, possessing confusion and difficulties to compare results [6].

In 2019, motivated by these challenges, Ascencio-Vásquez et al. [6] went a step further, devel-

oping a new worldwide classification based on temperature, precipitation, and solar irradiation. It is called the Köppen-Geiger-Photovoltaic (KGPV) climate classification, and the present project is inspired by and based on it. Consequently, a further discussion of the KGPV climate classification is pertinent. However, before entering into this extension of KG, it seems appropriate to examine first the KG climate classification itself.

1.2 KG climate classification

The Köppen Geiger (KG) climate classification was first formulated by Wladimir Köppen in 1900. It was updated and presented as the last version by Rudolf Geiger in 1961 [30]. Since then, it has been the most widely used reference for research in a wide variety of topics related to climate, hydrology, geography, and agriculture, and it has been commonly taught in schools and other educational institutions. Despite its longevity, the classification has not been changed significantly. Several authors have proposed new classifications, but these have not received enough acceptance to replace KG [38]. This is justified partially by its simplicity and ability to reproduce successfully the present and near-future climate, and partially by its popularity. However, different positions are found in the literature, and some authors argue that the current use of KG is mainly founded on historical inertia [38].

Köppen based his classification on his knowledge of plant sciences. The type of vegetation found in an environment is directly related to its climate. Therefore, Köppen developed a climate classification based on five vegetation groups determined previously by the botanist Alphonse De Candolle from the climate zones of the ancient Greeks [38]. These five vegetation groups consisted of plants from the equatorial zone (A), the arid zone (B), the warm temperate zone (C), the snow zone (D), and the polar zone (E). Two more letters were added to complete the classification: a second one for considering precipitation, and a third for air temperature [30]. Thus, a climate zone such as Afa means equatorial zone (A), fully humid (f), with hot summers (a).

Recently, numerous KG climate classification world maps have been redrawn from modern gridded data [38]. The most comprehensive is the one by Kottek et. al [30]. Based on temperature data provided by the Climate Research Unit (CRU) of the University of East Anglia [18] and precipitation data provided by the Global Precipitation Climatology Centre (GPCC) [40], Kottek developed a digital map with 31 climate types at a resolution of 0.5° latitude by 0.5° longitude for the period 1951-2000. While the exact criteria followed by the classification can be found in the original paper, it is convenient to summarise in Table 1.1 the parameters considered. For the sake of clarity, temperatures are measured in °C, and precipitations in mm/month except for P_{ann} and P_{th} which are in mm/year and mm, respectively. The resulting Köppen-Geiger climate classification map is depicted in Figure 1.1.

Table 1.1: Variables considered by the Köppen-Geiger climate classification.

Feature	Description
T_{ann}	Annual mean near-surface (2 m) temperature
T_{max}	Monthly mean temperature of the warmest month
T_{min}	Monthly mean temperature of the coldest month
T_{mon}	Monthly mean temperature
P_{ann}	Accumulated annual precipitation
P_{min}	Precipitation of the driest month
P_{th}	Dryness threshold
$P_{\text{smin}}/P_{\text{smax}}$	Lowest/Highest monthly precipitation value for the summer half-year
$P_{\text{wmin}}/P_{\text{wmax}}$	Lowest/Highest monthly precipitation value for the winter half-year

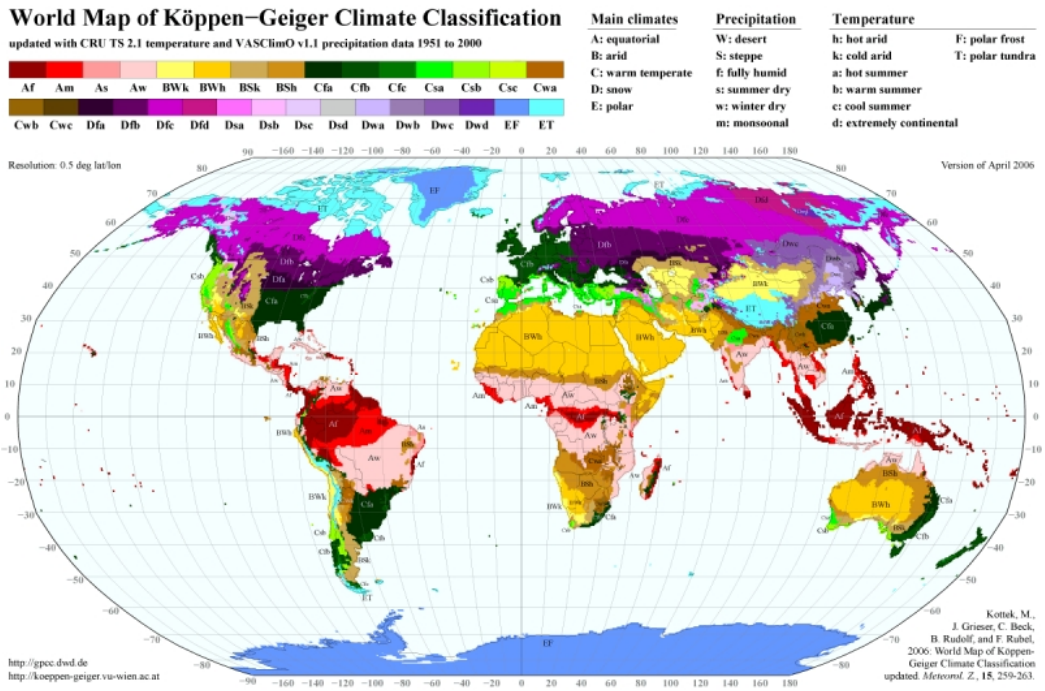


Figure 1.1: Köppen-Geiger climate classification [30].

1.3 KGPV climate classification

The Köppen-Geiger-Photovoltaic (KGPV) climate classification was proposed by Ascencio-Vásquez et al. [5] in 2018 to tackle the deficiencies that KG implies for photovoltaic studies. It was upgraded in a subsequent paper in 2019 [6]. This new classification complements KG by considering solar irradiation. It follows two criteria. First, based on the KG scheme, zones are classified in terms of temperature and precipitation, differentiating among Tropical (A), Desert (B), Steppe (C), Temperate (D), Cold (E), and Polar (F). Secondly, solar irradiation is considered to distinguish among Very High irradiation zones (K), High irradiation zones (H), Medium irradiation zones (M), and Low irradiation zones (L). Altogether, a zone could be, for instance, BK, indicating a desert climate with very high irradiation. Following this approach, it is possible to define 24 climate groups. However, half of them were neglected based on a land-surface ratio and population density criterion, resulting in a classification that divides the world into 12 zones.

Similar to Kottek, Ascencio-Vásquez et al. used the gridded data set (0.5° x 0.5°) provided by the University of East Anglia [18] and the GPCC [40]. More specifically, temperature data was taken from the dataset CRU TS4.01, whereas precipitation values were extracted from the GPCCv2018 dataset. On the other hand, the global horizontal irradiation (GHI) was taken from the reanalysis-based dataset ERA-Interim provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) [14]. Justifying that temperature evolution has presented a different behaviour after 1990, the data used for developing the classification corresponds to the period 1990 - 2016. The variables considered by the KGPV climate classification are summarised in the following table.

Table 1.2: Variables considered by the KGPV classification.

Feature	Description
T_{max}	Monthly mean temperature of the warmest month
T_{min}	Monthly mean temperature of the coldest month
P_{ann}	Accumulated annual precipitation
P_{th}	Dryness threshold
GHI_{ann}	Accumulated annual Global Horizontal Irradiation

Temperatures are recorded in °C, P_{ann} in mm/year, P_{th} in mm, and GHI_{ann} in kWh/m^2/year. Even though T_{ann} is not considered as a primary parameter, it is implicitly included in the determination of P_{th}. The Köppen-Geiger-Photovoltaic climate classification is illustrated in Figure 1.2.

To assess the classification, Ascencio-Vásquez et al. determined several PV performance indicators. Then, locations belonging to different PV-climate groups could be compared, pointing their differences out and gaining new insights. In the first paper [5], a correlation between the KGPV zones and the average expected annual specific energy yield (kWh/kW$_{\text{p}}$) in USA and Chile was analysed. Further, various PV technologies (mono c-Si, multi c-Si, CIGS, CdTe, and a-Si) were compared. It was concluded that a-Si at deserts with high irradiation (BH) results in the highest annual specific energy yield, whereas the lowest production takes place in temperate climates with

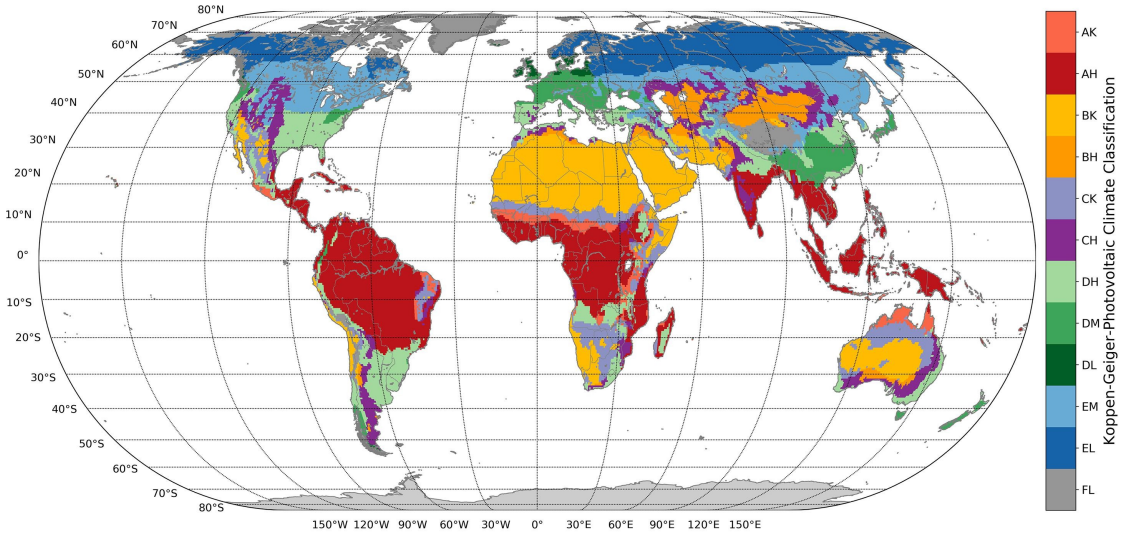


Figure 1.2: Köppen-Geiger-Photovoltaic climate classification [6].

medium irradiation (DM) using multi c-Si modules.

In the second work [6], besides the annual specific energy yield, other indicators such as the performance ratio (PR), the unit capacity factor (UCF), and the module operating temperature, were considered to assess the KGPV scheme worldwide. One location per KGPV zone was selected to analyse and compare the PV performance of a typical c-Si PV system. The most remarkable conclusions include that the best location for PV is the Atacama Desert in Chile, due to its high irradiation (high energy yield) and UCF, whereas PR is highest at Moscow, a cold climate with low irradiation. Furthermore, the evolution of PV performance over time and the impact of climate change were evaluated.

After the analysis, Ascencio-Vásquez et al. concluded that the results agree with expectations and that the KGPV climate classification is a convenient scheme to relate climate zones with PV performance. However, points of improvement were indicated too. The implementation of more climate variables such as wind speed, relative humidity, or ultra-violet irradiation was proposed. Lastly, it was noted that the quality of the model largely depends on the accuracy of the input data. In particular, irradiation data commonly poses a high uncertainty. To evaluate this critical point, Ascencio-Vásquez et al. compared the synthetic data calculated from ERA-Interim with ground measurements of 22 stations from the Baseline Surface Radiation Network (BSRN), obtaining a satisfying similarity.

1.4 Machine Learning

As seen above, the historically used KG climate classification turns out to be inconvenient for PV studies, since the relation between the climate variables considered and PV performance is limited. KGPV, by including the *GHI*, certainly made a significant improvement. However, this essential parameter was employed to distinguish, merely, among four basic levels of irradiation (low, medium, high, and very high). Moreover, the criteria based on temperature and precipitation are virtually equivalent to KG. Indeed, as Ascencio-Vásquez et al. concluded, additional parameters should be considered [6]. Therefore, even though KGPV stands out for its simplicity, the established relation between climate and PV performance might be suboptimal.

This work aims to develop a new PV-climate classification. However, in contrast to KG and KGPV, the objective is to derive this classification based on advanced statistical analysis techniques and state-of-the-art programming algorithms. In this way, an objective, precise, and comprehensive classification might be obtained, gaining new insights and enabling an assessment of KGPV.

In particular, to achieve these objectives, Machine Learning (ML) constitutes a fantastic tool. The term "Machine Learning" was introduced by the computer scientist Arthur Samuel, a professor at Stanford University and researcher in Bell Laboratories and IBM, among other positions [47]. IBM defines Machine Learning as "a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy" [21]. An alternative short description would be that ML gives computers the ability to learn without explicitly being programmed [9].

ML plays a key role in the increasingly important field of data science [21]. Sophisticated algorithms based on statistical methods are trained to make classifications or predictions, uncovering relevant information and extracting knowledge from the data. For this reason, ML is also known as predictive analytics or statistical learning [36]. These insights are commonly used to facilitate and support decision-making processes in business and engineering. With the continuous growth of big data, ML is expected to become more and more important in the next decades to identify the most relevant questions and the data to answer them [21].

ML finds applications in numerous fields: online advertising, face recognition, autopilot, space research, discovery of new particles, or medicine [36]. Likewise, it has been recently implemented in PV studies. A comprehensive review of the most recent and promising applications of ML in the field of PV systems was published by Tina et al. [44]. A trending application is Photovoltaic Power Forecast (PVPF) based on weather parameters. In this field, Essam et al. [16] conducted a thorough analysis of the most commonly used ML algorithms. They concluded that Artificial Neural Network (ANN) is the most suitable one. Khandakar et al. [28] applied ANN and several multiple regression models to study the impact on PVPF of relevant environmental variables in Qatar such as irradiance, relative humidity, ambient temperature, and wind speed. Furthermore, they used feature selection techniques to identify the most contributory variables. Alternatively, Ahmed et al. [4] proposed an ensemble-based Long Short-Term Memory (LSTM) algorithm reliant on adaptive weighting and data segmentation techniques for PVPF. In this case, global horizontal irradiation and weather temperature were selected as input features. Another widely studied application of ML is PV degradation analysis and prediction. For instance, Li et al. [31] performed

clustering based on environmental factors influencing field reliability to classify China into groups with similar degradation rates. Chantana et al. [10] conducted multiple regression analysis to determine a quantitative relation between environmental factors and the PV performance ratio for Si-based technologies. Liu et al. [32], after performing a regional clustering of mainland China based on relevant environmental factors for power generation, determined the optimum tilt angle under different cleaning cycles.

Broadly, Machine Learning models can be divided into three primary categories: supervised ML, unsupervised ML, and reinforcement ML [9]. Some authors might include a fourth category consisting of semi-supervised learning algorithms [21]. In this work, supervised and unsupervised learning algorithms are both implemented. More specifically, supervised learning is used for selecting the climate variables and calculating their weights, whereas unsupervised learning is conducted for developing the classification.

Supervised ML is, at the present time, the most successful kind of ML algorithm [9]. Its power resides in its ability to automate decision-making processes by generalizing from known examples [36]. Based on these known examples, called the labeled dataset, an algorithm is trained to classify data or predict outcomes accurately [21]. In particular, the algorithm is able to create an output for an input it has never seen before without any help from a human [36]. Popular supervised learning algorithms include nearest neighbors (KNN), linear regression, neural networks, ensemble methods, random forest, and support vector machine (SVM) ([21], [37]).

By contrast, in unsupervised ML an algorithm looks for patterns in unlabeled data [9]. Here, only the input data is known, and no known output data is given to the algorithm [36]. The objective is to extract knowledge from the data, discovering hidden patterns or data grouping without the need for human intervention. Unsupervised learning is commonly used for exploratory data analysis [21]. A popular unsupervised learning technique, and the one used in this work, is clustering. Clustering consists in partitioning data into distinct groups of similar items [36]. Examples of clustering algorithms are k-means, hierarchical clustering, and DBSCAN [37]. While there are many successful applications of these methods, they are usually harder to understand and evaluate [36], as it will be shown in Chapter 4.

There are numerous programming languages and computing platforms that enable performing data analysis and ML studies with relative ease. The most popular ones are Matlab, Python, and R. In this project, Python 3.10 was used. Even though making a comprehensive review and comparison between these platforms is out of the scope of this thesis, there are some points worth mentioning that justify this choice. From its creation in 1991 by Guido van Rossum, Python has emerged over the last couple of decades as a first-class tool for scientific computing tasks [45]. At present, Python is one of the most widely used languages for data analysis, and, especially, for modern Machine Learning, in industry and academia ([36], [43]). Python, in contrast with Matlab or R, is a general-purpose programming language. Therefore, an extensive collection of libraries is available, which provide tools to do a great variety of tasks, from data science to web development. This might constitute the main advantage of Python ([34], [45]). Moreover, in spite of this generality, it maintains a user-friendly language [36]. Another strong point in favour of Python is its large community [43]. Having said that, Matlab and R are great computing platforms too, and these outperform Python in some applications such as signal processing and causal inference research,

respectively [43]. In this regard, usually, the best option would be to use these platforms together [34]. Lastly, an special interest and motivation of the author to learn Python cannot be hidden.

As mentioned above, one of the strengths of Python is its third-party packages. Among the numerous packages used in this work, scikit-learn deserves to be highlighted [37]. Scikit-learn is the most prominent Python library for ML, containing numerous state-of-the-art ML algorithms, as well as a fantastic documentation. It is an open-source project, which has been widely used in industry and academia [36]. Other relevant libraries are Pandas, for data manipulation, NumPY, for arrays manipulation and calculations, and Matplotlib, for visualizations. To complete, the book "Introduction to Machine Learning with Python: A Guide for Data Scientists" by Andreas C. Müller and Sarah Guido [36] was used as the primary reference during the project.

1.5 Methodology

So far, it has been talked about finding a relationship between climate and PV performance. It is time to refine this a bit more. PV performance comprises several factors. In particular, the International Electrotechnical Commission (IEC) 61724 standard defines the following principal PV system performance indices: energy generated by PV systems (E_{ac}), reference yield (Y_r), final yield (Y_f), performance ratio (PR), capacity utilization factor (CUF), and PV system efficiency (η_{sys}) [8]. Among these parameters, Y_f is selected as the basis for the classification due to its relevance and comprehensibility. Furthermore, it enables easy comparison between systems with different capacities, as opposed to the E_{ac}. The final yield, or specific energy yield, is defined as the net daily, monthly, or annual electrical energy output of the PV plant divided by its rated power. It is given by the following expression [8]:

$$Y_f = \frac{Energy\ generated\ (E_{ac})}{Rated\ power\ of\ PV\ plant\ (P_{STC})} \tag{1.1}$$

It is measured in hours or, equivalently, kWh/kW. In this work, Y_f always refers to an annual basis.

On the other hand, in order to include an indicator of reliability and long-term performance, the degradation rate (k) might be utilised. The degradation rate is a measure of the power decline over time and is commonly expressed in relative percentage per year (percent/year) [24]. PV degradation is a complex topic. Indeed, climate-induced degradation is frequently poorly estimated, resulting in around eight percent inaccuracy in the levelized cost of electricity (LCOE) [26]. One of the main difficulties is that not only does degradation depend on several climate variables, but also on the particular interactions between them [31]. Some remarkable studies on degradation and climate are those conducted by Ascencio-Vásquez et al. [7], Jordan et al. [25], and Bansal et al. [8].

Since these are different PV performance indicators, two separate classifications might be distinguished: one for the specific energy yield, and one for the degradation rate. This work focuses on the former. A classification based on the specific energy yield is thoroughly developed and explained in the report, while degradation is simply explored in Chapter 5. Nevertheless, though short, the study of degradation yields relevant findings and should not be underestimated.

Thus, the main classification derived in this work is based on climate variables particularly relevant to the specific energy yield. Hence, the first question that must be answered is: Which are these climate variables? Moreover, to generate an objective classification via ML, not only is it essential to identify the climate variables, but also their levels of importance, or mathematically speaking, their weights. These two issues are solved with the help of supervised learning. A linear regression model is built to predict worldwide specific energy yields from the knowledge of several climate parameters. The predicted values can be compared with known data to obtain a measure of the error of the model. This provides a method to analyse the relevance of the climate variables on PV performance: the more relevant a parameter is, the lower the error of the prediction. Hence, the model can be optimised by selecting the most significant variables. Further, the algorithm provides the optimum weights. This is the subject of Chapter 3.

Once these key questions are successfully solved, the classification can be developed. Unsupervised learning, and more specifically, k-means, is implemented to find patterns and create clusters from the selected climate variables. A careful qualitative analysis is required to understand the results and present a final classification. This constitutes the most challenging task of the project. The clusters' centres and sizes, and several figures facilitate the exploration. Finally, a comparison with KGPV is made and conclusions are drawn. All this is treated in Chapter 4.

In principle, a similar method could be applied to degradation. However, in this case, assuming a linear dependence proves to be inappropriate. Therefore, an alternative approach is required. The use of random forests, a powerful supervised learning algorithm, to determine the most important variables for degradation and their weights is analysed. The results are presented in Chapter 5.

However, first of all, data must be collected. Without a comprehensive dataset, ML is useless. The success of the model and the validity of the results depend directly on the data used. For developing the classifications, data on three aspects is required: climate, specific energy yield, and degradation rates. This is the key to relating climate, on the one hand, to PV performance, on the other. Since the objective is to develop a worldwide classification, an extensive and accurate dataset is essential. In this project, a worldwide grid with resolution 0.5° latitude by 0.5° longitude is utilised[1]. The climate data is extracted from renowned climate research centres and institutions for the period 1991 to 2021. Specific energy yield and degradation rate values are provided by Ascencio-Vásquez et al. in [6] and [7], respectively. Chapter 2 is dedicated to describing and explaining the dataset and its construction in more detail. Furthermore, it includes an explanation of the two main algorithms implemented in this work: linear regression and k-means.

[1] Antarctica and most of Greenland are excluded.

1.6 Chapter summary

Chapter 1 covered the motivation for the project, previous works, tools, and methodology.

- A simple and effective procedure for assessing PV potential for a particular location is to analyse its climatic conditions.
- In general, climatic studies have been based on the Köppen-Geiger (KG) climate classification. However, this classification is solely based on temperature and precipitation, resulting in an unsatisfactory scheme for the PV field.
- In 2019, Ascencio-Vásquez et al. developed a new worldwide classification based on temperature, precipitation, and solar irradiation: the Köppen-Geiger-Photovoltaic (KGPV) climate classification. It follows two criteria. First, based on the KG scheme, zones are classified in terms of temperature and precipitation as Tropical, Desert, Steppe, Temperate, Cold, and Polar. Secondly, solar irradiation is considered to distinguish between Very High, High, Medium, and Low irradiation zones. From the 24 possible combinations, KGPV selected 12 to create the final classification, based on a land-surface ratio and population density criterion.
- An objective, precise, and comprehensive classification might be obtained using Machine Learning. In this work, supervised and unsupervised learning algorithms will be both implemented. Python 3.10 is the computing platform selected, being scikit-learn one of its most remarkable libraries for Machine Learning.
- The specific energy yield is selected as the basis for developing the classification. Therefore, the first step is to identify which climate variables are more relevant to the specific energy yield. A linear regression model is built to predict worldwide specific energy yields from the knowledge of several climate parameters. The predicted values can be compared with known data to obtain a measure of the error of the model. This provides a method to analyse the relevance of the climate variables on PV performance. This is the subject of Chapter 3.
- The classification is created in Chapter 4 using k-means. A qualitative analysis will be required to understand the results and present a final classification.
- The applicability of the methodology to degradation and an alternative approach is explored in Chapter 5.
- For developing the classifications, data on three aspects is required: climate, specific energy yield, and degradation rates. Chapter 2 is dedicated to describing and explaining the dataset and its construction. Furthermore, it includes an explanation of linear regression and k-means.

Chapter 2

Prelude

The first step in a Machine Learning problem-solving strategy consists in collecting data. To create the worldwide PV-climate classifications, a comprehensive dataset conformed by climate variables, specific energy yields, and degradation rates for the whole planet must be built. The first section of this chapter deals with this essential and challenging task. The result is a worldwide grid with a resolution of 0.5° latitude by 0.5° longitude which includes 12 climate variables, the specific energy yield, and the degradation rate for every point.

The chapter concludes with a second section where the algorithms used in Chapter 3 (linear regression) and Chapter 4 (k-means) are explained in detail. Furthermore, the strengths and weaknesses of these algorithms are analysed, and other alternatives are discussed. Linear regression stands out for its simplicity and ability to give a measure of the climate variables' importance. The main advantages of k-means are its low computational times and memory requirements, and that it always reaches a solution.

2.1 Data collection

This section describes the data collection process. It is divided into two subsections. First, the climate data is discussed. Then, in the second subsection, the data for the specific energy yield and degradation rates are presented.

2.1.1 Climate data

In principle, only climate parameters relevant to the specific energy yield and degradation rate are of interest. However, these climate parameters are not known a priori. Indeed, that is the question to answer in Chapter 3. Therefore, the climate dataset built here comprises all climate variables that *might* be relevant. This dataset serves as the baseline for the subsequent feature selection procedure.

The dataset consists of a matrix whereby each row corresponds to a particular location (sample), and each column contains the value of a climate variable (feature) for that location, except for the first two columns, which contain the latitude and longitude, respectively. The climate features are selected based on technical expertise. Moreover, KG and KGPV classifications are used as references. Ultimately, it is not desirable to deviate too much from these classifications if a comparison is aimed to be established later. As indicated by Ascencio-Vásquez et al. in KGPV, the evolution of temperature has been remarkably different since 1990 [6]. Therefore, for every climate feature, monthly average data is extracted from 1991 to 2021[1] and averaged, in turn, to obtain an average year.

Table 2.1 summarises the 12 final features included in the dataset. The following types of climate features might be distinguished: temperature, precipitation, humidity, irradiation, and wind.

Undoubtedly, temperature is a fundamental parameter for both climate and PV studies ([10], [28], [32]). As was seen in Chapter 1, it plays a principal role in KG and KGPV classifications. The decrease in performance with temperature is a well-known and thoroughly analysed fact [27]. Furthermore, temperature is known to trigger several degradation mechanisms ([7], [8]). Worldwide[2] temperature data, with a resolution of 0.5° x 0.5°, is provided by the Climate Research Unit (CRU) of the University of East Anglia. More specifically, the CRU TS V.4.06 climate dataset was used [18]. It is derived by the interpolation of monthly climate anomalies from extensive networks of weather station observations, and it is the same temperature dataset used in KG and KGPV. From this dataset, the variable "Mean 2 m temperature", *TMP*, is extracted. Once the average year is calculated, three climate features are added to the dataset: the annual mean temperature, T_{ann}, the monthly mean temperature of the warmest month, T_{max}, and the monthly mean temperature of the coldest month, T_{min}.

