Zarzo Castelló, M.; Martí Pérez, PC. (2011). Modeling the variability of solar radiation data among weather stations by means of principal components analysis. Applied Energy. 88(8):2775-2784. doi:10.1016/j.apenergy.2011.01.070.

Modeling the variability of solar radiation data among weather stations by means of principal components analysis

Manuel Zarzo [a], Pau Martí [b*]

[a]Departamento de Estadística e Investigación Operativa Aplicadas y Calidad. Universidad Politécnica de Valencia. Camino de Vera, s/n. 46022 Valencia. Spain.

[b]Departamento de Ingeniería Rural y Agroalimentaria. Universidad Politécnica de Valencia. Camino de Vera, s/n. 46022 Valencia. Spain.

*Corresponding author. Tel.: +34 963877490; Fax.: +34 963877499.

E-mail addresses: mazarcas@eio.upv.es (M. Zarzo), paumarpe@hotmail.es (P. Martí).

Abstract

Measurements of global terrestrial solar radiation (R_s) are commonly recorded in meteorological stations. Daily variability of R_s has to be taken into account for the design of photovoltaic systems and energy efficient buildings. Principal components analysis (PCA) was applied to R_s data recorded at 30 stations in the Mediterranean coast of Spain. Due to equipment failures and site operation problems, time series of R_s often present data gaps or discontinuities. The PCA approach copes with this problem and allows estimation of present and past values by taking advantage of R_s records from nearby stations. The gap infilling performance of this methodology is compared with alternative conventional approaches. A new method was also developed for R_s estimation if previous measurements are not available. Four principal components explain 66% of the data variability with respect to the average trajectory. By means of multiple linear regression, it was found that this variability can be fitted according to the latitude, longitude and altitude of the station where data were recorded from. Additional geographical or climatic variables did not increase the predictive goodness-of-fit. The resulting models allow the estimation of daily R_s values at any location in the area under study. The proposed methodology for estimating R_s based on geographical parameters would be of interest to design solar energy systems and to select their best location.

Keywords: solar radiation, missing data estimation, PCA, multivariate statistical monitoring

Introduction

Solar radiation plays a key role in evaporation, plant photosynthesis or crop growth and productivity [1-4]. In architecture, accurate estimates of long-term global solar radiation are required for the design and development of energy efficient buildings [5,6]. Solar radiation is also essential in biophysical models for risk assessment of forest fires, in hydrological simulation models of natural processes [7], in environmental and agrometeorological research as well as in atmospheric physics [8]. Nonetheless, the major interest of measuring solar radiation is for the simulation and design of solar energy systems [9].

The most common solar radiation measurements recorded in meteorological stations correspond to total radiation on a horizontal surface, R_s, also called global terrestrial solar radiation, which is normally given on an hourly or daily basis [6]. These data are required for the design of photovoltaic applications in remote or isolated areas where no connection to an electrical supply grid is available, for instance in rural or mountainous areas, natural parks, small islands and developing countries in general [10-14]. In these places, the daily R_s variability has to be taken into account for the design of photovoltaic systems in order to guarantee enough power generation for the essential electric devices [15]. In developed countries, solar energy systems are frequently implemented in buildings for water heating or electric power supply. The design of these systems is also based on R_s measurements. However, in many applications of solar energy, especially in the aforementioned isolated areas, projects are not supported by the required R_s data at the place of interest. Actually, solar radiation is measured at relatively few weather stations in comparison to other variables such as temperature or relative humidity (RH). This is generally due to the high cost, maintenance and calibration requirements of the measuring equipment [9,10,12,16,17].

Although suitable weather records have become more and more available in recent years, data reliability and quality is another problem. Even in automatic stations where solar radiation is measured, data records often lie outside the expected range [3-5] and are erroneous because of sensor calibration problems. According to Muneer et al. [9], another cause of errors is site operation problems such as instrument proximity to shading elements, electrical and magnetic fields, weather elements as well as bird or insect activity. Erroneous data need to be discarded, and equipment failures also cause missing data. As a result, time series of R_s often present data gaps or discontinuities.

