 doi:10.20944/preprints202103.0573.v1

Article

Modelling and prediction of monthly global irradiation using machine learning models

Cecilia Martínez-Castillo[1], Gonzalo Astray[1,2], Juan Carlos Mejuto[2]

[1] Universidade de Vigo, Grupo de Nutrición y Bromatología, Departamento de Química Analítica y Alimentaria, Facultade de Ciencias, 32004 Ourense, España
[2] Universidade de Vigo, Departamento de Química Física, Facultade de Ciencias, 32004 Ourense, España
[3] CITACA, Universidade de Vigo, Campus Auga, 32004 Ourense, España
* Correspondence: gastray@uvigo.es

Abstract: Different machine learning models (multiple linear regression, vector support machines, artificial neural networks and random forests) are applied to predict the monthly global irradiation (MGI) from different input variables (latitude, longitude and altitude of meteorological station, month, average temperatures, among others) of different areas of Galicia (Spain). The models were trained, validated and queried using data from three stations, and each best machine model was checked in two independent stations. The results obtained confirmed that the best ML methodology is the ANN model which presents the lowest RMSE value in the validation and querying phases 122.6·10kJ/(m^2·day) and 113.6·10kJ/(m^2·day), respectively, and predict conveniently for independent stations, 201.3·10kJ/(m^2·day) and 209.4·10kJ/(m^2·day), respectively. Given the good results obtained, it is convenient to continue with the design of artificial neural networks applied to the analysis of monthly global irradiation.

Keywords: prediction; solar irradiation; machine learning; artificial neural network; random forest; vector support machine.

1. Introduction

The introduction should briefly place the study in a broad context and highlight why it is important. It should define the purpose of the work and its significance. The current state of the research field should be carefully reviewed and key publications cited. Please highlight controversial and diverging hypotheses when necessary. Finally, briefly mention the main aim of the work and highlight the principal conclusions. As far as possible, please keep the introduction comprehensible to scientists outside your particular field of research. References should be numbered in order of appearance and indicated by a numeral or numerals in square brackets—e.g., [1] or [2,3], or [4–6]. See the end of the document for further details on references.

Solar radiation exerts its influence over all Earth's processes related to the environment, plant growing and even over the human activities development [1]. At ground level, the solar radiation data are important for a large number of applications related to agricultural hydrology, plant growth and others [1]. Besides these, global solar irradiation is a significant parameter in renewable energy applications (for example to determine size and model photovoltaic systems) [2].

To obtain global solar irradiation data remote measurements can be used using specific devices [3] which can be limited to a small number of meteorological stations [4], probably, due to their high cost and other inconveniences such as calibration and maintenance, [3,5]. Besides this, these data may not be accessible because the meteorological observatories that include measurement series of solar irradiation are still rarely distributed and these data present, sometimes, a problematic spatial interpolation in areas of intricate orography [6]. Even, according to some authors, only 1000 continental stations over the

world can measure the solar radiation [7,8] although, currently, this number can be probably higher.

According to different authors, the shortage, the difficulties and the uncertainties of these measurements can be corrected with estimations calculated from other more abundant variables (climatological properties) such as cloudiness, among others [3,6].

Take into account the increase in energy demand and consumption worldwide, and the search for alternatives to the decrease in fossil fuel reserves [9], it can be understood that the determination of the global solar irradiation can be very important in solar energy conversion systems. Solar photovoltaic energy has presented an important growth in the last years due to their cost reduction [10]. In this energetic context, the Spanish climatic conditions can obtain a high performance using photovoltaic solar energy [11]. Within this territory, the Galicia area is suitable for solar installations [9].

Taking into account all the above, to the design of solar energy conversion systems, it is necessary to have solar irradiation monthly average data, and these need to be reliability [6]. Due to this, different techniques have been developed to find a correlation between solar irradiation and other variables such as relative humidity, air temperature, among others [2]. Traditionally, the estimation has been carried out using parametric-empirical models which obtained a relatively high level of certainty [10].

Possibly, the three principal groups of models used to forecast solar radiation are machine learning, physical (or numerical weather prediction) and sky imaging [12]. Machine learning is an artificial intelligence subfield that studies and develop mathematical algorithms intended to comprehend data and obtain data without a prearranged model algorithm [13]. The machine learning models can find the relationship between inputs and outputs variables, which allow that these models can be used, sometimes, in classification problems, forecasting problems, among others [14].

Different machine learning models such as linear regression (MLR), support vector machines (SVM), artificial neural networks (ANN and random forest (RF) are available to prediction purposes. These models can be used in different areas such as:

- Environmental Science to determine fine particles (PM_1) concentration using environmental, meteorological and eventualities variables [15],
- in Bioinformatics to identify single-nucleotide polymorphisms [16] or to predict dihedral angle regions [17],
- in food chemistry to model and optimize the pectinase-assisted low-temperature extraction from cashew apple juice [18] or to optimize an enzymatic method to obtain modified artichoke pectin and pectic oligosaccharides [19] or
- in Economics to early prediction of university dropouts [20], among others.

Related to this study, this machine learning model can be used, together or separately, to different purposes.

- Multiple linear regression models can be used to predict the net radiation using weather data of solar radiation, temperature, among others [21].
- Artificial neural networks have been used to predict the solar irradiation at different time windows (hourly, daily and monthly) from different meteorological variables (temperature, atmospheric pressure, among others) or even including geographical coordinates such as latitude, longitude and altitude [1]. This kind of models can be used to determine daily, weekly and monthly global solar radiation in Fortaleza (Brazilian Northeast region) using 14-year-long data set to train three different ANNs models [3]. ANNs can also be used to determine different parameters such as the global horizontal irradiation (from meteorological data), the global tilted irradiation (from the horizontal global irradiation and others) and to forecast the hourly direct normal and the global horizontal irradiation from one to six hours horizon [7].
- Support vector machines models can be used to generate the daily global solar irradiation using a general (non-locally dependent) model [10]. The model (which used temperatures, relative humidity, wind speed and rainfall, among other variables) pre-

sented a high capacity of generalization for the different studied locations and improved, in terms of mean absolute error, the locally trained models in some locations [10]. SVM models can even be used to forecast of photovoltaic power [22].

- Random forest can be used to estimate the solar radiation using air pollution index in three different sites [23] or to forecast solar radiation and compared their result with other methods such as multivariate adaptive regression splines (MARS), classification and regression tree (CART) and M5 [24].

