ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 00 (2023) 1–18

NEURO

A Hybrid Intelligent Modeling approach for predicting the solar thermal panel energy production

Ángel Arroyo$^{0000-0002-1614-9075\text{a},*}$, Nuño Basurto$^{0000-0001-7289-4689\text{a}}$, Roberto Casado-Vara$^{0000-0003-0198-696X\text{a}}$, Míriam Timiraos$^{0000-0003-0982-6346\text{b}}$, José Luis Calvo-Rolle$^{0000-0002-2333-8405\text{b}}$

[a] *Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Digitalización, Escuela Politécnica Superior, Universidad de Burgos, Av. Cantabria s/n, 09006, Burgos, Spain.*
{aarroyo, nbasurto, rccasado}@ubu.es
[b] *University of A Coruña, CTC, CITIC, Department of Industrial Engineering. Avda. 19 de febrero s/n, 15405, Ferrol, A Coruña, Spain*
{miriam.timiraos.diaz, jlcalvo}@udc.es

Abstract

There is no doubt that the European Union is undergoing an ecological transition, with renewable energies accounting for an increasing share of energy consumption in the Member States. In Spain, solar energy is one of these rapidly expanding renewable sources. This study analyzes the solar energy production of a panel in the Spanish region of Galicia. It has been demonstrated that the solar energy produced by this panel can be predicted using a hybrid stepwise system. The missing value imputation is a key step in the process. This involves combining regression and clustering techniques on different subdivisions of the complete dataset, starting with a smaller and less complete dataset and performing appropriate imputations to create a larger and more complete collection. Finally, the dataset is divided into more relevant subsets for regression analysis to calculate the amount of solar energy generated. The imputing missing values using an Artificial Neural Network resulted in a more valid dataset for further processing than eliminating rows with corrupted or empty values. Also, properly applying clustering techniques gives better results than working on the whole dataset.

Keywords: Regression, Neural Network, Solar Energy, Renewable Energy, clustering

1. Introduction

Society and governments are concerned about protecting the environment. There are several reasons why environmental protection is essential. Although it may seem challenging to eliminate all harmful effects, striving towards sustainability and reducing environmental impact is crucial [1]. Renewable energy sources are necessary to reduce environmental damage and emissions when meeting energy needs [2]. However, it is important to consider the potential impacts of building a renewable energy plant, as some impact is usually present and it may not be possible to achieve complete harmlessness [3]. There is a legal requirement to optimize and design installations for maximum efficiency, as it is impossible to eliminate all impacts, even when using alternative and renewable energies [4]. Given text already follows principles or lacks context, it is crucial to measure the efficiency of installations using appropriate metrics and criteria to ensure the desired minimum impact is achieved [5]. It is also important to measure the efficiency of the institutions using appropriate ratios and criteria to ensure that the desired minimum impact is being obtained [5].

[*] Corresponding author

There is usually a significant environmental impact and economic investment during the construction phase of a power plant. Nevertheless, operational expenses generally decrease following the installation, dependent on technology [6]. Maintaining the plant in operation may be preferable to decommissioning it, depending on the circumstances [7]. Energy storage is strongly recommended in these cases [8]. Even though the end user does not generate energy, integrating renewable energy or other potential sources in different building types can pose significant energy management challenges. Developing mechanisms to ensure efficient management of energy generation and consumption across multiple facilities is therefore essential. The Smart Grid concept aims to optimize energy supply and demand by measuring and predicting energy generation and consumption, enabling efficient decision-making for the whole system. Regardless of the energy source, matching generation with demand remains challenging, so efforts must be made to minimize unnecessary energy purchases [9]. Some buildings with electricity requirements are not connected to the grid in specific geographic areas, where connection can be expensive and impractical. The implementation of energy storage systems and the isolation of energy generation, regardless of the nature of the source, is an alternative and more viable solution [10]. Many research and development projects have been carried out in recent years in response to the growing demand for energy generation and storage systems. Accurate forecasting can significantly improve the efficiency of buildings and installations, leading to the most favorable energy purchase and efficient energy storage [11]. Although many alternative solutions can be considered during the modeling process, performance variations may still occur.

Non-linearity presents a significant challenge in achieving optimal performance within the current problem. Despite the satisfactory results of simple intelligent systems in various applications, they struggle with non-linearity. The literature [12, 13, 14, 15, 16, 17, 18] demonstrates that soft computing techniques can frequently address non-linearity. To tackle non-linearity within a system, clustering techniques, like *k*-means [19, 20, 21, 22, 23, 24, 25], can be used to divide the problem. Furthermore, inference systems that apply other techniques dependent on fuzzy logic may also increase the system's precision [26]. This research focuses on simulating the solar thermal system of a bio-climatic home by utilizing real construction data. This research project focuses on modeling the solar thermal system of a bio-climatic home using real construction data. The research project aims to create a practical and efficient model for the solar thermal setup of bio-climatic homes. To address the non-linear nature of the problem, a clustering process was carried out before the regression stage [26].

This investigation aims to establish a regression model for a dataset comprising data on solar energy production through a solar panel. Various statistical techniques and Artificial Neural Network (ANN) models, alongside the frequently used *k*-means clustering technique, have been utilized to achieve this. A Hybrid Artificial Intelligent System (HAIS) was previously developed to anticipate solar energy production in a hybrid solar installation. This study employs advanced artificial neural network techniques to attain improved results. Furthermore, the application of *k*-means associated with regression techniques has been optimized, significantly improving the model's accuracy. Previous studies have utilized AI methodologies, such as phase-shifting material, to analyze one month's worth of data from solar collectors [27]. Recent research has proposed using artificial neural networks to predict the global horizontal, beam normal, and diffuse horizontal components of solar radiation [28]. However, regression techniques with clustering to minimize Mean Squared Error (MSE) in solar energy prediction have not been comprehensively addressed in any study to date. This paper addresses this research gap by presenting a case study on predicting solar energy and describing the methods implemented, such as the conflation of regression methods and clustering. It also presents the results, how these were analyzed and concluded, and future directions.

2. Case of Study

The solar thermal installation in a real bioclimatic residence was utilized to evaluate the methodology presented. Here is a succinct depiction of the physical system.

2.1. *Sotavento bioclimatic house*

The Sotavento bioclimatic house, shown in Figure 1, exemplifies the concept of bioclimatic design. It is located within the Sotavento Experimental Wind Park, which aims to promote energy conservation and the use of renewable energy. The park is located between the municipalities of Xermade (Lugo) and Monfero (A Coruña) in the Spanish autonomous community of Galicia.

Figure 1: Bioclimatic house

2.2. *Bioclimatic house facilities*

The thermal and electrical installations of the bioclimatic house have been augmented with renewable energy systems to complement the current facilities. Fig. 2 illustrates the systems and components related to thermal concerns. The thermal engineering is powered by three sources of renewable energy - solar, biomass, and geothermal, which provide Domestic Hot Water (DHW) and heating. The electrical installation consists of wind and photovoltaic energy sources and is connected to the electricity grid providing energy for the lighting and the house's electrical systems.

2.3. *Solar thermal system*

The solar thermal system of the bioclimatic house is exemplified in Figure 3. The diagram outlines the process of solar energy collection and storage. The circulation of hot liquid is depicted by the solid red line, while the dashed blue line signifies the circulation of cold liquid. The solar panels, located on the north facade inclined at an angle of 19° form the collection area of the solar thermal system. A closed ethylene-glycol hydraulic circuit links the solar panels and the solar storage tank.

3. Modeling approach

This section details the methods utilized for analyzing the dataset. Initially, data is acquired to obtain a dataset for the following phases. Subsequently, a data preprocessing stage is performed to prepare the data for further processing. The imputation of missing values is addressed through the use of an MLP and the specific Levenberg-Marquardt backpropagation algorithm. The subsequent action involved bifurcating the dataset into two distinct sets, namely training and test sets. The segregation was pivotal to implementing the pre-selected machine learning models. Two sets were essentially developed to train the models and evaluate their proficiency in prognosticating results. The training data is scheduled to be utilized in the ensuing stage, which is the Clustering stage, where the requisite clusters are to be obtained to construct regression models. Once the regression models become available, they will undergo testing using the test dataset in a specific phase. From the outcomes, metrics, and errors obtained in this stage, a model will be chosen that forms the selection phase and culminates in the final model. Please refer to Figure 4 for a visual representation of this process.