One alternative to cope with the lack of accurate R_s measurements is to use mathematical predictive models relying on climatic inputs. Several empirical, numerical and physically-based models have been proposed for R_s estimation based on different input combinations. They differ in sophistication from simple empirical equations based on common climatic data to more complex numerical models involving high computational costs and relying on numerous inputs. The most frequent inputs are sunshine duration, extraterrestrial radiation, mean temperature, maximum temperature, soil temperature, RH, number of rainy days, altitude, latitude, total precipitation, cloudiness, and evaporation [16]. Among the simplest methods for estimating solar radiation data, Hargreaves and Samani [18], Bristow and Campbell [19] and Allen [20] propose equations relying on maximum and minimum temperatures as well as on extraterrestrial radiation. These approaches, modified by other authors [17], take into account implicitly the geographical information of the studied locations by including theoretical extraterrestrial radiation values based on latitude, day of the year, sunset hour angle and relative distance earth-sun.

As an alternative to conventional approaches, artificial neural networks (ANNs) have been successfully applied for solar radiation estimation [1,2,5,8,10-13,21-31]. Techniques based on artificial intelligence have also been proposed particularly for isolated areas [14]. However, only a small part of these works present models fed by few easily measurable inputs such as temperature and/or RH records [11,28,29].

The development of R_s estimation methods not relying on local climatic records turns into a task of great relevance because even the simplest climatic parameters are not available in many cases given the limited number of available automatic weather stations. One approach hardly tackled in literature would be to develop models relying exclusively on exogenous R_s inputs from nearby locations with similar climatic conditions. These models are of relevant interest given the aforementioned ubiquitous problems such as data scarcity, equipment failures, maintenance and calibration as well as physical and biological constraints.

Data of R_s recorded daily at different stations can be regarded as a multivariate time series. Such type of data is common in the monitoring and control of industrial processes. For example, chemical reactors are usually monitored by means of electronic sensors that record the temperature at different points of the process. In this context of multivariate statistical process control (MSPC), principal components analysis (PCA) is a useful technique for process monitoring and diagnosis because it allows data estimation in case of faulty sensors [32]. PCA is one of the multivariate techniques wider spread [33]. In a recent study, PCA was used for modeling the spatial data variability from a set of RH sensors located at different positions [34]. The same PCA approach was applied here for the analysis of R_s values measured daily at 30 weather stations. PCA copes with gap infilling by taking advantage of R_s records from nearby stations. Multiple linear regression (MLR) was used to identify which geographical or climatic parameters are the ones that best explain the differences in R_s measurements recorded at the 30 stations. Once identified the key variables, a new methodology is proposed to estimate R_s when, apart from exogenous measurements, these parameters are also available.

Materials and methods

1. Data characterization

The database analyzed here consisted of daily R_s values from 30 weather stations located on the Mediterranean coast of Spain (Table 1). Data were obtained from the Valencian Institute of Agricultural Research (IVIA). The dataset was structured as a matrix of 30 rows (stations) by 2920 variables (in columns). Each variable corresponds to one day in the 8-year period under study (January 2000 to December 2007). The original R_s series contained missing data. In order to assess different procedures for gap infilling, it is convenient to work with a complete data matrix. Thus, all variables in the initial R_s dataset containing missing data were discarded, resulting a complete matrix with 1203 variables. Fig. 1 displays the average R_s for these 1203 days. A clear periodic trend can be observed due to R_s annual seasonability.

[FIGURE 1 NEAR HERE]

2. PCA models

2.1. PCA configuration and data pretreatment

Principal components are directions of maximum data variance obtained as linear combinations of the original variables. The projections of observations (weather stations, in this case) over these directions are called scores. The variable containing these projections over the first principal component (PC1) is called score vector ($\mathbf{t}_1$). Similarly, $\mathbf{t}_2$ contains the projections over PC2, and so on. The contributions of variables in the formation of a given component are called loadings, being $\mathbf{p}_1$ the loadings in the formation of PC1.