In many research articles, it can be also possible to see comparisons between this kind of models to predict solar irradiation and even other interesting variables related to the subject under study.

- MLR and ANN models can be compared in the estimation of monthly-average daily solar radiation over different locations in Turkey [25]. Different variables (latitude, longitude, altitude, land surface temperature and month) were used as input variables. According to the authors, the results showed that the ANN model could achieve a good performance compared to the MLR model.
- SVM and ANN models were used in a comparative study of different methods carried out by da Silva et al. (2017) to estimate the daily global solar irradiation. Four different kinds of architecture combining different input parameters were studied. The first model used the fractional daily sunshine duration and the solar radiation at the top of the atmosphere, while the other three models were developed adding other variables one by one (air temperature (maximum and minimum), precipitation and relative humidity) [26]. According to the authors, statistical indicators showed that SVM technique has better performance than ANN models for the study location (Botucatu/SP/Brazil). Neural models can be compared to random forest models to forecast the normal beam, horizontal diffuse and global components [27].
- SVM, ANN and deep neural network models can even be used to forecast of photovoltaic power [22] or to estimate electricity demand (using multiple linear regression, artificial neural network and support vector machine) [13].
- Random forest models to model the daily variability of solar irradiance can be compared to other methods such as multiple linear regression, obtaining the best results between both [28].

Therefore, the objective of this research was developed different machine learning prediction models (linear regression, artificial neural networks, support vector machines and random forest) to model the monthly global irradiation (MGI) from three meteorological station stations located in the Autonomous Community of Galicia (Spain) and then generalize the knowledge to other two nearby stations. This work is a summary of the final degree project developed by the first author of this research [29].

2. Materials and Methods

2.1. Study area

According to Vázquez [30], Galicia can be divided into four climatic zones based on their solar radiation. To carried out this research, five meteorological stations were selected, all of them belonging to climatic zone II. This zone is characterized to present global radiation values between 13.7 MJ/m^2·day (3.8 kWh/m^2·day) and 15.1 MJ/m^2·day (4.2 kWh/m^2·day) [30]. The selected meteorological stations were: i) Amiudal in the municipality of Avión, ii) Serra do Faro in Rodeiro, iii) Monte Medo in the municipality of Baños de Molgas, iv) Ourense-Estacións in the city of Ourense and v) Pazo de Fontefiz in Coles. The meteorological stations were selected taking into account the conditions and the quantity of available data to create useful and accurate models for the prediction of MGI.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

2.2. Database

The database was obtained from Meteogalicia website [31] which provides the meteorological data for the selected stations. The periodicity of the data was monthly which reduce the volume of handled data and, therefore, the computational cost of modelling.

The selected variables, in addition to the MGI (10 kJ/(m^2·day) were: i) latitude, ii) longitude and iii) altitude (m) of the station, iv) month order, v-vii) average, average of the maximum and average of the minimum temperatures (ºC); viii-xi) average, average of the maximum and average of the minimum relative humidities (%) and xii) precipitation (L/m^2).

Three meteorological stations, Amiudal, Serra do Faro and Monte Medo, were used to train (2005-2012), validate (2013-2015) and query (2016-2018) the models. The other two stations, Ourense-Estacións and Pazo de Fontefiz, were used to check the models' behaviour in different locations than the previous ones, that is, the knowledge generated in three stations is extrapolated to new locations. In these two stations, the data used includes the period between 2012 and 2018.

2.3. Implementation of models

As previously stated, four different kinds of models were developed: i) multiple linear regressions, ii) artificial neural networks, iii) support vector machines and iv) random forests. Different combinations of available variables (Table 1) were used to determine the MGI and study the influence of the average temperatures, the average relative humidity and precipitation. The geographic coordinates and the month of the year were selected for all the models.

Table 1. Variables, and their combination, used to develop the different models: i) latitude (Lat), ii) longitude (Long), iii) altitude (Alt), iv) month, v-vii) average (T_{av}), average of the maximum (T_{max}) and the average of the minimum temperature (T_{min}); viii-xi) average (HR_{av}), an average of the maximum (HR_{max}) and the average of the minimum (HR_{min}) relative humidity and xii) precipitation (P).

Combination type	Lat	Long	Alt	Month	T_{av}	T_{max}	T_{min}	HR_{av}	HR_{max}	HR_{min}	P
Type 1	■	■	■	■	■	■	■	■	■	■	■
Type 2	■	■	■	■	■	■	■				
Type 3	■	■	■	■	■	■	■				■
Type 4	■	■	■	■				■	■	■	
Type 5	■	■	■	■				■	■	■	■
Type 6	■	■	■	■	■	■	■	■	■	■	
Type 7	■	■	■	■							■

2.4. MLR models

Multiple linear regression analysis is a conventional method that relates different independent variables with a dependent one [32]. This method provides a linear input-output model for a specific data set [33]. Unlike the simple regression analysis, MLR analysis is closer to real situations because the phenomena are complex and must be explained using different variables that intervene in its existence [34].

It can be expressed mathematically as follows (equation 1):

 doi:10.20944/preprints202103.0573.v1

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

being y the desired variable, β0 the constant, β1-βn the regression coefficients, x1-xn the input variables and ε is the error.

2.5. ANN models

Artificial neural networks are a type of artificial intelligence (AI) model that simulates the human brain processes information [35]. ANNs presents different interesting aspects such as their error tolerant or their generalization capabilities, among others [36,37].

The most used artificial neural models are the multilayer feedforward neural network where the artificial neurons (also called nodes) are distributed into three different layers named as input, hidden and output layers [38]. The optimum number of neurons in the hidden layer can be determined by trial and error procedure [39–41]. The input layer receives the data provided by the user (in our case, the different variables from the Meteogalicia meteorological stations). During the model training, these information flow within the neural network and makes it possible to adapt the results to the desired value modifying weights and biases. This phase is finished when the desired error is reached.

The ANN model implemented in this research has been tested using different parameters combination such as i) the number of cycles (1 to 524288 in 19 steps with a logarithmic scale), ii) learning rate (0.1, 0.2 and 0.3), iii) momentum (0.1, 0.2 and 0.3) and decay (true or false).

2.6. SVM models

Support vector machines were developed in the 1990s, by different authors [42], to resolve classification problems and had a great reception and use due to due to its ability to deal non-linear data [10]. This method can be also used for regression purposes [13,42]. These approximations are a type of linear classifiers, which induce linear or hyperplane separators using a kernel function [42].