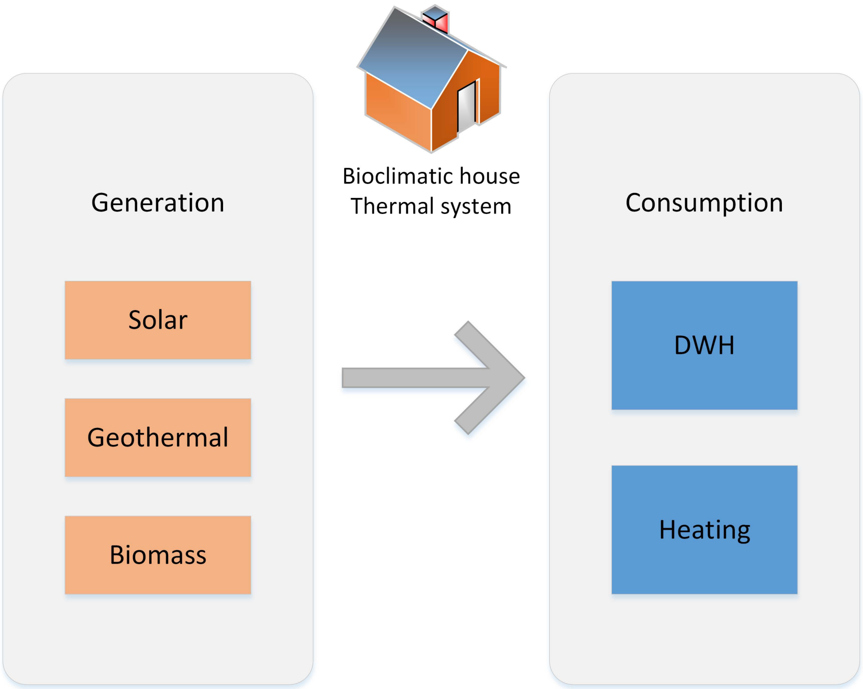

Figure 2: Thermal facilities scheme

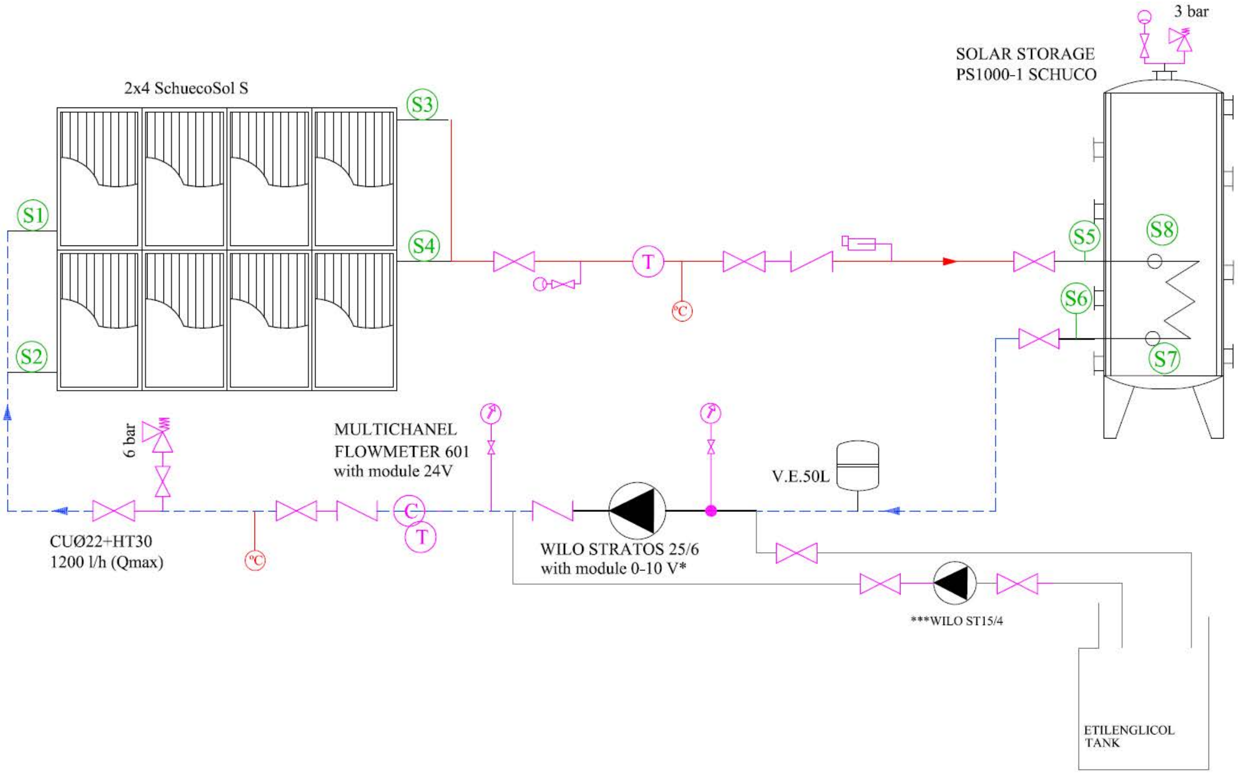

Figure 3: Solar thermal system scheme

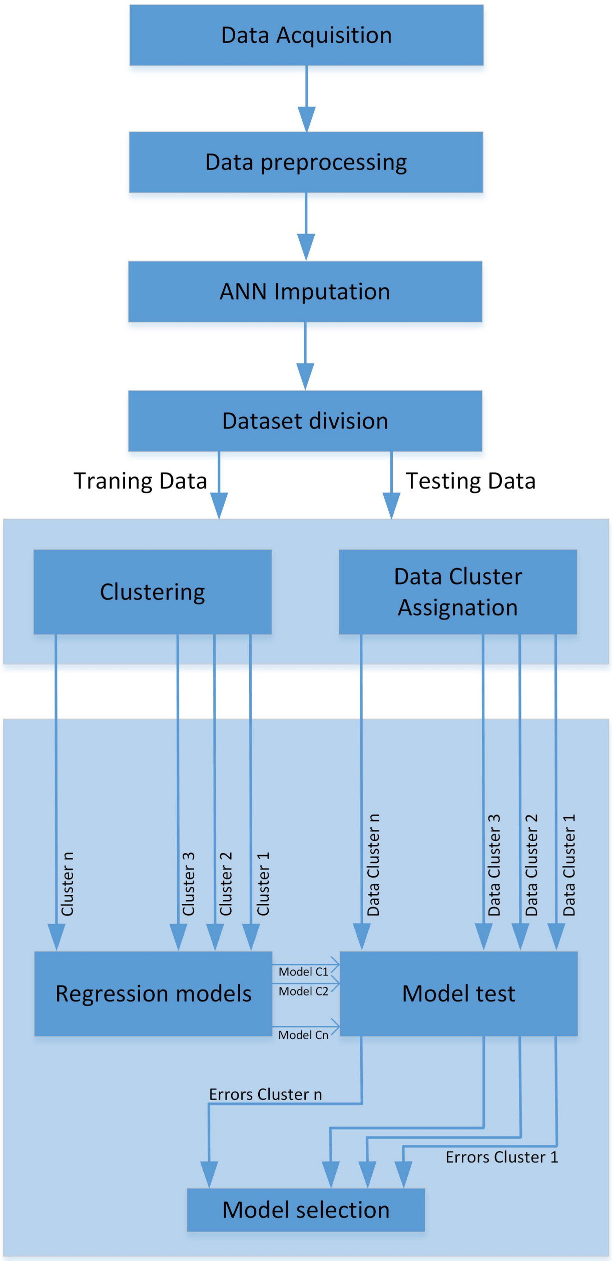

Figure 4: Modeling approach

3.1. Multiple Linear Regresion

Multiple linear regression (M-LR) is a generalization of simple linear regression because this approach makes it possible to relate a variable to more than one variable through a linear function in its parameters. Multiple linear regression can be used to estimate the relationship between two variables by allowing other variables to influence one another. The effect of other variables is canceled out to isolate and measure the relationship between the two variables of interest. The following equation defines MLR models:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{1}$$

