remote sensing

MDPI

Article

Multivariate Analysis for Solar Resource Assessment Using Unsupervised Learning on Images from the GOES-13 Satellite

Jared D. Salinas-González [1,†], Alejandra García-Hernández [1,*,†], David Riveros-Rosas [2,†], Gamaliel Moreno-Chávez [1], Luis F. Zarzalejo [3], Joaquín Alonso-Montesinos [4,5], Carlos E. Galván-Tejada [1], Alejandro Mauricio-González [1] and Adriana E. González-Cabrera [2]

[1] Academic Unit of Electrical Engineering, Autonomous University of Zacatecas, Jardín Juarez 147, Centro Historico, Zacatecas 98000, Mexico; jerad.salinas94@uaz.edu.mx (J.D.S.G.); gamalielmch@uaz.edu.mx (G.M.-C.); ericgalvan@uaz.edu.mx (C.E.G.-T.); amgdark@uaz.edu.mx (A.M.-G.)
[2] Geophysics Institute, Universidad Nacional Autónoma de México, Ciudad de Mexico 04150, Mexico; driveros@igeofisica.unam.mx (D.R.-R.); gonzalezc@igeofisica.unam.mx (A.E.G.-C.)
[3] Centro de Investigaciones Energéticas Medio Ambientales y Tecnológicas, Av. Complutense 40, 28040 Madrid, Spain; lf.zarzalejo@ciemat.es
[4] Department of Chemistry and Physics, University of Almería, 04120 Almeria, Spain; joaquin.alonso@ual.es
[5] Solar Energy Research Center, CIESOL, Joint Centre of the University of Almería-CIEMAT, 04120 Almeria, Spain
* Correspondence: alegarcia@uaz.edu.mx
† These authors contributed equally to this work.

Citation: Salinas-González, J.D.; García-Hernández, A.; Riveros-Rosas, D.; Moreno-Chávez, G.; Zarzalejo, L.F.; Alonso-Montesinos, J.; Galván-Tejada, C.E.; Mauricio-González, A.; González-Cabrera, A.E. Multivariate Analysis for Solar Resource Assessment Using Unsupervised Learning on Images from the GOES-13 Satellite. *Remote Sens.* **2022**, *14*, 2203. https://doi.org/10.3390/rs14092203

Academic Editor: Lunche Wang

Received: 28 January 2022
Accepted: 22 March 2022
Published: 5 May 2022

Abstract: Solar resource assessment is of paramount importance in the planning of solar energy applications. Solar resources are abundant and characterization is essential for the optimal design of a system. Solar energy is estimated, indirectly, by the processing of satellite images. Several analyses with satellite images have considered a single variable—cloudiness. Other variables, such as albedo, have been recognized as critical for estimating solar irradiance. In this work, a multivariate analysis was carried out, taking into account four variables: cloudy sky index, albedo, linke turbidity factor (TL2), and altitude in satellite image channels. To reduce the dimensionality of the database (satellite images), a principal component analysis (PCA) was done. To determine regions with a degree of homogeneity of solar irradiance, a cluster analysis with unsupervised learning was performed, and two clustering techniques were compared: k-means and Gaussian mixture models (GMMs). With respect to k-means, the GMM method obtained a smaller number of regions with a similar degree of homogeneity. The multivariate analysis was performed in Mexico, a country with an extended territory with multiple geographical conditions and great climatic complexity. The optimal number of regions was 17. These regions were compared for annual average values of daily irradiation data from ground stations using multiple linear regression. A comparison between the mean of each region and the ground station measurement showed a linear relationship with a R^2 score of 0.87. The multiple linear regression showed that the regions were strongly related to solar irradiance. The optimal sites found are shown on a map of Mexico.

Keywords: solar resource assessment; clustering analysis; satellite images; climatic features; unsupervised learning; solar energy

1. Introduction

In recent decades, renewable energies have gained importance worldwide. Renewable energies—in addition to reducing carbon dioxide emissions into the atmosphere, and preserving oil resources, which have a much greater diversity of uses—are tools for economic growth, job creation, and energy security [1,2]. Solar resource assessment is the basis of planning and installing solar power plants. In regions with great climatic diversity, the implementation of solar radiation models is complex and sources of error cannot be studied

due to the lack of surface measurements [3]. For these reasons, it is necessary to have solar radiance measurement networks whose distributions must be the most representative of the climatic characteristics that influence the solar radiance that reaches the surface [4]. In this way, the identification of geographic regions with similar climatic behaviors (in this case, those related to solar radiation) will allow optimization of the deployment, management, and maintenance of the network [5,6]. Therefore, the planning and installing of a measurement network requires a previous regional analysis that allows identifying and classifying the climatic diversity of the geographical areas to be evaluated; cluster analyses have been shown to work well for regionalization [7].

To place an optimal number of ground-based stations, a cluster analysis is critical to optimize the number of stations needed to cover the country. Thus, relevant works related to solar energy can be found, e.g., a time series of cloud modification factors (CMFs) in Greece; in this work, a methodology to regionalize geographical areas through a cluster analysis is proposed; satellite images with a resolution of 180 × 160 pixels were used in a time series of 730 days. For the cluster analysis, the k-means algorithm and the L method were applied, 22 optimal regions were obtained in the regionalization. Some of the limitations observed were: the images used had very small spatial resolutions, and the values of irradiation per pixel could not give a clear idea about the concentration of solar radiation in each region. Another limitation was that the validation of the regionalization was limited for homogeneity of the pixels with respect to the centroids of each class, instead of using external data sources [8]. Regarding the time series of global horizontal irradiation in Benelux, in this study, for the regionalization, the global horizontal irradiance (GHI) was analyzed using satellite images and data from solarimetric stations; the algorithms k-means and Ward were applied and the result was the regionalization throughout four classes [9]. Measures of variation in surface solar irradiance using ground-based observation data were applied in a regionalization of Japan, by performing cluster analysis using clear sky index data collected from 47 pyranometers; measurements were segmented into mean, variance, and entropy over a 5-year period. For cluster analysis, the Ward algorithm was applied, the study found 3 and 6 as the optimal regions evaluated by the CH index [10]. Other related works include solar radiation estimations from satellite images in Spain [11,12] or Vietnam [13–15], classifications of interannual and seasonal solar radiation variabilities in South America, geoclimatic variable analyses, such as isotherms, isohyets, evaporation, and humidity in Mexico [3], sunshine duration time series in Vietnam [16], etc. As can be seen, most of the regionalization works are based on direct measurements of solar radiation through a time series; geoclimatic variables have been used in very few. This can cause a problem when planning and installing a solarimetric measurement network, since a climatic criterion is required to determine representative measurement points. Regionalizing (a region's own) values of solar radiation, or variations of it, is not enough to identify the environmental diversity of the region to be evaluated.