A large combination of parameters to develop an SVM model is possible. However, to facilitate the development of these models, in this research it has been taken into account the combination of γ and C, in addition to the SVM type. The range values for γ and C were chosen taking into account the "A Practical Guide to Support Vector Classification" proposed by Hsu et al. for classification problems [43].

Therefore, the combination of parameters used for SVM models' development is i) SVM type (ε-SVR and ν-SVR), ii) γ (from 2-15 to 23 in 18 steps, with logarithmic scale) and iii) C (from 2-5 to 215 in 20 steps, with logarithmic scale).

2.7. RF models

Random forests are non-parametric method proposed in 2001 by Breiman [23,44]. A random forest model is a set of random trees that can be used for regression and classification [45].

RF models were implemented using combinations of i) number of trees (1 to 100 in 99 steps with linear scale), ii) criterion (least square), iii) maximum depth (-1 to 100 in 101 steps with linear scale) and iv) apply prepruning (true or false).

2.8. Statistics of the developed models

The statistics used to analyze the models were the squared correlation coefficient (r^2), the root mean square error (RMSE, equation 2) and the average absolute relative error (Error, equation 3). The best ANN, SVM and RF model was chosen according to the lowest RMSE in the validation phase.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \quad (2)$$

$$Error = \frac{\sum_{i=1}^{N}\left(\left|\frac{x_i - y_i}{y_i}\right|\right)}{N} \quad (3)$$

2.9. Equipment and software used

The different models were implemented in the server available at the Department of Physical Chemistry of the University of Vigo, Campus of Ourense (Intel® Core™ i7-8700 processor at 3.20 GHz, with 16GB of RAM). All models were run on Windows 10 Pro 64-bit operating system. Data were collected and processed using the software Microsoft Excel 2016, from Microsoft Office Professional Plus 2016 package, Microsoft, USA. MLR, ANN, SVM and RF models were developed using a Trial/Free version of RapidMiner Studio 9.0.993 software. Figures were made with SigmaPlot v. 13.0, Systat Software, Inc., USA.

3. Results and discussion

Table 2 show the bests models for each model type and the combination of variables used for that model. Next, the best models obtained for each of the studied approaches will be described.

3.1. MLR models

For the seven MLR models with different combination types, the one that presented the worst adjustment, based on the RMSE for the validation phase, was the model with combination 7. This model presented an RMSE of 567.4·10kJ/(m²·day) for the validation phase, which corresponds to a low r^2 (0.468). These bad adjustments for the validation phase are extensible to all phases of the model, training and querying. Thus, for these phases, the RMSE values are 521.5·10kJ/(m²·day) and 630.0·10kJ/(m²·day) which together with the low squared correlation, 0.426 and 0.343, make this model as a model that cannot be used for modelling the MGI. The rest of the models present better adjustments than the previous model, with RMSE values for the validation phase, between 292.4·10kJ/(m²·day) and 241.1·10kJ/(m²·day). These models offer for the querying phase some RMSE similar to those provided for the training and/or the validation phase and an average absolute relative errors between 18.2% and 19.5%. The best MLR model corresponds to a model with combination type 1 (Table 2), that is, an MLR that uses all the input available variables to model the behaviour of the MGI.

3.2. ANN models

The worst ANN model developed was the model with combination type 7. This model presented an RMSE value for the validation phase around 152.6·10kJ/(m²·day) which corresponds to an average absolute relative error of 12.1%. This value is close to the 10% that it is considered as a good error percentage for this kind of modelling. The training a querying phase present similar adjustments to the validation phase with squared correlation coefficients of 0.943 and 0.953 for training and validation, respectively. These adjustments make the worst ANN model an almost usable model for modelling the MGI, however, the other developed combination types clearly improve the worst ANN model, presenting RMSE values between 122.6·10kJ/(m²·day) and 149.4·10kJ/(m²·day) for the validation phase. The best ANN model (Table 2) corresponds to a model with combination type 4 (input variables; latitude, longitude, altitude, month and the three relative humidity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

3.3. SVM models

For the different SVM models developed, the model that presented the worst adjustment, based on the RMSE for the validation phase, was, again, the model with combination 7. It seems clear that in all the models seen, those models that only have the precipitation variable, in addition to the other four fixed variables, do not present good results. The combination type 7 SVM model presents for validation phase an RMSE of 170.4·10kJ/(m²·day) which corresponds to a good r^2 of 0.956. These adjustments for the validation phase are extensible to training and querying phase where the RMSE are 152.5·10kJ/(m²·day) and 174.3·10kJ/(m²·day) with high squared correlation values, 0.951 and 0.962 which make this model as a model that could be used for modelling the MGI. The rest of the models present better adjustments being the RMSE value in the validation phase dropped to 155.6 for the second-best model. The best SVM model corresponds to a model with combination type 5, that is, an SVM that uses eight input variables to model the MGI response (Table 2).

3.4. RF models

Finally, the last kind of ML models is the RF models. In this case, the worst model developed for was, unlike the other ML models, a model with combination type 2. This model presented an RMSE value for the validation phase around 212.4·10kJ/(m²·day) with an average absolute relative error of 15.0%. During the training and the querying phase present very different adjustments, 92.5·10kJ/(m²·day) and 165.1·10kJ/(m²·day), respectively. The other combination types slightly improve this model and a better model is obtained when the configuration 5 is used (Table 2).

Table 2. Adjustment parameters for each best approximation model developed according to its selected input variables. Latitude (Lat), longitude (Long), altitude (Alt), month, average (T_{av}), average of the maximum (T_{max}) and the average of the minimum temperature (T_{min}), average (HR_{av}), the average of the maximum (HR_{max}) and the average of the minimum (HR_{min}) relative humidity and precipitation (P). RMSE is the root mean square error and r^2 is the squared correlation coefficient.

													T		V		Q	
Combination type	**Model**	**Lat**	**Long**	**Alt**	**Month**	**T_{av}**	**Tmax**	**Tmin**	**HR_{av}**	**HRmax**	**HRmin**	**P**	**RMSE**	**r^2**	**RMSE**	**r^2**	**RMSE**	**r^2**
Type 1	MLR												226.3	0.892	241.1	0.904	245.8	0.885
Type 4	**ANN**												**127.1**	**0.967**	**122.6**	**0.975**	**113.6**	**0.980**
Type 5	SVM												105.6	0.977	153.1	0.961	156.7	0.967
Type 5	RF												94.8	0.982	159.5	0.962	227.9	0.933

3.5. Best models developed

Taking into account the previously chosen models (Table 2) we will now proceed to the analysis as a whole. It can be seen that the RMSE values obtained for the validation phase are included between 122.6·10kJ/(m²·day) and 241.1·10kJ/(m²·day).