where $i = 1, ..., p \in \mathbb{N}$ are the observations, Y is the dependent variable, X_i are explanatory variables, β_0 is the constant term, β_i are the slope coefficients for each explanatory variable and ϵ the model's error term [29]. Estimating the β_i parameters by the least squares method (LSM) is based on the same principle as simple linear regression but applied to p dimensions. Therefore, it is no longer a question of finding the best line but of finding the p-dimensional plane that passes as close as possible to the coordinate points. The sum of the squared deviations of the points on the plane X_i, X_j are minimized. An adjusted estimate of the coefficients β_i results from the least squares method. The term adjusted means after accounting for the linear effects of the other independent variables on the dependent variable and the predictor variable. The hypotheses are the same as for simple linear regression, that is:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \tag{2}$$

Testing $\beta_j = 0$ is equivalent to testing a hypothesis (i.e., is there an association between a dependent variable and independent variable being studied, all other things being equal, that is, other independent variables held constant) [30, 31].

3.2. Artificial Neural Networks

Artificial Neural Networks are simplified models of natural neural systems. In this research work, the following ones were applied:

3.2.1. Multilayer perceptron

The Multilayer Perceptron (MLP) is a common type of Artificial Neural Network (ANN) which is comprised of numerous layers of nodes. These nodes are linked by weights and generate output signals by computing the activation of the input sum. The MLP's structure includes an input layer that transfers the input vector to the remaining network layers [32]. The terms 'input vectors' and 'output vectors' refer to the inputs and outputs of the MLP and are represented as individual vectors. In addition to the output layer, a Multilayer Perceptron (MLP) will have one or more hidden layers in addition to the output layer. MLPs are fully connected, which means that each node is connected to all of the nodes in the layers that precede and follow it. The backpropagation algorithm is used to update weights during training. In this study, we have implemented the Levenberg-Marquardt (LM) algorithm [33, 34].

The Levenberg-Marquardt algorithm is an iterative optimization technique that decreases the residuals of a non-linear least squares issue [35]. It combines features of gradient descent and Gauss-Newton methods to provide a reliable approach to curve fitting and parameter estimation. The algorithm allows efficient navigation through parameter space by adjusting the step size at each iteration. It ensures convergence in various optimization scenarios by balancing the steepest descent and Gauss-Newton approaches (see algorithm 1).

This pseudo-code represents the necessary steps of the Levenberg-Marquardt algorithm. These steps include computing the Jacobian matrix, modifying the Hessian, and updating the parameters. The algorithm provides optimal fitting of nonlinear regression problems by progressively adjusting the parameters to reduce the difference between the predicted and observed values.

Algorithm 1 Levenberg-Marquardt Algorithm pseudo-code [36]

```
while not Convergence or not maximum iterations reached do
    Compute the Jacobian matrix and residuals
    Compute the approximate Hessian matrix
    Modify the Hessian with damping factor $\lambda$
    Solve the linearized system of equations
    Update parameters
    Evaluate new residuals
    if New residuals are reduced then
        Reduce damping factor $\lambda$
    else
        Increase damping factor $\lambda$
        Revert parameters and residuals
    end if
end while
```

3.2.2. Extreme Learning Machine

Extreme Learning Machine (ELM) tries to overcome some challenges faced by other techniques. ELM works for generalized Single-hidden Layer Feedforward Networks (SLFNs). The ELM concept is that the hidden layer of SLFNs need not be adjusted. Compared with other regression ANN techniques, ELM provides better generalization performance at a much faster learning speed and with the least parametric setting [37]. ELM can be summed up as follows: the hidden layer of ELM need not be iteratively adjusted [38]; According to Feedforward neural network theory [39], the training error and the norm of weights need to be minimized [40] and finally, the hidden layer feature mapping need to satisfy the universal approximation condition [40].

3.3. Clustering technique

Cluster analysis [41] organizes data by grouping data samples according to a criterion distance. Two individuals in a valid group will be much more similar than those in distinct groups.

3.3.1. k-means

k-means is an iterative algorithm that divides the dataset into K-predefined, non-overlapping clusters, with each data point belonging to only one cluster. It attempts to make the data points within a cluster as similar as possible while keeping the clusters as different (as far apart) as possible. The data points are allocated to a cluster so that the sum of the squared distances between the data points and the cluster centroid is minimal. The smaller the variation within a cluster, the more homogeneous (similar) the data points will be within the same cluster [42]. The *k*-means algorithm works as follows:

- The number of clusters is specified as K.
- Initialise the centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacing them.
- Continue iterating until the centroids do not change, i.e., the assignment of data points to clusters does not change.

To determine the optimal number of clusters (K), the following techniques were used in this study: (1) Calinski–Harabasz Index: Based on the mean between clusters and the covariance matrix within clusters, the Calinski-Harabasz index assesses the validity of clusters. It measures the separation in terms of the maximum distance between the centres of the clusters and the compactness as the sum of the distances between the objects and their cluster centre [43]. (2) Davies: Similar to the Calinski-Harabasz index, the Davies-Bouldin index gives the clusters the minimum distance within the cluster and the maximum distance between the centroids of the clusters. An adequate dataset partitioning is indicated by the minimum value of the index [44]. (3) Silhouette: The Silhouette Index evaluates the clustering performance based on the difference between the distances between and within clusters. Maximizing the value of this index also determines the optimal number of clusters [45].

The distance at which the points are measured is another important hyperparameter that allows tuning the performance of *k*-means to find the optimal model. The Cosine distance and the city block were used in this paper. The Cosine distance [46] is one of the most popular measures of textual similarity and is an essential computational burden in document comprehension tasks.

$$cos(x, y) = \frac{\sum_{i=1}^{m} x_i y_i}{\sqrt{\sum_{i=1}^{m} x_i^2} \sqrt{\sum_{i=1}^{m} y_i^2}} \tag{3}$$

On the other hand, the Cityblock distance [47] is the distance measure applied where each centroid is placed in the component-wise median of all the samples in the group. The distance from point *x* to each of the centroids is calculated as:

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - x_{tj}| \tag{4}$$

Where *j* is an instance of the vector *j*.

3.3.2. Experiments setup

First, data acquisition is carried out, which allows us to obtain a data set for the subsequent phases. Once the data set is available, it has to be prepared for further processing, so a data preprocessing stage is performed. With the data prepared, an MLP with the specific Levenberg-Marquardt backpropagation algorithm is used to deal with the imputation of missing values.

The next step was to divide our data set into two groups: training and test sets. This division was crucial for applying the machine learning models previously selected. Essentially, two sets have been created for training the models and test their performance and accuracy in predicting outcomes. The training data will be used in the next stage, the Clustering stage, in which a certain number of clusters will be obtained and used to build the regression models. Once these regression models are available, they will be tested with the test data set at a specific phase. With the achieved metrics and errors from this stage, a model will be selected, which will constitute the selection phase and lead to the final model.

4. Results and discussion

This section shows the results of applying the techniques detailed in Section 3 to the dataset described in detail in Section 2. The initial task involves addressing missing values in the dataset. Our dataset has undergone a transformation from the previously filtered dataset utilized in the previous work [48]. This earlier dataset had 34,645 rows and 7 columns, whereas our present dataset encompasses 52,685 rows and the same 7 columns as those in the outcome presented below. The missing data values are not evenly distributed across the twelve months of the analysis year. In particular, 90% of the missing values are found in the column related to the average radiation value, while 10% of the values are missing in the columns related to the average electricity production and the water flow rate of the solar thermal system. To address this issue, an MLP with the Levenberg-Marquardt Backpropagation training algorithm was implemented to impute the missing values. The imputation was an iterative process. The initial row with missing values was imputed by implementing regression on the prior group of rows. Once this group was imputed and became part of a new complete set of rows, any contiguous rows with missing values were then imputed, and the process restarted from the first row of this new dataset.