The software SIMCA-P 10.0 (Umetrics AB, Malmö, Sweden) was used to carry out all PCA models. It uses the NIPALS algorithm [35] which extracts components one by one. Given a matrix $\mathbf{X}$, this algorithm calculates $\mathbf{t}_1$ and $\mathbf{p}_1$, resulting a residual matrix $\mathbf{E}_1$ (eq. 1). PC2 is obtained by applying the NIPALS algorithm to $\mathbf{E}_1$. Thus, PC2 is the direction that explains the maximum data variability of $\mathbf{E}_1$ and remains orthogonal to PC1. Next, a residual matrix $\mathbf{E}_2$ is calculated (eq. 1). This procedure can be conducted sequentially until $\mathbf{E}$=0.

$$\left.\begin{aligned} \mathbf{E}_1 &= \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^{\mathsf{T}} \\ \mathbf{E}_2 &= \mathbf{E}_1 - \mathbf{t}_2\mathbf{p}_2^{\mathsf{T}} \\ &\dots \\ \mathbf{E}_k &= \mathbf{E}_{k-1} - \mathbf{t}_k\mathbf{p}_k^{\mathsf{T}} \end{aligned}\right\} \Rightarrow \ \mathbf{X} = \mathbf{t}_1\mathbf{p}_1^{\mathsf{T}} + \dots + \mathbf{t}_k\mathbf{p}_k^{\mathsf{T}} + \mathbf{E}_{k-1} \tag{1}$$

Each row of the R_s dataset is a time series that reflects the evolution of R_s recorded at one station. In the context of MSPC, such time series is often called ‘trajectory’ because of the trend observed when the parameter is plotted versus time. If a new row is obtained by averaging the values of each column, it could be regarded as the mean trajectory. In order to highlight the relationships (i.e. similarities and dissimilarities) among stations, data were mean-centered prior to PCA by subtracting the mean value of each column. As a result, the average of all centered variables becomes null. In the MSPC of batch chemical processes, the idea of subtracting the mean trajectory prior to PCA was first proposed by Nomikos and MacGregor [36]. The same methodology has also been successfully applied by other works [37,38]. Variables are also scaled to unitary variance prior to PCA when the variance among them is very different [38], but this is not the case here.

2.2. Number of relevant PCs and outlier detection

Different methods can be applied to decide how many components should be extracted (i.e. the value of *k* in eq. 1) for the purpose of modeling the systematic data variability of **X** [33]. Further components not calculated are included in a matrix of errors ($\mathbf{E}_{k\text{-}1}$ in eq. 1) that is assumed to account for random variation. One criterion implemented in the software SIMCA-P 10.0 is cross-validation [39]. It considers that one PC does not provide relevant information if it changes significantly when several observations are randomly removed.

Applying PCA to all R_s data assumes that the relationships among stations are basically maintained all the year round. In order to test this hypothesis, the R_s dataset was split in two subsets of about equal size, one containing those variables with an average value higher than 200 and another one containing the remaining variables. These subsets will be referred to as R_s>200 and R_s<200, respectively. The value of 200 is approximately the average value of all data in the R_s matrix (horizontal dotted line in Fig. 1). Next, two new PCA models were fitted, one with each submatrix. Results from both models were compared.

A scatterplot of the scores corresponding to two different components is referred to as a score plot. The score plot corresponding to PC1 and PC2 (i.e., $\mathbf{t}_2$ vs. $\mathbf{t}_1$), referred to here as the PC1/PC2 plot, usually highlights the basic similarities and dissimilarities among observations. Score plots with different combinations of PCs were visually inspected in order to detect outliers as well as to identify stations with a similar performance. The distance of observations to the PCA model was also checked.

2.3. PCA infilling approach for R_s estimation

Missing data due to sensor failures is a problem often encountered in MSPC. Different approaches have been proposed for PCA to deal with incomplete observations [40,41]. One of these algorithms is implemented in the software SIMCA-P 10.0 [42]. Starting from the complete R_s matrix, three new ones were obtained containing 5%, 10% and 15% randomly distributed gaps. In the four cases, data were mean-centered prior to PCA. After obtaining the score and loading vectors for each PC, they were used to reconstruct the **X** matrix (eq. 1). This procedure was conducted using Matlab version 7.4.0 (MathWorks Inc., Natick, MA, USA), considering an increasing number of PCs. Next, in order to assess the accuracy of the gap infilling method, the estimated missing values were compared with the original ones.

Four additional methods were tested for gap infilling: (i) by adopting as R_s estimations for a given station, the R_s records from the nearest station (1-neighbor); (ii) by obtaining the R_s average of two nearest stations with a similar altitude (2-neighbor); (iii) by adopting the R_s values from the nearest station in the score plot for PC1/PC2 (1-neighbor-SP); (iv) by assigning the R_s average of two neighboring stations in the score plot for PC1/PC2 (2-neighbor-SP). These methods will be referred to hereafter with the name indicated within brackets. In order to assess their efficiency for gap infilling, they were applied to the matrices with 5%, 10% and 15% of gaps. The neighboring stations in the 1- and 2-neighbor-SP methods were established only according to the score plot of the complete matrix.