In Mexico's case, pyrometers and radiometers have been installed in regions across the country through hydro-meteorological analyses instead of solar radiation analyses. Thus, it was relevant to conduct a cluster analysis of Mexico country based on solar energy characteristics. Identification of broad regions from geographic areas relevant to solar radiation is useful to the energy industry to optimally use solar resources, because sensors and tools can be installed in places that facilitate their exploitation throughout the day or year. Furthermore, thanks to solar energy use, the employment of fossil fuels can be reduced [8,10].

This paper presents a cluster analysis of Mexico through a multivariate analysis derived from satellite image features. The objectives of this work were to identify regions with common characteristics for the incidence of solar radiation, to establish a quantitative criterion for the planning of stations in extensive regions of territories. It was possible to identify geoclimatic variables with greatest influence on the identification of the regions, which establishes criteria to maintain (or not) the use of certain variables. Both objectives have (practically) not been used in previous works with this approach. A methodology

for clustering was used by applying machine learning (no supervised) algorithms and techniques for preprocessing, analyzing, and evaluating. Machine learning techniques were implemented for the first time in the regionalization of a country with great climatic complexity for the purposes of evaluating the solar resource. The paper's output is a map of regionalization of Mexico with the optimal regions with the best degrees of relationships between solar radiation and the features derived from the satellite images.

The paper is structured as follows: Section 2 describes the variables that were considered in the research, the data sources, and the methodology that was used; in Section 3 each algorithm that was used, in each step, is explained in detail: PCA, k-means, and GMM; and to validate the results, the L method with the Davies–Bouldin (DB) and Calinski–Harabasz (CH) indices. In Section 4, the results and the validations of the algorithms applied in each step are presented. The final section presents the discussion of the results and the main contributions of the research.

2. Materials and Methods

In this research, satellite images from the GOES-13 satellite were used. The GOES-13 is an artificial satellite for meteorological research developed by the National Oceanic and Atmospheric Administration (NOAA) and the National Environmental Satellite Data and Information Service (NEDIS). The images were processed in different bands with different physical and meteorological information. To carry out this research, the satellite images corresponding to the first band (cloud cover, daytime surface features, and radiance values) and fourth band (surface temperature or above the clouds) of the GOES-13 satellite were analyzed. The images were originally in network common data form (NetCDF), which is the format used to store multidimensional atmospheric, climatic, and oceanic data. The abbreviations in the back part of the paper show a list of abbreviations and acronyms that will be used in this paper.

The satellite images were processed with respect to four geoclimatic characteristics: albedo, altitude, TL2, and cloudy sky index. These four characteristics were used in the forecasting and calculation of solar radiation in the Heliosat-2 and ESRA methods [17,18] and they are also considered important characteristics to determine solar radiation in Mexico [3].

The parameters analyzed were the time series of albedo, TL2, cloudy sky index (η), and altitude (Z) from 2015.

The albedo is a dimensionless climatic value that is defined as the amount of sunlight of all wavelengths, reflected from an object, substance, or surface [19]. The albedo parameter was obtained from GOES13 satellite images; the images were processed with Heliosat-2 routines. From the albedo algorithm, the minimum radiance values were obtained two hours prior to and two hours after solar noon. The TL2 factor provides an approximation to the model of the atmospheric absorption and scattering of solar radiation as it passes through the atmosphere. Moreover, it can be viewed as the number of dry and clean atmospheres, which would produce the attenuation and dispersion observed in the real atmosphere by the clear sky atmosphere [20]. The TL2 values have mean values typically in a range from 1 to 6. A low value of TL2 represents a clear atmosphere (with a low suspended particles), a high value means a wet and warm atmosphere, and values between 6 to 7, a polluted atmosphere [17]. The altitude is the vertical distance between a place or point on earth and the sea level. It is measured in height above sea level. The cloudy sky index is a dimensionless climate variable representing each pixel's cloud cover percentage in the image. The values are normalized in a [0:1] scale, where 0 represents a clear sky and 1 a completely cloudy sky [3]. The cloudy sky index parameter was also obtained from GOES13 images and the values correspond to monthly mean values. The images also were processed with the Heliosat-2 process and this process gives relative values valid for solar irradiance estimations.

For this work, 365 images of albedo, 1095 images of cloudy sky index (taken in the range of 12:15 to 1:15 p.m. in 30 min jumps, were analyzed; these images were averaged by

day and then per month), 12 images of TL2 per month, and an altitude image. All images had the same height–width resolution (2127 × 3066 pixels).

To carry out the regionalization of Mexico using unsupervised learning algorithms, the methodology shown in Figure 1 was followed.

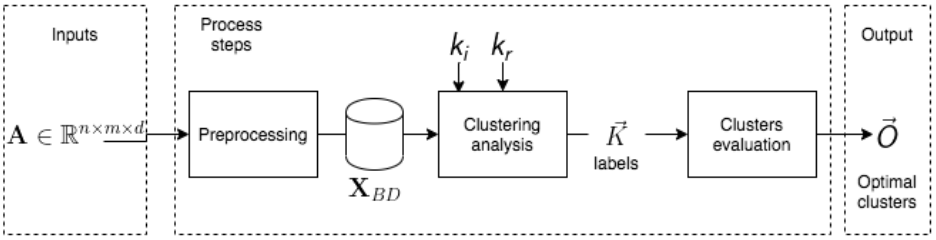


Figure 1. Diagram of the methodology for regionalization of Mexico.

The inputs were the satellite images analyzed for the four geoclimatic characteristics represented as a float matrix $\mathbf{A} \in R^{n \times m \times d}$, where n is the number of rows and m the number of columns (related to the *height* × *width* of the pixels), and d is the time series considered (days/months).

To prepare the data, a preprocessing step had to be carried out that consisted of cleaning and transforming the data in order to implement the machine learning algorithms. The preprocessing step in this study involves reducing the number of pixels to be analyzed and the computational complexity of the algorithms. Images with very large spatial resolutions and time series analyses produce very computationally heavy results. For example, to be able to analyze the albedo images of Mexico throughout the 365 days of the year, there are 2127 × 3066 × 365 pixels, which would give a total of 2.380304430×10^9 pixels, which translates into a quantity of data that is computationally difficult to process. To reduce the number of pixels and create a database that allows the analysis and evaluation of clusters, Figure 2 describes the activities that were carried out.

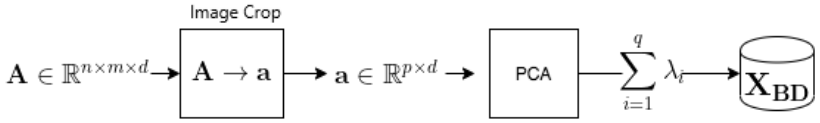


Figure 2. Preprocessing step diagram.