Once the complete dataset, consisting of the aforementioned 52,685 rows and 7 columns, is generated, the case study under examination will enable reliable analyses. Subsequently, two aspects can be demonstrated through the combination of regression tests and clustering techniques:

- Regression analysis on the whole dataset produces superior findings to that of a dataset where missing values have been removed.

- Regression results improve when the dataset is divided into appropriate subsets, which are subsequently clustered into data clusters, as compared to using the entire dataset.

Regression techniques were applied to various subsets of data in order to predict the thermal power generated by the solar system. The Cross-Validation (CV) [49] validation scheme was utilized for all regressions. CV is a method that separates the data into two subsets: training and testing. The training samples were used to train each algorithm whilst the testing samples were used to validate the results. The parameter *n* was consistently set to 10 for all experiments conducted in this study to partition the data.

Table 1 demonstrates the effects of employing the two regression methods, M-LR and ELM, on the entire dataset following the imputation of missing values and the incomplete set after removing missing values rows.

Table 1: Regression results on the full and incomplete dataset.

Dataset	Technique	Time Execution	Std Time	MSE	Std MSE
Complete Dataset	*M-LR*	1.80E-01	3.00E-01	4.31E-27	4.19E-29
	ELM	4.31E-02	3.36E-02	**2.32E-29**	1.37E-31
Incomplete Dataset	*M-LR*	1.34E-01	1.51E-01	2.24E-05	8.11E-08
	ELM	**3.22E-02**	3.11E-02	1.80E-08	2.06E-10

To achieve the entire dataset, missing values were estimated in nearly 20,000 rows, as previously mentioned in this section. Table 1 illustrates that the ELM technique significantly enhances results in terms of MSE; however, the M-LR scenario remains almost identical. Execution times, conversely, see a slight deterioration due to the larger dataset, but it is not significant compared to the substantial gain achieved in the MSE calculation. The process of data imputation leads to a significantly more extensive dataset comprising of high-quality information. This dataset can be effectively used in other procedures involving data processing.

Table 2 shows the results of applying the three indices to the completed dataset for calculating the optimal value of parameter k (number of clusters), described in Section 2.

Table 2: Optimum value for parameter K.

	Calinski	Davies	Silhouette
Time	2.35	2.49	39.29
Optimal K	6	3	3

To ascertain the most favourable K value, we utilized the three indices described in Section3, illustrated in Table 2. Each value within the range necessitates evaluation with the provided indices. Clarification of technical terminologies should be included upon first usage. Notably, two of these indices have delivered optimal parameter values of 3 and 6, for K and Calinski, respectively, within the range of 2 to 10. It is worth noting that among the three, the Silhouette index exhibited the slowest performance, displaying significant variability in its execution time.

Table3 shows the distribution of samples per cluster for the complete dataset, applying *k*-means and the two selected distance measures: Cityblock and Cosine.

Table 3: Samples distribution for the complete Dataset.

	Measure	
Cluster	Cosine	City
C1	42735	21220
C2	6423	7418
C3	3527	24047

Using *k*-means, the distance of Cityblock and Cosine, and the dataset segmented into the twelve months and the four weather seasons, Table 4 shows the distribution of samples per cluster.

Tables 3 and 4 provide an overview of how the samples were distributed among the three clusters created for the entire dataset, the complete dataset divided into the four meteorological stations, and finally, the dataset divided into

Table 4: Sample distribution by cluster.

	Measure					
Month	**City**	**Cosine**	**City**	**Cosine**	**Season**	**Cluster**
1	3826	473	10898	5605	Winter	*C1*
2	3469	1765	10735	4226	Spring	*C1*
3	3544	2115	9783	2459	Summer	*C1*
4	3191	714	9837	5896	Autumn	*C1*
5	3253	1427				*C1*
6	3429	804				*C1*
7	3329	779				*C1*
8	3268	866				*C1*
9	559	2653				*C1*
10	695	2774				*C1*
11	3013	2379				*C1*
12	3970	1930				*C1*
1	181	2100	1571	5349	Winter	*C2*
2	154	668	963	2498	Spring	*C2*
3	347	1482	1843	7710	Summer	*C2*
4	441	3384	1731	1342	Autumn	*C2*
5	464	693				*C2*
6	384	1208				*C2*
7	581	2555				*C2*
8	787	1115				*C2*
9	741	967				*C2*
10	2668	1119				*C2*
11	254	331				*C2*
12	42	2002				*C2*
1	457	1891	491	2006	Winter	*C3*
2	409	1599	1371	6345	Spring	*C3*
3	573	867	1752	3209	Summer	*C3*
4	652	186	1673	6003	Autumn	*C3*
5	748	2345				*C3*
6	507	2308				*C3*
7	554	1130				*C3*
8	408	2482				*C3*
9	3150	830				*C3*
10	1095	565				*C3*
11	1053	1610				*C3*
12	452	532				*C3*

the twelve months of the year. In contrast to the previous post, Cityblock distance measurement was used along with Cosine. Using Cosine is justified since it leads to a more equitable distribution of samples among the developed data clusters. The datasets presented in these tables were utilized to generate the results in the subsequent tables.

Table 5 shows the MSE and runtime results of applying the regression techniques to the subsets of data produced by *k*-means with the two distance measures to the complete dataset.

Table 5: Regressions on the full clustered dataset.

Measure	**Cluster**	**Time M-LR**	**MSE M-LR**	**Time ELM**	**MSE ELM**
City	*C1*	1.29E-01	1.67E-05	3.56E-02	1.02E-27
City	*C2*	2.58E-02	1.53E-05	1.12E-02	4.16E-33
City	*C3*	2.47E-02	1.15E-31	**9.61E-03**	**3.95E-34**
Cosine	*C1*	2.62E-02	3.21E-05	1.55E-02	1.86E-29
Cosine	*C2*	2.25E-02	2.61E-30	1.03E-02	4.22E-27
Cosine	*C3*	3.08E-02	7.81E-29	1.95E-02	3.76E-31

Table 5 examines the performance of the M-LR and ELM methods when applied to the complete dataset and segmented into three data clusters. It was observed that both techniques yield favorable MSE values, particularly the ELM approach. Furthermore, clustering the data before regression resulted in better MSE values than working with ungrouped data. These findings demonstrate the utility of data clustering as a pre-processing step for regression. Upon analyzing the results for each distance measure, it was found that Cosine produces the best MSE for both M-LR and ELM, as well as C3. The ELM method was also observed to exhibit the fastest execution times. It should be noted that the calculation of MSE by M-LR exhibits high variability in its results.

Table 6 shows the results of applying the M-LR and MSE to the subsets of data produced by *k*-means with the two distance measures to the four meteorological seasons subsets of data. in Table 6, the whole dataset has been split into four subsets, which correspond to the four meteorological seasons of the year. Thereafter, each of these four groups has been subjected to a *k*-means analysis. One can see how the MSE metrics produced in Table 5 and in Table 1 are improved by M-LR and, in particular, ELM. The four seasons of the year produce incredibly low values for MSE. Which of the two distance measurements produces superior outcomes cannot be determined with certainty. The run times for the ELM are shorter than those for the M-LR, which is consistent with the earlier findings.

It is worth noting that for cluster C2, with the Cityblock measure and for the summer season, an MSE of 0 is achieved for the two regression techniques used. This is a cluster with 1,843 rows that did not appear in the filtered file, which again shows the effectiveness of the first step of the imputation of missing values. As a not-so-positive part of the results presented, it can be mentioned the high variability in the results obtained depends mainly on the cluster treated. When dividing the samples into clusters, there will always be one with highly related samples and the other two to a lesser extent.