According to this, the multiple linear regression model is the one that obtains the worst RMSE value in the validation phase with a value of 241.1·10kJ/(m²·day) and the worst squared correlation coefficient value (0.904). This model obtained an average absolute relative error around 19.2%. Regarding the training phase, the RMSE presents lower value 226.3·10kJ/(m²·day) compared with the validation phase, nevertheless, the Error and the r^2 values also present lower values (17.3% and 0.892).

Figure 1-A shows the experimental and modelled MGI values by the MLR model. It can be seen how both the training and validation phase cases follow the line with slope

one (red line), however, a great dispersion is observed in them, this fact can be intuited by the high values of absolute average relative error for both phases (17.3% and 19.2% for training and validation, respectively). These high errors are increased by the existence of some points that are clearly distant from the line with slope one.

Given the results shown for both phases, it is expected that the results for the querying phase will also be the worst compared to the rest of the models. The RMSE is greater than invalidation phase (245.8·10kJ/(m²·day)) and the adjustments, in terms of squared correlation, was the lowest for the three phases (0.885).

In Figure 1-A it can be seen that the querying cases also follow the line with slope one, however, as happened with the cases of the training and validation phases, these do not adjust the line, observing the existence of some point that is far away.

As expected, the MLR model is not capable of learning correctly and then generalizing that knowledge afterwards. Given the results shown in the three phases, it can be concluded that the MLR model is not a suitable model for MGI modelling. A possible explanation for the poor adjustments of the MLR model may be based on the use of the month variable, which does not present a linear relationship with the MGI.

The next model in terms of the RMSE value in the validation phase is the RF model that presents a value of 159.5·10kJ/(m²·day). This value is clearly improved in the model training phase (94.8·10kJ/(m²·day)). In both phases, the RF model clearly improves the MLR model, both in RMSE values and in its squared correlation values (0.982 and 0.962 vs. 0.892 and 0.904, for the training and validation phase, respectively). Besides this, the model presents a good behaviour in terms of average absolute relative error.

Figure 1-B shows the experimental and modelled MGI values by the RF model. It can be seen how the training phase; the cases follow better the line with slope one then the cases predicted by the MLR. This behaviour is similar for the validation cases. The behaviour of both phases is good and reaches the average absolute relative error values of 5.9% and 10.5% for training and validation, respectively.

If we analyze the adjustments for the querying phase it can be seen how the RF model presents, for this phase, the worst adjustments in terms of RMSE (227.9·10kJ/(m²·day)) although the average absolute relative error remains at similar levels to those of the validation phase (10.7%).

In Figure 1-B it can be seen that the querying cases also follow the line with slope one, however, a similar dispersion than provided by the MLR model is observed. It can be seen some cases that deviate more from the trend line one, although in the area of low MGI values it can be seen that the RF model adjusts much better than the MLR model. Due to this the average absolute relative error is good (around 10.7%).

In view of the results shown in the three phases, it can be concluded that the RF model is a suitable approach for MGI modelling.

The second-best model, taking into account the RMSE values in the validation phase, is the model developed based on support vector machines. The adjustments for the validation phase are kept close to the RF model, in fact, the RMSE value for the SVM model is 153.1·10kJ/(m²·day) compared to 159.5·10kJ/(m²·day) for the RF model and the squared correlation values are the same (0.961 vs. 0962 for SVM and RF, respectively). The same happens with the error, which remains for both around 11%. For the training phase, a slight worsening of the fit for the SVM model reaching an RMSE of 105.6·10kJ/(m²·day) is observed.

Figure 1-C shows the experimental and modelled MGI values by the SVM model. Training and validation phase cases follow the line with slope one (red line) nevertheless the model provided worse adjustment for the training cases in comparison with the RF model. This may be due to the existence of some points in the middle area that clearly move away from the line with slope one. Both phases showed good adjustments is in terms of average absolute relative error reaching values of 4.9% and 11.0% for training and validation, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

Given the results provided by the SVM model, it can be assuming that for the querying phase the model will work well. According to the adjustment parameters for the querying phase, it can be said that both the RMSE and the r^2 values remain close to the RMSE of the validation phase.