Another relevant temperature-related feature, especially for degradation, is the daily temperature difference or thermal cycling ([7], [26]). Similarly, this feature is extracted from the CRU TS dataset. The variable is present with the name "Diurnal 2 m temperature range" (DTR). In this case, only the annual mean value is considered (DTR_{ann}).

[1] The period used for precipitation was from 1991 to 2020, since the year 2021 was not yet available.
[2] Antarctica is not included.

Table 2.1: Climate features included in the dataset.

Feature	Description
T_{ann}	Annual mean near-surface (2 m) temperature
T_{max}	Monthly mean temperature of the warmest month
T_{min}	Monthly mean temperature of the coldest month
DTR_{ann}	Annual mean daily temperature difference
P_{ann}	Accumulated annual precipitation
P_{min}	Accumulated precipitation of the driest month
RH_{ann}	Annual mean relative humidity
GHI_{ann}	Accumulated annual Global Horizontal Irradiation
GHI_{max}	Maximum accumulated monthly Global Horizontal Irradiation
GHI_{min}	Minimum accumulated monthly Global Horizontal Irradiation
UV_{ann}	Accumulated annual UV irradiation
WS_{ann}	Annual mean near-surface (2 m) wind speed

Regarding precipitation, two features are considered: the accumulated annual precipitation, P_{ann}, and the precipitation of the driest month, P_{min}. Precipitation data is provided by the Global Precipitation Climatology Centre (GPCC), an organisation operated by the German Weather Service (DWD) [40]. In particular, the dataset used in this work is the GPCC Full Data Reanalysis Version 5, which provides high-quality gridded (0.5° x 0.5° resolution) monthly precipitation data. Although less frequently, accumulated daily precipitation has been used in PV forecasting studies [16]. Precipitation by itself might not be a variable directly related to PV performance, but it indirectly affects other factors. For instance, strong rainfall helps remove dust deposition [32]. Other important features affected by precipitation include temperature, humidity, and solar irradiation [33]. Furthermore, precipitation is essential in climate studies such as the KG climate classification.

However, precipitation does not properly reflect the distribution in humidity, which is a more relevant PV stressor [26]. Certainly, humidity is recognised as one of the main causes of degradation ([7], [8], [29]). Among the several existing ways of expressing humidity, relative humidity (RH) is frequently used in PV degradation studies ([7], [29], [31]). This variable is provided by the Copernicus Climate Change Service Data Store (CDS), a service implemented by the European Centre for Medium-Range Weather Forecasts (ECMWF), in the dataset "Essential climate variables for assessment of climate variability from 1979 to present" ([12], [20]). In particular, the variable is called "Surface air relative humidity" and the resolution is 0.25° x 0.25°. Hence, the data must be further processed before adding it to the dataset as the annual mean relative humidity RH_{ann} with resolution 0.5° x 0.5°. Alternatively, relative humidity can be calculated from the dew point temperature and ambient temperature, as illustrated in [7].

It is evident that solar irradiation is the most relevant feature for PV performance studies. How-

ever, there exist numerous measures and varieties of solar irradiation. In this work, the Global Horizontal Irradiation (GHI) and the Ultra-Violet Irradiation (UV) are considered. GHI data is also taken from the CDS, although this time from the dataset "ERA5 monthly averaged data on single levels from 1940 to present" ([11], [19]). The variable is called "Surface solar radiation downwards", and it was downloaded as a monthly averaged reanalysis form. This parameter is defined as the "amount of solar radiation that reaches a horizontal plane at the surface of the Earth" and it compromises both direct and diffuse solar radiation [11]. A validation of this variable can be found in [7]. Similar to RH, the initial resolution is 0.25° x 0.25° and extra processing is required. Eventually, three features are added to the dataset: the accumulated annual Global Horizontal Irradiation, GHI_{ann}, the maximum accumulated monthly Global Horizontal Irradiation, $GHI_{\max}$, and the minimum accumulated monthly Global Horizontal Irradiation, $GHI_{\min}$.

UV exposure is directly related to several degradation processes [17]. Furthermore, it is an indicator of the solar spectrum. Therefore, the accumulated annual Ultra-Violet irradiation is considered, UV_{ann}. The ERA5 dataset contains a variable called "Downward UV radiation at the surface" which might be used to calculate UV_{ann}. However, as discussed in [7], UV is typically referred to for wavelengths below 400 nm, while the variable given in ERA5 covers the range from 200 to 440 nm. As a consequence, the latter significantly overestimates the UV irradiation. Following the alternative proposed in the same paper, UV is calculated using the approach given in [46]:

$$UV_A = (7.210 - 2.365 \cdot k_t^*) \cdot 10^{-2} \cdot GHI \tag{2.1}$$

$$UV_B = (1.897 - 0.860 \cdot k_t^*) \cdot 10^{-3} \cdot GHI \tag{2.2}$$

$$UV = UV_A + UV_B \tag{2.3}$$

where

$$k_t^* = max(0.1, min(0.7, k_t)) \tag{2.4}$$

and k_t is the clearness index, i.e., the GHI divided by the solar radiation at the top of the atmosphere, both variables available in the ERA5 dataset.

The last feature considered is the wind speed (WS). Wind affects the temperature of the solar module and consequently has an impact on the specific energy yield [7]. Wind also has an impact on the degradation rate, causing mechanical load or big pressures, and is directly related to soiling effects ([8], [29]). It has been widely considered in previous classification studies ([26], [32]). Wind speed data is taken from the ERA5 dataset. It is available as the horizontal speed of the wind at a height of ten meters above the surface of the Earth (10 m wind speed). However, temperature data is known at a height of two meters, so the following correction is applied to obtain the wind speed at the same height [7]:

$$WS_{2m} = \left(\frac{2}{10}\right)^{0.2} \cdot WS_{10m} \tag{2.5}$$

Hence, the final feature added to the dataset corresponds to the annual mean wind speed at a two meters height (WS_{ann}).

2.1.2 Specific energy yield and degradation data

Theoretical worldwide specific energy yield values for crystalline silicon (c-Si) modules were calculated by Ascencio-Vásquez et al. to assess the KGPV climate classification [6]. They simulated a typical day for each month, multiplied by the number of days in each month, and summed up to the annual value. The impact of temperature, balance-of-system efficiency, and spectral and angular-reflection losses were considered. On the other hand, shading, soiling, and snow losses were neglected. This data is available with a 0.5° x 0.5° resolution. Latitude values range from -55° to 70°, so Antarctica and most of Greenland are excluded. Figure 2.1 illustrates the data.

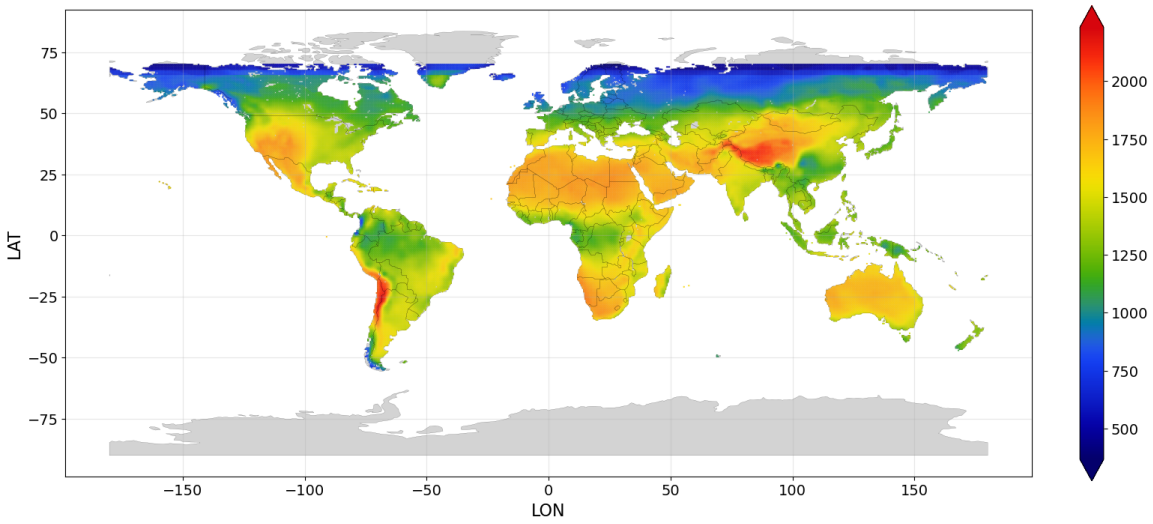


Figure 2.1: Specific energy yield (kWh/kW) calculated by Ascencio-Vásquez et al. [6].

The specific energy yield (Y_f) is added as another column to the climate dataset. Therefore, for every location (rows), the dataset contains its climate features and specific energy yield. This enables finding a relation between climate and Y_f.

As a final remark, it is worth mentioning that, even though both the climate and specific energy yield data have a resolution of 0.5° x 0.5°, the actual range of values is different. For instance, latitude in the climate dataset starts counting at -55.25°, while in the Y_f, it starts at -55°. Therefore, it is necessary to interpolate to have both datasets referred to exactly the same locations. The same issue is found regarding longitude.

Similarly, theoretical degradation rates were calculated by Ascencio-Vásquez et al. for monocrys-

talline silicon PV modules installed in an open-rack mounting configuration [7]. Three degradation mechanisms were modelled: hydrolysis-degradation, related to the effect of temperature and humidity, thermo-mechanical-degradation, related to high temperature and temperature differences, and photo-degradation, due to temperature, humidity, and UV irradiation. The effects of these degradation mechanisms were added to obtain a total degradation rate. The data is available with a resolution of 0.25° x 0.25°. Worldwide degradation rates are shown in Figure 2.2.

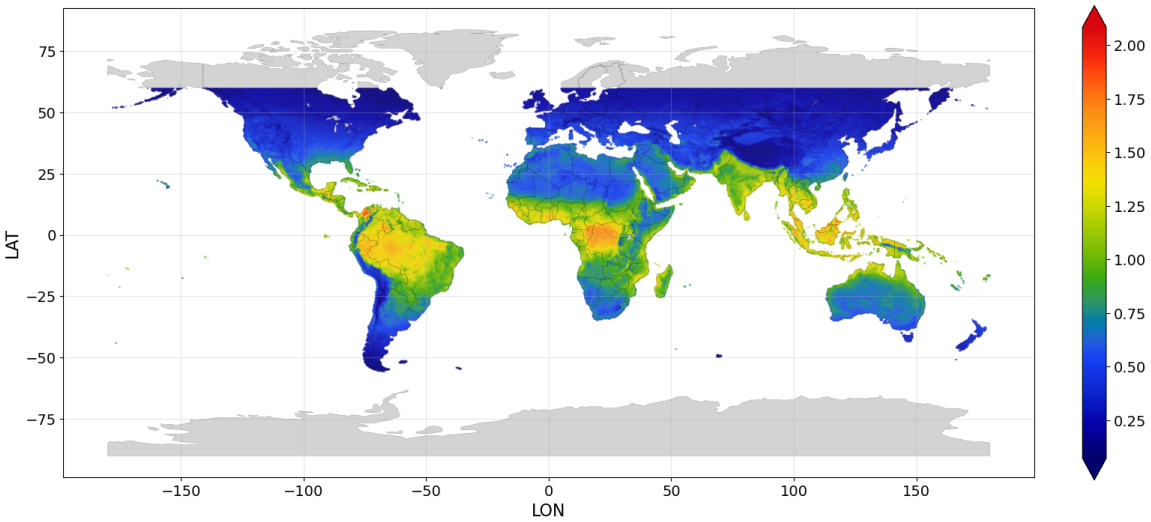


Figure 2.2: Degradation rates (percent/year) calculated by Ascencio-Vásquez et al. [7].

The degradation rate is added as a new column to the climate dataset, but in a second copy, so the specific energy yield values are not in the same dataset as degradation. The main reason is that the latitude range is different, ending in 54.5° in this case. Furthermore, the specific energy yield and degradation rate are studied separately in this project, so this is also more convenient. Lastly, the resolution must be adjusted to 0.5° x 0.5°.

2.2 Two algorithms

The created dataset is used in Chapter 3 to develop a Linear Regression model and select the most relevant climate variables. Then, in Chapter 4, k-means is implemented to create a new PV-climate classification based on those climate variables. This section introduces these two algorithms.[3]

Linear Regression

One of the most successful and widely used types of supervised learning algorithms is linear models [36]. Several authors have implemented linear models in PV studies in the past [44]. In short, linear models make a prediction using a linear function of the input features [36]. In mathematical notation:

$$y_p = w_0 \cdot x_0 + w_1 \cdot x_1 + ... + w_n \cdot x_n + b \tag{2.6}$$

where y_p denotes the prediction, and x_i denotes the feature i. The parameters learnt by the model are the weights associated with each feature, w_i, and the interception, b.

The basic routine followed when implementing the model, and supervised learning algorithms in general, consists of the following steps. First, data is collected so that points in the form $(x_0, x_1, ..., x_n, y)$ are established. Secondly, the data is partitioned into two groups, called the training data and the test data. The training data is fed into the model, and the parameters are adjusted so that the error between the predictions, y_p, and the known values, y, is minimised. Finally, once the model has been created, it is evaluated using the test data, which enables knowing how accurate the Machine Learning model is for data it has not seen before [9].

There exists numerous linear models for regression. In this work, Linear Regression will be implemented. Linear Regression, or ordinary least squares (OLS), is the simplest and most classic linear method for regression [36]. Linear Regression fits the parameters, w_i and b, so that the mean squared error (MSE) is minimised. The mean squared error is the sum of the squared differences between the predictions and the true values [4]:

$$MSE = \frac{1}{N} \cdot \sum_{j=1}^{N} (y_{p,j} - y_j)^2 \tag{2.7}$$

where N is the total number of points.

Furthermore, in order to evaluate the quality of the predictions and establish comparisons between different models, several error measures might be used. A comprehensive summary of commonly used evaluation indexes is given in [4]. The most sensible metric might be the regression score function, R^2. This is a measure of the ability of a model to predict or explain an outcome based on linear regression. The best possible score is 1.0, while a constant model that always predicts the average y disregarding the input features would get an R^2 of 0.0 [37]. It is the default score function

[3] This is the scheme and algorithms implemented for studying the specific energy yield. Degradation is treated separately in Chapter 5.

used in scikit-learn and the first metric to consider. Other indexes used in this work are the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE).

Linear Regression presents several advantages. Its simplicity implies low computational times and memory requirements, both for training and predicting. Moreover, Linear Regression makes it relatively easy to understand how a prediction is made [36]. A mathematical model is built that assigns to every feature a weight, which is an indicator of its influence on the output. This point is essential for the objectives of this work. Lastly, Linear Regression enables making predictions for new data not necessarily in the same range as the training dataset. In other words, it enables extrapolation. This might be relevant when the data collected is not uniform, and it is a drawback, besides the level of complexity, of other sophisticated algorithms such as Random Forests.

On the other hand, Linear Regression has some limitations. A linear dependence between the target and the features is assumed, which might not always be realistic. Nevertheless, it is important to note that usually several features are used, resulting in a multi-dimensional model. Hence, linear models can be very powerful. In this sense, simply imagining a straight line when talking about a linear dependence might be misleading [36]. Of course, in some applications, the model might still be inappropriate. A solution would consist in applying a non-linear transformation to the features. For instance, a logarithmic or an exponential function might be applied [36]. In this case, the model can be written in the following form:

$$y_p = w_0 \cdot f_0(x_0) + w_1 \cdot f_1(x_1) + ... + w_n \cdot f_n(x_n) + b \tag{2.8}$$

where f_i denotes the function applied to the feature x_i.

Further limitations of Linear Regression in this project are discussed in the last section of Chapter 3.

K-means

K-means might be the simplest and most commonly used clustering algorithm [36]. The objective of k-means is to group data points with certain similarities to discover underlying patterns. Each of these groups is called a cluster and is described by the mean, μ, of the samples constituting it. The mean of a cluster is also known as its centroid or cluster center [37].

The algorithm requires the number of clusters, k, to be specified by the user. Then, the clusters are formed aiming to minimise the within-cluster sum-of-squares criterion, also known as the inertia. Inertia can be recognized as a measure of how internally coherent clusters are [37]. This is achieved following the next steps [36]:

1. Initially, the centroids are randomly generated.
2. Each data point is assigned to the closest cluster center.
3. Centroids are recalculated using the data points assigned to them in step 2.
4. Steps 2 and 3 are repeated until the change is less than a threshold.

Figure 2.3 illustrates the procedure for two clusters.

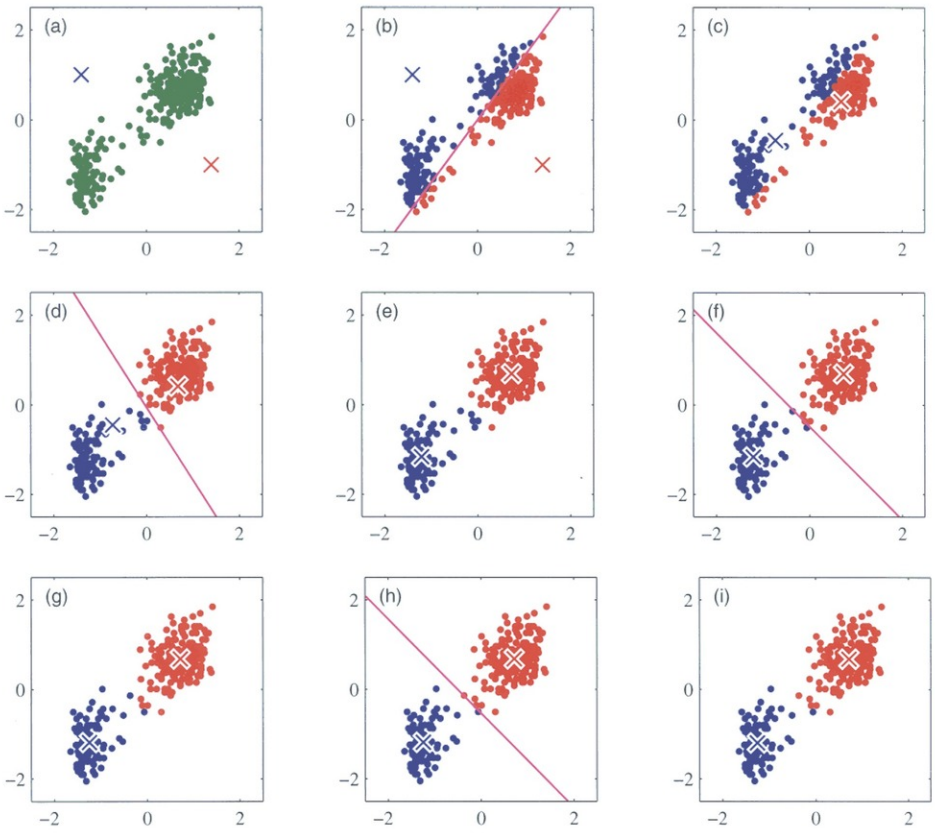


Figure 2.3: Illustration of k-means clustering algorithm. [15]

Given enough time, k-means will always converge. However, due to the randomness of step 1, running the algorithm again might result in different clusters [36]. In other words, different local minimums might be achieved. For that reason, the procedure is repeated several times and the algorithm selects the solution with the lowest inertia. To reach an optimum solution with an acceptable number of attempts, usually, the k-means++ initialization scheme is implemented. In this case, the initial random centroids are generated distant from each other, so the probability of obtaining a local minimum is reduced [37].

K-means scales well to large datasets and has been successfully used in many fields. A characteristic of k-means is that the desired number of clusters must be specified as an input parameter. In general, this is inconvenient, since in many applications the optimum number of clusters is unknown. Therefore, a tedious optimisation procedure might be required. On the other hand, the possibility of fixing the number of clusters is an advantage in some applications since the algorithm is forced to give a solution. Other clustering algorithms, such as DBSCAN, calculate the number

of clusters, but some points can be classified as "noise". In this project, this is not a valid result. Furthermore, DBSCAN requires two input parameters, whose optimisation can be even more challenging.

The main limitation of k-means is that it can only capture relatively simple shapes [36]. Inertia assumes that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes [37]. Also, k-means tries to form clusters of relatively similar size. Again, the influence of these disadvantages must be assessed according to the particular application. The data used in this work is quite homogeneous, so the first point is not a serious issue. On the other hand, very small or unique clusters are not interesting for developing a worldwide classification. Instead, the aim is to obtain a pragmatic classification, able to recognise the most relevant features. For this purpose, the preciseness of k-means might be considered sufficient.

To conclude, the virtue of k-means lies in its simplicity. It demands low computational times and memory requirements, which allows the development of several classifications. This contrasts with other popular clustering algorithms such as hierarchical clustering, whose limitations in terms of speed and memory make its implementation into this project impossible.

2.3 Chapter summary

In this chapter, a dataset containing the climate variables, specific energy yield values, and degradation rates has been built. Furthermore, the main algorithms used in this work have been explained.

- A worldwide grid with a resolution of 0.5° latitude by 0.5° longitude which includes 12 climate variables has been created. More specifically, the climate variables included in the dataset are: T_{ann}, T_{max}, T_{min}, DTR_{ann}, P_{ann}, P_{min}, RH_{ann}, GHI_{ann}, GHI_{max}, GHI_{min}, UV_{ann}, and WS_{ann}.
- Theoretical worldwide specific energy yield values for crystalline silicon (c-Si) modules are provided by Ascencio-Vásquez et al. [6]. Similarly, theoretical degradation rates were calculated by Ascencio-Vásquez et al. for monocrystalline silicon PV modules [7]. The specific energy yield and the degradation rate are added as another column to the climate dataset.
- Linear Regression will be implemented in Chapter 3 to select the most relevant climate variables. It is a simple model with low computational times and memory requirements. Furthermore, it assigns a logical weight to every feature. As the main limitation, assuming a linear dependence might be inappropriate in some applications.
- K-means groups data points with certain similarities to discover underlying patterns and will be used in Chapter 4 to create the classification. A characteristic of k-means is that the desired number of clusters must be specified as an input parameter. Thus, the algorithm is forced to give a solution. The main limitation of k-means is that it can only capture relatively simple shapes. On the other hand, it demands low computational times and memory requirements.

Chapter 3

Selection and Weighing

At this point, a worldwide dataset composed of several climate features and the specific energy yield has been obtained. Now, the dependency between climate and PV performance can be studied. In this chapter, the most relevant climate features are selected and weighted for developing the classification in Chapter 4.

As previously discussed, the approach consists in implementing Linear Regression to analyse the suitability of each feature for predicting accurately the specific energy yield. The first section, "Calculations", explains how the model is built and the operations performed. Feature selection proves to be a challenging issue. Pearson coefficients, automated feature selection, and technical expertise are combined to choose 79 possible sets of features. Then, these are evaluated using Linear Regression. In the second section, the results are carefully analysed and a decision regarding the features to use in Chapter 4 is made. Four combinations, each with a different number of features, are proposed. The MAPEs of these models are around six percent. Finally, the limitations of this approach are discussed in the last section. In particular, non-linear relationships are explored and optimisation algorithms are recommended.

3.1 Calculations

First of all, following the typical routine in supervised learning, the dataset is partitioned into two subsets: the training data and the test data. Here, the training data corresponds to 75 percent of the original dataset, while the test data consists of the remaining 25 percent. This is just a general good rule of thumb, and similar partitions are also applicable [36]. Secondly, it is necessary to rescale the data. Data scaling is a common preprocessing method applied before implementing the actual ML algorithm [36]. Data scaling consists in reconstructing the dataset to reduce the impact of different orders of magnitude [4]. This is typically caused by the different units employed among the features. For instance, irradiation has an order of magnitude of 10^8, clearly much higher than other variables like humidity or temperature. These discrepancies in scale and range affect directly the weights' calculation. Therefore, in order to obtain more reliable and comprehensive results, all data points must be transformed to the same scale [4].

There are numerous transformations available for data scaling in scikit-learn. StandardScaler, RobustScaler, MinMaxScaler, and Normalizer might be the most useful ones [37]. In this work, StandardScaler is applied. It transforms the data so that every feature has a mean equal to 0 and a variance of 1 [36]. Besides being easy to understand, this technique has proved successful in the optimisation of Machine Learning algorithms [4]. Thus, StandardScaler is used to scale the data. It should be stressed that it is very important to apply exactly the same transformation to the training data and the test data [36].

Finally, the Linear Regression model is built using the training data. Following the nomenclature introduced in Chapter 2, here the features, x_i, are the climate variables, while the target, y, is the specific energy yield. After fitting the weights, the model can be evaluated using the test data. In other words, the specific energy yield for each sample of the test data is predicted using the corresponding climate features, and the prediction is compared to the known value. Thus, the performance of the model can be measured. The lower the error, the higher the correlation between the climate features and the specific energy yield. Figure 3.1 summarises the procedure.

In principle, all features could be fed into the model and the algorithm would calculate their optimum weights. Then, these weights could be used for developing the classification. However, this would result in a complex classification, difficult to analyse and understand. Moreover, many climate variables are related to each other or have minor importance, so these can be discarded for the classification criteria. On the other hand, an insufficient number of features would result in poor accuracy. Therefore, it is essential to make a wise selection.

Knowing the number and which features should be selected is not a trivial issue. The climate features are not independent of each other, and there exist particular combinations which remarkably improve the performance of the model. For that reason, the importance of an individual feature depends on the other features with which it is combined. Consequently, the optimum combination can vary significantly when changing the number of features selected.

Certain statistical parameters and tools can facilitate the analysis. A first insight into the relevance of a feature is provided by the Pearson correlation coefficient or Pearson's r. It evaluates the linear correlation between the feature and the target [37]. More specifically, it is defined as the ratio

between the covariance and the product of the standard deviations, and its value ranges between -1 and 1. A Pearson's r equal to 1 or -1 means a perfect linear correlation, the sign just indicating the direction. In contrast, a coefficient of 0 is an indicator of no linear dependency [49].

It is important to note that Pearson's r measures the dependence of the individual variable, so no information is gained regarding the interactions between the climate features. Consequently, feature selection cannot be based solely on these numbers. Nevertheless, it provides a sense of the importance of each feature and can be utilised to make some decisions. For instance, Ahmed et al. [4] considered significant only those variables with a Pearson coefficient higher than 0.4. Table 3.1 illustrates the calculated Pearson coefficients.