The preprocessing step consists of three activities: image crop, normalization of pixel variables, and PCA per variable.

To reduce the number of pixels or aberrant pixels in the image, a crop is applied and is notated as **a** matrix. Then, the image is segmented to ensure no pixels in the time series without values. Subsequently, a normalization of the features permits that values in the images with high variance do not dominate the rest of the values. In this paper, the min–max scale normalization was used and is denoted by Equation (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

where x is a pixel value from the **a** matrix, x_{min} is the smallest value and x_{max} is the higher value of the **a** matrix . The result of this task is an **â** matrix, which values are defined in the [0, 1] range. With the **â** matrix, PCA reduces the dimension of the time series d, keeping the significant quantity of information. PCA is an orthogonal linear transformation for mapping high-dimensional data to lower dimensional space that retains the maximum of the data variance. The first coordinate axis preserves the higher variance and is called the first principal component, and so on, the variance is preserved decreasingly. PCA

was carried out based on the Equations (2) and (3). Where w are the eigenvectors of the covariance matrix Σ, λ the eigenvalues, I the identity matrix, and det the determinant matrix.

$$\Sigma w = \lambda w \tag{2}$$

$$det(\Sigma - \lambda I) = 0 \tag{3}$$

The result for applying PCA is the sum of the eigenvalues where the first axis contains the highest variance [21]. The preprocessing step's general output is a dataset with the principal components of each variable (albedo, TL2, cloudy sky index, altitude), the explained variance ratio of each variable is greater or equals to the 90% of the temporal variance.

3. Theory/Calculation

In this section, the methodology steps of clustering analysis and cluster evaluation are discussed in detail.

3.1. Clustering Analysis

The clustering analysis is an unsupervised machine learning task that separates the pixels in clusters (regions) that meet some similarity criteria [8,22]. It is recommended to analyze within a range of possible regions, in order to have a great number of clusters to evaluate. The algorithms proposed in this paper are k-means and GMM. This process step's output is a vector of labels that associates a pixel to a region for each k number of clusters in the range.

K-means is the most common clustering method that assigns each pixel of a set of n pixels to one of the k desired regions. The objective is to minimize the differences within-cluster pixels of each group and maximize the differences intra-cluster. The algorithm works given a set of n pixels, $X = \{x_1, x_2, x_3, \ldots, x_n\}$ and a priori number of k clusters centers $C = \{c_1, c_2, c_3, \ldots, c_k\}$, in the initial place, it classifies each pixel to a region with the nearest distance to its cluster center. Under an iterative procedure, the within-cluster sum of squares is calculated, and the centers of the clusters are reassigned to result in a final partition that optimizes the clustering quality by minimizing the within-cluster sum of square distances of any pixel point to its nearest cluster center, as it is defined in the Equation (11). Where $\|.\|$ is the L^2 distance's norm [8].

GMM is a soft clustering algorithm, which means that each pixel has the probability of belonging to one or more clusters from a finite set of Gaussian distributions of unknown parameters. A Gaussian multivariate probability distribution is given in Equation (4).

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} EXP(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{4}$$

where the x is the dataset of pixels, D is the number of dimensions of each datum, μ is the mean that defines the center of the group, Σ is the covariate matrix that defines its width, and π is a mixing probability that defines the size of the Gaussian function. The probability that a pixel comes from a Gaussian k is defined by Equation (5). Where z is a latent variable that takes two possible values, 1 when the pixel x came from Gaussian k and 0 otherwise [23,24].

$$p(z_{nk} = 1|x_n) = \pi_k \tag{5}$$

GMM supposes that there are clusters with parameters Σ, π, and μ. From Bayes rule, the probability that a pixel belongs to a Gaussian is given by Equation (6).

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)} \tag{6}$$

In order to determine the parameters, the expectation–maximization (EM) algorithm is used:

1. Expectation step
   (a) Initialize μ_k, Σ_k and π_k with random values.
   (b) Estimate with the parameters $\gamma(z_{nk})$.
2. Maximization step
   (a) Update the parameters μ_k, Σ_k and π_k using the followings Equations (7)–(9).

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{N} \tag{7}$$

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_n}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{8}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{9}$$

The algorithm converges after a defined number of iterations or the parameters stop updating with new values.

Clustering Evaluation

Clustering validation and evaluation are techniques used to determine the number of classes of a clustering algorithm; the external validations are based on previous knowledge about the data, and the internal validation uses the intrinsic information of the geometrical structure of the data [8].

The *DB* and *CH* indices were employed in Zagouras works to determine the appropriate number of classes. The *DB* index is based on the within-group and between-group distance ratios, where a smaller *DB* value indicates compact and separated clusters. In contrast, the *CH* index is based on the cluster center positions in the dataset to define all data point proximities. The highest *CH* index is related to cluster partitions composed of well-separated clusters.

To determine a reasonable number of clusters, the L method searches a critical point (c) in the evaluation graph to determine the appropriate number of clusters. It is based on the intersection of two best-fit lines on the left and the right side of the critical point (c). Each fit line should cover most of the data points, the root mean squared error in the critical point c ($RMSE_c$) is a vector of the evaluation and the optimal number of classes (O) is the class with the lower *RMSE_c* [8,25].

The $RMSE_c$ is described in Equation (10).

$$RMSE_c = \frac{c-1}{k-1} \cdot RMSE_{L_c} + \frac{k-c}{k-1} \cdot RMSE_{R_c} \tag{10}$$

where $RMSE_{L_c}$ is the root mean squared error in the left side of the critical point c and $RMSE_{R_c}$ in the right side of c. The L method is a well-suited procedure for obtaining an optimal number of regions (classes). Because the work is analyzed with different climatic features, it is hard to know the relationship between these features in the classes with solar radiation data. For that reason, it was proposed to evaluate the center classes with irradiation data.

The centroids of the optimal classes can be seen as a matrix that contains the annual mean of each feature, and for comparing, the ground-based irradiation data were used.

Figure 3 presents the methodology used to relate the centroids data of each cluster to daily solar radiation.

The inputs are the centroids for each cluster of the optimal class. As the centroids were in PCAs, it was necessary to transform the centroids to the original dimension values, c_{rec} contains each cluster's normalized centroids. Then the centroids must be reconstructed to its original values. The features values reconstructed with a length of the original time

series space are denoted by x_{rec} , then they were averaged for obtaining the annual mean for each feature in a cluster.