2.4. R_s estimation from geographic parameters

Principal component regression (PCR) was used to study if score vectors can be predicted according to geographic and climatic parameters. The proposed methodology comprises two steps. First, score and loading vectors of the *k* relevant PCs were extracted from the R_s matrix. Second, step-wise MLR was applied to fit each score vector according to the following independent variables: *latitude, longitude, altitude, minimum distance to the sea, temperature* (*average, maximum, minimum*), *RH*, *wind speed*, *Gorezynski continentality index* [43] and *cumulated rain.* Climatic parameters correspond to the average values for the 1203 days of study. The software Statgraphics plus 5.1 (StatPoint Technologies Inc., Warrenton, VA, USA) was used to conduct all regression models.

Once obtained the *k* predictive equations, they might be applied to estimate the t_1, t_2, ... t_k scores of a new station according to its geographic and climatic data. Next, the R_s estimation for the *j*-th day would be obtained based on these predicted scores and the loadings calculated in the first step for the *j*-th day (eq. 2), being *k* the number of relevant PCs and μ_j the *j*-th column average of the R_s matrix.

$$(R_s)_j = \mu_j + \hat{t}_1\, p_{1j} + \hat{t}_2\, p_{2j} + ... + \hat{t}_k\, p_{kj} \qquad (2)$$

In order to assess the performance of the proposed method, MLR equations were applied using the geographic and climatic data of each station and, next, the R_s matrix was reconstructed by applying eq. 2 for the 1203 days. Predicted values were compared with the original ones. The 1- and 2- neighbor-SP approaches described in the previous section were also tested. The first one consisted of adopting for a given station, the R_s records from the station with most similar t_1 and t_2 scores. Similarly, the estimation according to the 2-neighbor-SP method was obtained as the R_s average of the two nearest stations in the PC1/PC2 score plot. The t_1 and t_2 scores of the target station were previously estimated by applying the MLR equations based on its geographic and climatic parameters.

3 Performance indicators

Several error parameters were calculated to assess the performance accuracy of the proposed estimation methods. The average absolute relative error (AARE), the mean absolute error (MAE) and the mean squared error (MSE), which are commonly used in time series analysis, were obtained according to eqs. 3, 4 and 5, respectively, being x_i the observed R_s value, $\hat{x}_i$ the prediction, and n the number of missing data randomly created in the R_s matrix.

$$AARE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{x_i - \hat{x}_i}{x_i} \right| \quad (3)$$

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| x_i - \hat{x}_i \right| \quad (4)$$

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \hat{x}_i \right)^2 \quad (5)$$

Results and discussion

1. PCA of the R_s matrix: relevant PCs and outlier detection

Three PCA models were carried out, one with all 1203 variables of the R_s dataset, another using the set of R_s>200 variables and a third one with R_s<200. In the three models, the score plot for PC3/PC4 reveals that station s19 presents abnormal values in both components (figures not shown). This station is the most northern one, which might explain its different performance. However, the position of s19 in the PC1/PC2 score plot is not abnormal (Fig. 2). Taking into account that PC3 and PC4 provide relevant information, station s19 was discarded and the three models were repeated. A summary overview of these six models is shown in Table 2.

The software SIMCA-P 10.0 considers that a certain component explains systematic data variability if the goodness-of-fit for that component obtained by cross-validation (Q^2) is higher than a certain threshold [42]. In five of the six models, the cross-validation criterion is satisfied up to PC4 (Table 2). In order to further investigate the number of relevant PCs, it was checked that the $\mathbf{t}_1$ score vector obtained from the R_s<200 model with 29 stations is strongly correlated with that from the R_s>200 model ($r = 0.964$, $p < 0.0001$). The correlation is also statistically significant for the $\mathbf{t}_2$, $\mathbf{t}_3$ and $\mathbf{t}_4$ vectors ($p < 0.0001$) but not in the case of $\mathbf{t}_5$ ($r = 0.267$, $p = 0.161$) nor $\mathbf{t}_6$ ($r = 0.291$, $p = 0.125$). Again, this result suggests that 4 components should be used to describe the systematic variability of the R_s matrix.