Figure 5 provides a visual representation of the runtimes shown in Table 6. It is evident that ELM has shorter run times than M-LR across all four seasons. The summer season exhibits the greatest variability, while the autumn season has the least, with similar run times in between. There seems to be no obvious cause for the variation in running times between seasons as the numerical disparities are negligible, and there are no significant distinctions in the distribution of samples in datasets.

Figure 6 corresponds to the mean MSE values shown in the table Table 6. It is observed that except for the spring, which offers exceptionally high values for the measurement of the Cityblock distance using the M-LR technique, the other three meteorological stations offer average values of MSE without great differences between them in the two techniques applied for Cosine and Cityblock. At a glance, the best values in calculating the MSE by applying MSE and, in the case of Cityblock, are shown. This is because Cityblock performs a very unbalanced grouping of the samples in groups so that in small but highly homogeneous groups, the best values in the calculation of the MSE are obtained.

Table 7 shows the results of applying the regression techniques to the subsets of data produced by *k*-means to the twelve months of the year.

Table 6: MSE and run times for the 4 weather seasons clustered by *k*-means.

Season	**Distance**	**Cluster**	**Time M-LR**	**MSE M-LR**	**Time ELM**	**MSE ELM**
Winter	*City*	*C1*	2.45E-02	1.63E-24	2.10E-02	1.90E-26
Winter	*City*	*C2*	2.68E-02	3.22E-32	1.04E-02	1.11E-33
Winter	*City*	*C3*	2.31E-02	3.58E-33	8.39E-03	1.32E-34
Winter	*Cosine*	*C1*	2.56E-02	2.04E-30	1.06E-02	2.13E-31
Winter	*Cosine*	*C2*	2.42E-02	3.85E-08	9.65E-03	1.92E-08
Winter	*Cosine*	*C3*	2.66E-02	1.75E-04	9.56E-03	2.01E-07
Spring	*City*	*C1*	3.05E-02	8.58E-05	1.28E-02	1.66E-28
Spring	*City*	*C2*	2.25E-02	8.77E-31	8.50E-03	2.13E-33
Spring	*City*	*C3*	2.31E-02	1.05E-32	8.85E-03	9.27E-34
Spring	*Cosine*	*C1*	2.09E-02	1.88E-04	8.29E-03	2.77E-28
Spring	*Cosine*	*C2*	2.22E-02	3.06E-06	8.03E-03	6.25E-10
Spring	*Cosine*	*C3*	2.36E-02	1.20E-30	9.43E-03	1.70E-31
Summer	*City*	*C1*	2.41E-02	5.34E-25	1.01E-02	7.24E-27
Summer	*City*	*C2*	1.99E-02	**0**	7.53E-03	**0**
Summer	*City*	*C3*	2.64E-02	1.29E-29	9.82E-03	7.95E-33
Summer	*Cosine*	*C1*	2.22E-02	**0**	8.69E-03	4.00E-09
Summer	*Cosine*	*C2*	2.33E-02	2.82E-30	9.59E-03	1.63E-31
Summer	*Cosine*	*C3*	2.72E-02	9.06E-05	9.71E-03	1.61E-07
Autumn	*City*	*C1*	2.48E-02	4.07E-25	1.08E-02	1.45E-27
Autumn	*City*	*C2*	2.11E-02	1.23E-25	7.59E-03	3.43E-28
Autumn	*City*	*C3*	2.08E-02	6.69E-25	8.14E-03	2.04E-27
Autumn	*Cosine*	*C1*	2.17E-02	7.04E-26	9.11E-03	2.03E-28
Autumn	*Cosine*	*C2*	2.06E-02	4.29E-04	*7.25E-03*	3.36E-06
Autumn	*Cosine*	*C3*	2.15E-02	5.61E-25	9.04E-03	1.92E-27

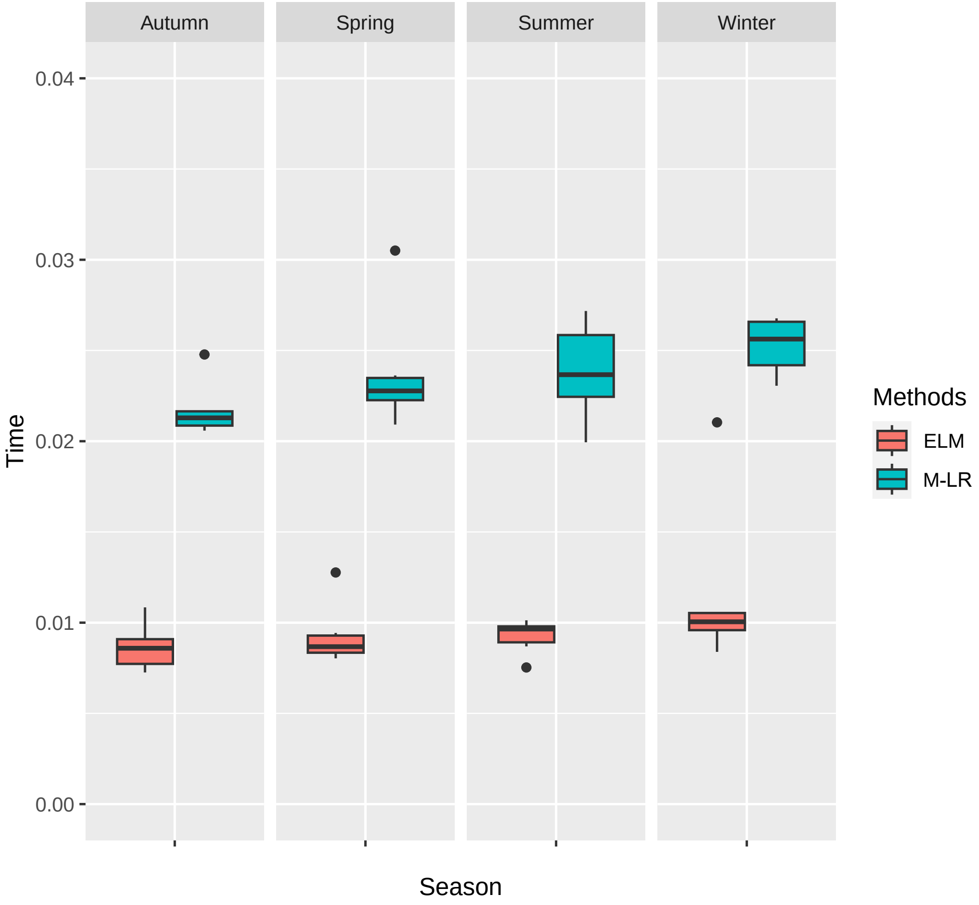

Figure 5: Boxplot depicting the execution time per method and season.

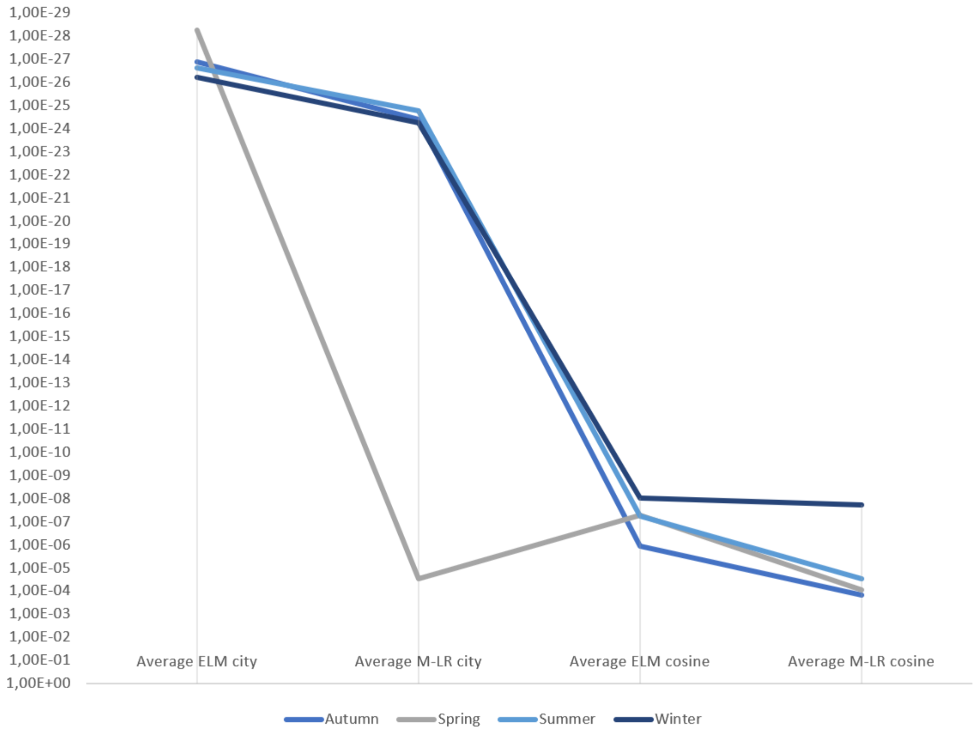

Figure 6: Line chart depicting the MSE per season and distance measure.