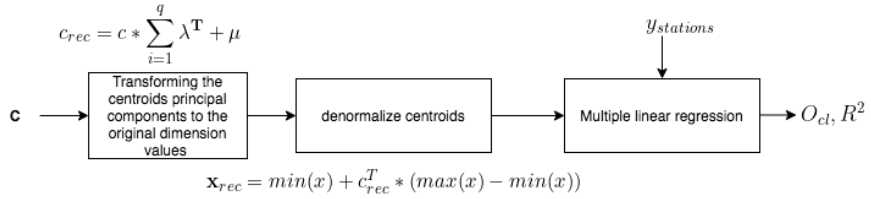


Figure 3. Diagram of the methodology for regionalization of Mexico Country.

The multiple linear regression was used for evaluating the relationship between the $y_{stations}$ (annual daily irradiation) data of each cluster and the features. The R^2 was used between a dependent variable and a set of independent variables to measure the correlation. A value close to 1 means a strong positive relationship, a value closer to 0 means that it is not a relationship. In comparison, a negative relationship close to -1 means that they are inversely related [26]. The output is the number of classes with the clusters O_{cl} with the higher R^2 and lower $RMSE$.

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \left\| x_j - c_i \right\|^2 \tag{11}$$

The algorithm has been used in different regionalization works [8,9,17,27].

4. Results

The results of data preprocessing of each feature for the PCA algorithm and cluster analysis are reported in this section, as well as the correlation analysis between regionalization and solar radiation data measured on the surface.

4.1. Preprocessing Results

The preprocessing consisted of different steps: the first involved reading the images as matrices (height $\times$ width $\times$ d (time series)), the second involved cropping the images, and the third involved applying PCA to reduce the number of pixels and mapping the images to lower dimensions to create the dataset.

Images had -1 values when there was no value and albedo images had an atypical noise at the top. The images were cropped as observed in Figure 4.

The dimensions of the cropped images were 1689 $\times$ 2007 pixels. To reduce the number of pixels for each image and evaluate only the pixels that belonged to the Mexican surface, a second cut was made. The same cut was made to all of the images to keep the homogeneity. If a pixel in any of the images (time series) had a value with -1, the pixel was removed.

With this change, a temporal dataset was made with 1,130,253 rows (pixels) and (d) columns where d was the (day/month) series. Moreover, to reduce the dimensionality of the columns, a normalization was made, and then the PCA algorithm. The results of applying PCA are presented in Table 1.

Table 1. Dataset per variable after applying the PCA algorithm (the cloudy sky index was averaged per day and then per month before applying PCA due to the high variances of these images).

Feature	Number of Pixels	Number of PCAs	Explained Variance
Albedo	1,130,253	6	90.06%
TL2	1,130,253	3	95.53%
Cloudy Sky index	1,130,253	6	93.05%
Altitude	1,130,253	1	100.0%

The result was the X_{BD} dataset of 1,130,253 pixels and 16 PCA's.

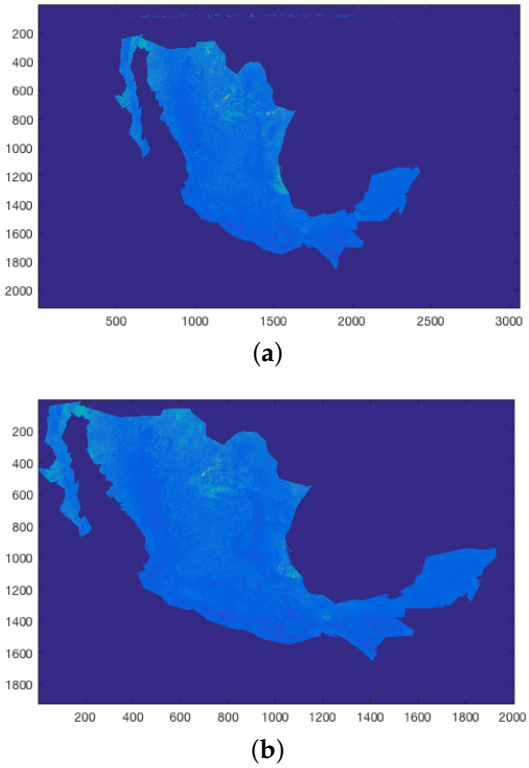


Figure 4. Albedo images. (**a**) Image with noises. (**b**) Cropped images without noises.

4.1.1. Clustering Analysis and Validation (Results)

The k-means and GMM algorithms were applied to the dataset $\mathbf{X_{BD}}$ in a range of $2 \leq k \leq 50$ classes. The L method was then used with the CH and DB indices for each class. The k-means and GMM evaluations are presented in Figures 5 and 6.

Using the k-means algorithm gives the optimal number of classes 4 and 17 (Figure 5), while using the GMM algorithm, the optimal number of classes is 10 (Figure 6). In this research the algorithms were applied in a range of $2 \leq k \leq 50$ classes to find the lines of best fit; the greater the number of classes, the lines of best fit can be smoothed more, but this also requires greater computational capacity to apply it.

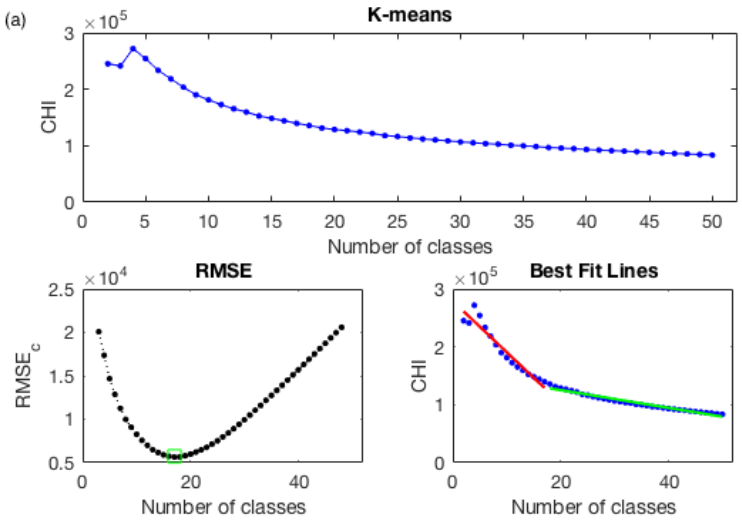


Figure 5. *Cont.*

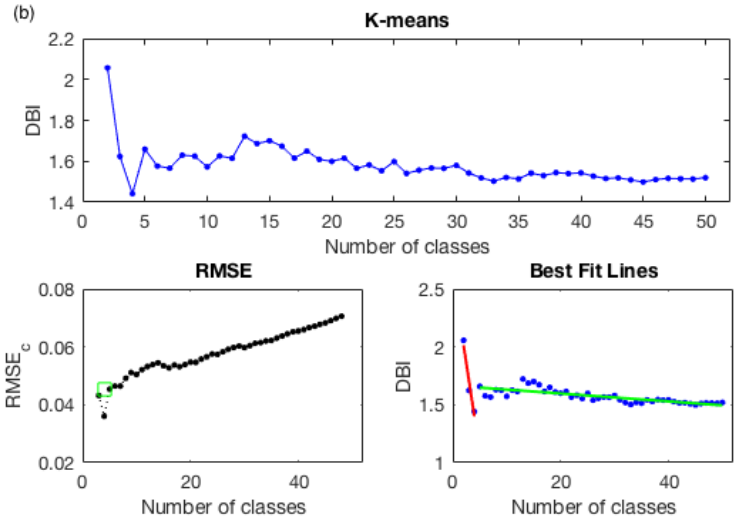


Figure 5. L method for CH and DB indices using k-means algorithm. (**a**) L method with CH index. (**b**) L method with DB index.