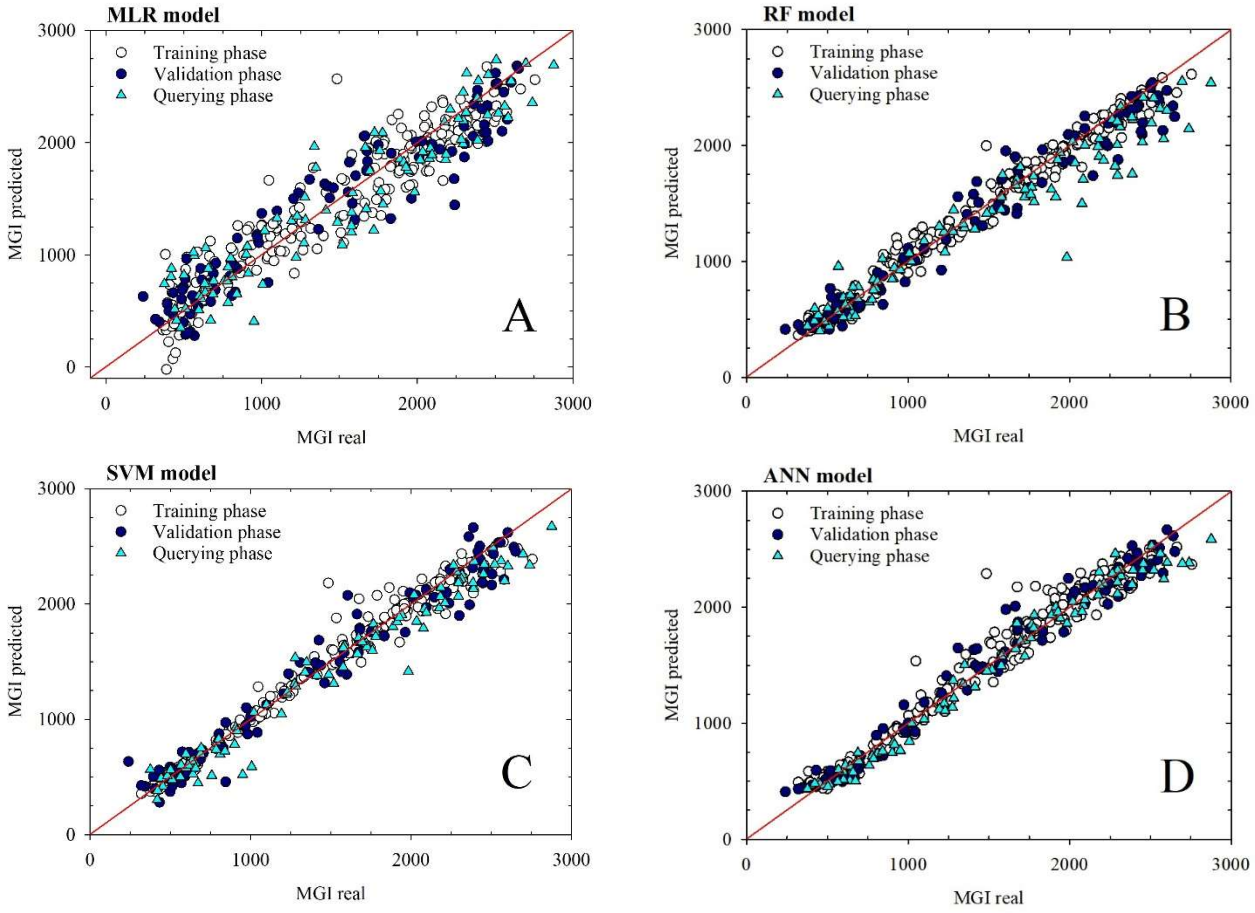

Figure 1. Graphical representation for the real and modelled values of MGI during the training phase (white dots), validation phase (black dots) and querying phase (turquoise triangles) for each select model: A) multiple linear regression -MLR-, B) random forest -RF-, C) support vector machine -SVM- and D) artificial neural network -ANN-. Redline is the line with slope one.

Figure 1-C shows the querying behaviour and it can be seen that the cases are close to the line with slope one, however, this model provided better fits then the RF model, although it can be seen some points in the lower and upper area that stray from the line with slope one.

Taking into account that the model offers 8.7% of absolute average relative error for the querying phase, it can be affirmed that the SVM model is a suitable model for MGI modelling.

Finally, for all the models designed, the best model is the ANN model taking into account the criterion of the lowest RMSE value in the validation phase.

This model obtained for validation phase the lower RMSE with a value of 122.6·10kJ/(m²·day) that corresponds with the highest squared correlation coefficient (0.975). Regarding the training phase, the RMSE value is the second lower value 127.1·10kJ/(m²·day) which supposes an absolute average relative error of 7.3%.

Figure 1-D shows the experimental and modelled MGI values by the ANN model. It can be seen how for the training phase some points distance from the line with slope one.

This fact can be explained that the ANN model did not obtain the best adjustments for the training phase, in comparison with the SVM and the RF model. This behaviour is reversed for the validation phase where it can be seen how this model is the one with the best fits to line with slope one. This behaviour is reversed for the validation phase where it can be seen how this model is the one with the best fits to line with slope one (obtaining an average absolute relative error of 8.8%).

Given the good results provided by the ANN model for both phases, good results for the querying are expected. In this case, the RMSE value is lower than in both training and validation phases (113.6·10kJ/(m²·day)) and corresponds with a squared correlation of 0.980 (the highest for all the models in this phase).

In Figure 1-D it can be seen that the querying behaviour for this model. It can be seen, as unlike the other models, the validation cases are very close to the line with slope one. Some small dispersion is observed in the area with high MGI values, but this behaviour is an exception in the model.

Finally, taking into account all the adjustments provided by the model and the low absolute average relative error for querying phase (6.6%) it can be affirmed that the ANN model is a suitable model for the MGI modelling.

Regarding the variables used by each of the best models, it can be seen in Table 2 that all the selected models have as input variables (apart from the latitude, longitude, altitude and month) all the variables of relative humidity and precipitation. This fact is only broken by the ANN model that does not use the precipitation variable. Regarding the MLR model, it can be seen how it includes temperature variables among its input variables. The inclusion of these variables may be due to the fact that the MLR model, being a linear model, does not work properly with non-linear variables, as is the case of the month variable. Due to this fact, the authors understand that this variable can be counteracted by the MLR model with the inclusion of temperature variables in the model.

3.6. ANN generalization to different locations

After analyzing all the machine learning models developed in the previous section it will proceed to check how the best models work in the two reserved stations (Pazo de Fontefiz in Coles and Ourense-Estacións in Ourense) which have not been used any of the previous phases. The adjustments for the best models applied to these stations are presented in Table 3.

It can be seen how the support vector machines model is the one that offers worse modelling values for both stations; in fact, it presents errors in terms of RMSE much higher than the other selected models (Table 3). It can be seen how for the Pazo de Fontefiz station the error, in terms of root mean square error, is practically double (402.9·10kJ/(m²·day)) that the error presented by the best-selected model; while for the Ourense-Estacións station, the error (807.9·10kJ/(m²·day)) is almost four times greater than that presented by the best model. As expected, these high errors affect the average relative absolute error presented by each station, so the Pazo de Fontefiz station presents an error of 24.8%, being overcome by the error obtained in Ourense-Estacións, 47.2%.

The SVM model presents good adjustments in terms of squared correlation (upper than 0.940); however, taking into account the adjustments of the root mean square error and the average absolute relative error it can be concluded that the SVM model is not a suitable model for the MGI modelling.

The remaining three models have better adjustments than the SVM model. The one with the worst fit is the MLR model that presents an error, in terms of root mean square error, of 285.2·10kJ/(m²·day) and 233.4·10kJ/(m²·day) for the stations of Pazo de Fontefiz and Ourense-Estacións, respectively (Table 3). Compared to the SVM model, this model improves its adjustments in terms of RMSE and error, although not in terms of squared correlation. The errors of this model are around 19% for each station. According to this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

error level, we can say that the model shows good behaviour, but shows a higher error percentage than desired, especially for the Pazo de Fontefiz station.

Table 3. Adjustment parameters for each of the best models applied to the stations of Pazo de Fontefiz in Coles and Ourense-Estacións in Ourense. RMSE is the root mean square error, Error is the average absolute relative error (%) and r^2 is the squared correlation coefficient.