Table 7: MSE and runtimes for the twelve months of the year clustered using *k*-means.

Month	Measure	Cluster	Time M-LR	MSE M-LR	Time ELM	MSE ELM
1	*City*	*C1*	2.74E-01	2.27E-24	5.00E-02	4.48E-26
1	*City*	*C2*	5.85E-02	2.63E-32	1.17E-02	5.78E-34
1	*City*	*C3*	2.23E-02	5.58E-31	9.78E-03	1.33E-32
1	*Cosine*	*C1*	2.08E-02	8.80E-04	**8.71E-03**	1.91E-06
1	*Cosine*	*C2*	2.12E-02	5.82E-08	1.03E-02	1.45E-07
1	*Cosine*	*C3*	2.08E-02	3.98E-30	9.74E-03	4.94E-31
2	*City*	*C1*	2.08E-02	3.45E-24	9.98E-03	4.69E-26
2	*City*	*C2*	2.11E-02	**0**	9.86E-03	**0**
2	*City*	*C3*	2.13E-02	**0**	9.63E-03	**0**
2	*Cosine*	*C1*	2.24E-02	2.17E-08	9.77E-03	2.45E-08
2	*Cosine*	*C2*	2.23E-02	5.12E-04	1.06E-02	7.39E-07
2	*Cosine*	*C3*	2.03E-02	7.52E-30	9.87E-03	8.29E-31
3	*City*	*C1*	2.14E-02	2.15E-23	9.86E-03	1.11E-25
3	*City*	*C2*	2.16E-02	**0**	9.60E-03	**0**
3	*City*	*C3*	2.06E-02	**0**	9.42E-03	**0**
3	*Cosine*	*C1*	2.41E-02	5.81E-30	9.56E-03	5.54E-31
3	*Cosine*	*C2*	2.18E-02	2.41E-07	1.00E-02	6.71E-10
3	*Cosine*	*C3*	2.21E-02	3.08E-04	1.00E-02	2.28E-07
4	*City*	*C1*	2.15E-02	5.35E-25	9.66E-03	4.71E-27
4	*City*	*C2*	2.03E-02	**0**	9.42E-03	**0**
4	*City*	*C3*	2.02E-02	**0**	1.00E-02	**0**
4	*Cosine*	*C1*	2.24E-02	**0**	1.06E-02	8.22E-09

4	*Cosine*	*C2*	2.01E-02	3.20E-31	9.31E-03	4.38E-32
4	*Cosine*	*C3*	2.07E-02	1.15E-03	9.71E-03	7.09E-07
5	*City*	*C1*	2.18E-02	3.76E-24	9.86E-03	4.41E-26
5	*City*	*C2*	2.49E-02	3.73E-30	1.08E-02	8.68E-33
5	*City*	*C3*	2.15E-02	2.51E-32	9.85E-03	2.54E-33
5	*Cosine*	*C1*	2.21E-02	2.20E-04	9.95E-03	3.05E-07
5	*Cosine*	*C2*	2.24E-02	**0**	1.06E-02	2.07E-08
5	*Cosine*	*C3*	2.00E-02	5.83E-30	9.54E-03	6.82E-31
6	*City*	*C1*	2.24E-02	1.66E-04	1.00E-02	1.34E-27
6	*City*	*C2*	2.60E-02	9.26E-05	1.14E-02	2.13E-32
6	*City*	*C3*	2.02E-02	1.08E-04	9.91E-03	2.54E-32
6	*Cosine*	*C1*	1.99E-02	6.85E-29	9.39E-03	9.31E-30
6	*Cosine*	*C2*	2.39E-02	2.16E-06	1.01E-02	6.53E-11
6	*Cosine*	*C3*	2.14E-02	2.70E-04	9.66E-03	4.33E-29
7	*City*	*C1*	2.32E-02	1.95E-24	1.13E-02	2.45E-26
7	*City*	*C2*	2.46E-02	1.14E-28	1.24E-02	5.32E-32
7	*City*	*C3*	1.99E-02	**0**	9.13E-03	**0**
7	*Cosine*	*C1*	2.03E-02	8.00E-08	9.50E-03	2.84E-08
7	*Cosine*	*C2*	2.01E-02	2.01E-29	9.98E-03	7.18E-31
7	*Cosine*	*C3*	2.40E-02	2.66E-04	1.02E-02	4.21E-07
8	*City*	*C1*	2.37E-02	2.70E-24	1.11E-02	3.95E-26
8	*City*	*C2*	1.96E-02	**0**	9.30E-03	**0**
8	*City*	*C3*	2.38E-02	**0**	1.12E-02	**0**
8	*Cosine*	*C1*	2.04E-02	**0**	9.97E-03	5.27E-10
8	*Cosine*	*C2*	2.05E-02	2.43E-04	1.02E-02	4.48E-07
8	*Cosine*	*C3*	2.05E-02	2.89E-30	9.92E-03	6.56E-31
9	*City*	*C1*	2.04E-02	**0**	9.54E-03	**0**
9	*City*	*C2*	2.04E-02	**0**	9.74E-03	**0**
9	*City*	*C3*	2.13E-02	1.31E-24	1.10E-02	1.76E-26
9	*Cosine*	*C1*	2.27E-02	2.67E-30	1.12E-02	2.07E-31
9	*Cosine*	*C2*	2.11E-02	3.13E-04	9.21E-03	5.30E-07
9	*Cosine*	*C3*	1.99E-02	**0**	9.43E-03	0
10	*City*	*C1*	2.13E-02	8.15E-25	1.05E-02	9.68E-28
10	*City*	*C2*	2.03E-02	1.13E-22	9.54E-03	4.63E-26
10	*City*	*C3*	1.97E-02	1.95E-25	8.96E-03	6.16E-28
10	*Cosine*	*C1*	2.08E-02	1.10E-25	1.08E-02	3.44E-28
10	*Cosine*	*C2*	2.15E-02	3.74E-04	9.14E-03	6.44E-08
10	*Cosine*	*C3*	2.05E-02	**0**	9.04E-03	**0**
11	*City*	*C1*	2.10E-02	1.21E-22	9.88E-03	7.44E-27
11	*City*	*C2*	2.01E-02	7.93E-25	9.26E-03	7.11E-30
11	*City*	*C3*	2.19E-02	9.43E-25	1.09E-02	7.43E-28
11	*Cosine*	*C1*	2.28E-02	2.74E-25	1.05E-02	7.10E-28
11	*Cosine*	*C2*	2.03E-02	2.76E-30	9.46E-03	1.96E-31
11	*Cosine*	*C3*	2.15E-02	3.94E-04	9.84E-03	6.66E-09
12	*City*	*C1*	2.02E-02	1.70E-27	9.35E-03	3.57E-29
12	*City*	*C2*	2.19E-02	**0**	9.82E-03	**0**
12	*City*	*C3*	2.28E-02	2.66E-31	1.08E-02	3.94E-31
12	*Cosine*	*C1*	2.07E-02	9.18E-31	9.08E-03	1.03E-30
12	*Cosine*	*C2*	2.06E-02	3.47E-27	9.01E-03	6.88E-29
12	*Cosine*	*C3*	2.21E-02	4.66E-27	1.00E-02	1.34E-27

Finally, the dataset was divided into twelve subsets, corresponding to the twelve months of the year, as shown in Table 7. Following this, the *k*-means clustering was conducted on each of these twelve subsets of data, utilizing both Cosine and Cityblock distance measures. Consistent with the findings outlined in Table 5, both regression methods produced notably low mean squared error (MSE) scores. Specifically, MLE occurred in 16 cases and M-LR in 17 cases, all yielding a value of 0.