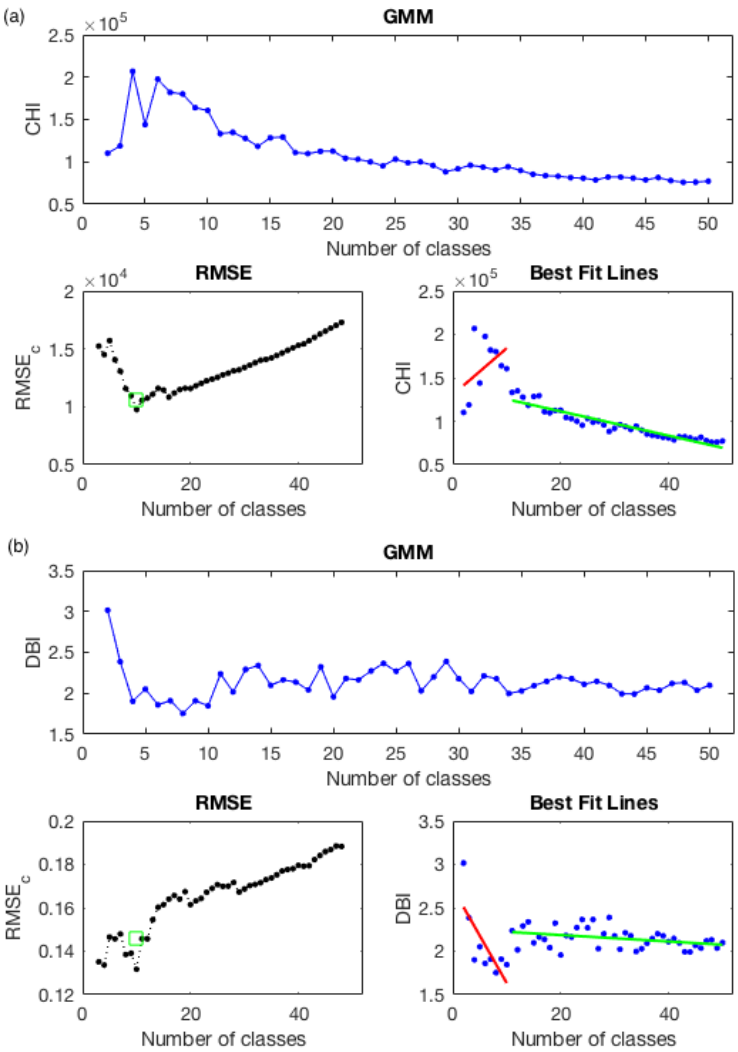


Figure 6. L method for CH and DB indices using GMM algorithm. (**a**) L method with CH index. (**b**) L method with DB index.

4.1.2. Relationship between Clusters and the Solar Radiation

The National Weather Service (SMN due to the Spanish acronym) monitors a network of 180 automatic weather stations (called EMAS due to the Spanish acronym), each with a pyranometer Kipp & Zonen (R) model CMP11 to measure global solar irradiance. However, the stations were installed with hydrometeorological criteria, and most stations have the pyranometers installed in inappropriate conditions [28]. An analysis was carried out to identify measuring issues in the database, and 26 stations were selected to calculate the daily irradiation for the year 2015. Each station records a 10 min average irradiance using a Campbell Scientific data logger. Data quality criteria, described in previous work [28], were applied to ensure the precision of the used data. Therefore, the stations were labeled by their belonging to the classes via their latitude (*Lat.*) and longitude (*Lon.*) geographic positions. Then the annual daily irradiation of stations was averaged per class, and the results were sorted from the lowest to the highest irradiation values. The cluster centroids were transformed to the original values to have the annual values of albedo, TL2, cloudy sky index, and altitude for each class.

The stations with their coordinates in *Lat.* and *Lon.*, their annual average daily irradiation values, and their belonging classes are presented in Table 2.

Table 2. Stations with their annual average daily irradiation and cluster class

Station	*Lat.* °N	*Lon.* °E	Annual Average Daily Irradiation (Wh/m^2)	K-Means: 17 Cl.	K-Means: 4 Cl.	GMM 10 Cl.
Nueva Rosita	27.92	101.33	4736.95	14	3	9
Matías Romero	16.88	95.03	4744.03	1	4	3
Paraíso	18.42	93.15	5348.72	1	4	3
Centla	18.40	92.64	4899.53	1	4	3
Mexicali	32.66	115.29	5759.59	15	1	7
Presa Abelardo	32.44	116.90	5953.55	15	1	7
Ocampo	28.82	102.52	5478.52	2	1	5
Maguarachi	27.85	107.99	5440.13	17	1	5
Obispo	24.25	107.18	5378.4	11	4	4
Monclova	18.05	90.82	5242.85	4	4	8
Acaponeta	22.46	105.38	5297.43	7	4	1
Agustín Melgar	25.26	104.00	5197.85	12	1	5
Metehuala	23.64	100.65	5649.75	12	1	2
Oxktzcab	20.29	89.39	5250.9	4	4	8
Petacalco	17.98	102.12	5402.63	7	4	10
Nevados Toluca	19.12	99.77	4390.92	16	2	10
Apatzingan	19.08	102.37	5797.92	7	4	10
Angamacutiro	20.12	101.72	5913.77	10	2	10
Atoyac	17.20	100.44	5471.69	7	4	10
Ixtla	19.09	98.64	5060.64	16	2	10
Atlacomulco	19.79	98.87	5405.35	5	2	2
Perote	19.54	97.26	5607.01	16	2	10
Altzomonil	19.11	98.65	4747.28	16	2	10
Miahuatlan	16.34	96.57	5636.19	7	4	10
Nochistlan	17.43	97.24	5636.27	10	2	10
Nogales	31.29	110.91	5959.9	8	1	7

Table 3 describes the annual average daily irradiation for classes and their corresponding annual averages of albedo, TL2, cloudy sky index, and altitude.