Table 3.1: Pearson correlation coefficient for every climate feature. Values close to +1 or -1 indicate a high linear dependence between the feature and the specific energy yield.

Feature	**Pearson correlation coefficient**
T_{ann}	+ 0.66
T_{max}	+ 0.63
T_{min}	+ 0.60
DTR_{ann}	+ 0.76
P_{ann}	- 0.13
P_{min}	- 0.27
RH_{ann}	- 0.72
GHI_{ann}	+ 0.92
GHI_{max}	+ 0.78
GHI_{min}	+ 0.77
UV_{ann}	+ 0.91
WS_{ann}	+ 0.01

It is clear from Table 3.1 that GHI_{ann} and UV_{ann} have a strong linear correlation with the specific energy yield. Therefore, these variables are expected to play a fundamental role in the classification. On the other hand, the low Pearson coefficients for precipitation and wind speed suggest that these variables will not be so relevant. In particular, WS_{ann} has a coefficient of almost zero, so it will be disregarded for the rest of the analysis. It is important to clarify this conclusion. This does not mean that wind speed does not have an impact on the specific energy yield. It simply indicates that, globally, there is no correlation between the WS_{ann} and the Y_f. In the last section of this chapter, it is shown that this conclusion applies even considering nonlinear dependencies.

Another tool that might be helpful to guide decision-making is automatic feature selection [28]. Among the several methods available for automatic feature selection in scikit-learn [37], recursive feature selection (RFE) is one of the most powerful options. RFE builds a series of models varying

the number of features. As a result, based on the weights calculated, it enumerates the features based on their importance to the algorithm [36]. RFE was implemented using two different algorithms: Linear Regression and Random Forests. The choice of Linear Regression is evident: it is the algorithm used in this work. The motivation for including Random Forests is that this is a very powerful algorithm whose results might be interesting. Table 3.2 illustrates the positions assigned to each variable.

Table 3.2: RFE ranks the features in order of importance, from most (1) to least (11).

Feature	**Linear Regression**	**Random Forests**
T_{ann}	3	9
T_{max}	4	5
T_{min}	2	6
DTR_{ann}	9	7
P_{ann}	10	8
P_{min}	8	11
RH_{ann}	11	4
GHI_{ann}	6	1
GHI_{max}	7	2
GHI_{min}	5	3
UV_{ann}	1	10

The result is completely different for Linear Regression and Random Forests. Considering that these are different algorithms, this is not necessarily a surprise. It might then be concluded that the results for Linear Regression are more significant, since this is the algorithm used in this work. However, when making the actual calculations, it is shown that the results given by RFE are not accurate[1]. Moreover, the order of importance determined by Random Forests is closer to the optimum. Thus, RFE is not performing satisfactorily and the given order of importance should not be simply accepted. Nevertheless, it can be used as a starting point for searching for the optimum combination of features.

Overall, there is no straightforward way of selecting the features. It is necessary to try several combinations to find the optimum set. Therefore, the following methodology has been finally applied. For every possible number of features (ranging from 1 to 11), the model has been evaluated using different combinations. The possible combinations have been chosen based on the Pearson coefficients, RFE results, and technical expertise[2]. In total, 79 options have been evaluated. Table 3.3 summarises the optimum combination found for each number of features, with their associated errors. Error measures include the R^2, RMSE, MAE, and MAPE. Furthermore, a case consisting of a random variable is included to show the validity of the model and establish a reference. The

[1] This can be seen in Appendix I.

[2] Frequently, expert knowledge greatly simplifies feature selection procedures [36].

rest of the calculations can be found in Appendix I.

It is worth mentioning that the errors shown here are calculated using the test data. There is also an error associated with the training data, which can provide further insights. In this work, since theoretical values are being used, the data is very homogeneous and both errors are identical. Furthermore, since the data is randomly divided into training and test data, running the algorithm again results in slightly different values. To reduce this effect, the algorithm was executed three times, and the error and weights were averaged.

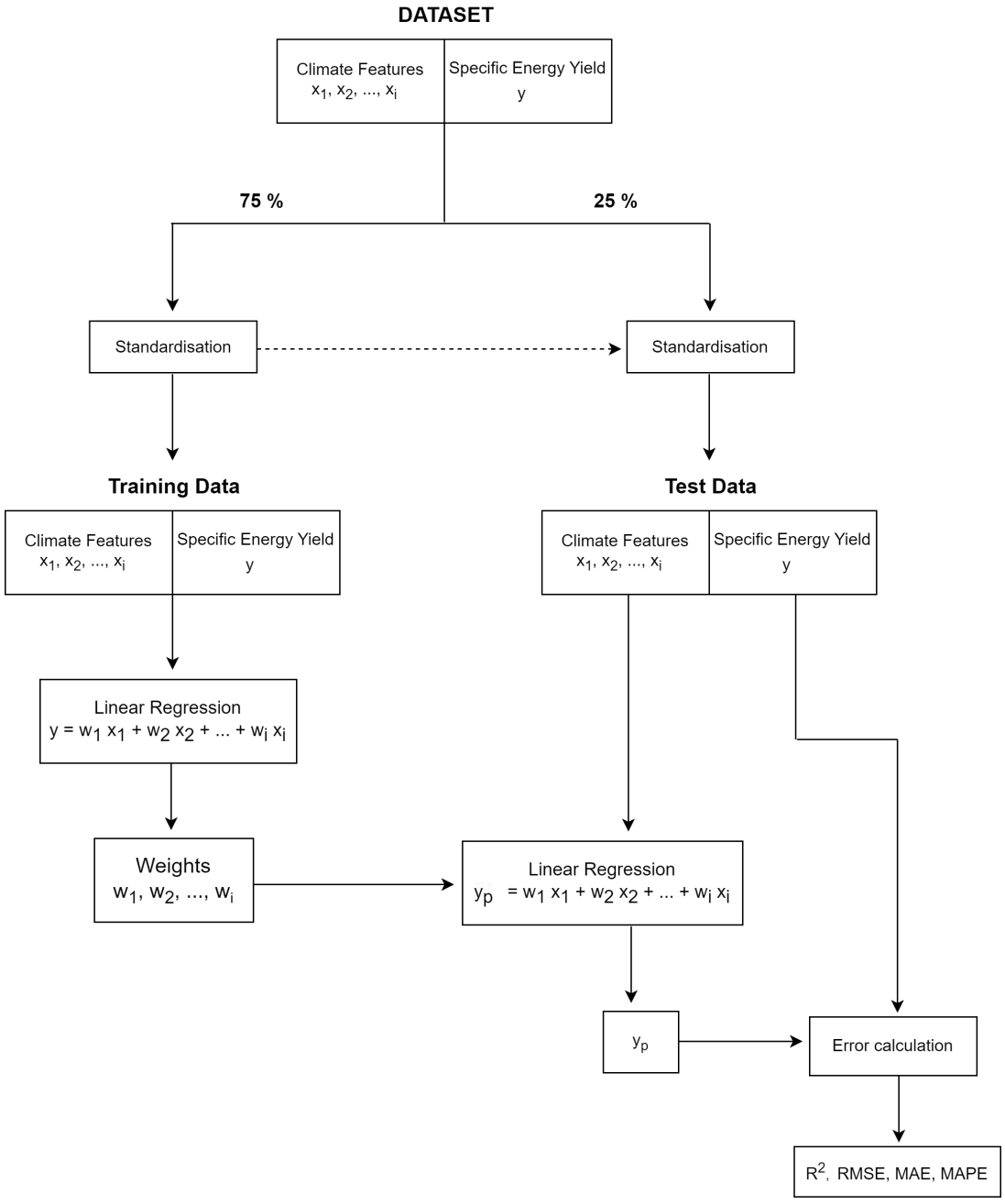


Figure 3.1: Flowchart of the feature selection procedure. First, the dataset is partitioned into the training data and the test data. Secondly, the data is standardised. The test data must be standardised based on the training data. Then, the Linear Regression model is built using the training data. Finally, after fitting the weights, the model is evaluated using the test data.

Table 3.3: Optimum combination for each possible number of features. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	**Random**	$\mathbf{R^2}$	**RMSE**	**MAE**	**MAPE (%)**
0												1.61	0	360.4	301.28	28.3
1								332.3					0.852	139.61	109.14	9.9
2								533.28		-216.81			0.904	112.37	84.05	7.9
3								894.04	-165.25	-471.95			0.923	99.71	77.88	6.8
4	-65.77							964.93	-200.28	-463.66			0.931	95.27	73.84	6.7
5			-104.31					169.91	-133	-435.09	748.66		0.937	90.77	70.31	6.5
6	-98.57					-18.39		266.6	-169.9	-497.71	732.93		0.938	89.45	68.86	6.3
7	500.46	-183.59	-423.15					269.94	-121.51	-426.79	610.23		0.941	87.31	67.64	6.2
8	501.44	-191.16	-409.65			-17.56		252.12	-134.75	-452.43	652.1		0.943	86.03	66.77	6.1
9	495.06	-190.9	-404.93			-17.2	-4.19	210.69	-132.36	-450.36	688.52		0.943	86.06	66.86	6.1
10	500.14	-192.01	-408.37		-1.4	-16.28	-4.01	209.25	-132.57	-450.56	690.07		0.943	85.7	66.49	6.0
11	491.87	-189.94	-397.83	7.48	-2.3	-16.24	-1.13	217.01	-132.73	-447.14	671.82		0.944	85.33	66.75	6.1

3.2 Analysis of the results

Table 3.3 provides, for a fixed number of features, the optimum set, the weights, and the error associated with that Linear Regression model. Based on these results, the criteria for developing the classification in the next chapter can be defined.

The first remark is the importance of the irradiation features. Indeed, when fixing the number of features to three, the optimum combination consists of the three measures of *GHI*. Only using these variables, an R^2 of 0.923 and a MAPE of 6.8 percent are achieved. From this point, the temperature starts to play a fundamental role too, appearing T_{ann} when adding the fourth feature. It is interesting to note that T_{ann} and T_{min} seem to provide more information than T_{max} or DTR_{ann}. From the five features, UV_{ann} becomes the variable with the highest weight, a sign of its importance to the model. The next variable to appear is P_{min}, which after showing for the first time with six features and being disregarded with seven, it is always selected. Finally, with very low weights, RH_{ann}, P_{ann}, and DTR_{ann} are added, in that order.

Besides the features, the errors must be carefully analysed. In particular, the regression score function, R^2. The first thing to notice is that, from eight features, the R^2 takes a virtually constant value of 0.943. This suggests that selecting more than eight features is inconvenient, since the model's complexity would increase without tangible improvement. Following the same reasoning, it might be concluded that seven features are preferable to eight, or that five is better than six, due to the small difference in the error measures. However, in principle, it is not clear the impact of these differences on the classification. Therefore, some caution is required. On the other hand, making the classification with a number of features less than three can be discarded. Besides the suboptimal error, using only irradiation variables for developing the classification might not be too insightful.

Overall, at this point, it is possible to discard some options, but making here a final decision about the optimum number of features would not be very convincing. Certainly, the most promising combinations seem to be those for four, five, seven, and eight features. However, how different would the classification be using seven features instead of four? This question cannot be answered at the moment. Consequently, it has been decided to proceed with these four possible candidates. Their effect in the final classification is analysed and compared in the next chapter, and only then, is a final decision made.

To conclude this section, it is worth analysing the quality of the predictions in more detail. From Table 3.3, it is known that MAE values are around 70 while MAPEs are a bit higher than 6 percent. Alternatively, a visual assessment of the predictions might be made based on Figure 3.2. In this figure, the specific energy yield predictions (y-axis) are directly compared to the known values or targets (x-axis). Ideally, a straight line (red line in the figure) should be obtained. In particular, these results correspond to the model built using seven features.

In general, the predictions follow a similar behaviour to the actual values, except for regions characterised by a very low specific energy yield, where the discrepancy is significantly higher. Therefore, for these regions, the classification might be less accurate. On the other hand, it is typically assumed that any algorithm which can predict the output with at least 80 percent of the error lying

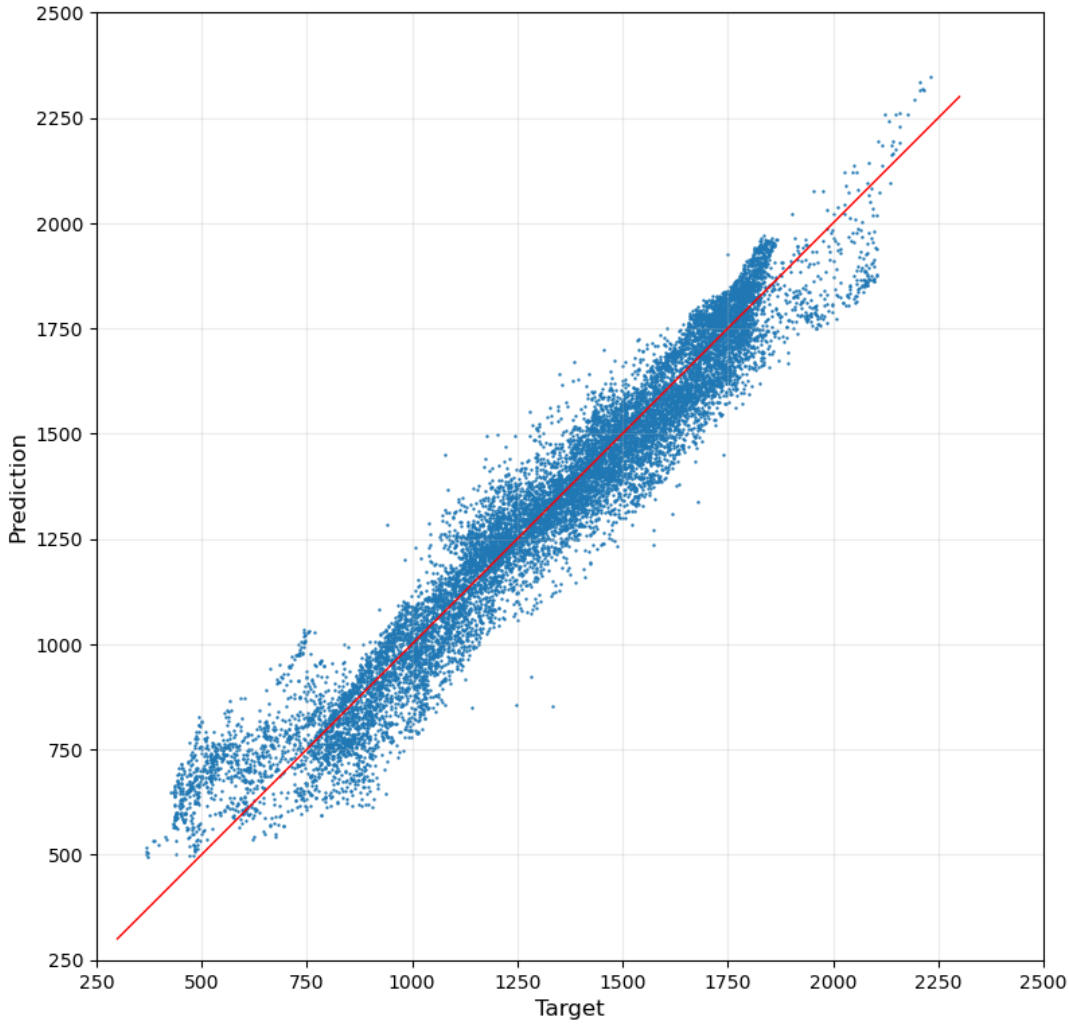


Figure 3.2: Comparison between the targets and the predictions made by the Linear Regresssion model using seven features.

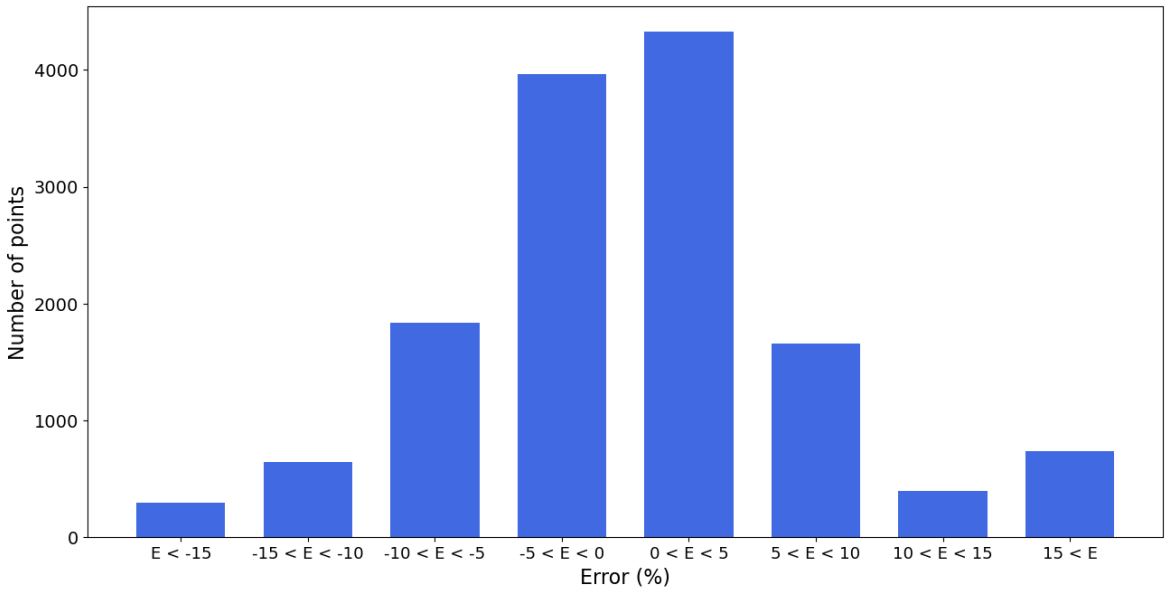


Figure 3.3: Predictions classified into groups according to their relative error. The 85 percent have an error lower than 10 percent.

within 10 percent of the target is a good predictive model [28]. The MAPE already suggests this is the case, but even a clearer view might be provided by Figure 3.3. Here, the points are classified according to their relative error. Again, these results are obtained using seven features.

To be more specific, it is determined that around 85 percent of the points have an error lower than 10 percent. In this figure, it also stands out that a significant number of points have an error higher than 15 percent. These, presumably, correspond to the region of very low specific energy yields. Although not shown here to avoid repetition, these conclusions are equally drawn for the cases of four, five, and eight features.

There is no point in further examining the quality of the predictions. It should be stressed that the objective of this chapter is to identify the most relevant features and associate a logical weight to them so that a rational classification can be developed. Linear Regression allows for conducting this process with ease and provides acceptable predictions. Therefore, Linear Regression has proved successful. In contrast, a simple algorithm such as KNN would make almost perfect predictions for this particular dataset, but it would hardly provide insights into the importance of the features and their weights.

3.3 Limitations

Linear Regression has shown many strengths, but there is a weakness that cannot be ignored: it only considers linear dependencies. Even though the use of many features results in a multi-dimensional model which might overcome this deficiency, linear models might miss important information. For instance, wind speed and precipitation show a low linear correlation with the specific energy yield, but, could they have a strong non-linear dependency that the model fails to identify? This is a fair question that should be tackled.

Previously, Pearson's r coefficients were calculated to illustrate the linear dependency between an individual feature and the specific energy yield. There exists a statistical parameter that can be used to measure *any* kind of correlation. This is the Mutual Information (MI) coefficient [37]. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency [37]. Hence, the MI for every feature was calculated. Table 3.4 presents the results.

Table 3.4: Mutual Information coefficients.

Feature	**Mutual Information coefficient**
T_{ann}	0.67
T_{max}	0.68
T_{min}	0.66
DTR_{ann}	0.60
P_{ann}	0.60
P_{min}	0.56
RH_{ann}	0.80
GHI_{ann}	1.35
GHI_{max}	0.81
GHI_{min}	1.08
UV_{ann}	1.28
WS_{ann}	0.28

Logically, the variables with a high Pearson coefficient, like GHI_{ann} and UV_{ann}, have a high MI too. Similarly, it is seen that WS_{ann}, which had a Pearson's r of almost zero, has a very low MI as well. Perhaps more interesting is the case of precipitation. Precipitation features, both P_{ann} and P_{min}, show significant MI values, while the Pearson coefficients were low. This suggests that their importance in the model could become higher by applying an appropriate transformation[3]. This, in turn, could result in better predictions and a different feature selection.

To illustrate this idea, the logarithmic and square root functions were applied to P_{ann}, P_{min},

[3]This method was explained in the second section of Chapter 2

and, to provide another example, $GHI_{\min}$. Then, the Pearson coefficients for these new variables were calculated. Results are shown in the following table. The original variables are also included to facilitate the comparison (Linear).

Table 3.5: Pearson coefficients for different transformations.

Function / **Feature**	Linear	Logarithm	Square Root
P_{ann}	-0.13	-0.3	-0.2
$P_{\min}$	-0.27	-0.54	-0.45
$GHI_{\min}$	0.77	0.75	0.84

It is clear that these transformations, especially the logarithm, improve the linear dependency of precipitation. Taking the square root of $GHI_{\min}$ improves its correlation too. In this way, precipitation could play a more relevant role in the model and the classification. However, besides still presenting a relatively low dependency, it should be considered that the significance and convenience of a classification based on these new variables would be less clear. Therefore, it has been decided to continue working only with the linear dependencies. Nevertheless, the implementation of non-linear dependencies, either by this method or just by changing the algorithm, might be a future point of improvement. The application of other algorithms is explored in Chapter 5 for degradation, a strong non-linear phenomenon.

Lastly, the procedure described poses another difficulty. Independently of the algorithm, it is necessary to study each possible combination of climate features separately. Unfortunately, the total number of possible combinations is too high. For that reason, it was decided to try 79 cases, selected based on the Pearson coefficients, RFE results, and technical expertise. Thus, although unlikely, some interesting combinations might have been missed. In this regard, an automated optimisation model such as Particle Swarm Optimization (PSO), or Genetic Algorithm (GA) might be explored.

3.4 Chapter summary

Chapter 3 covered the feature selection procedure required to identify the most relevant climate variables for creating the classification in the following chapter.

- Linear Regression was implemented to analyse the suitability of the climate variables for predicting accurately the specific energy yield.
- The importance of an individual climate parameter depends on the other features with which it is combined. Consequently, the optimum combination can vary significantly when changing the number of features selected.
- There is no straightforward way of selecting the climate variables. It is necessary to try several combinations to find the optimum set. Based on the Pearson coefficients, RFE, and technical expertise, 79 options have been evaluated.
- Irradiation (GHI_{ann}, GHI_{max}, GHI_{min}, UV_{ann}) and temperature (T_{ann}, T_{min}) are the most relevant climate variables.
- From eight features, the R_2 takes a virtually constant value of 0.943. Making the classification with a number of features less than three is discarded due to poor performance. The most promising combinations are those for four, five, seven, and eight features. MAPEs are a bit higher than 6 percent, and the predictions follow a similar behaviour to the known values except for regions characterised by very low specific energy yields, where the discrepancy is higher.
- Applying a nonlinear transformation to some climate variables, like P_{min}, might improve the model's accuracy. However, the complexity of the model would increase significantly.
- An automated optimisation model, such as Particle Swarm Optimisation (PSO) or Genetic Algorithm (GA), might be implemented to consider more possible combinations of climate features.

Chapter 4

Clustering

In Chapter 2, a comprehensive dataset was built. Then, in Chapter 3, the most relevant features were identified and, more specifically, four promising combinations, each for a different number of features (four, five, seven, and eight), were proposed. All is ready for developing the Machine Learning driven PV-climate classification.

This chapter consists of three sections. In the first section, "Exploration", a method to select and optimise the final classification is illustrated. It consists in creating several classifications using k-means, followed by a careful qualitative analysis based on various parameters and tools. In this way, the optimum number of features and clusters is fixed as seven and 20, respectively. Furthermore, this exploration procedure paves the way to understanding the properties of the clusters.

In the second section, "ML driven PV-climate classification", the final classification is presented and described in detail. First, clusters are divided into six possible climate types: Tropical (Tro), Desert (Des), Mountainous (Mon), Temperate (Tem), Cold (Col), and Polar (Pol). Secondly, clusters inside the same climate type are ordered based on the level of irradiation. Several figures are included to illustrate the reasoning.

Finally, the chapter concludes with an assessment of the classification. The clusters are analysed from the specific energy yield standpoint, and the proposed classification is compared with KGPV.

4.1 Exploration

As previously discussed, k-means is implemented in this project to develop the classification. The procedure consists of the following steps. First, the features selected in Chapter 3 are standardised (using StandardScaler), and multiplied by their corresponding weights. Then, as discussed in Chapter 2, k-means is applied to create the clusters from these features. The desired number of clusters is fixed beforehand. In this way, a classification can be obtained. Figure 4.1 illustrates the procedure.

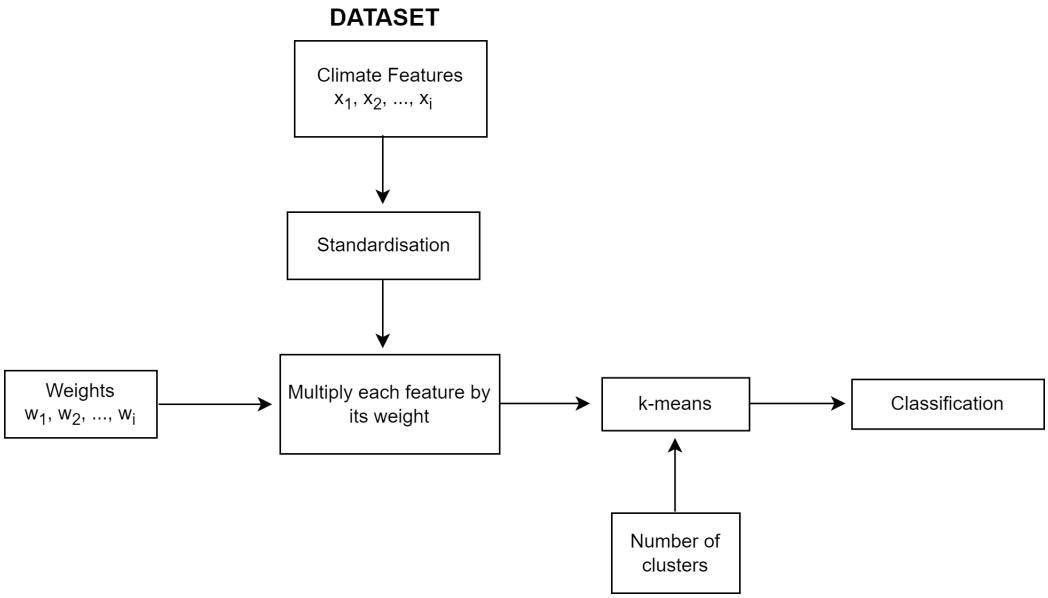


Figure 4.1: Flowchart of the clustering procedure. The weights were calculated in Chapter 3 and the number of clusters is fixed by the user.