	Q_{co}			Q_{Ou}		
Model	**RMSE**	**Error**	**r^2**	**RMSE**	**Error**	**r^2**
MLR	285.2	19.5	0.865	233.4	18.1	0.915
ANN	**201.3**	**13.1**	**0.935**	**209.4**	**14.7**	**0.971**
SVM	402.9	24.8	0.949	807.9	47.2	0.971
RF	246.1	21.2	0.920	216.5	19.6	0.950

The second-best model is the random forest model. This model improves the modelling's of the previous models in terms of RMSE, for each of the analyzed stations, but the percentage errors remain high (21.2% and 19.6% for Pazo de Fontefiz and Ourense-Estacións, respectively) (Table 3). Despite that the model presents high squared correlation (greater than 0.91), the use of the model should be limited due to the possibility of presenting poor modelling, especially in areas with low MGI.

Finally, the ANN model, which had been chosen in the previous section as the model with the best adjustments for each development phases, has emerged as the model with the best predictions for these two independent stations (Table 3). The model improves each of the statistics (except r^2 for the Pazo de Fontefiz station) for each of the stations, reporting errors, in terms of RMSE, around 201.3·10kJ/(m²·day) and 209.4·10kJ/(m²·day) for the stations of Pazo de Fontefiz and Ourense-Estacións, respectively. Likewise, for this model, the squared correlation is high (0.935 and 0.971) and the average absolute relative error remain close to 10% error for each of the stations (which is considered as a good error percentage). These good adjustments are reflected in Figure 2. The first thing to note is the different size in the database between the two stations. The Pazo de Fontefiz station has data from July 2012 to August 2018 (a total of 73 months), while the Ourense-Estacións station has data from June 2014 to August 2018 (a total of 49 monthly measurements). Figure 2 shows the time series for the real MGI values (olive colour) and the values modelled by the ANN model.

Figure 2-A shows the time series for the Pazo de Fontefiz station. It can be seen the IGM`s cycles time series with their maximums in the summer months and their minimums in the winter months (range from 334·10kJ/(m²·day) to 2605·10kJ/(m²·day)). The ANN modellings are shown in the figure as a black line. It can be seen how the modellings fit, almost perfectly, to the real-time series, which means (as we have already seen in the adjustments) that the ANN model can accurately model the behaviour of the MGI for Pazo de Fontefiz station. It can be seen how for the low-value areas of MGI the modelling overestimates the values (this happens in the areas of a lower value) while the model behaves, in general, well for high-value areas of MGI (although it also some underestimation is observed). Given the good adjustments, and from the Figure 2-A modelling time series, provided by the ANN model for the Pazo de Fontefiz station, it can be said that this model is capable of generalizing the knowledge of the previous phase to other nearby geographical stations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2021 doi:10.20944/preprints202103.0573.v1

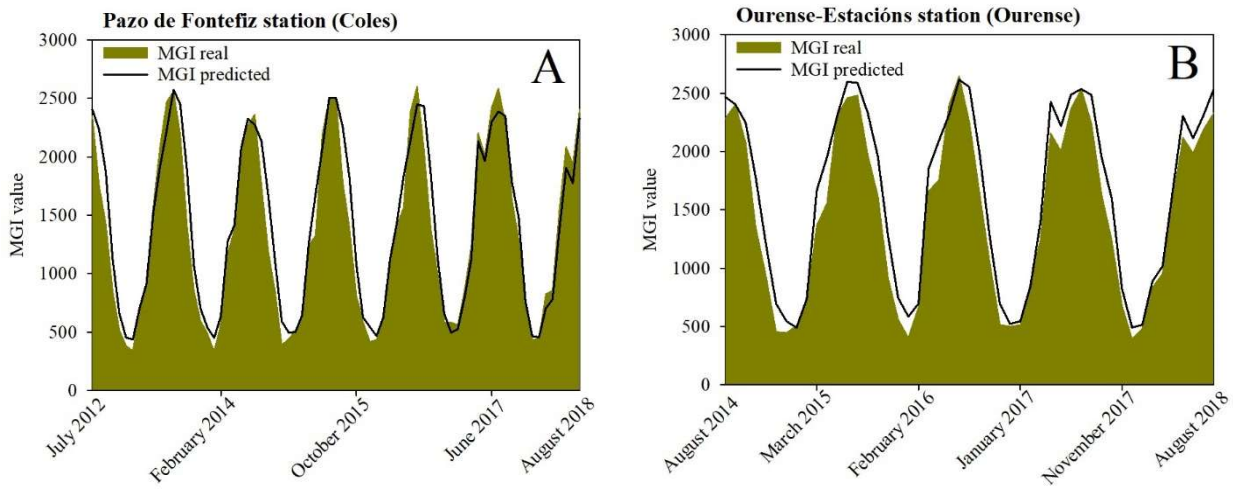

Figure 2. Real and modelled time series for Pazo de Fontefiz and Ourense-Estacións stations. The olive shade corresponds to the actual values, and the black line corresponds to the values modelled by the ANN model.

Figure 2-B shows the time series for the Ourense-Estacións station. In this case, the time series has a range from 393·10kJ/(m²·day) to 2647·10kJ/(m²·day). It can be seen how the modelling fits the real-time series, however in this case the settings show a worse behaviour than in the case of the Pazo de Fontefiz station. Again, it can be seen how for the low areas of MGI the modelling overestimate, in general, the MGI values (can even see how this behaviour is observed in some measurement in the maximum area) although for high MGI the ANN model generally behaves well. Because of the adjustments and the Figure 2-B time series, it can be said that the ANN model is usable on other nearby geographical stations.

4. Conclusions

Based on the goodness of statistics, the modelling carried out by MLR, ANN, SVM and RF methodologies can predict the variable MGI, generally, in an appropriate way for the stations used for its implementation (Amiudal in Avión, Serra do Faro in Rodeiro and Monte Medo in Baños de Molgas). The results vary where these models are applied to other locations, Pazo de Fontefiz in Coles and Ourense-Estacións in Ourense. Attending to the adjustments obtained for each station it can be affirmed that the best model is the ANN which presents the lowest RMSE value in the validation and querying phases 122.6·10kJ/(m²·day) and 113.6·10kJ/(m²·day), respectively, and predict conveniently for Coles and Ourense station 201.3·10kJ/(m²·day) and 209.4·10kJ/(m²·day), respectively.