Table 3. Annual daily irradiation, annual averages of albedo, TL2, cloudy sky index, and altitude per cluster classes.

Evaluation: *k*-Means-17 Classes					
Class	**Annual Daily Irradiation (Wh/m^2)**	**Albedo**	***TL2***	**Cloudy Sky Index**	**Altitude (*mAMSL*)**
16	4952.0	0.7651	3.7766	0.0706	2010
14	4737.0	1.5362	4.1138	0.0797	279
1	4997.4	0.9692	4.1138	0.0768	282
12	5423.8	1.1008	3.1486	0.0458	1890
4	5246.9	0.9216	4.2178	0.0662	83
11	5378.4	1.407	3.8554	0.049	259
5	5405.4	0.9852	3.2987	0.0456	2190
17	5440.1	0.8627	3.488	0.0515	2050
2	5478.5	1.5647	3.6405	0.0448	1.340
7	5521.2	0.9344	3.9526	0.0435	616
10	5775.0	0.9273	3.792	0.039	1.450
15	5856.6	3.0128	3.4441	0.041321	211
8	5959.9	1.7008	2.8913	0.0386	660
Evaluation: *k*-Means-4 Classes					
Class	**Annual Daily Irradiation (Wh/m^2)**	**Albedo**	***TL2***	**Cloudy Sky Index**	**Altitude (*mAMSL*)**
3	4736.95	1.4228	3.9373	0.0724	417
2	5251.6	0.9089	3.5908	0.0493	1880
4	5315.5	1.0929	4.0504	0.0597	300
1	5634.2	1.3156	3.3587	0.0467	1410
Evaluation: GMM-10 Classes					
Class	**Annual Daily Irradiation (Wh/m^2)**	**Albedo**	***TL2***	**Cloudy Sky Index**	**Altitude (*mAMSL*)**
9	4736.95	1.3981	3.9828	0.0758	412
3	4997.43	1.0493	3.1392	0.045	1.900
8	5246.9	0.9156	4.2207	0.0662	66
1	5297.4	0.9402	3.4437	0.0467	1670
10	5366.4	0.8934	3.7703	0.0455	1350
5	5372.2	1.5972	3.6213	0.0458	1540
4	5378.4	1.2612	3.8461	0.0501	590
2	5528	1.0493	3.1392	0.045	1900
7	5891.0	1.9808	3.1757	0.0398	528

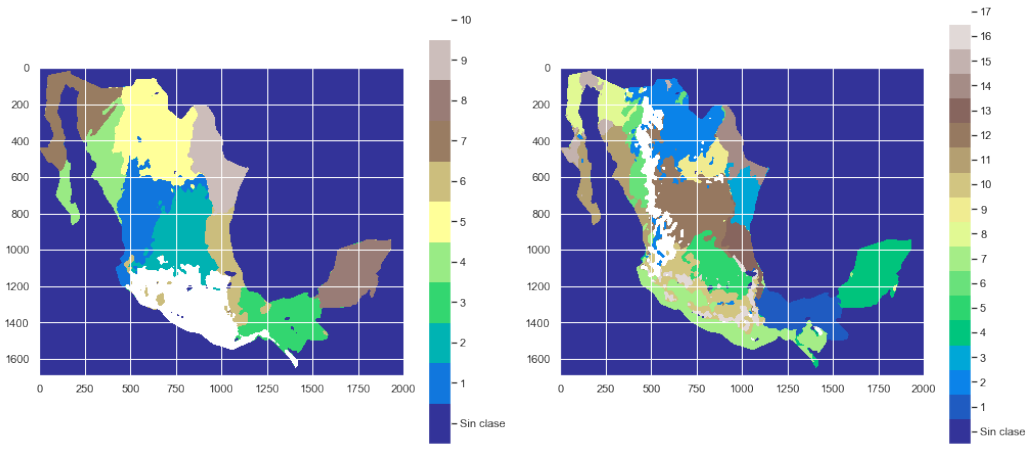
The linear relationship evaluated through multiple linear regression is described in Figure 7.

Evaluation	RMSE	R^2
K-means-17 Clusters	127.781	0.87
K-means-4 Clusters	6.4311×10^{-13}	1
GMM-10 Clusters	107.8616	0.85

Figure 7. *RMSE* and R^2 values of the relationship between geoclimatic variables and solar radiation.

As can be concluded from Figure 7, the k-means–four cluster evaluation has the best *RMSE* and R^2, compared with the cases with a larger number of classes. Still, it is more likely that the model was overfitting due to a low quantity of classes and centroid data. That is why this class does not represent a viable relationship with respect to solar radiation. However, the evaluations with 17 and 10 clusters give a better approximation to a good relationship between the variables (annual features) and the yearly daily irradiation of each cluster. The 17-class evaluation offers a better correlation with the score, while the 10-class evaluation offers the lower *RMSE*. The maps of the regionalization of Mexico—to 10 and 17—clusters are shown in Figure 8.

The regionalization obtained corresponds with the climatic characteristics that can be seen in the country. Such is the case with the warm subhumid zone of the Yucatan Peninsula, which corresponds to class 4 in Figure 8b, or the warm humid region of the state of Tabasco, north of Chiapas and south of the state of Veracruz, which corresponds to the class 1 of Figure 8b. The very dry regions of the north of the states of Sonora and Baja California are identified, corresponding to region 9 of Figure 8b. In both maps of Figure 8, it can be seen that, depending on the number of classes, the regionalization identifies the climatic zones in more detail when the number of classes increases, always based on the variables used for the cluster analysis.



(**a**) (**b**)

Figure 8. Mexico regionalization. (**a**) GMM 10 classes. (**b**) K-means 17 classes.

In all of the features, the best relationships between the annual daily irradiation is with the cloudy sky index shown in Figure 9.

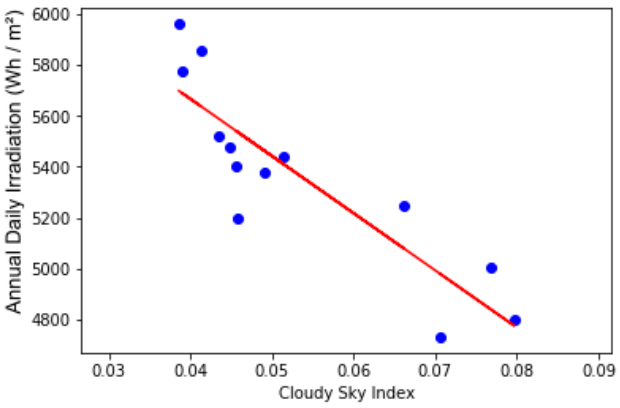


Figure 9. Relationship between the annual average daily irradiation and cloudy sky index in the evaluation of 17 classes.