However, would this be a proper classification? Moreover, which combination of the four proposed in the previous chapter should be finally utilised? And, how many clusters should be formed? If the result is to be meaningful, one must be able to answer these questions effectively. In this section, the last two issues are addressed, while the final assessment of the classification is left to a subsequent stage (section 4.3).

The approach consists in fixing first an optimum number of clusters for each possible combination of features. Thus, four classifications are obtained, one for each combination. Then, these classifications are compared and a final decision is made. Nevertheless, selecting an appropriate number of clusters is not a trivial issue. Indeed, the first step to building an optimum classification is to define an evaluation method. This essential requirement is usually the most challenging part of clustering studies [36]. In unsupervised learning, in contrast to supervised, there are no known values to compare and assess the results of the algorithm. Consequently, it might be difficult to

evaluate a particular classification as good or bad.

To guide this decision, several mathematical strategies have been proposed. The Elbow method and the Silhouette coefficient stand out due to their simplicity and previous success. However, unfortunately, these pure quantitative analyses rarely work for the datasets found in practice [36]. Indeed, these methods were applied in this work with poor performance. The results can be found in Appendix II.

The reality is that there is not a clear and unique solution to this challenge. Instead, a careful and tedious qualitative exploration procedure is required. This is a common approach in clustering algorithms [36]. Fortunately, numerous tools can facilitate the study and help to make objective decisions. Data visualization is especially relevant for this project. This includes several types of plots, which will be explained as they appear. Other insightful parameters are the clusters' centers and sizes.

The exploration method is best explained using an example. The combination corresponding to four features is utilized hereunder. This is the simplest case and will facilitate the exposition. Having fixed the number of features, the question to answer now is how many clusters must be selected to obtain the optimum classification. Perhaps, a good starting point would be creating the same number of zones as KGPV, that is, 12. Figure 4.2 illustrates the classification obtained in this case.

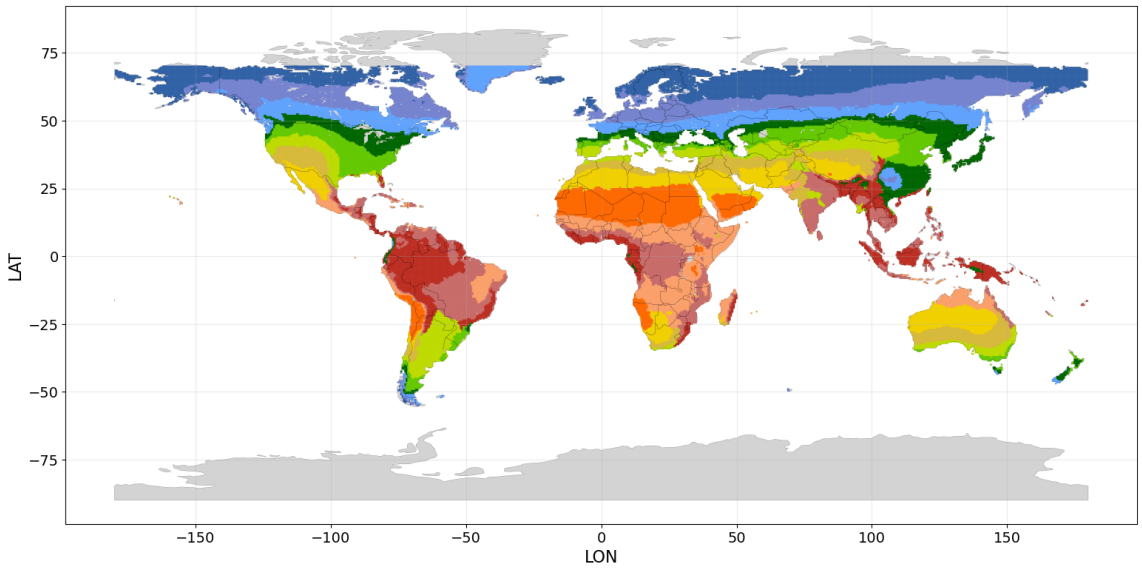


Figure 4.2: Classification created using four features and specifying 12 clusters.

Independently of how well the classification might look, this is not a convincing way of fixing the number of clusters. The methodology applied in this project is completely different from KGPV. Consequently, the optimum number of groups may not be the same. The first idea that comes to mind might be increasing the number of clusters and observing the changes in the classification. Thus, Figure 4.3 presents the classifications associated with 12, 13, 14, and 15 clusters.

A visual comparison of these maps enables a rough analysis of the impact of increasing the number of clusters. As shown in Figure 4.3, when proceeding from 12 to 13, some regions in China, Japan, and South Korea are distinguished as a new cluster (dark red in the figure). When adding the 14^{th} cluster, a new group appears mainly in central Africa and northern Australia (white). Lastly, if 15 clusters are considered, Greenland is classified separately (purple). Certainly, all these changes seem to be improving the classification. However, are they truly significant? Why are these new clusters formed? What characterizes them? These fundamental questions are essential for making an objective decision. However, it is not possible to answer them just by observing the classifications.

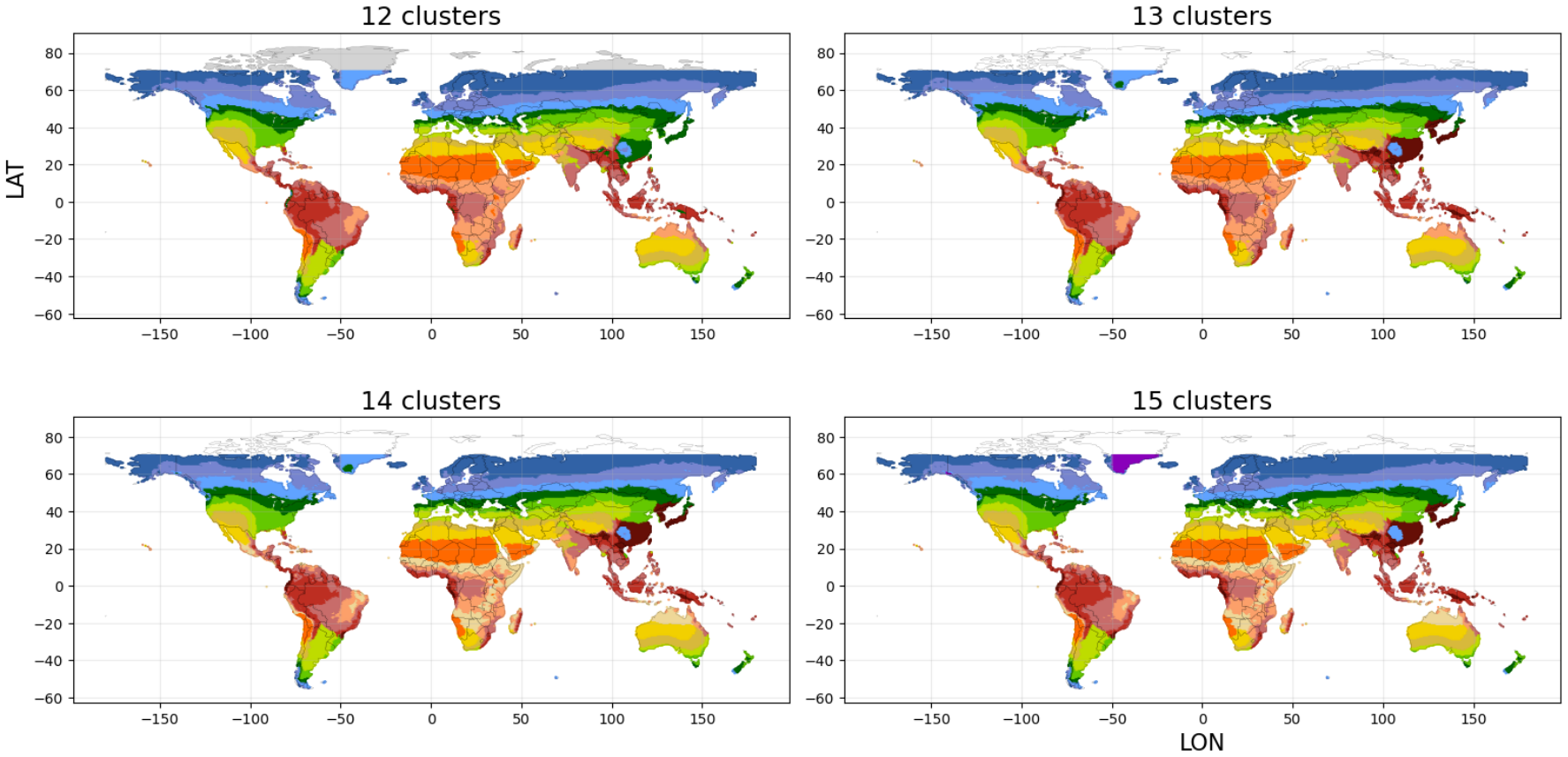


Figure 4.3: Classifications obtained using four features and specifying the number of clusters to 12, 13, 14, and 15. When proceeding from 12 to 13, some regions in China, Japan, and South Korea are distinguished as a new cluster (dark red). When adding the 14^{th} cluster, a new group appears mainly in central Africa and northern Australia (white). Lastly, if 15 clusters are considered, Greenland is classified separately (purple).

Fortunately, there exists a tool that proves very insightful: the pair plot. A pair plot consists of a matrix of scatter plots, each representing the points for every possible pair of features. The diagonal of this matrix is filled with a histogram of each feature [36]. The data points are colored according to the clusters they belong to. This figure enables an understanding of how the clusters are formed, and their main properties. Furthermore, it can be used to predict the formation of new groups. The pair plot associated with four features and 12 clusters is illustrated in Figure 4.4.

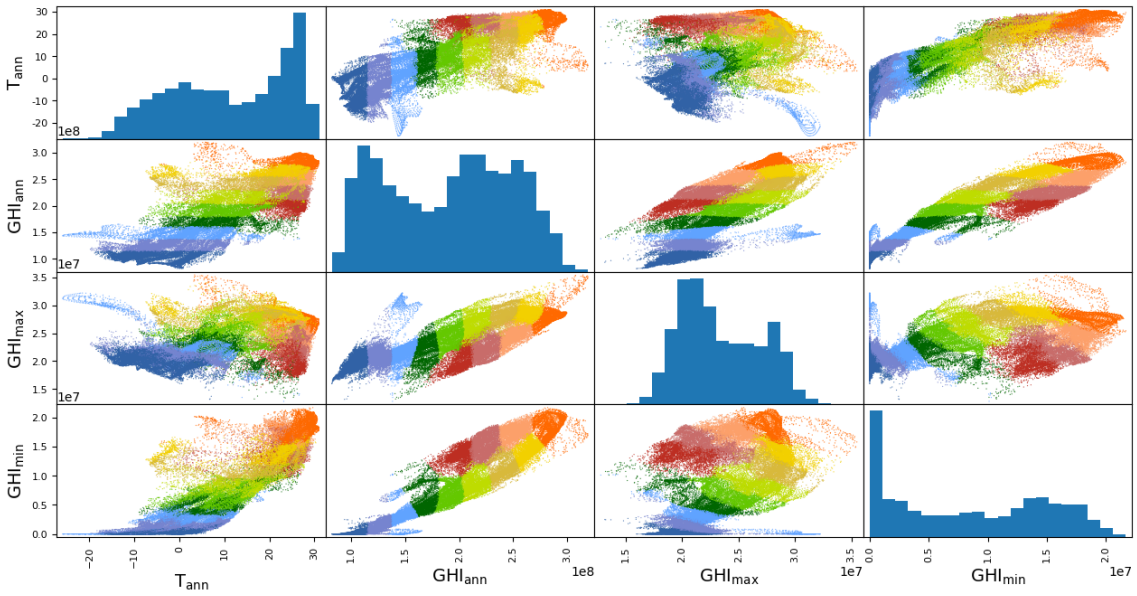


Figure 4.4: Pair plot associated with the classification based on four features and 12 clusters. The units of T_{ann} and GHI, not shown in the figure for the sake of clarity, are °C and J/m^2, respectively.

Even though Figure 4.4 contains a lot of interesting information, it is difficult to analyse, especially when a higher number of features is considered. Therefore, it is usually more convenient to use the pair plot only to identify the most enlightening pair of features. Then, this particular pair can be illustrated separately in more detail to conduct the analysis. For instance, the pair GHI_{ann} - T_{ann} is very informative. Figure 4.5 illustrates it.

From this figure, the following remarks can be made. First, the high weight given to the GHI_{ann} is here evident. The clusters are mainly formed based on this criterion, as the vertical divisions indicate. However, there is a point at which temperature starts playing a role too, and, as a consequence, the reddish clusters are formed. These are characterized by a high T_{ann}. More specifically, their cluster centers indicate an average T_{ann} of 24 °C. Overall, it is seen that these reddish clusters have a similar level of irradiation to the greenish ones but a higher temperature.

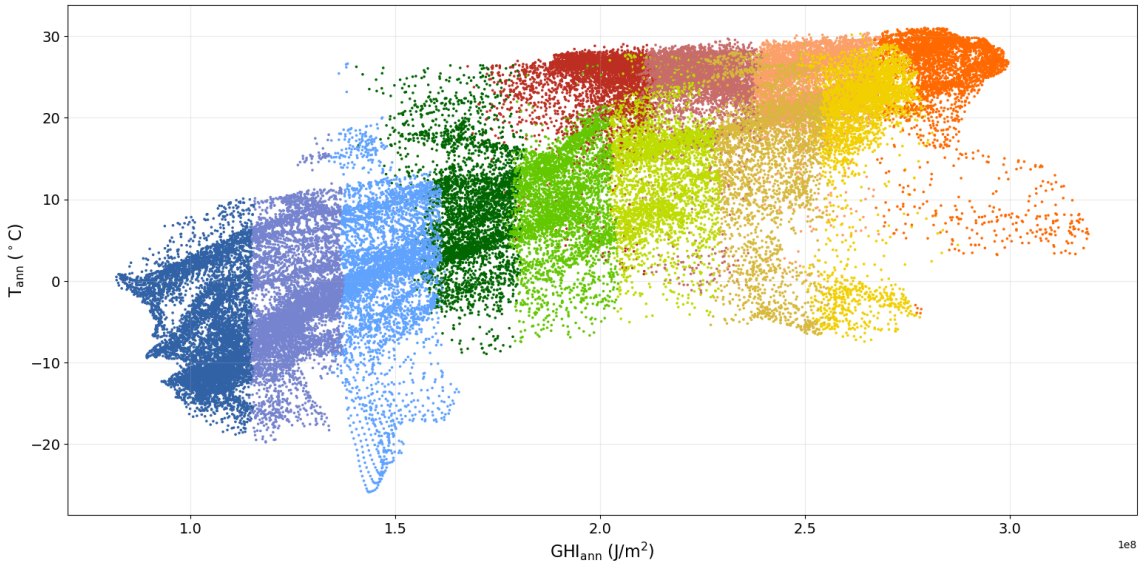


Figure 4.5: Scatter plot for the pair GHI_{ann} - T_{ann}. Result obtained using four features and 12 clusters.

Following the previous reasoning, it could be expected that clusters covering a wide range of temperatures, such as the bluish ones, will be partitioned if a higher number of clusters is fixed. This could result in a better classification. To corroborate this hypothesis, Figure 4.6 shows the pair GHI_{ann} - T_{ann} for 12, 13, 14, and 15 clusters.

Indeed, as Figure 4.6 shows, when increasing the number of clusters to 13, the effect of temperature partitions the dark green cloud. A new dark red group is formed, with the same level of irradiation but a higher T_{ann}. The 14^{th} cluster is the high-temperature version of the golden group. However, this time is created from the light salmon cluster, which is divided into two groups with the same T_{ann} but different GHI_{ann}. Finally, when going from 14 to 15 clusters, a purple group stems from the light blue one. Again, temperature is the reason.

Based on the previous paragraph, 15 clusters seem a better choice than 12. The new clusters identify relevant climate regions, significantly improving the classification's accuracy. They constitute independent clouds of points and have clearly defined properties. Even the white cluster, although it might look small and irrelevant, has a size of 3038 points[1]. This is comparable to other big groups like the dark green, whose size is 3742 points. In this regard, the size of a cluster cannot be estimated based on these plots. Furthermore, the white region might be seen as a necessary step to achieve the purple climate region.

[1]It is reminded that here a point is a space of the grid with resolution 0.5° latitude by 0.5° longitude. The surface (km^2) of this square depends on the location of the planet, being higher at the equator and lower at the poles.

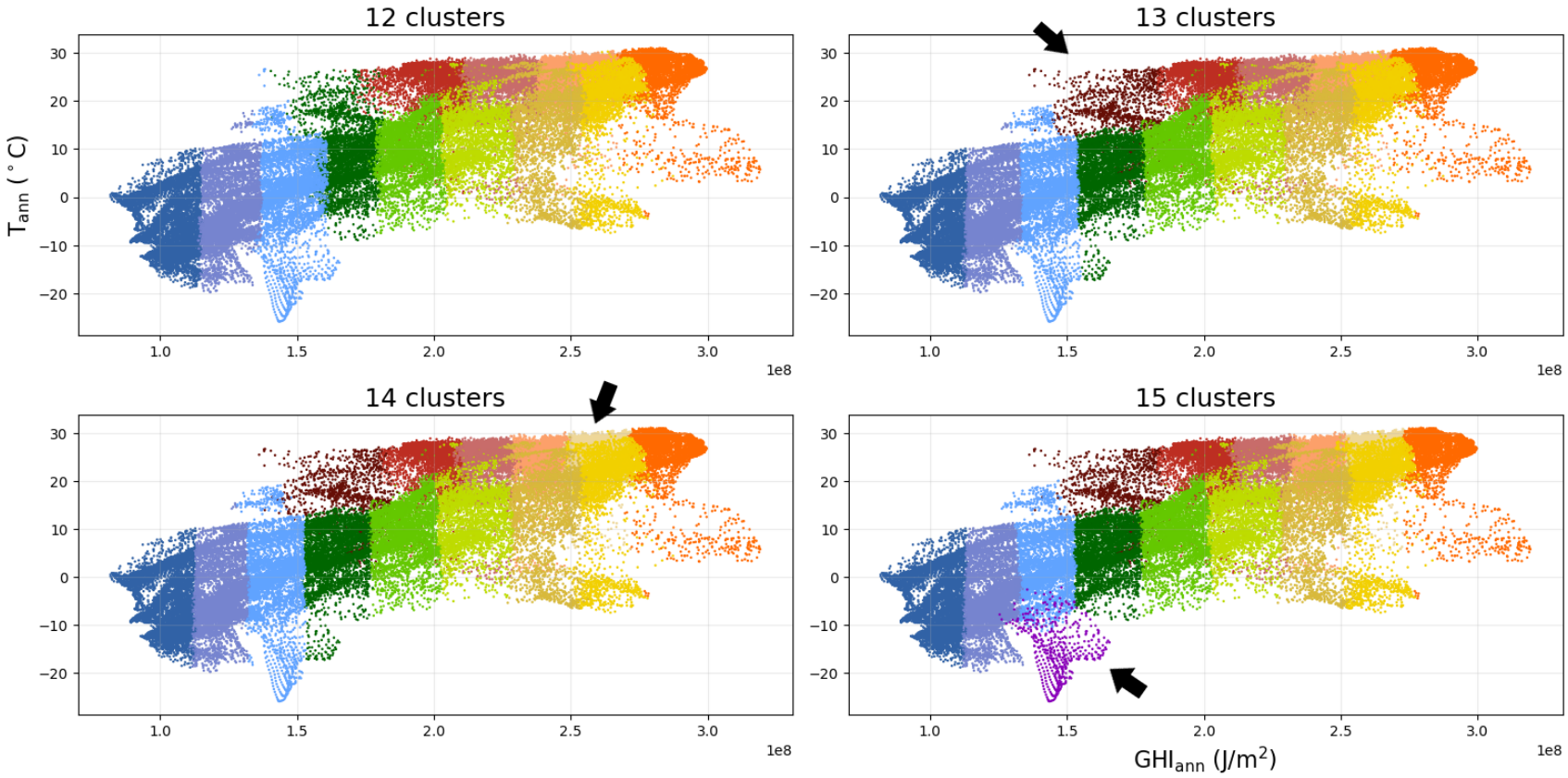


Figure 4.6: Scatter plot GHI_{ann} - T_{ann} for 12, 13, 14, and 15 clusters. When increasing the number of clusters to 13, the effect of temperature partitions the dark green cloud. A new dark red group is formed, with the same level of irradiation but a higher T_{ann}. The 14^{th} cluster is the high-temperature version of the golden group. However, this time is created from the light salmon cluster, which is divided into two groups with the same T_{ann} but different GHI_{ann}. Finally, when going from 14 to 15 clusters, a purple group stems from the light blue one. Again, temperature is the reason. Result obtained using four features.

To conclude this qualitative analysis demonstration, Figure 4.7 shows the world classifications and scatter plots for 16 and 17 clusters. With 16 clusters, another yellow group appears. It could provide a higher level of detail and might be interesting. However, when increasing to 17 clusters, a grey cluster is formed whose significance is very unclear. It overlaps other clusters and does not show clearly defined properties. This suggests that forming 17 clusters or more is not appropriate and that 16 clusters is the ceiling for four features.[2]

The qualitative analysis has been illustrated for four features. A similar procedure can be applied to five, seven, and eight. For instance, Figure 4.8 illustrates similar scatter plots for seven features. Thus, an adequate number of clusters for each combination is found. The classifications proposed are presented in Figure 4.9. For four features, it was decided to select 15 clusters. When using five features, the grey cluster mentioned above is delayed until the 18^{th} cluster, so 17 clearly defined climate regions can be identified. Lastly, 20 groups are formed when using either seven or eight features.

In principle, the higher the number of features, the higher the classification's accuracy. This was already suggested in Table 3.3. More significant clusters can be found when using more variables. On the other hand, using more features makes the qualitative analysis much more challenging. Therefore, there exists a trade-off between accuracy and complexity. Figure 4.9 shows that the classifications based on seven and eight features are almost identical. Thus, seven climate variables are preferable. Conversely, the way the clusters are created using four or five features is remarkably different than for seven. This is very clear for the northern hemisphere. When using four or five features, the clusters are very dependent on the latitude, resulting in an oversimplification for some parts of the planet. In this regard, seven features improve the classification's accuracy substantially in comparison to four or five. Overall, the classification corresponding to seven features and 20 clusters is selected as the final classification. The following section is devoted to analysing the clusters and their properties in depth.

[2]The formation of this grey cluster is explained by the other two features: $GHI_{\max}$, and $GHI_{\min}$. The usefulness of these two features will be shown later.

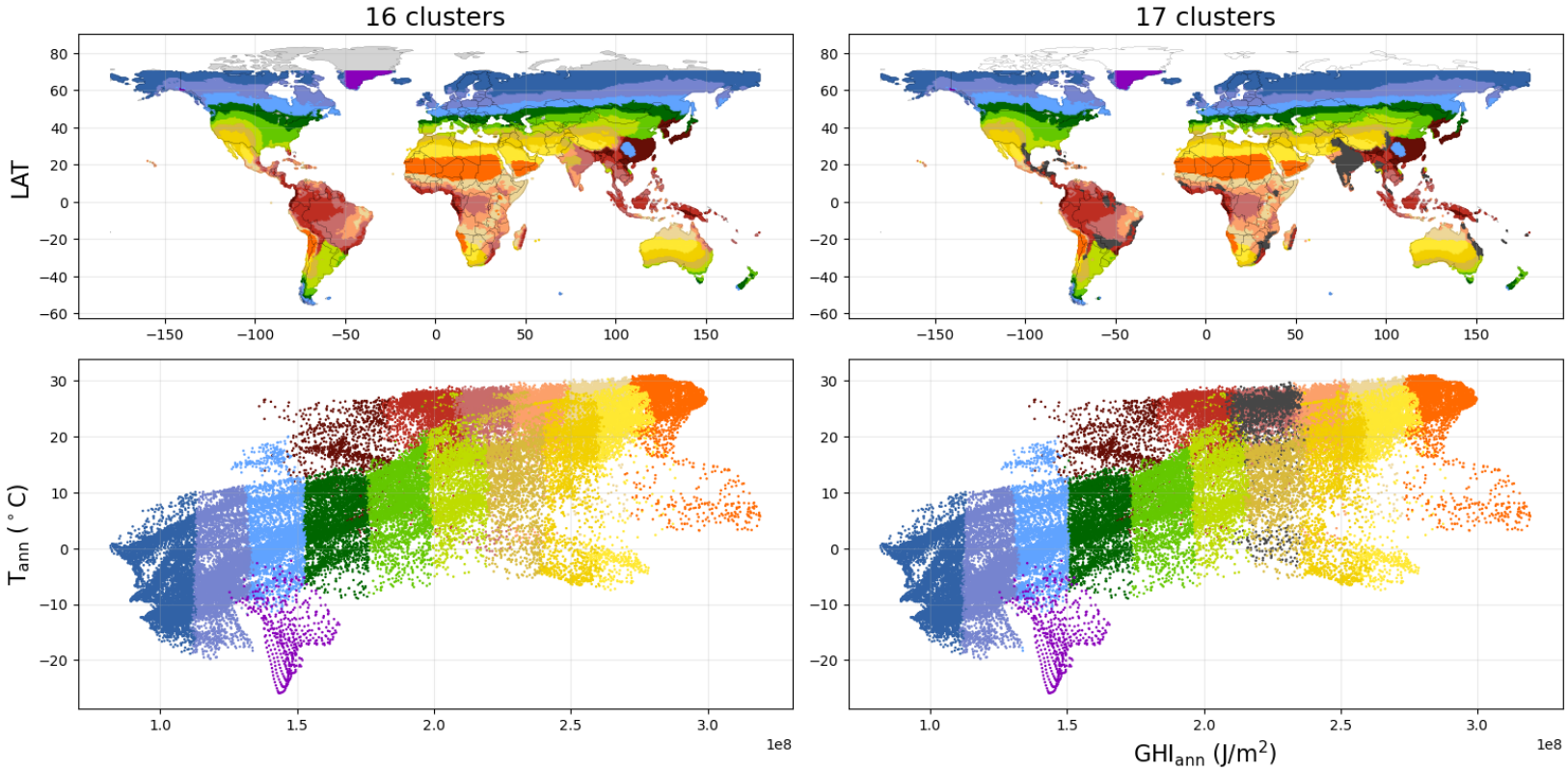


Figure 4.7: World classifications and scatter plots (GHI_{ann} - T_{ann}) for 16 and 17 clusters. With 16 clusters, another yellow group appears. It could provide a higher level of detail and might be interesting. However, when increasing to 17 clusters, a grey cluster is formed whose significance is very unclear. It overlaps other clusters and does not show clearly defined properties. Results obtained using four features.