For all this, it can be concluded that, given the good results obtained, it is convenient to continue with the design of artificial neural networks applied to the analysis of monthly global irradiation in different areas of the autonomous community of Galicia to obtain a general model for the entire region.

All the models developed in this research could be improved with the inclusion of more stations, using different random split datasets, taking into account new meteorological input variables, among others.

Author Contributions: Conceptualization, C. Martínez-Castillo and G. Astray; methodology, C. Martínez-Castillo; formal analysis, C. Martínez-Castillo and G. Astray; writing—original draft preparation, C. Martínez-Castillo and G. Astray; writing—review and editing, C. Martínez-Castillo, G.

Astray and J.C. Mejuto; supervision, G. Astray and J.C. Mejuto. All authors have read and agreed to the published version of the manuscript.

Acknowledgements: G.A. thanks to the University of Vigo for his contract supported by "Programa de retención de talento investigador da Universidade de Vigo para o 2018" budget application 0000 131H TAL 641. Authors thank Meteogalicia and the Consellería de Medio Ambiente, Territorio e Vivenda of Xunta de Galicia for the database used in this research. Astray G. thanks Xunta de Galicia, Consellería de Cultura, Educación e Ordenación Universitaria, for the computer equipment financed in 2017 from his postdoctoral grant B, POS-B/2016/001, K645 P.P.0000 421S 140.08. Authors thank RapidMiner Inc. for the Trial/Free license of RapidMiner Studio 9.0.993 software. This work is a summary of the final degree project developed by the first author of this research.

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

References

1. Diez, F.J.; Navas-Gracia, L.M.; Chico-Santamarta, L.; Correa-Guimaraes, A.; Martínez-Rodríguez, A. Prediction of horizontal daily global solar irradiation using artificial neural networks (ANNs) in the Castile and León region, Spain. *Agronomy* **2020**, *10*, 96, doi:10.3390/agronomy10010096.
2. Yacef, R.; Benghanem, M.; Mellit, A. Prediction of daily global solar irradiation data using Bayesian neural network: A comparative study. *Renew. Energy* **2012**, *48*, 146–154, doi:https://doi.org/10.1016/j.renene.2012.04.036.
3. Rocha, P.A.C.; Fernandes, J.L.; Modolo, A.B.; Lima, R.J.P.; da Silva, M.E.V.; Bezerra, C.A.D. Estimation of daily, weekly and monthly global solar radiation using ANNs and a long data set: a case study of Fortaleza, in Brazilian Northeast region. *Int. J. Energy Environ. Eng.* **2019**, *10*, 319–334, doi:10.1007/s40095-019-0313-0.
4. Nonhebel, S. The importance of weather data in crop growth simulation models and assessment of climatic change effects, 1993.
5. Hunt, L.A.; Kuchar, L.; Swanton, C.J. Estimation of solar radiation for use in crop modelling. *Agric. For. Meteorol.* **1998**, *91*, 293–300, doi:https://doi.org/10.1016/S0168-1923(98)00055-0.
6. Prieto, J.I.; Martínez-García, J.C.; García, D. Correlation between global solar irradiation and air temperature in Asturias, Spain. *Sol. Energy* **2009**, *83*, 1076–1085, doi:https://doi.org/10.1016/j.solener.2009.01.012.
7. Notton, G.; Voyant, C.; Fouilloy, A.; Duchaud, J.L.; Nivet, M.L. Some applications of ANN to solar radiation estimation and forecasting for energy applications. *Appl. Sci.* **2019**, *9*, 209, doi:10.3390/app9010209.
8. WRDC No Title Available online: https://www.re3data.org.
9. Vázquez Vázquez, M. *Atlas de radiación solar de Galicia*; Vázquez Vázquez, M., Ed.; Universidade de Vigo: Vigo, 2005; ISBN 84-609-7101-5.
10. Antonanzas-Torres, F.; Urraca, R.; Antonanzas, J.; Fernandez-Ceniceros, J.; Martinez-de-Pison, F.J. Generation of daily global solar irradiation with support vector machines for regression. *Energy Convers. Manag.* **2015**, *96*, 277–286, doi:https://doi.org/10.1016/j.enconman.2015.02.086.
11. Espejo Marín, C. La energía solar fotovoltaica en España. *Nimbus Rev. Climatol. Meteorol. y paisaje* **2004**, *1–14*, 5–31.
12. Fouilloy, A.; Voyant, C.; Notton, G.; Motte, F.; Paoli, C.; Nivet, M.-L.; Guillot, E.; Duchaud, J.-L. Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability. *Energy* **2018**, *165*, 620–629, doi:https://doi.org/10.1016/j.energy.2018.09.116.
13. Solyali, D. A comparative analysis of machine learning approaches for short-/long-term electricity load forecasting in Cyprus. *Sustainability* **2020**, *12*, 3612, doi:10.3390/SU12093612.
14. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.-L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582, doi:https://doi.org/10.1016/j.renene.2016.12.095.
15. Morantes-Quintana, G.R.; Rincón-Polo, G.; Pérez-Santodomingo, N.A. Multiple linear regression model to estimate PM1 concentration | Modelo de regresión lineal múltiple para estimar concentración de PM1. *Rev. Int. Contam. Ambient.* **2019**, *35*,

179–194, doi:10.20937/RICA.2019.35.01.13.