The relationship shown in Figure 9 has a R^2 of 0.76. Still, thanks to the linear model with the ensemble with the other variables, it was possible to get a better relationship instead of only using the cloudy sky index.

5. Discussion

A map with optimal regions with the best correlations between solar radiation and related parameters derived from satellite images was obtained. Those parameters are relevant for the solar radiation intensity at the surface level through the Heliosat-2 and ESRA models to distribute a solar network according to climatic parameters related to solar

radiation at the surface level. Several algorithms were used and explained in detail: PCA, k-means, and GMM, and to validate the L method, DB and CH indices were applied. A preprocessing procedure was successfully applied to reduce the computational complexity of the algorithms. Based on the proposed methodology, it was possible to regionalize Mexico with 10 and 17 classes with two types of evaluations: an internal validation using the L method and a relationship evaluation between (non-satellite) solar energy data and the annual averages of albedo, TL2, the cloudy sky index, and altitude per cluster (derived from satellite images), with help from the multiple linear regression. It allowed obtaining a 0.87 score. The correlation shows a good agreement between the daily global solar irradiation obtained from pyranometers belonging to the EMAS network of the SMN and the corresponding class where each pyranometer is located. This fact is useful because the reliability of the regions is assured. Moreover, as can be expected, the cloudy sky index is the most relevant feature in this clustering work, and the altitude is the least.

K-Means and GMM are both unsupervised clustering techniques, but work differently. K-means groups data points using Euclidean distance for cluster membership. K-means is widely used due to its simplicity and speed. GMM uses a probabilistic assignment of data points to clusters, GMM uses the mean and the standard deviation, and each cluster is described by a separate Gaussian distribution. GMM is much slower as compared to clustering with k-means but works better in resolving the membership ambiguities that arise with overlapping clusters. Because the two algorithms work differently, it is to be expected that, when the boundaries between the clusters are not very clear, different results can be obtained. Mexico is one of the countries with the greatest diversity of ecosystems; between one place and another, the climatic conditions can vary considerably. This paper's maps allowed us to use this information to install solar networks through Mexico and as a reference to other solar energy studies. The evaluation of the clusters with 10 and 17 classes in this work can be used for the same purposes without having a problem between each; moreover, we could use any of the evaluation classes as the number of well defined stations in Mexico due to the high accuracy in the evaluation. With this, it can help decision making in order to maximize the number of stations that can be installed according to the financial resources of the solar network management. In this paper machine learning techniques were implemented for the first time in the regionalization of a country with great climatic complexity for the purposes of evaluating the solar resources. We should note that Mexico presents great climatic diversity, with mountain areas, coastal areas, highlands, deserts, tropical rain-forests, and temperate forests. For this reason, this methodology could be applied in other countries or geographic regions with vast geographic diversity. In the results, after applying some algorithms of machine learning, it was possible to observe the most optimal way to regionalize the country according to different climatic features, but also when performing the correlation of the cluster evaluation with the solar radiation data measured in surface stations, a good correlation was observed, which gave greater validity to the results obtained.

The main objective of making the correlation between irradiation and the cloudy sky index was to observe the strength of the relationship between the two variables; the results show that there is a relationship between the two variables, and the fact that it is not so strong means that there are other variables that would also correlate with irradiation apart from the cloudy sky index. This justifies our multivariate research and the following multivariate studies in solar irradiation.

For future works, it will be possible to evaluate the $RMSE$ and R^2 scores of the evaluation clusters above and below the optimal number of classes in order to evaluate the behaviors of the non-optimal number of classes near the optimal. Other algorithms can be applied for a better modeling of the relationship of the data, such as random forest, or support vector machines (SVMs). To determine the optimal number of classes, other indices, such as the silhouette index (SI), could be used—it gives a score between $[1, -1]$, a number near to 1 can be expressed as a better clustering, but the problem with this index is the quantity of data due to the complexity of the calculation for obtaining this metric.

For future works, other climatic features could be used for regionalization. Ultimately, methods for selecting features can be applied; these methods help to reduce the number of features in order to determine the best combinations of features that give better relationships between the data. Feature selection methods give better results of the relationships between features, selecting the best climatic features for using—and also reducing—the complexity of the algorithms.

Author Contributions: Conceptualization, J.D.S.-G., A.G.-H. and D.R.-R.; data curation, J.D.S.-G., A.G.-H., D.R.-R., G.M.-C. and A.E.G.-C.; formal analysis, J.D.S.-G., A.G.-H., D.R.-R. and C.E.G.-T.; investigation, J.D.S.-G., A.G.-H., D.R.-R., G.M.-C., L.F.Z., J.A.-M., C.E.G.-T. and A.M.-G.; methodology, J.D.S.-G., A.G.-H., D.R.-R., G.M.-C., C.E.G.-T. and A.M.-G.; project administration, A.G.-H. and D.R.-R.; resources, A.G.-H., D.R.-R., L.F.Z., J.A.-M. and A.E.G.-C.; software, A.M.-G.; supervision, A.G.-H. and D.R.-R.; validation, J.D.S.-G., A.G.-H., D.R.-R., G.M.-C., L.F.Z., J.A.-M., C.E.G.-T. and A.E.G.-C.; visualization, J.D.S.-G., A.G.-H., D.R.-R., G.M.-C., A.M.-G. and A.E.G.-C.; writing—original draft, J.D.S.-G.; writing—review and editing, A.G.-H., D.R.-R., G.M.-C., L.F.Z., J.A.-M., C.E.G.-T., A.M.-G. and A.E.G.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by General Direction of Personal Academic (DGAPA-UNAM) through the PAPIIT IN112320 project, by Ministerio de Ciencia e Innovación through the MAPV Spain Project (PID2020-118239RJ-I00) and also a special thanks to the Consejo Zacatecano de Ciencia, Tecnología e Innovación for their support of the research carried out and the results obtained.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CH	Calinski Harabasz
DB	Davis Bouldin
EM	expectation–maximization
EMAS	automatic weather station
GHI	global horizontal irradiance
GMM	Gaussian mixture models
GOES-13	Geostationary Operational Environmental Satellite-13
Lat	latitude
Lon	longitude
mAMSL	meters above mean sea level
NEDIS	National Environmental Satellite Data and Information Service
NetCDF	network common data form
NOAA	National Oceanic and Atmospheric Administration
PCA	principal component analysis
R^2	coefficient of determination
$RMSE_c$	root mean squared error of a critical point c
$RMSE_{L_c}$	root mean squared error on the left side of the critical point c
$RMSE_{R_c}$	root mean squared error on the right side of the critical point c
SI	silhouette index
SMN	National Weather Service
SVM	support vector machine
TL2	Linke turbidity
UNAM	National Autonomous University of Mexico
X_{DB}	database of the variables