Moreover, the number of occurrences where a value of 0 was obtained in the MSE calculation was more frequent with the Cityblock measure than with Cosine. Even after excluding these occurrences, the MSE value for ELM remained lower than for M-LR in most clusters.

It is important to note the considerable variability in the results obtained, as the same month may yield significantly different results depending on the distance measure and the cluster. This suggests that there may be one or two clusters with highly related samples and a third cluster with the most dissimilar samples, resulting in poorer outcomes. Such variability was not as pronounced in Table 6, where the results did not exhibit such high variability.

The best configuration is to work on the imputed data set, not on the filtered set of missing values. Then, apply the Extreme Linear Machine by dividing the data set by months or by meteorological quarters on which we will apply k-means with a value of k=3 and preferably using the Cityblock distance measure, which gives slightly better results than the Cosine distance. As to whether it is better to divide the year into 12 subsets or three subsets, there is not a strong answer, but I believe that the division by weather seasons offers the ideal compromise between good regression results, size of the clusters, and number of data sets.

5. Conclusions and Future Work

This study extensively explores predicting and imputing missing values in an actual dataset containing information on the power generated by a solar panel in the Galicia region of Spain. Prior research has utilized different regression methods to predict the power produced by the solar panel; however, these were conducted on a preprocessed dataset, which had rows containing missing or corrupt data excluded. Approximately 60% of the entire dataset was lost due to the exclusion of these rows. As a result, critical information was omitted, affecting all the months analyzed and three attributes. Therefore, imputation was required to resolve the problem.

The first step in the iterative process on the initial dataset was to impute the deleted rows, resulting in a complete dataset. This was accomplished through the use of an Artificial Neural Network, specifically the Multilayer Perceptron with the Levenberg-Marquardt Backpropagation training algorithm. As indicated in Table 1, the imputation method employing an Artificial Neural Network was successful in lowering regression errors on a new complete dataset compared to a filtered dataset.

Subsequently, the solar energy produced by the solar energy panel was predicted by employing the Multiple Linear Regression algorithm and the novel Artificial Neural Network method of Extreme Learning Machine, in combination with the application of the *k*-means clustering technique and two distance measures: Cosine and Cityblock. In all instances, the superiority of the Extreme Learning Machine in the task of approximating the regression on the solar energy panel previously discussed was demonstrated in the various regression tasks presented in Tables 1, 5, 6, and 7, as evidenced by the lower error rates and faster execution times. Moreover, the dataset was subdivided into subsets on which the *k*-means clustering technique was applied to obtain even more optimal results. The Cityblock and Cosine distance measures have been used for the *k*-means algorithm and a value of K=3 (returned as the optimal value by the corresponding algorithms). The Cityblock measure has returned slightly better results in the Mean Square Error calculation. The *k*-means has been applied to the whole year, see Table 5, to the dataset divided into months, see Table 7 and to meteorological quarters in Table 6.

The best results have been on the meteorological quarters since there is a good compromise between the MSE values, the number of clusters treated, and the size of these clusters. As part of future research, we aim to enhance the missing value imputation process by utilizing modern Neural Models specifically tailored to each case study.

As part of future research, we aim to enhance the missing value imputation process by utilizing modern Neural Models that are specifically tailored to each case study.

6. Acknowledgement

Míriam Timiraos's research was supported by the "Xunta de Galicia" (Regional Government of Galicia) through grants to industrial PhD (`http://gain.xunta.gal/`), under the "Doutoramento Industrial 2022" grant with reference: 04_IN606D_2022_2692965. CITIC, as a Research Center of the University System of Galicia, is funded by Consellería de Educación, Universidade e Formación Profesional of the Xunta de Galicia through the European Regional Development Fund (ERDF) and the Secretaría Xeral de Universidades (Ref. ED431G 2019/01).

References

[1] T. Kuwae, M. Hori, The Future of Blue Carbon: Addressing Global Environmental Issues, Springer Singapore, Singapore, 2019, Ch. 1, pp. 347–373. doi:10.1007/978-981-13-1295-3_13.

[2] H. Karunathilake, K. Hewage, W. Mérida, R. Sadiq, Renewable energy selection for net-zero energy communities: Life cycle based decision making under uncertainty, Renewable energy 130 (2019) 558–573. doi:10.1016/j.renene.2018.06.086.

[3] R. Prakash, I. K. Bhat, et al., Energy, economics and environmental impacts of renewable energy systems, Renewable and sustainable energy reviews 13 (9) (2009) 2716–2721. doi:10.1016/j.rser.2009.05.007.

[4] M. Wei, S. Patadia, D. M. Kammen, Putting renewables and energy efficiency to work: How many jobs can the clean energy industry generate in the us?, Energy policy 38 (2) (2010) 919–931. doi:10.1016/j.enpol.2009.10.044.

[5] E. Giacone, S. Mancò, Energy efficiency measurement in industrial processes, Energy 38 (1) (2012) 331–345. doi:10.1016/j.energy.2011.11.054.

[6] S. Peake, et al., Renewable Energy. Power for a Sustainable Future., no. Ed. 4, OXFORD university press, 2018.

[7] R. L. Keeney, Siting energy facilities, Academic Press, 2013. doi:10.1016/C2013-0-10951-8.

[8] B. Dunn, H. Kamath, J.-M. Tarascon, Electrical energy storage for the grid: a battery of choices, Science 334 (6058) (2011) 928–935. doi:10.1126/science.1212741.

[9] M. Amin, The smart-grid solution, Nature 499 (7457) (2013) 145–147.

[10] L. A. Fernandez-Serantes, J. A. Montero-Sousa, J. L. Casteleiro-Roca, X. M. Vilar-Martinez, J. L. Calvo-Rolle, Gestión de almacenamiento energético para instalaciones de generación-distribución, DYNA Ingenieria e Industria 92 (2) (2017) 140–141.

[11] C. W. Potter, A. Archambault, K. Westrick, Building a smarter smart grid through better renewable energy information, in: Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES, IEEE, 2009, pp. 1–5. doi:10.1109/PSCE.2009.4840110.

[12] O. Fontenla-Romero, J. L. Calvo-Rolle, Artificial intelligenge in engineering: past, present and future, DYNA 93 (4) (2018) 350–352.

[13] E. Jove, J. M. Gonzalez-Cava, J.-L. Casteleiro-Roca, H. Quintián, J. A. Méndez Pérez, R. Vega Vega, F. Zayas-Gato, F. J. de Cos Juez, A. León, M. MartÍn, J. A. Reboso, M. Wozniak, J. Luis Calvo-Rolle, Hybrid Intelligent Model to Predict the Remifentanil Infusion Rate in Patients Under General Anesthesia, Logic Journal of the IGPL 29 (2) (2020) 193–206. doi:10.1093/jigpal/jzaa046.

[14] E. Jove, J.-L. Casteleiro-Roca, H. Quintián, J.-A. Méndez-Pérez, J. L. Calvo-Rolle, A new method for anomaly detection based on non-convex boundaries with random two-dimensional projections, Information Fusion 65 (2021) 50–57. doi:10.1016/j.inffus.2020.08.011.

[15] M. T. García-Ordás, H. Alaiz-Moretón, J.-L. Casteleiro-Roca, E. Jove, J. A. Benítez-Andrades, I. García-Rodríguez, H. Quintián, J. L. Calvo-Rolle, Clustering techniques selection for a hybrid regression model: A case study based on a solar thermal system, Cybernetics and Systems 0 (0) (2022) 1–20. doi:10.1080/01969722.2022.2030006.