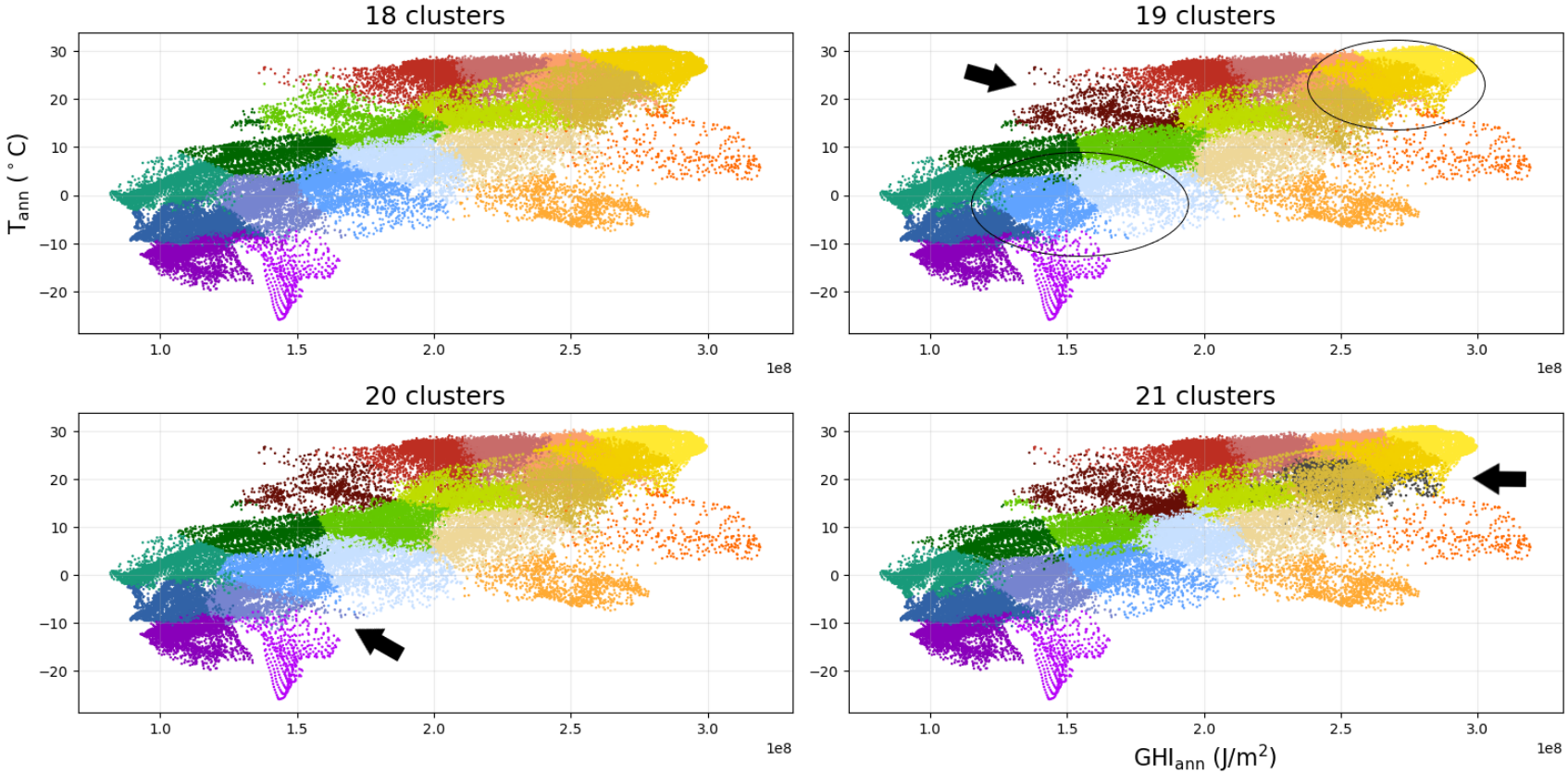


Figure 4.8: Decision procedure for seven features based on the scatter plot GHI_{ann} - T_{ann} for 18, 19, 20, and 21 clusters. When going from 18 to 19, two new clusters appear (reddish and yellowish) while one disappears (bluish). Then this cluster appears again with 20 clusters. The 21$^{\text{st}}$ group (grey) overlaps with other clusters and has an unclear significance.

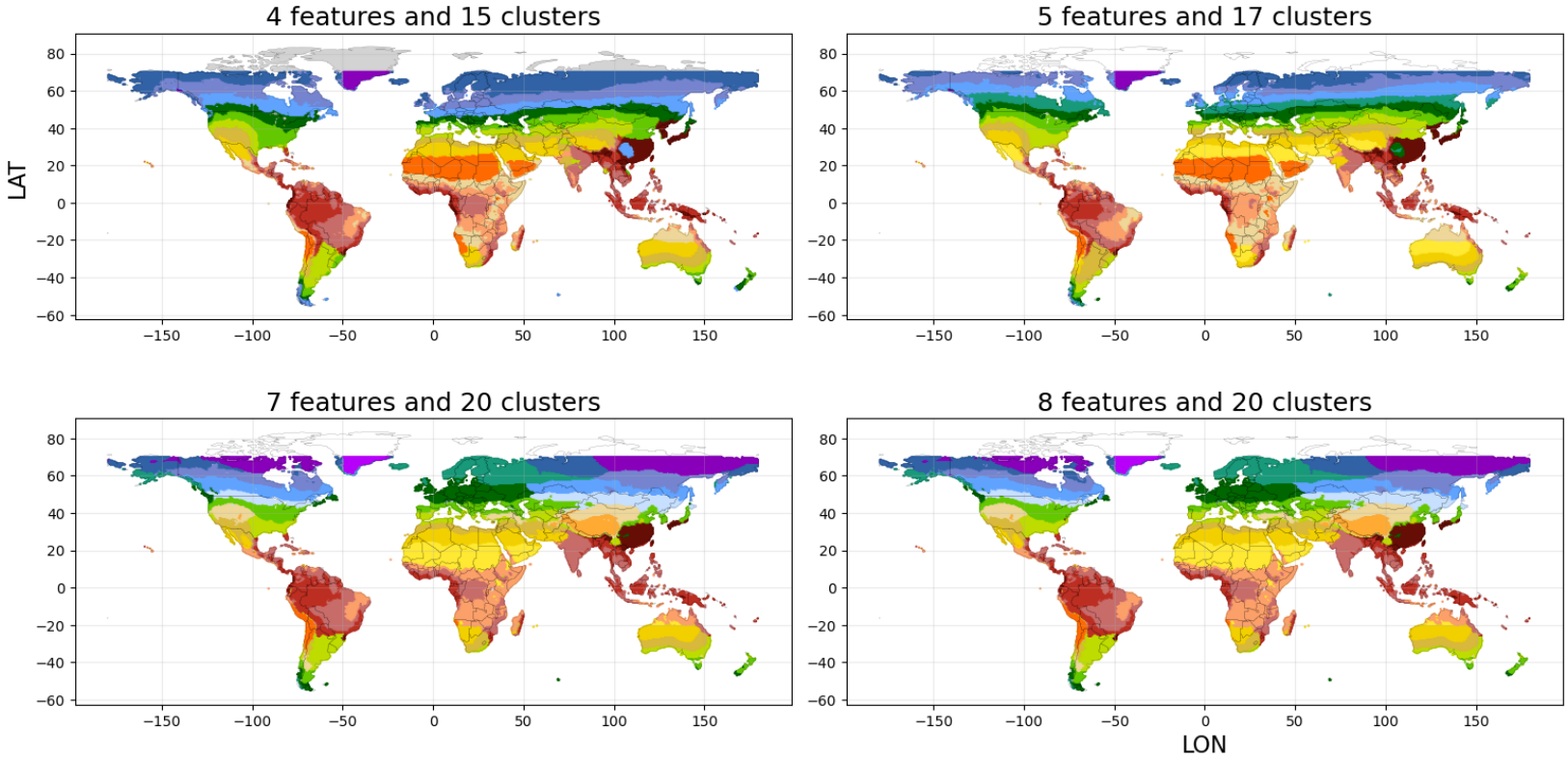


Figure 4.9: Classification proposed for each possible combination of features. The classifications based on seven and eight features are almost identical. Thus, seven climate variables are preferable. Conversely, the way the clusters are created using four or five features is remarkably different than for seven. This is very clear for the northern hemisphere. When using four or five features, the clusters are very dependent on the latitude, resulting in an oversimplification for some parts of the planet. In this regard, seven features improve the classification's accuracy substantially in comparison to four or five.

4.2 ML driven PV-climate classification

The final classification is illustrated in Figure 4.10. As indicated above, it comprises 20 climate regions based on seven features: T_{ann}, T_{max}, T_{min}, GHI_{ann}, GHI_{max}, GHI_{min}, and UV_{ann}.

The names and colors associated with the clusters are indicated by the bar at the right of the figure. They are inspired by the approach followed in KGPV. First, clusters are divided into six climate types: Tropical (Tro), Desert (Des), Mountainous (Mou), Temperate (Tem), Cold (Col), and Polar (Pol). Then, the clusters inside each of these climate types are ordered from minor to greater irradiation. Therefore, both Tro1 and Tro4 are tropical climates, but Tro4 has a higher level of irradiation than Tro1. It is important to note that these numbers only apply inside the same climate type. Hence, even though Pol1 and Tro1 have the same number, they do not have the same level of irradiation. Table 4.1 summarises the cluster's names, centers, and sizes.

Table 4.1: Cluster's names, centers, and sizes. Temperatures are given in °C and irradiations in J/m^2. The size is the number of points constituting the cluster.

Name	$\boldsymbol{T}_{\text{ann}}$	$\boldsymbol{T}_{\text{max}}$	$\boldsymbol{T}_{\text{min}}$	$\boldsymbol{GHI}_{\text{ann}}$ ($\cdot 10^8$)	$\boldsymbol{GHI}_{\text{max}}$ ($\cdot 10^7$)	$\boldsymbol{GHI}_{\text{min}}$ ($\cdot 10^7$)	$\boldsymbol{UV}_{\text{ann}}$ ($\cdot 10^6$)	**Size**
Tro4	24.52	27.08	21.58	2.49	2.43	1.75	14.83	4069
Tro3	25.64	28.1	22.85	2.23	2.23	1.51	13.59	4782
Tro2	25.03	26.68	22.8	2	2	1.33	12.5	3658
Tro1	17.57	26.34	7.5	1.63	1.85	0.88	10.33	1009
Des3	27.87	33.81	20.19	2.81	2.73	1.85	16.09	3341
Des2	23.34	31.41	13.84	2.66	2.84	1.44	15.24	3714
Des1	18.53	27.48	9.16	2.46	2.83	1.15	14.22	2913
Mou3	9.43	12.14	5.84	2.86	2.98	1.76	16.4	383
Mou2	-1.38	9.74	-13.55	2.48	2.81	1.28	14.25	1126
Mou1	8.47	21.77	-5.77	2.17	2.67	0.85	12.68	2220
Tem4	17.2	25.72	8.38	2.09	2.57	0.87	12.52	3029
Tem3	10.53	22.71	-2.11	1.81	2.39	0.59	10.95	3274
Tem2	8.27	19.16	-2.54	1.38	2.07	0.25	8.57	2433
Tem1	1.83	15.5	-11.13	1.06	1.91	0.05	6.67	2893
Col4	2.8	19.47	-16.08	1.74	2.31	0.50	10.45	2840
Col3	0.24	16.98	-18.22	1.42	2.13	0.25	8.7	3691
Col2	-5.68	14.45	-28.86	1.28	2.11	0.13	7.85	3201
Col1	-5.81	13.35	-23.43	1.05	1.98	0.02	6.54	2811
Pol2	-16.14	-3.74	-26.17	1.47	3	0.01	8.4	451
Pol1	-12.1	-11.7	-34.83	1.09	2.08	0.01	6.72	3580

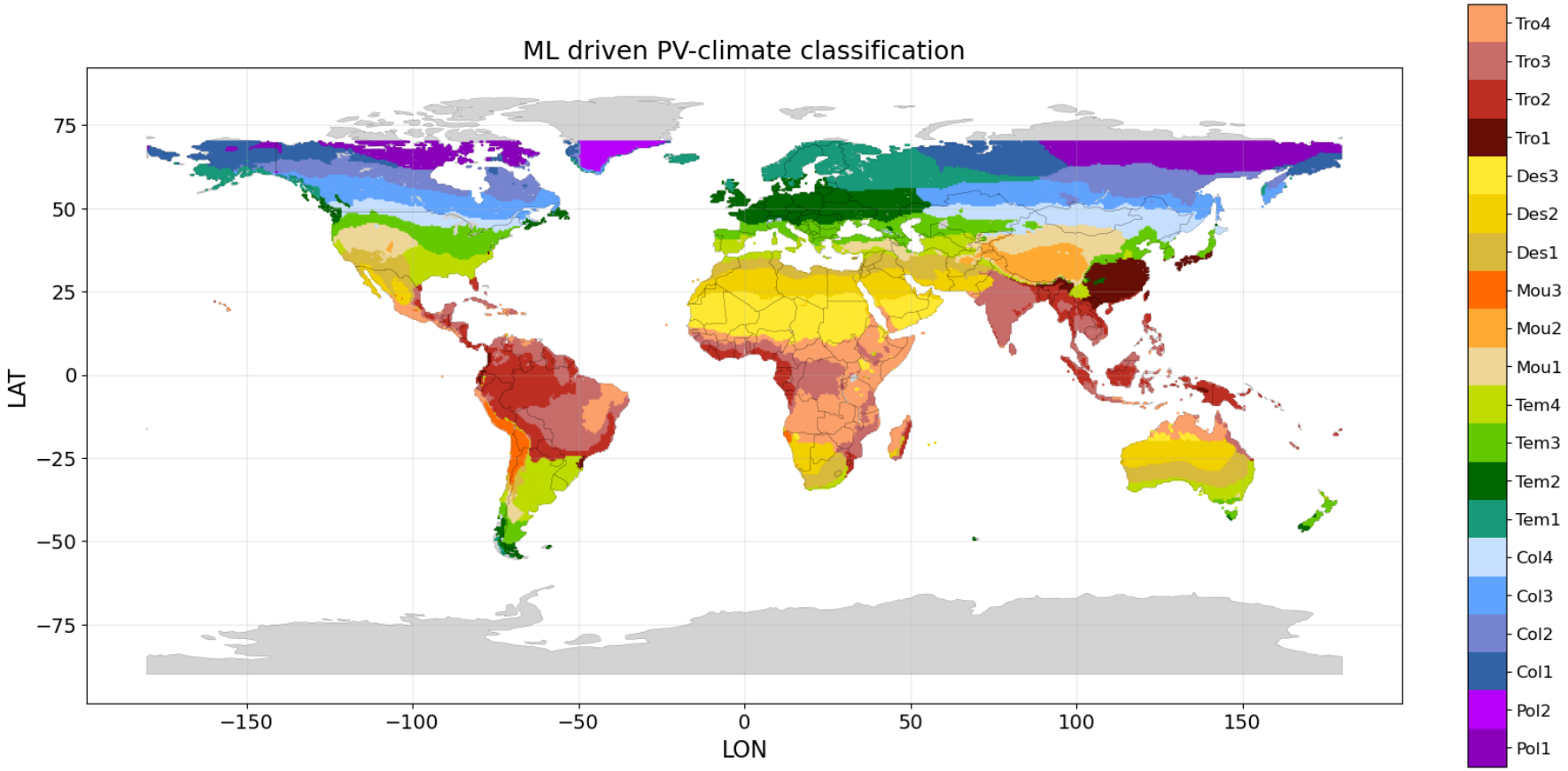


Figure 4.10: ML driven PV-climate classification. First index: Tro-Tropical, Des-Desert, Mou-Mountainous, Tem-Temperate, Col-Cold, Pol-Polar. The second index orders from minor to greater irradiation the clusters inside a particular climate type.

Even though the names are inspired by the KGPV scheme, they are ultimately justified by the properties of the clusters. These are understood with the help of the cluster's centers and pair plots. First, Table 4.2 summarises the most relevant points of each climate type, similar to those considered in KGPV.

Table 4.2: The climate types considered by the classification with their key points.

Climate type	**Description**
Tropical	High T_{min}. Low seasonal dependence.
Desert	High T_{max}. High GHI_{ann} and UV_{ann}.
Mountainous	Moderate temperatures. High GHI_{ann} and UV_{ann}.
Temperate	Moderate temperatures. High seasonal dependence.
Cold	Low T_{min}. High seasonal dependence.
Polar	Low T_{ann}. Extreme seasonal dependence.

These properties are readily apparent by observing the following scatter plots. Figure 4.11 illustrates the pairs GHI_{ann} - T_{ann} and GHI_{ann} - T_{min}. From the first pair, it is concluded that Polar climates are characterized by a low T_{ann} or that Mountainous regions present a high GHI_{ann} but moderate T_{ann}. On the other hand, the bottom plot shows the low T_{min} of the Cold climates, comparable to that of the Polar regions. Furthermore, the high T_{min} of the Tropical climates can be observed.

To illustrate the seasonal dependence, Figure 4.12 plots GHI_{ann} against GHI_{max} and GHI_{min}. Tropical regions have a high GHI_{min} in comparison to other climates such as Temperate. However, regarding GHI_{max}, the scenario is reversed. This shows the high and low seasonal dependence of the Temperate and Tropical regions, respectively. Similar conclusions apply to Cold and Polar climates. In particular, it is remarkable the extreme seasonal dependence of Pol2.

Lastly, some specific clusters might demand further discussion. These are Tro1, Des3, and Tem1. The first one corresponds to the Tropical climate with the lowest irradiation. However, its T_{ann} and T_{min} are remarkably different than for the other tropical clusters. In this regard, it could seem appropriate to consider it as a Temperate cluster. Indeed, looking at the map, KGPV classifies these regions as Temperate. However, the scatter plots suggest that Tropical is a more logical decision. In this sense, this cluster might be considered a transition from Temperate to Tropical. A similar case occurs with Tem1, which might be understood as a transition from Cold to Temperate. The dark red and greenish-blue colours assigned to these clusters make these ideas more visual.

On the other hand, Des3 is classified as a Desert climate, but its temperature and seasonal dependence are not that different from Tropical. To clarify the classification of this cluster as Desert, precipitation might be considered. Figure 4.12 shows the scatter plot for the pair GHI - P_{ann}. Since P_{ann} is not a feature directly considered by the clustering algorithm, the clusters are less clearly defined in this figure. However, it can be seen that Tropical regions certainly have a level of precipitation higher than Des3. This completely justifies the desert climate type. It is worth

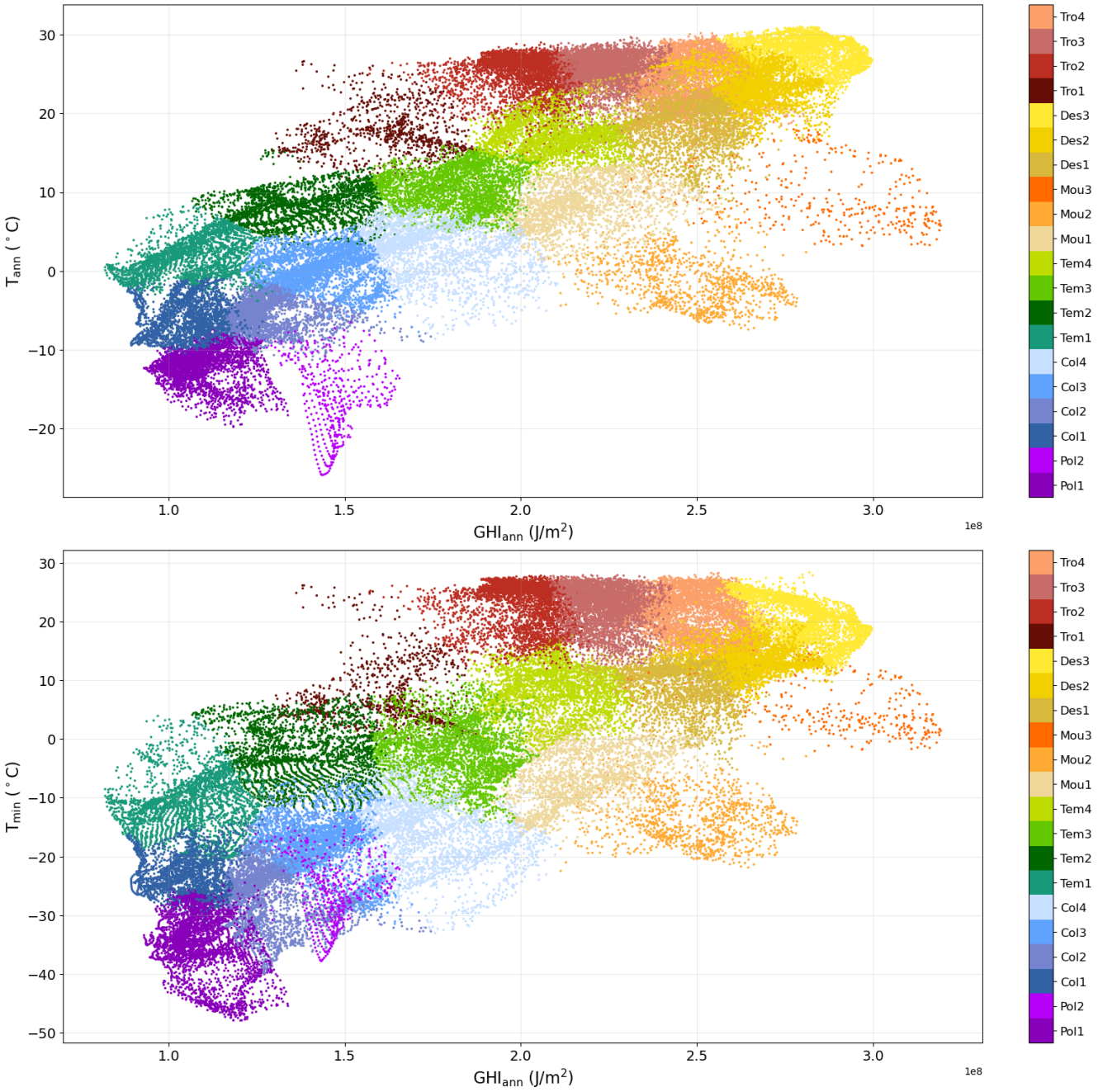


Figure 4.11: Scatter plots for the pairs GHI_{ann} - T_{ann} (above) and GHI_{ann} - T_{min} (below). Polar climates are characterized by a low T_{ann} while Mountainous regions present a high GHI_{ann} but moderate T_{ann}. The bottom plot shows the low T_{min} of the Cold climates, comparable to that of the Polar regions. The high T_{min} of the Tropical climates can be observed.

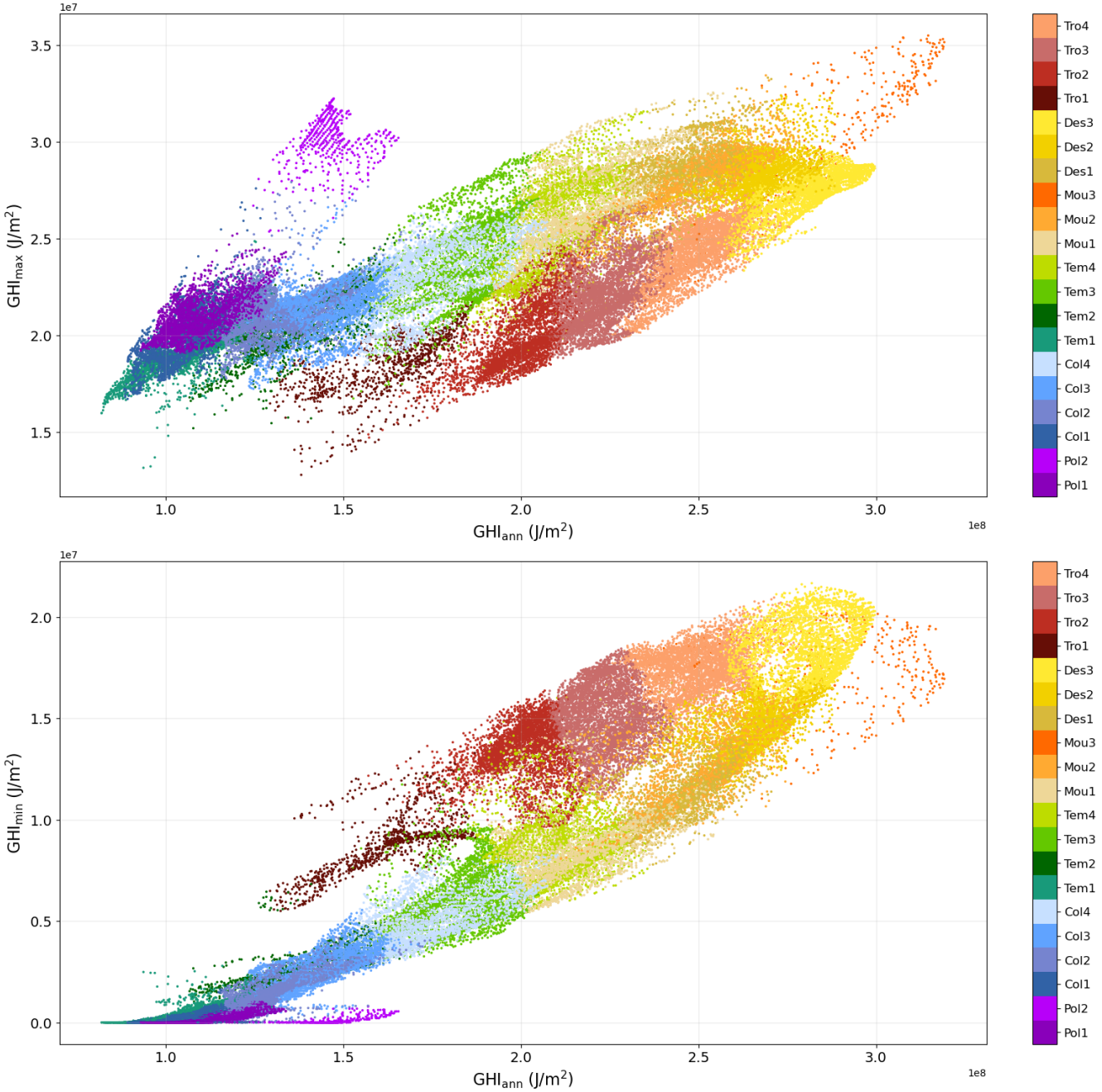


Figure 4.12: Scatter plots for the pairs GHI_{ann} - GHI_{max} (above) and GHI_{ann} - GHI_{min} (below). Tropical regions have a high GHI_{min} in comparison to other climates such as Temperate. However, regarding GHI_{max}, the scenario is reversed. This shows the high and low seasonal dependence of the Temperate and Tropical regions, respectively. Similar conclusions apply to Cold and Polar climates.