References

1. Aitken, D. *Transitioning to a Renewable Energy Future*; ISES White Paper; International Solar Energy Society: Freiburg, Germany, 2003.
2. Holm, D.; McIntosh, J. Renewable energy–the future for the developing world. *Renew. Energy Focus* **2008**, *9*, 56–61. [CrossRef]
3. Riveros-Rosas, D.; Bonifaz, R.; Valdes, M.; Rivas, R. Análisis por Región de Información Solarimétrica en la República Mexicana. In Proceedings of the XI Congreso Iberoamericano de Energía Solar y XXXVIII Semana Nacional de Energía Solar, Querétaro, México, 6–10 October 2014.
4. Sengupta, M.; Habte, A.; Wilbert, S.; Gueymard, C.; Remund, J. *Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications*; Technical Report; National Renewable Energy Lab. (NREL): Golden, CO, USA, 2021.

5. Zagouras, A.; Kolovos, A.; Coimbra, C.F. Objective framework for optimal distribution of solar irradiance monitoring networks. *Renew. Energy* **2015**, *80*, 153–165. [CrossRef]
6. Martín-Pomares, L.; Romeo, M.G.; Polo, J.; Frías-Paredes, L.; Fernández-Peruchena, C. Sampling Design Optimization of Ground Radiometric Stations. In *Solar Resources Mapping*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 253–281.
7. Carvalho, M.; Melo-Gonçalves, P.; Teixeira, J.; Rocha, A. Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of temperatures and precipitation. *Phys. Chem. Earth Parts A/B/C* **2016**, *94*, 22–28. [CrossRef]
8. Zagouras, A.; Kazantzidis, A.; Nikitidou, E.; Argiriou, A. Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations. *Sol. Energy* **2013**, *97*, 1–11. [CrossRef]
9. Journée, M.; Müller, R.; Bertrand, C. Solar resource assessment in the Benelux by merging Meteosat-derived climate data and ground measurements. *Sol. Energy* **2012**, *86*, 3561–3574. [CrossRef]
10. Watanabe, T.; Takamatsu, T.; Nakajima, T.Y. Evaluation of variation in surface solar irradiance and clustering of observation stations in Japan. *J. Appl. Meteorol. Climatol.* **2016**, *55*, 2165–2180. [CrossRef]
11. Vindel, J.M.; Valenzuela, R.X.; Navarro, A.A.; Zarzalejo, L.F. Methodology for optimizing a photosynthetically active radiation monitoring network from satellite-derived estimations: A case study over mainland Spain. *Atmos. Res.* **2018**, *212*, 227–239. [CrossRef]
12. Vindel, J.M.; Valenzuela, R.; Navarro, A.A.; Zarzalejo, L.F.; Paz-Gallardo, A.; Souto, J.A.; Méndez-Gómez, R.; Cartelle, D.; Casares, J.J. Modeling Photosynthetically Active Radiation from Satellite-Derived Estimations over Mainland Spain. *Remote Sens.* **2018**, *10*, 849–862. [CrossRef]
13. Thanh Nga, P.T.; Ha, P.T.; Hang, V.T. Satellite-Based Regionalization of Solar Irradiation in Vietnam by k-Means Clustering. *J. Appl. Meteorol. Climatol.* **2021**, *60*, 391–402. [CrossRef]
14. Laguarda, A.; Alonso-Suárez, R.; Terra, R. Solar irradiation regionalization in Uruguay: Understanding the interannual variability and its relation to El Niño climatic phenomena. *Renew. Energy* **2020**, *158*, 444–452. [CrossRef]
15. De Lima, F.J.L.; Martins, F.R.; Costa, R.S.; Gonçalves, A.R.; Dos Santos, A.P.P.; Pereira, E.B. The seasonal variability and trends for the surface solar irradiation in northeastern region of Brazil. *Sustain. Energy Technol. Assess.* **2019**, *35*, 335–346.
16. Polo, J.; Gastón, M.; Vindel, J.; Pagola, I. Spatial variability and clustering of global solar irradiation in Vietnam from sunshine duration measurements. *Renew. Sustain. Energy Rev.* **2015**, *42*, 1326–1334. [CrossRef]
17. Olcoz Larraéyoz, A. *Implementación del Método Heliosat para la Estimación de la Radiación Solar a Partir de Imágenes de Satélite*; Technical Report; Universidad Pública de Navarra: Pamplona, Spain, 2014.
18. Rigollier, C.; Lefèvre, M.; Wald, L. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. *Sol. Energy* **2004**, *77*, 159–169. [CrossRef]
19. Gueymard, C.A.; Lara-fanego, V.; Sengupta, M.; Xie, Y. Surface albedo and reflectance: Review of definitions, angular and spectral effects, and intercomparison of major data sources in support of advanced solar irradiance modeling over the Americas. *Sol. Energy* **2019**, *182*, 194–212. [CrossRef]
20. Laguarda, A.; Abal, G. Índice de turbidez de Linke a partir de irradiación solar global en Uruguay. *Avances en Energias Renovables y Medio Ambiente* **2016**, *20*, 35–46.
21. Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media: Sebastopol, CA, USA, 2017.
22. Lantz, B. *Machine Learning with R: Expert Techniques for Predictive Modeling*; Packt Publishing Ltd.: Birmingham, England, 2019.
23. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: Berlin/Heidelberg, Germany, 2006.
24. Murphy, K.P. *Machine Learning a Probabilistic Perspective*, 1st ed.; The MIT Press: Cambridge, MA, USA, 2012; p. 1067.
25. Zagouras, A.; Pedro, H.T.; Coimbra, C.F. Clustering the solar resource for grid management in island mode. *Sol. Energy* **2014**, *110*, 507–518. [CrossRef]
26. Chi, Y. *R Tutorial with Bayesian Statistics Using Stan*, 1st ed.; R Tutorials: Cupertino, CA, USA, 2009; p. 563.
27. Govender, P.; Brooks, M.J.; Matthews, A.P. Cluster analysis for classification and forecasting of solar irradiance in Durban, South Africa. *J. Energy S. Afr.* **2018**, *29*, 51–62.
28. Riveros-Rosas, D.; Arancibia-Bulnes, C.; Bonifaz, R.; Medina, M.; Peón, R.; Valdés, M. Analysis of a solarimetric database for Mexico and comparison with the CSR Model. *Renew. Energy* **2015**, *75*, 21–29. [CrossRef]