[16] E. Jove, J.-L. Casteleiro-Roca, R. Casado-Vara, H. Quintián, J. A. M. Pérez, M. S. Mohamad, J. L. Calvo-Rolle, Comparative study of one-class based anomaly detection techniques for a bicomponent mixing machine monitoring, Cybernetics and Systems 51 (7) (2020) 649–667. doi:10.1080/01969722.2020.1798641.

[17] J. L. Calvo-Rolle, H. Quintian-Pardo, E. Corchado, M. del Carmen Meizoso-López, R. F. García, Simplified method based on an intelligent model to obtain the extinction angle of the current for a single-phase half wave controlled rectifier with resistive and inductive load, Journal of Applied Logic 13 (1) (2015) 37–47. doi:10.1016/j.jal.2014.11.010.

[18] I. Machón-González, H. López-García, J. L. Calvo-Rolle, A hybrid batch som-ng algorithm, in: The 2010 international joint conference on neural networks (IJCNN), IEEE, 2010, pp. 1–5. doi:10.1109/IJCNN.2010.5596812.

[19] J.-L. Casteleiro-Roca, J. L. Calvo-Rolle, J. A. Méndez Pérez, N. Roqueñí Gutiérrez, F. J. de Cos Juez, Hybrid intelligent system to perform fault detection on bis sensor during surgeries, Sensors 17 (1) (2017) 179. doi:10.3390/s17010179.

[20] H. Quintián, J. L. Calvo-Rolle, E. Corchado, A hybrid regression system based on local models for solar energy prediction, Informatica 25 (2) (2014) 265–282.

[21] R. Casado-Vara, I. Sittón-Candanedo, F. D. la Prieta, S. Rodríguez, J. L. Calvo-Rolle, G. K. Venayagamoorthy, P. Vega, J. Prieto, Edge computing and adaptive fault-tolerant tracking control algorithm for smart buildings: A case study, Cybernetics and Systems 51 (7) (2020) 685–697. doi:10.1080/01969722.2020.1798643.

[22] A. Leira, E. Jove, J. M. Gonzalez-Cava, J.-L. Casteleiro-Roca, H. Quintián, F. Zayas-Gato, S. T. Álvarez, S. Simic, J.-A. Méndez-Pérez, J. Luis Calvo-Rolle, One-Class-Based Intelligent Classifier for Detecting Anomalous Situations During the Anesthetic Process, Logic Journal of the IGPL (11 2020). doi:10.1093/jigpal/jzaa065.

[23] L. A. Fernandez-Serantes, J. L. Casteleiro-Roca, J. L. Calvo-Rolle, Hybrid intelligent system for a half-bridge converter control and soft switching ensurement, Revista Iberoamericana de Automática e Informática industrial (2022). doi:10.4995/riai.2022.16656.

[24] J. M. Gonzalez-Cava, R. Arnay, J. A. Mendez-Perez, A. León, M. Martín, J. A. Reboso, E. Jove-Perez, J. L. Calvo-Rolle, Machine learning techniques for computer-based decision systems in the operating theatre: application to analgesia delivery, Logic Journal of the IGPL 29 (2) (2020) 236–250. doi:10.1093/jigpal/jzaa049.

[25] L. A. Fernandez-Serantes, J.-L. Casteleiro-Roca, H. Berger, J.-L. Calvo-Rolle, Hybrid intelligent system for a synchronous rectifier converter control and soft switching ensurement, Engineering Science and Technology, an International Journal (2022) 101189doi:10.1016/j.jestch.2022.101189.

[26] J. M. Gonzalez-Cava, J. A. Reboso, J. L. Casteleiro-Roca, J. L. Calvo-Rolle, J. A. Méndez Pérez, A novel fuzzy algorithm to introduce new variables in the drug supply decision-making process in medicine, Complexity 2018 (2018) 1–15. doi:10.1155/2018/9012720.

[27] Y. Varol, A. Koca, H. F. Oztop, E. Avci, Forecasting of thermal energy storage performance of phase change material in a solar collector using soft computing techniques, Expert Systems with Applications 37 (4) (2010) 2724–2732. doi:10.1016/j.eswa.2009.08.007.

[28] L. Benali, G. Notton, A. Fouilloy, C. Voyant, R. Dizene, Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components, Renewable energy 132 (2019) 871–884. doi:10.1016/j.renene.2018.08.044.

[29] D. F. Andrews, A robust method for multiple linear regression, Technometrics 16 (4) (1974) 523–531. doi:10.1080/00401706.1974.10489233.

[30] G. Ciulla, A. D'Amico, Building energy performance forecasting: A multiple linear regression approach, Applied Energy 253 (2019) 113500. doi:10.1016/j.apenergy.2019.113500.

[31] G. K. Uyanık, N. Güler, A study on multiple linear regression analysis, Procedia-Social and Behavioral Sciences 106 (2013) 234–240. doi:10.1016/j.sbspro.2013.12.027.

[32] S. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, and classification, IEEE Transactions on Neural Networks 3 (5) (1992) 683–697. doi:10.1109/72.159058.

[33] J. Bilski, J. Smoląg, B. Kowalczyk, K. Grzanek, I. Izonin, Fast computational approach to the levenberg-marquardt algorithm for training feedforward neural networks, Journal of Artificial Intelligence and Soft Computing Research 13 (2023).

[34] Z. Yan, S. Zhong, L. Lin, Z. Cui, Adaptive levenberg–marquardt algorithm: A new optimization strategy for levenberg–marquardt neural networks, Mathematics 9 (17) (2021) 2176.

[35] A. Ranganathan, The levenberg-marquardt algorithm, Tutoral on LM algorithm 11 (1) (2004) 101–110.

[36] M. I. Lourakis, et al., A brief description of the levenberg-marquardt algorithm implemented by levmar, Foundation of Research and Technology 4 (1) (2005) 1–6.

[37] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1-3) (2006) 489–501. doi:10.1016/j.neucom.2005.12.126.

[38] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: a new learning scheme of feedforward neural networks, in: 2004 IEEE international joint conference on neural networks (IEEE Cat. No. 04CH37541), Vol. 2, Ieee, 2004, pp. 985–990. doi:10.1109/IJCNN.2004.1380068.

[39] P. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, IEEE Transactions on Information Theory 44 (2) (1998) 525–536. doi:10.1109/18.661502.

[40] G.-B. Huang, L. Chen, C. K. Siew, et al., Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Networks 17 (4) (2006) 879–892.

[41] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM computing surveys (CSUR) 31 (3) (1999) 264–323.

[42] M. Ahmed, R. Seraj, S. M. S. Islam, The k-means algorithm: A comprehensive survey and performance evaluation, Electronics 9 (8) (2020) 1295. doi:10.3390/electronics9081295.

[43] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 3 (1) (2007) 1–27. doi:10.1080/03610927408827101.

[44] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 224–227. doi:10.1109/TPAMI.1979.4766909.

[45] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.

[46] S. Sohangir, D. Wang, Improved sqrt-cosine similarity measurement, Journal of Big Data 4 (1) (2017) 1–13. doi:10.1186/s40537-017-0083-6.

[47] A. Chakraborty, N. Faujdar, A. Punhani, S. Saraswat, Comparative study of k-means clustering using iris data set for various distances, in: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 332–335. doi:10.1109/Confluence47617.2020.9058328.

[48] Á. Arroyo, H. Quintian, J. L. Calvo-Rolle, N. Basurto, Á. Herrero, A hais approach to predict the energy produced by a solar panel, in: Hybrid Artificial Intelligent Systems: 17th International Conference, HAIS 2022, Salamanca, Spain, September 5–7, 2022, Proceedings, Springer, 2022, pp. 195–207. doi:10.1007/978-3-031-15471-3_18.

[49] S. Arlot, A. Celisse, A survey of cross-validation procedures for model selection, Statistics surveys 4 (2010) 40–79. doi:10.1214/09-SS054.