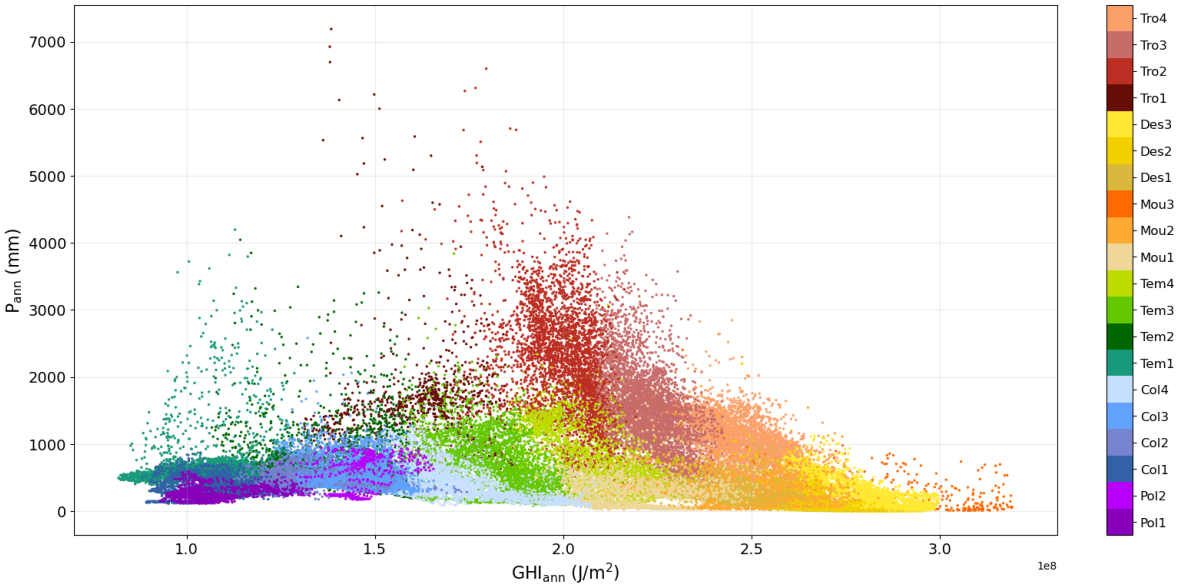


Figure 4.13: Scatter plot for the pair GHI_{ann} - P_{ann}. Tropical regions certainly have a level of precipitation higher than Des3.

pointing out that the algorithm does not need precipitation for making this distinction.

4.3 Assessment

This chapter concludes with a reflection on the validity of the model and a comparison with KGPV.

The objective was to create a classification based on the most relevant climate variables to the specific energy yield. Has this objective been achieved? In principle, the feature selection procedure developed in Chapter 3 assures it. Nevertheless, a final check never hurts. For this purpose, Figure 4.14 illustrates the scatter plot for the pair GHI_{ann} - Y_f. If the climate variables were effectively related to the specific energy yield, the clusters would be expected to show this relationship too. Certainly, this seems to be true. Clusters Mou2 and Mou3 have the highest specific energy yield, followed by Mou1 and the Desert regions. It is interesting to note that even though the Desert clusters have a higher GHI_{ann} than Mou1, they have a similar specific energy yield. The impact of temperature is here evident.

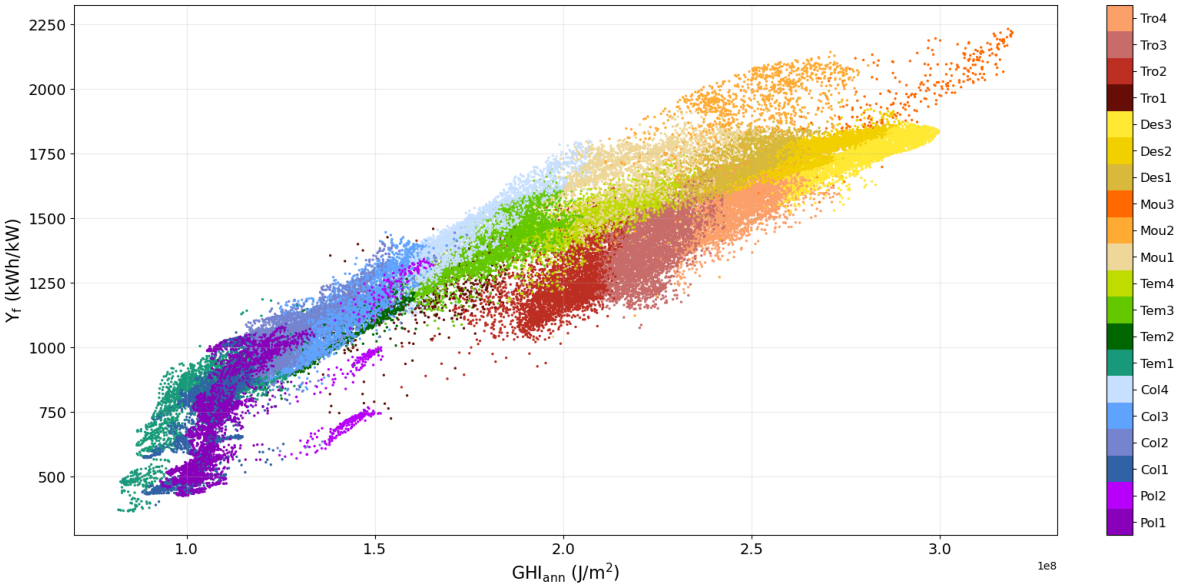


Figure 4.14: Scatter plot for the pair GHI_{ann} - Y_f.

Following the previous reasoning, the impact of temperature on Tropical climates is very severe. For instance, Tro4 has a similar level of irradiation to Des1 and Mou2, but a remarkably lower Y_f. Of course, considering regions with an equivalent range of temperatures, the higher the irradiation, the higher the specific energy yield. This is clearly illustrated by comparing Tro3 and Tro4. Lastly, the relationship between the clusters and the specific energy yield seems less clear for regions with low values. This is in accordance with the conclusions drawn in Chapter 3, where it was seen that the predictions for these regions are less accurate.

A glance at KGPV (Figure 1.2) and ML driven PV-climate classification (Figure 4.10) makes

it evident that both classifications present several similarities and differences. To compare both schemes, besides observing directly the world classifications, it is insightful to analyse their scatter plots for the pair GHI_{ann} - T_{ann}. These are illustrated in Figure 4.15.

Broadly, the same types of climates are distinguished: Tropical, Desert, Temperate, Cold, and Polar. However, the third climate is different. KGPV defines a group called Steppe (C), which is not clearly identified in this work. By contrast, a Montaunious climate type has been proposed instead. KGPV classifies these regions as Polar, because of the low temperatures. However, their high irradiation strongly suggest considering an independent group. These Mountainous regions are characterised by a high GHI_{ann} and low temperatures, which make them the regions with the highest PV performance and should not be mixed with the Polar regions of the northern hemisphere.

It is important to note that three Mountainous regions have been proposed: Mou1, Mou2, and Mou3. Mou2 is what KGPV considers as Polar with low irradiation (FL). However, it has been shown that the GHI_{ann} for this region is relatively high. Mou3 regions are classified in KGPV either as FL or BK (desert and very high irradiation). Since precipitation is not a criterion for developing the classification in the present work, this distinction is not made. Mou1 is a bit trickier. The cluster's centers and plots suggest that this cluster is not a pure Mountainous region. In particular, the high T_{max} stands out. Indeed, this group can be understood as an intermediate step between Desert and Mountainous. In fact, in KG these regions are classified as cold arid. In KGPV, this group is not clearly identified.

On the other hand, KGPV proposes a total of 12 groups[3], while in this work 20 clusters have been identified. Consequently, more subdivisions are considered inside each climate type and a higher level of detail is achieved. For instance, KGPV distinguishes solely between two Tropical climates (AH and AK), in contrast to the four regions found here. Furthermore, since the methodologies are totally different, even equivalent clusters present disparate shapes. Hence, even though both classifications distinguish similar Temperate clusters, they are drawn very differently.

Finally, there is a fundamental difference between KGPV and ML driven PV-climate classification. It is seen that the latter presents homogeneous and even clusters, with their borders clearly defined. By contrast, in KGPV, particular and small regions might be classified separately inside a big cluster, resulting in more complex borders. This can be observed both in the scatter plots and the world classifications. In this regard, the west coast of North America constitutes a great example. It is important to remember here that KGPV applies to every point a criteria defined beforehand while in this work the classification is created using a clustering algorithm (k-means). It might be thought that k-means fails to identify fine details. However, this is not necessarily the case, as shown in Chapter 5, being the nature of the climate variables used the principal factor. Ultimately, even clusters in the scatter plot do not necessarily result in homogeneous climate zones in the world classification. In any case, this difference might be seen as a benefit or a drawback. On the one hand, clean divisions are easier to understand and apply in practice. On the other, this way of creating the clusters might overlook relevant details. This point might require further study and other clustering algorithms may be investigated.

[3]To be fair, it considers 24 groups, but 12 of them are neglected based on a land-surface ratio and population density criterion. Nevertheless, this does not affect the point of the paragraph.

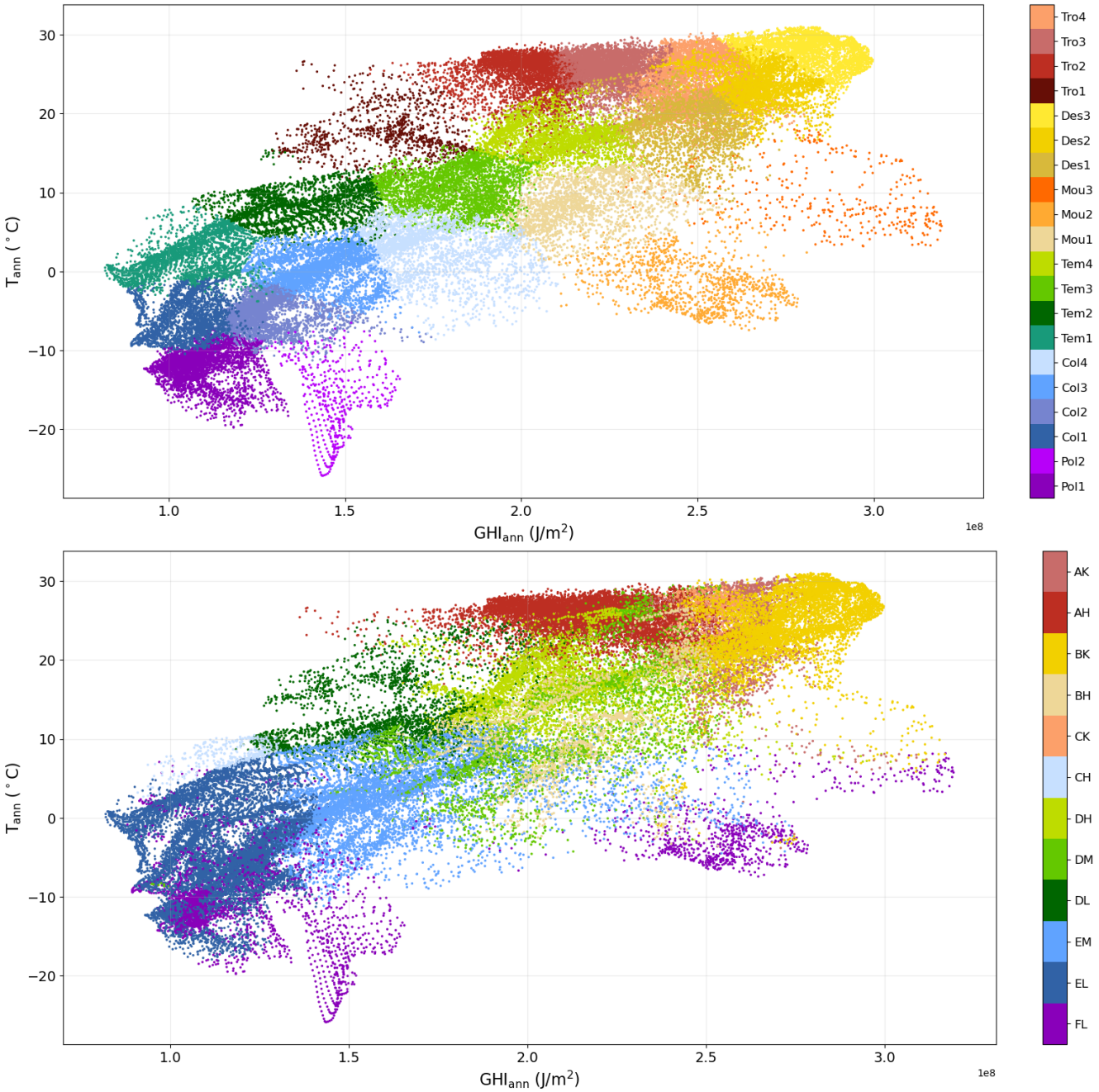


Figure 4.15: Scatter plots for the pair GHI_{ann} - T_{ann} for ML driven PV climate classification (above) and KGPV (below).

4.4 Chapter summary

This chapter explained the clustering method and the final classification proposed in this work to relate climate and specific energy yield.

- K-means is implemented to create the classification.
- The number of clusters and features is optimised following a qualitative analysis. The pair plot proves to be a very insightful tool.
- First, an optimum number of clusters is fixed for each number of features: 15 clusters for four features, 17 clusters for five, and 20 clusters for either seven or eight features.
- Secondly, a trade-off between accuracy and complexity is required to select the final number of features. Seven features offer a remarkable improvement in comparison to four or five features, while the difference with eight features is negligible. Therefore, the classification corresponding to seven features and 20 clusters is selected.
- The clusters are divided into six climate types: Tropical (Tro), Desert (Des), Mountainous (Mon), Temperate (Tem), Cold (Col), and Polar (Pol). Then, the clusters inside each of these climate types are ordered from minor to greater irradiation.
- The classification shows a satisfactory correlation with the specific energy yield. Mou2 and Mou3 have the highest values. The impact of temperature on Tropical climates is very severe. The relationship between the clusters and the specific energy yield seems less clear for regions with low values.
- The introduction of the Mountainous region is the main difference to the KGPV scheme. Moreover, 20 groups are identified, in contrast to the 12 final zones proposed by KGPV. Lastly, since the methodologies are totally different, even equivalent clusters present disparate shapes. The level of detail of the classification might require further study.

Chapter 5

Reflections on degradation

A new PV-climate classification has been proposed based on the relationship between climate and specific energy yield. Nevertheless, other criteria might be applied. In this chapter, the relationship between climate and degradation is explored. For this purpose, the dataset developed in Chapter 2 with the degradation rates is utilised.

The procedure conducted in Chapters 3 and 4 for the specific energy yield is repeated for degradation. However, this chapter does not aim to propose a final classification as was done in Chapter 4. Instead, the objective is to illustrate some tentative results. These pose interesting questions and challenges which must be considered when dealing with degradation. Furthermore, it provides further insights into the methodology applied in this project.

The chapter is divided into two sections. The first section deals with the feature selection step, similar to Chapter 3. It is shown that Linear Regression is not applicable to studying degradation. The search for an alternative demands a further understanding of the methodology. Random Forests, an algorithm able to predict nonlinear behaviours, is analysed. In the second section, an exploratory classification for degradation is shown. The ability of k-means to identify fine details is discussed. Finally, the inadequacy of using Random Forests for weighting the features and the complexity of degradation are exposed.

5.1 Feature selection

In principle, the same method implemented for the specific energy yield could be applied to selecting the most relevant climate features for degradation. Hence, the idea is to propose a set of possible combinations and evaluate them using a Linear Regression model. The dataset with the climate variables (features) and the known degradation rate values (targets) was created in Chapter 2. Again, the combinations are based on the Pearson coefficients, RFE results, and technical expertise. Table 5.1 shows the Pearson coefficient and RFE index for each feature.

Table 5.1: Pearson coefficients and RFE indexes for the degradation rate. Random Forest was used for the RFE procedure.

Feature	**Pearson coefficient**	**RFE index**
T_{ann}	0.84	2
T_{max}	0.62	7
T_{min}	0.87	1
DTR_{ann}	-0.03	10
P_{ann}	0.56	3
P_{min}	0.18	12
RH_{ann}	0.14	4
GHI_{ann}	0.48	11
GHI_{max}	-0.07	9
GHI_{min}	0.73	8
UV_{ann}	0.56	6
WS_{ann}	-0.4	5

However, when the models are evaluated, the errors in the predictions turn out to be too high for any combination or number of features. To illustrate this, Figure 5.1 compares the predictions of the model with the known values, or targets, when using all features (which is always the most accurate combination). This figure strongly suggests that the model is not performing properly. Indeed, the MAPE for this case is almost 20 percent, which might be considered unacceptable even for this feature selection procedure. Therefore, it is concluded that implementing Linear Regression to select the most relevant features for degradation is not an adequate approach.

This result is a consequence of the complex and nonlinear behaviour of degradation. Different combinations of factors (PV technology, climate, system size, tilt, orientation, and outdoor exposure period) trigger different degradation mechanisms that are difficult to model ([7], [8]). Indeed, usually, the lifetime of the PV modules is shorter than expected when exposed to outdoor conditions [32]. In this regard, it is important to remember that only three degradation mechanisms for

mono-crystalline silicon are being considered in this work.

The previous setback imposes the search for an alternative approach. A supervised learning algorithm able to determine the importance of the features (weights) and work with nonlinear dependencies is needed. To the best knowledge of the author, Random Forests could be a promising option. Random Forests is a powerful supervised learning algorithm that essentially consists of a collection of decision trees [36]. It has the ability to make accurate predictions for very complex datasets. Furthermore, it determines the importance of each variable giving them a weight. This weight is a number between 0 and 1, where 0 means nil importance while 1 indicates a perfect ability for making predictions. The sum of all weights, or feature importances, always equals 1 [36].

Random Forests has already been used in this work, via RFE, to guide the decision-making process regarding the possible climate feature combinations. Nevertheless, what is being contemplated here is that Random Forest could also be implemented to make the predictions, calculate the weights, and, in the end, select the climate features. It would simply replace Linear Regression.

Of course, implementing Random Forests has some disadvantages. It is a much more complex algorithm than Linear Regression, and, consequently, it has higher computational times. Moreover, there is a subtle point that requires particular attention. The weights calculated by Random Forests are a measure of the feature importance in the sense that they rate how important are for the decision trees. The meaning of these weights and their suitability for making the classification using k-means is doubtful. In this regard, Linear Regression seems more appropriate since the weights have a clear mathematical significance, the "slope" of each feature, and their applicability to k-means, which is based on Euclidean distances, seems more logical. This point requires further analysis and should be carefully considered for selecting other algorithms, not only for Random Forests.

Unfortunately, studying these questions in depth would be out of the scope of this project. Nevertheless, at least, the concern for the particular case of Random Forest is explored. Thus, it is decided to continue with Random Forests to select the most relevant climate variables for degradation. Then, a classification is created in the following section, providing further insights.

Since Random Forests requires higher computational times, the number of possible combinations studied is reduced with respect to the specific energy yield. In this case, 39 combinations are evaluated, in contrast to the 79 combinations for the latter. Table 5.2 summarises the best combination for each number of features, together with their respective errors (R^2, RMSE, MAE, and MAPE). The rest of the combinations can be found in Appendix III.

Random Forest improves the quality of the predictions significantly. Now, using all features, the MAPE is 4.4 percent. Moreover, considering only five features, a MAPE of 4.9 percent is achieved. The targets and predictions for five features are illustrated in Figure 5.2. The accuracy of the model is remarkable. In order to present a tentative classification for degradation and continue with these reflections, the five-feature combination is selected. It offers a low MAPE and a convenient number of features.

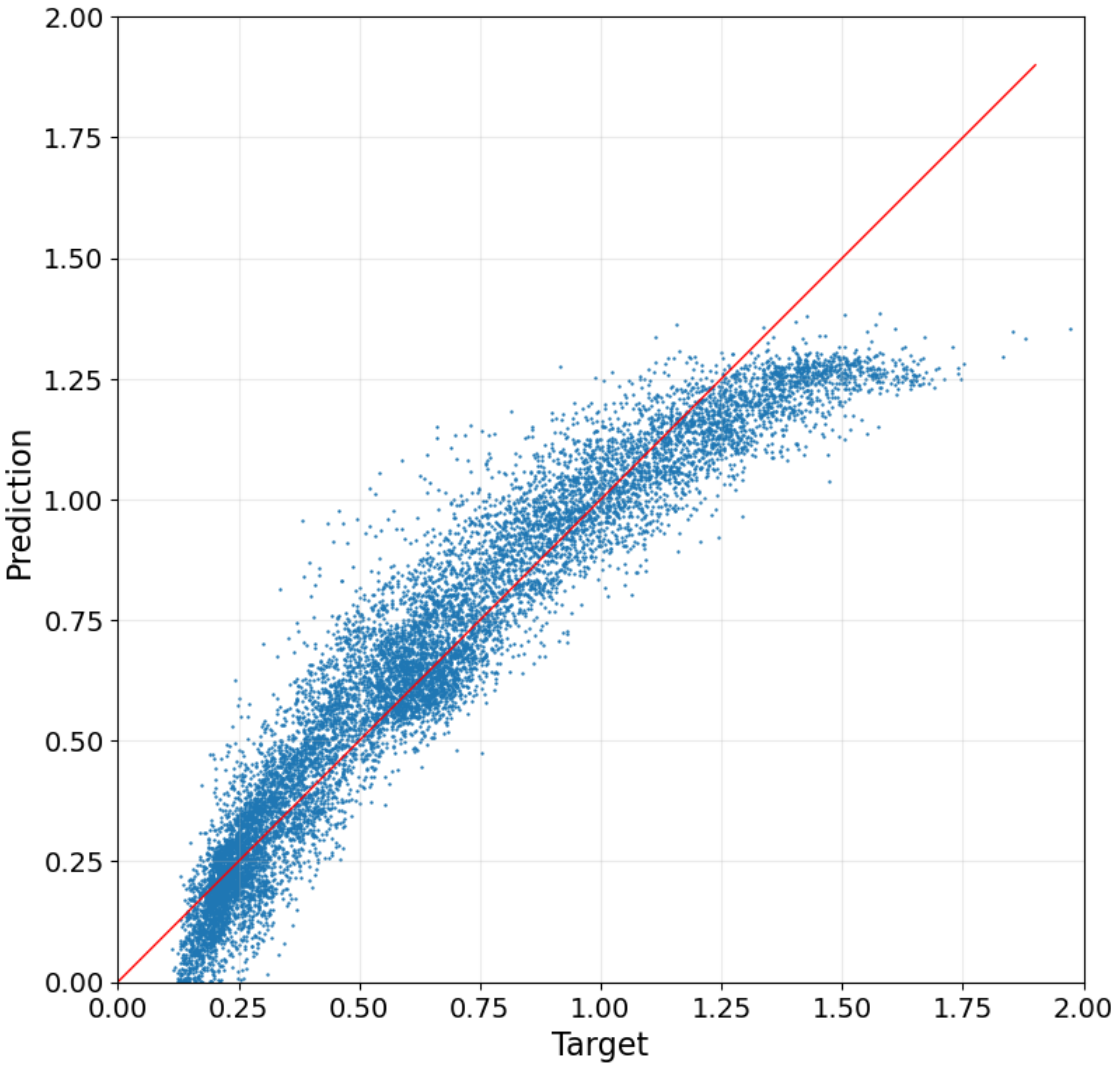


Figure 5.1: Comparison between the targets and predictions of degradation rates using a Linear Regression model based on all features.

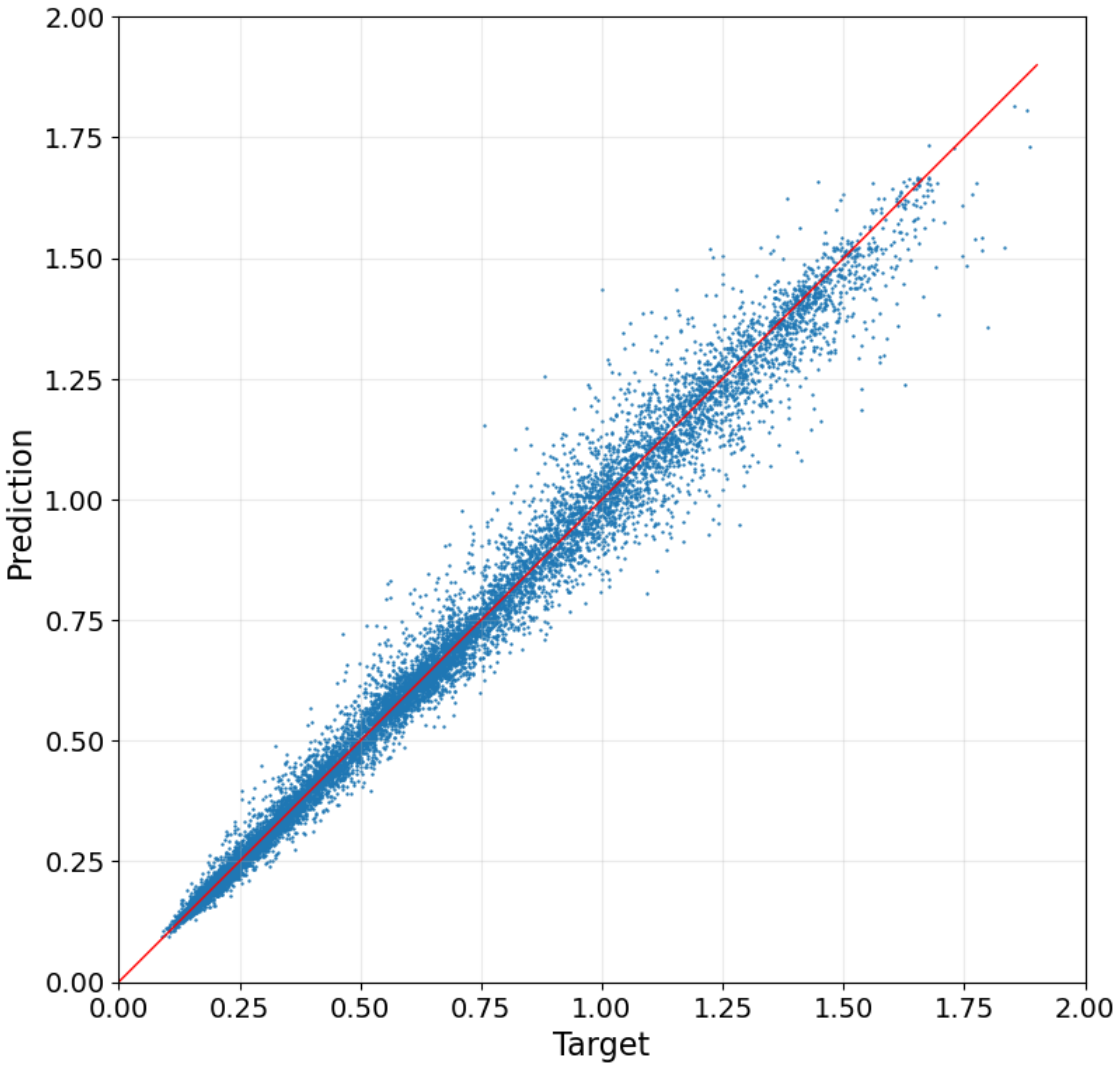


Figure 5.2: Comparison between the degradation rates targets and predictions made by Random Forests using five features.