16. O'Fallon, B.D.; Wooderchak-Donahue, W.; Crockett, D.K. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics* **2013**, *29*, 1361–1366, doi:10.1093/bioinformatics/btt172.
17. Zimmermann, O.; Hansmann, U.H.E. Support vector machines for prediction of dihedral angle regions. *Bioinformatics* **2006**, *22*, 3009–3015, doi:10.1093/bioinformatics/btl489.
18. Abdullah, S.; Pradhan, R.C.; Pradhan, D.; Mishra, S. Modeling and optimization of pectinase-assisted low-temperature extraction of cashew apple juice using artificial neural network coupled with genetic algorithm. *Food Chem.* **2021**, *339*, 127862, doi:https://doi.org/10.1016/j.foodchem.2020.127862.
19. Sabater, C.; Blanco-Doval, A.; Montilla, A.; Corzo, N. Optimisation of an enzymatic method to obtain modified artichoke pectin and pectic oligosaccharides using artificial neural network tools. In silico and in vitro assessment of the antioxidant activity. *Food Hydrocoll.* **2021**, *110*, 106161, doi:https://doi.org/10.1016/j.foodhyd.2020.106161.
20. Behr, A.; Giese, M.; Teguim K., H.D.; Theune, K. Early Prediction of University Dropouts-A Random Forest Approach. *Jahrb. Natl. Okon. Stat.* **2020**, *240*, 743–789, doi:10.1515/jbnst-2019-0006.
21. Ocampo, D.; Rivas, R. Estimating daily net radiation from Multiple Linear Regression Models | Estimación de la radiación neta diaria a partir de Modelos de Regresión Lineal Múltiple. *Rev. Chapingo, Ser. Ciencias For. y del Ambient.* **2013**, *19*, 263–271, doi:10.5154/r.rchscfa.2012.04.031.
22. Kim, M.; Song, H.; Kim, Y. Direct short-term forecast of photovoltaic power through a comparative study between coms and himawari-8 meteorological satellite images in a deep neural network. *Remote Sens.* **2020**, *12*, 2357, doi:10.3390/rs12152357.
23. Sun, H.; Gui, D.; Yan, B.; Liu, Y.; Liao, W.; Zhu, Y.; Lu, C.; Zhao, N. Assessing the potential of random forest method for estimating solar radiation using air pollution index. *Energy Convers. Manag.* **2016**, *119*, 121–129, doi:https://doi.org/10.1016/j.enconman.2016.04.051.
24. Srivastava, R.; Tiwari, A.N.; Giri, V.K. Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India. *Heliyon* **2019**, *5*, e02692, doi:https://doi.org/10.1016/j.heliyon.2019.e02692.
25. Şahin, M.; Kaya, Y.; Uyar, M. Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data. *Adv. Sp. Res.* **2013**, *51*, 891–904, doi:https://doi.org/10.1016/j.asr.2012.10.010.
26. da Silva, M.B.P.; Francisco Escobedo, J.; Juliana Rossi, T.; dos Santos, C.M.; da Silva, S.H.M.G. Performance of the Angstrom-Prescott Model (A-P) and SVM and ANN techniques to estimate daily global solar irradiation in Botucatu/SP/Brazil. *J. Atmos. Solar-Terrestrial Phys.* **2017**, *160*, 11–23, doi:https://doi.org/10.1016/j.jastp.2017.04.001.
27. Benali, L.; Notton, G.; Fouilloy, A.; Voyant, C.; Dizene, R. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renew. Energy* **2019**, *132*, 871–884, doi:10.1016/j.renene.2018.08.044.
28. Huang, J.; Troccoli, A.; Coppin, P. An analytical comparison of four approaches to modelling the daily variability of solar irradiance using meteorological records. *Renew. Energy* **2014**, *72*, 195–202, doi:https://doi.org/10.1016/j.renene.2014.07.015.
29. Martínez Castillo, C.A. Modelado de la irraciación global mensual usando estaciones de la red de Meteogalicia., Universidad de Vigo, 2019.
30. Vázquez Vázquez, M. *Radiación solar e severidade climática en Galicia*; Vázquez Vázquez, M., Ed.; Universidade de Vigo: Vigo, 2008; ISBN 978-84-612-4469-0.
31. Meteogalicia Observacións. Rede meteorolóxica.
32. Elbayoumi, M.; Ramli, N.A.; Fitri Md Yusof, N.F. Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM2.5–10 and PM2.5 concentrations in naturally ventilated schools. *Atmos. Pollut. Res.* **2015**, *6*, 1013–1023, doi:https://doi.org/10.1016/j.apr.2015.09.001.
33. Al-Alawi, S.M.; Abdul-Wahab, S.A.; Bakheit, C.S. Combining principal component regression and artificial neural networks for

more accurate predictions of ground-level ozone. *Environ. Model. Softw.* **2008**, *23*, 396–403, doi:https://doi.org/10.1016/j.envsoft.2006.08.007.

34. Rodríguez-Jaume, M.-J.; Mora Catalá, R. Análisis de regresión múltiple. In *Estadística informática: casos y ejemplos con el SPSS*; Rodríguez-Jaume, M.-J., Mora Catalá, R., Eds.; Publicaciones de la Universidad de Alicante: Alicante, 2001; pp. 109–123 ISBN 84-7908-638-6.
35. Agatonovic-Kustrin, S.; Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **2000**, *22*, 717–727, doi:10.1016/S0731-7085(99)00272-1.
36. Balas, C.E.; Koç, M.L.; Tür, R. Artificial neural networks based on principal component analysis, fuzzy systems and fuzzy neural networks for preliminary design of rubble mound breakwaters. *Appl. Ocean Res.* **2010**, *32*, 425–433, doi:10.1016/j.apor.2010.09.005.
37. Basheer, I.A.; Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *J. Microbiol. Methods* **2000**, *43*, 3–31, doi:https://doi.org/10.1016/S0167-7012(00)00201-3.
38. Yolmeh, M.; Habibi Najafi, M.B.; Salehi, F. Genetic algorithm-artificial neural network and adaptive neuro-fuzzy inference system modeling of antibacterial activity of annatto dye on Salmonella enteritidis. *Microb. Pathog.* **2014**, *67–68*, 36–40, doi:10.1016/j.micpath.2014.02.003.
39. Lee, K.Y.; Chung, N.; Hwang, S. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. *Ecol. Inform.* **2016**, *36*, 172–180, doi:https://doi.org/10.1016/j.ecoinf.2015.08.011.
40. Zhang, L.; Li, H.; Kong, X.-G. Evolving feedforward artificial neural networks using a two-stage approach. *Neurocomputing* **2019**, *360*, 25–36, doi:https://doi.org/10.1016/j.neucom.2019.03.097.
41. Sutariya, V.; Groshev, A.; Sadana, P.; Bhatia, D.; Pathak, Y. Artificial neural network in drug delivery and pharmaceutical research. *Open Bioinforma. J.* **2013**, *7*, 49–62, doi:10.2174/1875036201307010049.
42. Carmona Suárez, E.J. *Tutorial sobre Máquinas de Vectores Soporte (SVM)*; Universidad Nacional de Educación a Distancia (UNED): Madrid, Spain, 2016;
43. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification Available online: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed on Nov 9, 2020).
44. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32, doi:10.1023/A:1010933404324.
45. RapidMiner Documentation. Random Forest Available online: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/parallel_random_forest.html (accessed on Nov 3, 2020).