Table 5.2: Optimum combination for each possible number of features. Results obtained using Random Forests. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	WS_{ann}	**Random**	**R^2**	**RMSE**	**MAE**	**MAPE (%)**
0													1.000	-0.456	0.465	0.375	88.4
1			1.000											0.818	0.163	0.116	21.5
2	0.811						0.189							0.935	0.098	0.064	10.6
3	0.790						0.094					0.116		0.969	0.068	0.043	7.2
4	0.784						0.083				0.018	0.116		0.981	0.053	0.034	5.5
5		0.848	0.038				0.046				0.014	0.054		0.983	0.050	0.031	4.9
6		0.035	0.849				0.043			0.011	0.011	0.052		0.985	0.048	0.029	4.7
7	0.134		0.696		0.100		0.026			0.006	0.010	0.027		0.986	0.045	0.028	4.7
8	0.129	0.006	0.697		0.101		0.025			0.006	0.009	0.026		0.986	0.046	0.028	4.5
9	0.129	0.006	0.699		0.096		0.025		0.004	0.006	0.007	0.028		0.986	0.045	0.027	4.5
10	0.129	0.006	0.696	0.003	0.098		0.026		0.004	0.005	0.008	0.026		0.987	0.044	0.027	4.4
11	0.130	0.006	0.698	0.003	0.096		0.025	0.003	0.003	0.005	0.005	0.025		0.987	0.044	0.027	4.4
12	0.128	0.005	0.698	0.003	0.097	0.002	0.025	0.003	0.003	0.005	0.005	0.026		0.987	0.044	0.027	4.4

5.2 Clustering

The objective of this section is to reflect on some challenges found when dealing with degradation. Therefore, in contrast to the specific energy yield, where an extensive qualitative analysis was performed to optimise the number of clusters and features, here just one classification is presented. This is enough to illustrate the desired ideas. Thus, Figure 5.3 shows the classification created using five features and eight clusters. There is no specific reason to choose eight clusters. It is just a number that meets the needs of the explanation and keeps the classification relatively simple.

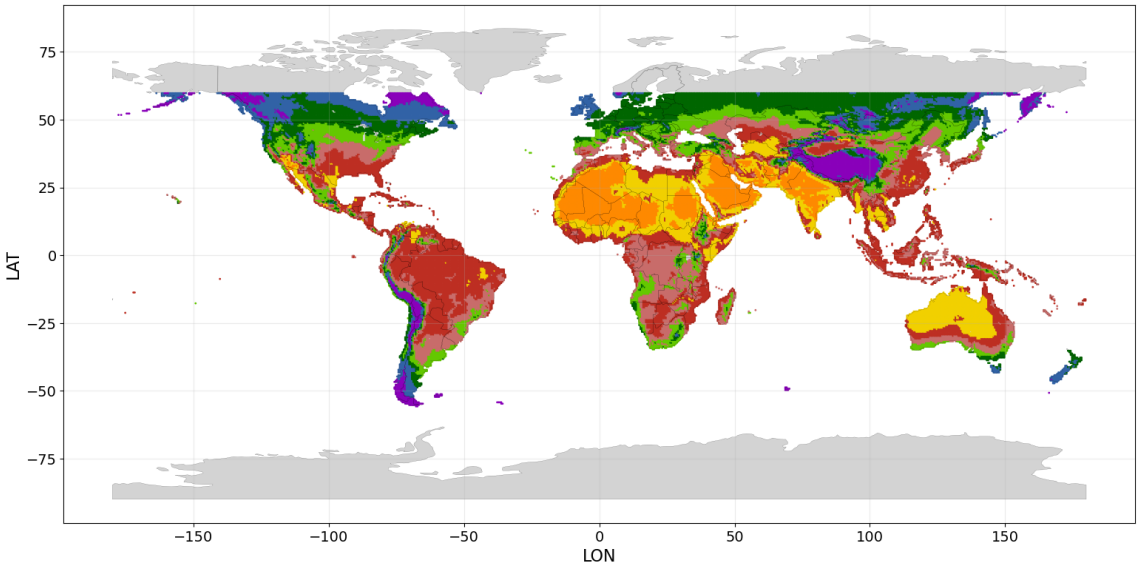


Figure 5.3: Worldwide classification for degradation based on five features and eight clusters.

There is one remarkable difference between this classification and the one developed for the specific energy yield. In Chapter 4, it was mentioned that using k-means might result in homogeneous and even clusters. This was a difference between the classification developed and KGPV. The west coast of North America was put as an example. Now, interestingly, this classification for degradation is identifying very small particularities, just as KGPV does. And only eight clusters have been created! Indeed, if the number of clusters is increased, the level of detail becomes untenable. What is happening here?

To answer this question, it is necessary to understand how the classification is being formed. Table 5.2 presents the weights associated with each of the five climate variables. $T_{\max}$ has a weight much higher than the rest. More specifically, its weight is 0.848, while the weight for the WS_{ann}, the second most important feature, is 0.054. This suggests that the clusters are being formed only considering $T_{\max}$. The following plot shows the scatter plot for the pair $T_{\max}$ - WS_{ann}. The vertical lines confirm the hypothesis.

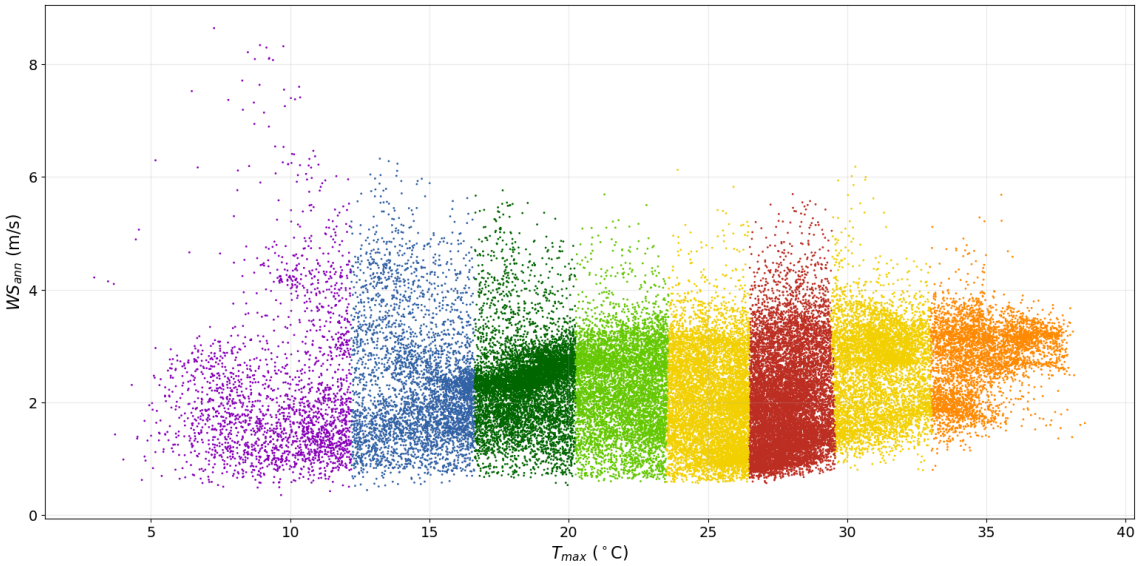


Figure 5.4: Scatter plot for the pair $T_{\max}$ - WS_{ann}.

Using only one climate feature to create the clusters makes the classification very susceptible to small details. Furthermore, $T_{\max}$ is a variable with a significantly higher variability than the UV_{ann}, which was the most relevant feature for the specific energy yield. These two points explain the observed behaviour. Moreover, it is a demonstration of the complexity of the degradation mechanisms and the variables involved in them. Lastly, it is concluded that k-means is able to produce a high-detailed classification in some circumstances. The number of features, their weights, and the nature of the climate variables might be as important for this point as the algorithm itself. Ultimately, even clusters do not necessarily result in homogeneous climate regions in the world classification.

Finally, the validity of this classification might be questioned. Considering such a high weight for the $T_{\max}$ seems suspicious. To assess the classification, Figure 5.5 shows the relationship between the clusters, the $T_{\max}$, and the degradation rate, k. Again, it is seen that the clusters have been determined exclusively by the $T_{\max}$. Moreover, the relationship between the clusters and the degradation rate is almost nil. Therefore, the classification does not seem to have achieved its objective.

It might be thought that the problem is the number of clusters. Perhaps, increasing the number of clusters results in a higher correlation with the degradation rate. To discard this scenario, Figure 5.6 shows the same plot but using 20 clusters. Again, the classification is not very informative.

The weight given to $T_{\max}$ is simply too high. In other words, the weights calculated do not seem appropriate for creating the classification. Therefore, the climate zones found for degradation are not adequate. This result might confirm the concern regarding the use of Random Forests for this purpose. It seems that the algorithm implemented in the feature selection procedure must be

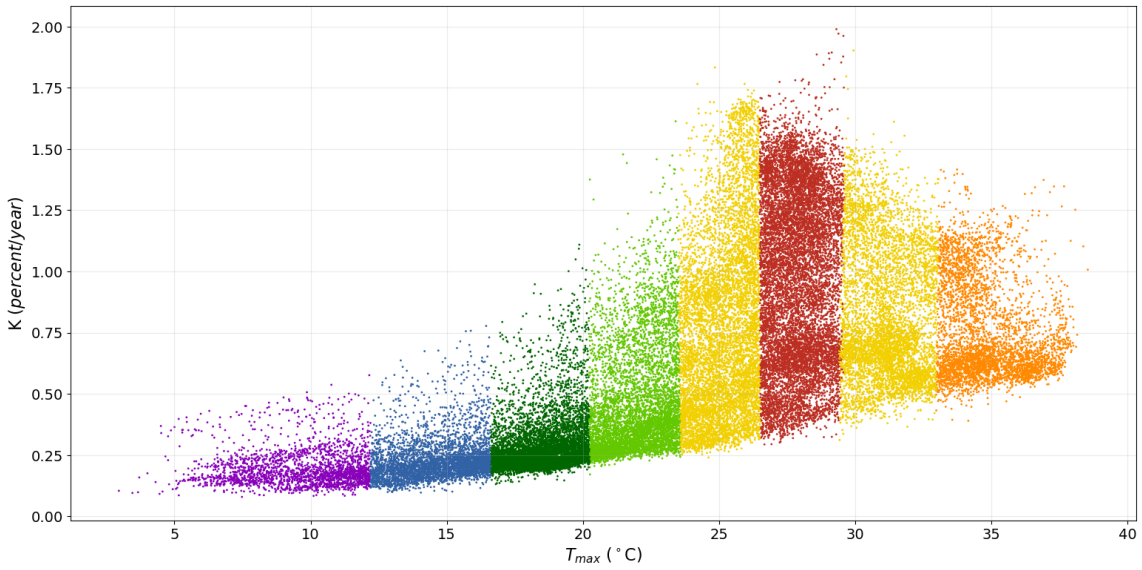


Figure 5.5: Scatter plot showing the relationship between the eight clusters, the degradation rate, and the $T_{\max}$.

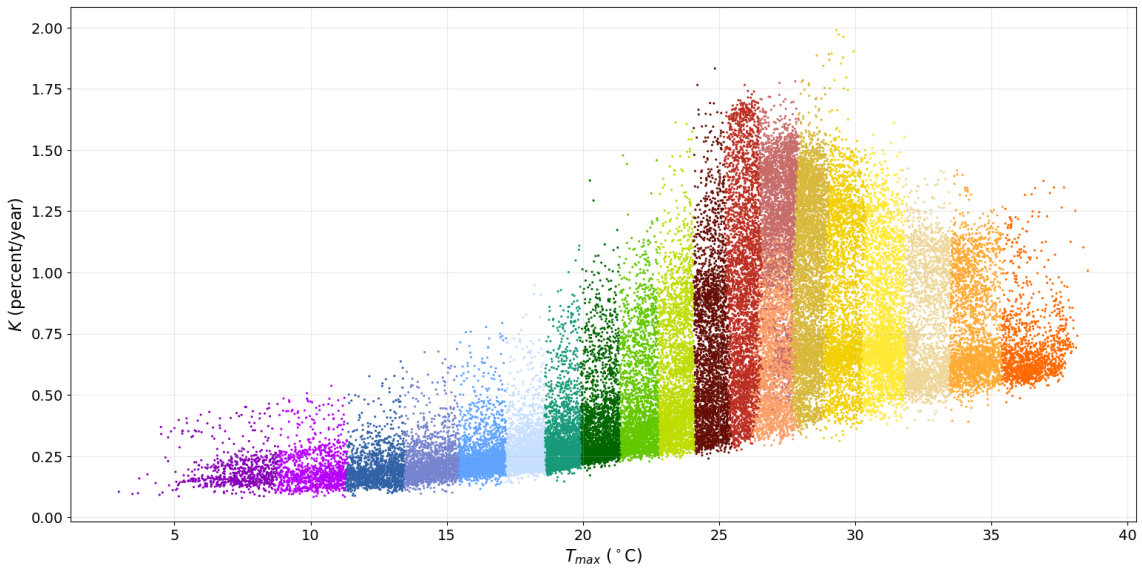


Figure 5.6: Scatter plot showing the relationship between 20 clusters, the degradation rate, and the $T_{\max}$.

mathematically linked to the clustering algorithm used for creating the classification. In this regard, Linear Regression was a good choice for k-means. However, the non-linearity of degradation, demands a different approach.

A promising alternative might be Multivariate Adaptive Regression Splines (MARS). This is an extension of Linear Regression that enables modelling non-linearities [48]. The idea is that the weight of each variable can vary depending on the region of the space. Since the meaning of the weights is the same as in Linear Regression, this algorithm could be implemented together with k-means.

5.3 Chapter summary

Chapter 5 explored the relationship between climate and degradation. Moreover, the methodology developed in Chapters 3 and 4 was tested.

- Implementing Linear Regression to select the most relevant features for degradation is not an adequate approach since MAPEs of 20 percent are obtained. This is a consequence of the complex and nonlinear behaviour of degradation.
- As an alternative, Random Forest was proposed. It was applied to 39 combinations of climate features. The MAPE can be then reduced to 4.4 percent. A five-feature combination was selected to present a tentative classification..
- K-means can produce a high-detailed classification. The number of features, their weights, and the nature of climate variables might be as relevant to this point as the algorithm itself.
- The weights calculated by Random Forest prove to be inadequate for creating a classification using k-means. In particular, the weight given to $T_{\max}$ is too high, resulting in a poor correlation between the clusters and the degradation rate. It seems that the algorithm implemented in the feature selection procedure must be mathematically linked to the clustering algorithm used for creating the classification.
- A promising alternative might be Multivariate Adaptive Regression Splines (MARS): an extension of Linear Regression that enables modelling non-linearities.

Chapter 6

Conclusions

This project has proposed a new PV-climate classification, inspired by Ascencio-Vásquez et al. in KGPV [6]. To achieve this objective, it was necessary to select the most relevant climate variables for PV performance and to establish a classification criterion. Machine Learning was the technique proposed to tackle these issues. In particular, supervised learning was used to identify and weigh the climate variables more correlated to the specific energy yield, while unsupervised learning to create the classification. Lastly, the applicability of this methodology to degradation was explored.

Data collection is the first step in any Machine Learning project. This work required data on three aspects: climate, specific energy yield, and degradation rate. A worldwide grid with a resolution of 0.5° latitude by 0.5° longitude was created. The climate data was extracted from renowned climate research centres and institutions for the period 1991 to 2021. More specifically, 12 climate variables were included in the dataset: T_{ann}, T_{max}, T_{min}, DTR_{ann}, P_{ann}, P_{min}, RH_{ann}, GHI_{ann}, GHI_{max}, GHI_{min}, UV_{ann}, and WS_{ann}. Specific energy yield and degradation rate values, used as targets, are provided by Ascencio-Vásquez et al. in [6] and [7], respectively.

The feature selection procedure consisted in implementing Linear Regression to predict the specific energy yield from the knowledge of the climate variables. The combinations of climate variables that make more accurate predictions are considered more suitable for developing the classification. Furthermore, the weights calculated by Linear Regression provide a logical measure of the features' importance to the model. The high number of possible combinations forces to simplify the procedure. Therefore, 79 possible combinations were proposed based on the Pearson coefficients, recursive feature elimination (RFE), and technical expertise. Then, these were evaluated by comparing their predictions with the known specific energy yield values.

A trade-off between accuracy and simplicity is required to make a final decision. Furthermore, the impact of the error differences on the final classification is difficult to predict. Therefore, four potential candidates were selected. These correspond to the optimums for four, five, seven, and eight features. With MAPEs between 6.1 and 6.7 percent, they perform proper predictions of the specific energy yield, except for very low values, where the error might be higher.

K-means can be implemented to create a classification from the climate variables selected and

their weights. However, an evaluation method is required to assess the results and decide the optimum number of features and clusters. An approach consisting of a qualitative analysis based on several plots, the clusters' centres, and sizes, was followed. First, an optimum number of clusters was found for each possible combination: 15 clusters for four features, 17 clusters for five, and 20 for both seven and eight features. The objective was to obtain clearly-defined and comprehensive groups. Then, the optimum combination can be selected. Again, there exists a trade-off between accuracy and complexity. Since the classifications for seven and eight features are similar, the latter case was discarded. On the other hand, the improvement in accuracy between five and seven clusters is remarkable. Overall, the final classification consists of seven features and 20 clusters.

The names of the groups follow a scheme inspired by KGPV. First, an index distinguishes among six different climate types: Tropical (Tro), Desert (Des), Mountainous (Mou), Temperate (Tem), Cold (Col), and Polar (Pol). Secondly, the clusters inside each of these climate types are ordered from minor to greater irradiation. The classification presents a satisfactory correlation with the specific energy yield. Clusters Mou2 and Mou3 have the highest values, followed by Mou1 and the Desert regions. The impact of temperature in Tropical climates is very severe. For instance, Tro4, despite having the same level of irradiation as Des1 and Mou2, presents a significantly lower specific energy yield. As expected, the classification is less accurate for regions with very low values. Overall, new insights have been found in comparison to KGPV, which only considers 12 clusters. The main difference is the distinction of the Mountainous region, characterised by low temperatures and high irradiation. Lastly, since the methodologies are totally different, even equivalent clusters present disparate shapes.

The methodology implemented for the specific energy yield fails for degradation due to its complexity and non-linear behaviour. In particular, MAPEs around 20 percent are obtained. Therefore, the use of Random Forests instead of Linear Regression was analysed. For this case, the MAPEs are reduced to values lower than 5 percent. Nevertheless, the meaning of the weights calculated by Random Forests, and their applicability to k-means, are doubtful. In this regard, Linear Regression seems more appropriate since the weights have a clear mathematical significance, the "slope" of each feature, and their applicability to k-means, which is based on Euclidean distances, seems more logical. Indeed, a classification created using the results of Random Forests proves to be inconvenient. An inappropriate weight was given to $T_{\max}$, resulting in a classification with a low relationship with the degradation rate.

To improve the performance of the model, this study recommends two things. The first recommendation deals with the feature selection procedure. Here, the analysis was simplified by choosing 79 combinations. Even though this approach produced a satisfactory result, some relevant combinations might have been missed. Thus, implementing an optimisation algorithm such as Particle Swarm Optimization or the Genetic Algorithm is recommended to consider all possibilities. Secondly, a promising approach to integrate non-linear dependencies and, at the same time, keep the logical weights, might be using Multivariate Adaptive Regression Spline (MARS). This is especially recommended for degradation, although it could improve the results for the specific energy yield too.

Acknowledgments

The results contain modified Copernicus Climate Change Service information 2023. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

Appendix

This Appendix consists of three sections. In Appendix I, the feature selection results for the specific energy yield, discussed in Chapter 3, are shown. More specifically, the 79 combinations are illustrated with their corresponding errors. Appendix II explains two quantitative methods studied to optimise the classification: the elbow method and the silhouette coefficient. These approaches are commonly used to select the appropriate number of clusters [39] and might be an alternative to the qualitative analysis developed in Chapter 4. However, as shown, they perform poorly in this particular dataset. Lastly, Appendix III summarises the feature selection results for degradation, discussed in Chapter 5.

Appendix I: Feature selection results for the specific energy yield

Table A.1: Feature selection results for the specific energy yield. A Linear Regression model was implemented. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	Random	R^2	RMSE	MAE	MAPE (%)
1												1.61	0	360.4	301.28	28.3
1								332.3					0.852	139.61	109.14	9.9
1											328.14		0.827	150.54	116.99	10.3
2								279.12	71.77				0.868	131.07	97.66	9.2
2								781.65			-450.6		0.863	133.66	104.06	9.6
2								533.28		-216.81			0.904	112.37	84.05	7.9
2							-50.99	297.04					0.861	134.68	106.07	9.6
2				72.6				281.19					0.869	130.34	102.09	9.3
2	-96.23							410.93					0.875	127.49	97.52	9.4
2		-40.08						362.31					0.856	137.22	106.67	9.8
2			-104.8					414.62					0.885	122.34	94.17	9.0
2					-47.02			333.68					0.869	130.24	101.46	9.3
2						-25.13		327.12					0.856	137.32	107.84	9.8
2			-154.7								454.19		0.888	121	92.4	9.0
3								894.04	-165.25	-471.95			0.923	99.71	77.88	6.8
3	86.03		-228.57								442.21		0.891	119.25	92.08	8.8
3			-32.38					524.92		-180.6			0.904	112.19	83.3	8.0
3	-107.51							318.36			102.16		0.877	126.33	97.21	9.3
3			-149.32					47.16			403.18		0.888	121.15	92.72	9.0
3							30.73	588.14		-251.76			0.905	110.64	82.22	7.6
3								28.99		-321.36	603.38		0.913	106.77	81.3	7.3
3	-18.23							532.06		-199.79			0.905	111.96	83.32	7.9
3				25.26				497.38		-197.07			0.906	110.74	82.93	7.8
3								297.18	69.77		-16.63		0.869	130.94	97.59	9.2
4							24.47	930.17	-162.92	-494.98			0.925	99.05	77.71	6.8
4	658.86	-197.55	-620.13								426.7		0.896	116.56	90.32	8.5
4			-58.42					919.09	-183.74	-435.04			0.929	96.1	74.48	6.7
4	-65.77							964.93	-200.28	-463.66			0.931	95.27	73.84	6.7
4		-44.84						981.22	-199.52	-503.96			0.93	96.08	74.81	6.7
4			-114.58					-456.37		-288.22	1149.22		0.926	98.08	73.31	7.0
4				21.89				859.66	-164.83	-453.14			0.926	98.24	76.41	6.7
4								665.54	-145.59	-479.36	221.36		0.924	99.45	77.95	6.8
4					-3.59			894.5	-167.85	-470.33			0.924	99.15	77.61	6.8
4						-16.81		909.41	-179.83	-480.94			0.926	98.1	76.65	6.7
5		-42.1					9.93	993.16	-197.06	-512.2			0.93	95.44	74.52	6.6
5	-59.47		-6.26					959.63	-198.33	-459.68			0.93	95.74	74.26	6.7
5	-42.93	-21.84						984.48	-204.75	-482.86			0.93	95.1	73.74	6.6
5		-31.83	-40.07					971.94	-201.52	-468.68			0.932	94.23	72.99	6.6
5	826.82	-270.75	-641.9							-268.68	612.7		0.933	93.06	70.18	6.5
5	774.63	-235.26	-670.6					260.8			139.09		0.898	115.6	90.32	8.4
5	1041.16	-338.02	-748.96					515.21		-210.71			0.923	100.32	75.7	7.0
5	85.83		-167.82				29.11			-261.86	659.74		0.924	99.9	74.67	7.1
5			-104.31					169.91	-133	-435.09	748.66		0.937	90.77	70.31	6.5
5	-102.18							340.67	-163.17	-484.34	648.3		0.936	91.61	70.39	6.4

Table 6.1: Table A.1: *Continuation.* Feature selection results for the specific energy yield. A Linear Regression model was implemented. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	Random	R^2	RMSE	MAE	MAPE (%)
6		-11.29	-67.31				38.37	1002.15	-190.15	-474.33			0.933	93.08	71.84	6.5
6	742.31	-243.31	-603.89					-237.39		-286.11	885.15		0.935	92.36	69.56	6.5
6			-102.98	3.83				156.67	-130.94	-430.67	752.5		0.937	91.06	70.26	6.5
6	-45.42		-64.68					207.85	-144.43	-451.6	741.05		0.937	90.8	70.03	6.5
6			-103.86				9.99	260.72	-137.24	-440.04	672.59		0.937	90.86	70.03	6.5
6		-25.08	-87.4					247.86	-149.06	-461.4	713.37		0.938	89.88	69.1	6.4
6			-102.59		-9.92			80.92	-130.6	-429.59	830.46		0.937	90.56	69.99	6.5
6			-99.12			-12.13		139.62	-138.45	-445.33	785.54		0.938	90.48	69.81	6.5
6	-98.57					-18.39		266.6	-169.9	-497.71	732.93		0.938	89.45	68.86	6.3
6		-37.09	-26.96			-13.49		990.03	-213.25	-489.23			0.932	94.25	72.88	6.6
7		-14.06	-57.27	16.18			40.11	979.81	-188.28	-469.38			0.935	92.33	71.21	6.5
7	500.46	-183.59	-423.15					269.94	-121.51	-426.79	610.23		0.941	87.31	67.64	6.2
7	-66.53		-37.28			-15.82		199.54	-157.86	-477.53	775.96		0.938	89.8	69.21	6.4
7		-32.18	-73.15			-17.9		231.76	-162.65	-487.26	753.53		0.94	88.65	68.1	6.3
7		-34.87	-83.75				-20.2	75.43	-143.88	-457.29	868.09		0.939	89.68	68.79	6.4
7	-56.78		-55.53				-6.09	162.05	-144.68	-454.2	787.7		0.938	90.07	69.39	6.4
7	750.12	-247.31	-604.55			-8.26		-268.62		-290.27	916.06		0.935	91.86	69.12	6.5
7	704.53	-243.21	-573.36				-25.99	-439.49		-289.02	1079.36		0.936	91.76	69.47	6.5
7	604.21	-202.64	-478.74				36.1	899.47	-148.26	-430.26			0.939	88.91	69.42	6.2
7	617.98	-232	-449.7			-13.64		888.21	-169.67	-444.4			0.938	89.76	70.06	6.3
8	501.44	-191.16	-409.65			-17.56		252.12	-134.75	-452.43	652.1		0.943	86.03	66.77	6.1
8	498.25	-187.28	-420				-10.35	175.13	-119.78	-427.15	698.84		0.942	87.06	67.64	6.1
8	493.39	-184.06	-412.35	7.91				263.59	-122.97	-427.15	609.38		0.942	86.75	67.45	6.1
8	494.82	-187.25	-411.4		-18.99			123.2	-123.18	-426.73	756.39		0.943	86.09	66.77	6.1
8		-13.71	-54.47	16.69	-6.93		46.08	983.18	-189.49	-469.23			0.935	91.74	70.74	6.4
9	495.06	-190.9	-404.93			-17.2	-4.19	210.69	-132.36	-450.36	688.52		0.943	86.06	66.86	6.1
9	504.13	-191.39	-411.67		-2.37	-16.52		231.74	-133.52	-450.7	669.39		0.943	86.48	67.18	6.1
9	494.99	-190.27	-399.69	7.52		-17.38		243.35	-133.53	-449.12	648.33		0.943	86.26	67.08	6.1
9	602.14	-204.77	-464.63	12.63	-9.66		43.38	884.93	-149.72	-426.57			0.94	88.49	69.34	6.2
9	495.88	-190.17	-405.77	7.57	-18.77			124.61	-123.14	-425.24	744.86		0.943	86.19	67.06	6.1
9	488.29	-186.17	-407.9	7.23			-8.41	190.97	-120.31	-425.05	675.09		0.942	87.24	67.62	6.2
9	495.29	-189.36	-411.05		-18.61		-3.11	98.51	-122.26	-427.05	780.08		0.942	86.57	67.11	6.1
10	500.14	-192.01	-408.37		-1.4	-16.28	-4.01	209.25	-132.57	-450.56	690.07		0.943	85.7	66.49	6.0
10	489.21	-189.17	-396.15	7.36		-17.23	-1.48	228.8	-133.07	-448.78	662.54		0.943	86.21	67.09	6.1
11	491.87	-189.94	-397.83	7.48	-2.3	-16.24	-1.13	217.01	-132.73	-447.14	671.82		0.944	85.33	66.75	6.0

Appendix II: The elbow method and silhouette coefficient

In this project, a qualitative analysis was conducted to select the number of clusters for the classification. However, as mentioned in Chapter 4, there are scoring metrics for clustering which might be used to optimise the classification on a quantitative basis. It was claimed that these methods do not perform adequately for this application, which is a common circumstance [36]. In this section, the results of two quantitative analyses are illustrated.

First, the elbow method is explored. To perform the elbow method, the algorithm is run several times, increasing the number of clusters. For each case, the inertia is recorded. Then, it is plotted as a function of the number of clusters. The inertia decreases when increasing the number of clusters since the distance from each point to its closes cluster center decreases. However, typically, there is a point where the slope of the curve stabilizes. This bending point, or elbow point, might be considered the optimum number of clusters [39]. Figure A.1 illustrates the elbow curve for the four-features case analysed in Chapter 4.

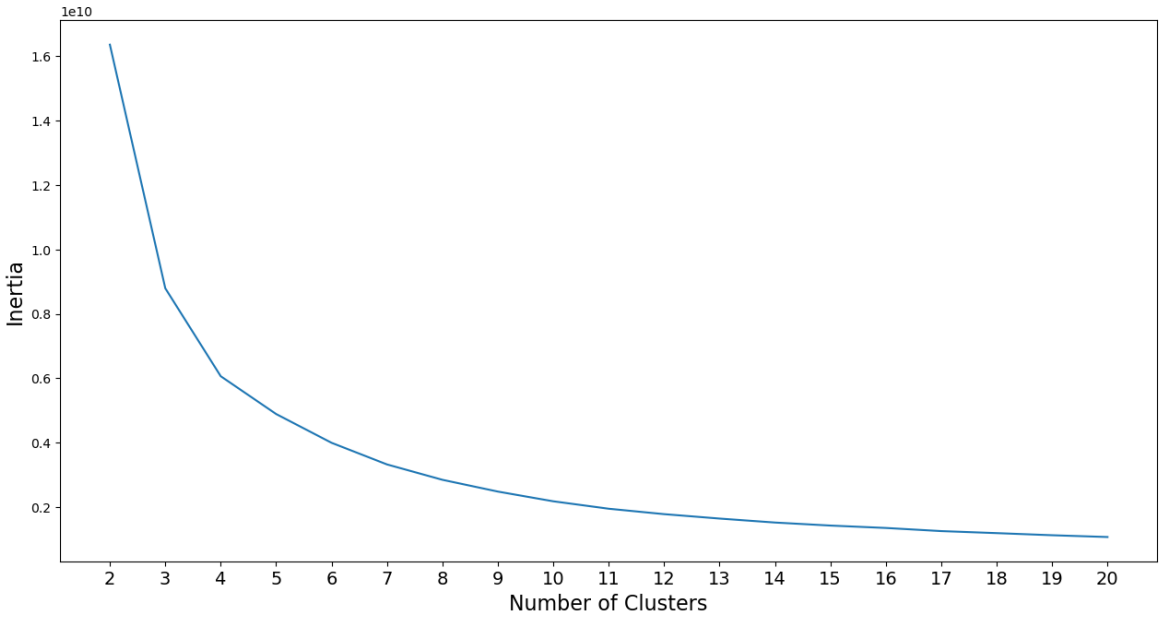


Figure A.1: The elbow method: the inertia is calculated for different number of clusters. The point where the curve bends might be considered the optimum number. Result obtained for the specific energy yield using four features.

In this figure, the number of clusters ranges from 2 to 20. Indeed, the higher the number, the lower the inertia. However, the elbow point is not clearly defined. The figure might suggest that it is around 5 clusters. The KneeLocator algorithm indicates that the actual number is 6. Therefore,

based on the elbow method, the optimum number of clusters would be 6, in contrast to the 15 clusters selected via the qualitative analysis. Nevertheless, using 6 clusters clearly result in a poor classification. Overall, this is a rather unconvincing method.

Another quantitative approach commonly used to evaluate the appropriate number of clusters is the silhouette coefficient. The silhouette coefficient is a measure of cluster cohesion and separation [39]. It quantifies how well a data point fits into its assigned cluster based on the mean intra-cluster distance and the mean nearest-cluster distance. Its value ranges between -1 (poor clustering performance) and 1 (excellent clustering performance). Values near 0 indicate overlapping clusters while negative coefficients generally indicate that a sample has been assigned to the wrong cluster [37].

Again, k-means is run several times increasing the number of clusters from 2 to 20. However, this time, the silhouette coefficients are recorded. Figure A.2 shows the result. In principle, the number with the highest coefficient would be the optimum one. However, this corresponds to only two clusters, which clearly shows the inadequacy of this approach for this dataset.

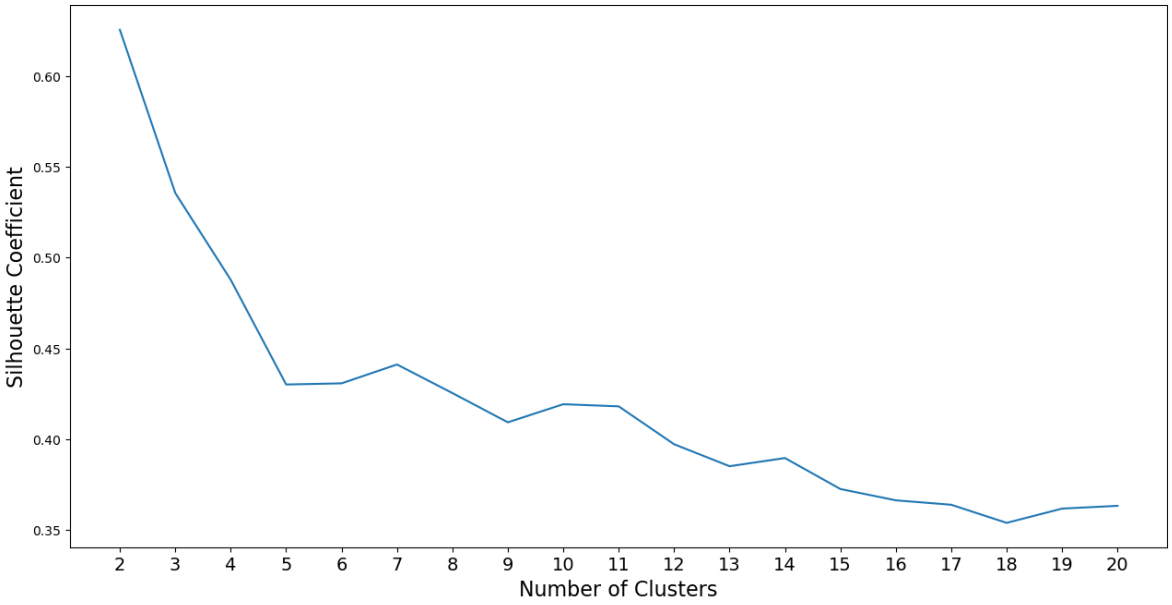


Figure A.2: The silhouette coefficient is plotted against the number of clusters. The higher the coefficient, the better the clustering performance. Result obtained for the specific energy yield using four features.

In summary, the available quantitative approaches to optimise the number of clusters, such as the elbow method or the silhouette coefficient, present a poor performance for the dataset used in this project. Thus, the qualitative analysis approach is justified.

Appendix III: Feature selection results for degradation

Table A.3: Feature selection results for degradation. Random Forests was implemented. The weights given to the climate variables and the error of the model are shown.

Features	T_{ann}	T_{max}	T_{min}	DTR_{ann}	P_{ann}	P_{min}	RH_{ann}	GHI_{ann}	GHI_{max}	GHI_{min}	UV_{ann}	WS_{ann}	Random	R^2	RMSE	MAE	MAPE (%)
1													1	-0.456	0.465	0.375	88.4
1			1											0.818	0.163	0.116	21.5
1	1													0.701	0.212	0.138	21.8
1							1							-0.173	0.418	0.315	76.8
1					1									0.158	0.352	0.269	64.5
1												1		-0.106	0.403	0.319	76.2
2	0.800				0.204									0.924	0.105	0.071	12.3
2	0.811						0.189							0.935	0.098	0.064	10.6
2			0.889				0.111							0.924	0.105	0.070	12.5
2	0.870									0.130				0.816	0.163	0.102	15.2
2	0.806											0.194		0.925	0.106	0.070	12.0
2	0.191		0.809											0.900	0.121	0.078	12.6
3	0.146		0.718		0.136									0.950	0.085	0.054	8.7
3	0.799						0.172			0.029				0.958	0.079	0.049	7.8
3	0.798						0-173				0.029			0.956	0.081	0.049	7.8
3	0.790						0.094					0.116		0.969	0.068	0.043	7.2
3			0.869				0.064					0.067		0.961	0.076	0.049	8.9
3	0.777				0.174							0.049		0.958	0.079	0.050	8.4
4	0.786						0.115			0.016		0.082		0.980	0.055	0.034	5.6
4	0.136		0.754				0.050					0.060		0.977	0.058	0.036	5.8
4	0.778				0.154		0.029					0.040		0.973	0.062	0.039	6.5
4	0.138		0.707		0.112		0.042							0.966	0.071	0.043	6.8
4	0.784						0.083				0.018	0.116		0.981	0.053	0.034	5.5
5	0.132		0.702		0.107		0.030					0.030		0.979	0.056	0.034	5.4
5	0.132		0.753				0.050				0.015	0.055		0.983	0.050	0.031	5.0
5	0.774				0.148		0.025				0.016	0.037		0.983	0.050	0.032	5.2
5	0.784						0.080			0.009	0.013	0.119		0.983	0.050	0.031	5.1
5		0.848	0.038				0.046				0.014	0.054		0.983	0.050	0.031	4.9
6	0.133		0.697		0.103		0.030				0.013	0.028		0.984	0.048	0.030	4.9
6		0.038	0.800		0.092		0.028				0.012	0.030		0.985	0.047	0.029	4.7
6	0.125	0.007	0.754				0.045				0.014	0.055		0.985	0.048	0.029	4.8
6		0.035	0.849				0.043			0.011	0.011	0.052		0.985	0.048	0.029	4.7
6	0.773				0.148		0.024			0.007	0.012	0.036		0.984	0.049	0.030	5.0
7	0.131	0.007	0.700		0.095		0.027				0.012	0.028		0.984	0.048	0.029	4.7
7	0.134		0.696		0.100		0.026			0.006	0.010	0.027		0.986	0.045	0.028	4.7
8	0.129	0.006	0.697		0.101		0.025			0.006	0.009	0.026		0.986	0.046	0.028	4.5
9	0.129	0.006	0.699		0.096		0.025		0.004	0.006	0.007	0.028		0.986	0.045	0.027	4.5
10	0.129	0.006	0.696	0.003	0.098		0.026		0.004	0.005	0.008	0.026		0.987	0.044	0.027	4.4
11	0.130	0.006	0.698	0.003	0.096		0.025	0.003	0.003	0.005	0.005	0.025		0.987	0.044	0.027	4.4
12	0.128	0.005	0.698	0.003	0.097	0.002	0.025	0.003	0.003	0.005	0.005	0.026		0.987	0.044	0.027	4.4

Bibliography

[1] International Energy Agency. *Guidelines for Operation and Maintenance of Photovoltaic Power Plants in Different Climates*. 2022. ISBN: 978-3-907281-13-0. URL: https://iea-pvps.org/key-topics/guidelines-for-operation-and-maintenance-of-photovoltaic-power-plants-in-different-climates/.

[2] International Renewable Energy Agency. *Renewable capacity statistics 2023*. IRENA, 2023. URL: https://mc-cd8320d4-36a1-40ac-83cc-3389-cdn-endpoint.azureedge.net/-/media/Files/IRENA/Agency/Publication/2023/Mar/IRENA_RE_Capacity_Statistics_2023.pdf?rev=d2949151ee6a4625b65c82881403c2a7.

[3] International Renewable Energy Agency. *Renewable power generation costs in 2021*. 2022. ISBN: 978-92-9260-452-3. URL: www.irena.org.

[4] Razin Ahmed et al. "Computationally expedient Photovoltaic power Forecasting: A LSTM ensemble method augmented with adaptive weighting and data segmentation technique". In: *Energy Conversion and Management* 258 (2022), p. 115563. ISSN: 0196-8904. DOI: https://doi.org/10.1016/j.enconman.2022.115563. URL: https://www.sciencedirect.com/science/article/pii/S0196890422003594.

[5] Julián Ascencio-Vásquez, Kristijan Brecl, and Marko Topi. "Köppen-Geiger-Photovoltaic Climate Classification". In: *2018 IEEE 7th World Conference on Photovoltaic Energy Conversion (WCPEC) (A Joint Conference of 45th IEEE PVSC, 28th PVSEC 34th EU PVSEC)*. 2018, pp. 2270–2275. DOI: 10.1109/PVSC.2018.8547952.

[6] Julián Ascencio-Vásquez, Kristijan Brecl, and Marko Topič. "Methodology of Köppen-Geiger-Photovoltaic climate classification and implications to worldwide mapping of PV system performance". In: *Solar Energy* 191 (2019), pp. 672–685. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2019.08.072. URL: https://www.sciencedirect.com/science/article/pii/S0038092X19308527.

[7] Julián Ascencio-Vásquez et al. "Global Climate Data Processing and Mapping of Degradation Mechanisms and Degradation Rates of PV Modules". In: *Energies* 12.24 (2019). ISSN: 1996-1073. DOI: 10.3390/en12244749. URL: https://www.mdpi.com/1996-1073/12/24/4749.

[8] Neha Bansal, Shiva Pujan Jaiswal, and Gajendra Singh. "Comparative investigation of performance evaluation, degradation causes, impact and corrective measures for ground mount and rooftop solar PV plants – A review". In: *Sustainable Energy Technologies and Assessments* 47 (2021), p. 101526. ISSN: 2213-1388. DOI: https://doi.org/10.1016/j.seta.2021.101526. URL: https://www.sciencedirect.com/science/article/pii/S2213138821005373.

[9] Sara Brown. *Machine Learning, explained*. Apr. 2021. URL: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[10] Jakapan Chantana, Aika Kamei, and Takashi Minemoto. "Influences of environmental factors on Si-based photovoltaic modules after longtime outdoor exposure by multiple regression analysis". In: *Renewable Energy* 101 (2017), pp. 10–15. ISSN: 0960-1481. DOI: https://doi.org/10.1016/j.renene.2016.08.037. URL: https://www.sciencedirect.com/science/article/pii/S0960148116307388.

[11] Copernicus Climate Change Service, Climate Data Store. *ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. [Accessed 1-March-2023]. 2023. DOI: 10.24381/cds.f17050d7.

[12] Copernicus Climate Change Service, Climate Data Store. *Essential climate variables for assessment of climate variability from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. [Accessed 15-March-2023]. 2018. URL: https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecv-for-climate-change?tab=overview.

[13] P.K. Dash et al. "A novel climate classification criterion based on the performance of solar photovoltaic technologies". In: *Solar Energy* 144 (2017), pp. 392–398. ISSN: 0038-092X. DOI: https://doi.org/10.1016/j.solener.2017.01.046. URL: https://www.sciencedirect.com/science/article/pii/S0038092X17300658.

[14] D. P. Dee et al. "The ERA-Interim reanalysis: configuration and performance of the data assimilation system". In: *Quarterly Journal of the Royal Meteorological Society* 137.656 (2011), pp. 553–597. DOI: https://doi.org/10.1002/qj.828. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.828. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828.

[15] Daniel Dupal. *K-means clustering algorithm explained*. Feb. 2016. URL: http://dendroid.sk/2011/05/09/k-means-clustering/.

[16] Yusuf Essam et al. "Investigating photovoltaic solar power output forecasting using machine learning algorithms". In: *Engineering Applications of Computational Fluid Mechanics* 16.1 (2022), pp. 2002–2034. DOI: 10.1080/19942060.2022.2126528. eprint: https://doi.org/10.1080/19942060.2022.2126528. URL: https://doi.org/10.1080/19942060.2022.2126528.

[17] Hamed Hanifi et al. "Loss analysis and optimization of PV module components and design to achieve higher energy yield and longer service life in desert regions". In: *Applied Energy* 280 (Dec. 2020), p. 116028. DOI: 10.1016/j.apenergy.2020.116028.

[18] Ian Harris et al. "Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset". In: *Scientific Data* 7 (Apr. 2020). DOI: 10.1038/s41597-020-0453-3.

[19] H. Hersbach et al. "ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS)". In: (2023). [Accessed 1-March-2023]. DOI: 10.24381/cds.f17050d7.

[20] H. Hersbach et al. "Essential climate variables for assessment of climate variability from 1979 to present. Copernicus Climate Change Service (C3S) Data Store (CDS)". In: (2018). [Accessed 15-March-2023].

[21] IBM. *What is machine learning?* URL: https://www.ibm.com/topics/machine-learning.

[22] IEA. *Solar PV – Analysis*. Sept. 2022. URL: https://www.iea.org/reports/solar-pv.

[23] D Jordan, J Wohlgemuth, and Sarah Kurtz. "Technology and Climate Trends in PV Module Degradation". In: Jan. 2012, pp. 3118–3124. DOI: 10.4229/27thEUPVSEC2012-4DO.5.1.

[24] Dirk C. Jordan et al. "Compendium of photovoltaic degradation rates". In: *Progress in Photovoltaics: Research and Applications* 24.7 (2016), pp. 978–989. DOI: https://doi.org/10.1002/pip.2744. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.2744. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2744.

[25] Dirk C. Jordan et al. "Photovoltaic fleet degradation insights". In: *Progress in Photovoltaics: Research and Applications* 30.10 (2022), pp. 1166–1175. DOI: https://doi.org/10.1002/pip.3566. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pip.3566. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.3566.

[26] Todd Karin, C. Birk Jones, and Anubhav Jain. "Photovoltaic Degradation Climate Zones". In: *2019 IEEE 46th Photovoltaic Specialists Conference (PVSC)*. 2019, pp. 0687–0694. DOI: 10.1109/PVSC40753.2019.8980831.

[27] Kotaro Kawajiri, Takashi Oozeki, and Yutaka Genchi. "Effect of Temperature on PV Potential in the World". In: *Environmental science & technology* 45 (Aug. 2011), pp. 9030–5. DOI: 10.1021/es200635x.

[28] Amith Khandakar et al. "Machine Learning Based Photovoltaics (PV) Power Prediction Using Different Environmental Parameters of Qatar". In: *Energies* 12.14 (2019). ISSN: 1996-1073. DOI: 10.3390/en12142782. URL: https://www.mdpi.com/1996-1073/12/14/2782.

[29] Michael Koehl, Markus Heck, and Stefan Wiesmeier. "Categorization of weathering stresses for photovoltaic modules". In: *Energy Science & Engineering* 6.2 (2018), pp. 93–111. DOI: https://doi.org/10.1002/ese3.189. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ese3.189. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ese3.189.

[30] Markus Kottek et al. "World Map of the Köppen-Geiger Climate Classification Updated". In: *Meteorologische Zeitschrift* 15 (May 2006), pp. 259–263. DOI: 10.1127/0941-2948/2006/0130.

[31] Shaoshuai Li et al. "A method for determining the applicable geographical regions of PV modules field reliability assessment results based on regional clustering of environmental factors and their weights". In: *Sustainable Energy Technologies and Assessments* 53 (2022), p. 102620. ISSN: 2213-1388. DOI: https://doi.org/10.1016/j.seta.2022.102620. URL: https://www.sciencedirect.com/science/article/pii/S2213138822006701.

[32] Wei Liu et al. "Research on optimum tilt angle of photovoltaic module based on regional clustering of influencing factors of power generation". In: *International Journal of Energy Research* 45 (Feb. 2021). DOI: 10.1002/er.6584.

[33] Weidong Liu et al. "Photovoltaic module regional clustering in mainland China and application based on factors influencing field reliability". In: *Renewable and Sustainable Energy Reviews* 133 (2020), p. 110339. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2020.110339. URL: https://www.sciencedirect.com/science/article/pii/S1364032120306274.

[34] MathWorks. *Matlab vs. python: Which one is right for you?* URL: https://www.mathworks.com/products/matlab/matlab-vs-python.html.

[35] Leonardo Micheli, Matthew Muller, and Sarah Kurtz. "Determining the effects of environment and atmospheric parameters on PV field performance". In: *2016 IEEE 43rd Photovoltaic Specialists Conference (PVSC)*. 2016, pp. 1724–1729. DOI: 10.1109/PVSC.2016.7749919.

[36] Andreas Christian Müller and Sarah Guido. *Introduction to machine learning with python: A guide for data scientists*. First Edition. O'Reilly Media, 2016.

[37] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[38] Murray Peel, Brian Finlayson, and Thomas Mcmahon. "Updated World Map of the Koppen-Geiger Climate Classification". In: *Hydrology and Earth System Sciences Discussions* 4 (Oct. 2007). DOI: 10.5194/hess-11-1633-2007.

[39] Real Python. *K-means clustering in Python: A practical guide*. Jan. 2023. URL: https://realpython.com/k-means-clustering-python/.

[40] Bruno Rudolf et al. "New GPCC Full Data Reanalysis Version 5 Provides High-Quality Gridded Monthly Precipitation Data". In: *GEWEX News* 21 (Jan. 2011).

[41] Nikolaos Skandalos and Dimitris Karamanis. "An optimization approach to photovoltaic building integration towards low energy buildings in different climate zones". In: *Applied Energy* 295 (2021), p. 117017. ISSN: 0306-2619. DOI: https://doi.org/10.1016/j.apenergy.2021.117017. URL: https://www.sciencedirect.com/science/article/pii/S0306261921004839.

[42] Nikolaos Skandalos et al. "Building PV integration according to regional climate conditions: BIPV regional adaptability extending Köppen-Geiger climate classification against urban and climate-related temperature increases". In: *Renewable and Sustainable Energy Reviews* 169 (2022), p. 112950. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2022.112950. URL: https://www.sciencedirect.com/science/article/pii/S1364032122008310.

[43] Manie Tadayon. *Python vs R vs Matlab for machine learning, causal inference, signal processing, and more.* Dec. 2020. URL: https://medium.com/swlh/python-vs-r-vs-matlab-for-machine-learning-causal-inference-signal-processing-and-more-b837a988c674.

[44] Giuseppe Marco Tina et al. "A State-of-Art-Review on Machine-Learning Based Methods for PV". In: *Applied Sciences* 11.16 (2021). ISSN: 2076-3417. DOI: 10.3390/app11167550. URL: https://www.mdpi.com/2076-3417/11/16/7550.

[45] Jake VanderPlas. *Python Data Science Handbook: Essential Tools For Working With Data*. First Edition. O'Reilly, 2016.

[46] Lucien Wald. "A simple algorithm for the computation of the spectral distribution of the solar irradiance at surface". In: (Jan. 2018). DOI: 10.13140/RG.2.2.17025.76648.

[47] Wikipedia contributors. *Arthur Samuel (computer scientist) — Wikipedia, The Free Encyclopedia*. [Online; accessed 25-March-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Arthur_Samuel_(computer_scientist)&oldid=1135114644.

[48] Wikipedia contributors. *Multivariate adaptive regression spline — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Multivariate_adaptive_regression_spline&oldid=1130941885. [Online; accessed 28-May-2023]. 2023.

[49] Wikipedia contributors. *Pearson correlation coefficient — Wikipedia, The Free Encyclopedia*. [Online; accessed 26-April-2023]. 2023. URL: https://en.wikipedia.org/w/index.php?title=Pearson_correlation_coefficient&oldid=1146097966.