In the R_s>200 model with 29 observations, station s27 has abnormal values of PC5 and PC7 and appears as an outlier in the PC5/PC7 score plot (Fig. not shown). The same result was obtained with the R_s<200 model and the one with 1203 variables. The

R_s pattern of station s27 is slightly different to the rest probably because it has the highest distance to the sea and the most continental climate. Actually, it presents the lowest average and minimum temperature among the 30 stations. Nonetheless, s27 was not discarded because its performance is not abnormal in the four relevant PCs. Different score plots were visually inspected, but no additional outliers were identified.

2. Similarities among stations based on the score plots

R^2_X is usually called goodness-of-fit because it indicates how good is a given PC to fit the observed values. PC1 explains about 37% of the mean-centered data variability (Table 2). The coordinate position of stations in the PC1/PC2 score plot, if properly rotated, is strikingly similar to their geographic position (Fig. 2). The rotation was achieved by plotting ($-2\ \mathbf{t}_1 + \mathbf{t}_2$) vs ($-\ \mathbf{t}_1 - 2\mathbf{t}_2$).

[FIGURE 2 NEAR HERE]

In order to assess if the differences among stations are relevant in practice, the average R_s value was calculated for all stations. Averages follow approximately a normal distribution, being 225.3 the maximum value (station s29) and 187.4, the minimum value (station s6). Thus, the average R_s of s29 is 20.2% higher than in the case of station s6, which highlights the importance of choosing correctly the location for a solar energy system. It was found that average R_s values were correlated with t_1 scores (r = 0,756). Thus, PC1 will highlight which stations provide higher or lower values than the average trajectory. Further PCs will describe changes in the shape with respect to the mean trajectory. Moreover, t_1 scores are also correlated with latitude (r = - 0,939), as reflected in Fig. 2. The PC3/PC4 score plot for the R_s<200 and R_s>200 models are quite similar (Fig. 3), which again indicates that PC3 and PC4 provide relevant information.

[FIGURE 3 NEAR HERE]

The PCA model with 4 components accounts for about 66% of the mean-centered data variability (Table 2). Thus, the PC1/PC2 and PC3/PC4 score plots will highlight the most relevant similarities and dissimilarities. Stations close to each other in both plots will present a similar performance, i.e., a trajectory of R_s recordings with a similar average value and shape. After visually inspecting both score plots (Figs. 2 and 3), four clusters of stations with a similar R_s pattern were established: cluster A (stations s1, s4, s6, and s15), B (s2, s5, s9, and s10), C (s26, s29, and s30) and D (s11, s12, and s14). They basically differ in latitude (Fig. 2) while C is the cluster with highest altitude, which implies a more continental climate. Trajectories of stations belonging to the same clusters were averaged, centered with respect to the mean trajectory and smoothed using a moving average of order 50 (Fig. 4). Cluster C yields the trajectory with highest average values and, moreover, its pattern is somewhat different. This distinctive performance is basically explained by PC2. Clusters C and D correspond to southern stations and their R_s values are higher than clusters A and B, which reflects the negative correlation between latitude and R_s. Fig. 3 shows that stations in clusters A and B are discriminated by PC4 which implies that their trajectories are somewhat different, as reflected by Fig. 4.

[FIGURE 4 NEAR HERE]

3. Gap infilling results

Taking into account that station s19 is an outlier, it was disregarded for the gap infilling study. PCA was applied to the complete R_s matrix (29 stations by 1203 variables) and to the three matrices containing 5%, 10% and 15% of missing data (i.e., 1744, 3489 and 5233 gaps, respectively). Table 3 shows that R^2_X and Q^2 values of these models are nearly the same regardless of the amount of missing data. Despite the presence of gaps, PC4 satisfies the cross-validation criterion. MSE, MAE and AARE (eqs. 3 to 5) were calculated by comparing discarded data with the predictions obtained by eq. 1 with an increasing number of PCs. In the complete R_s matrix, error parameters become null using 28 components (see Fig. 5), which implies a perfect fit if the maximum number of possible PCs is used. Only a slight increase of MAE and AARE is observed as the amount of missing data increases. Fig. 5 also shows that PCA models built with 4 or 5 PCs lead to similar errors. Additional components just provide a slight decrease of the error indicators, which is consistent with the cross-validation results suggesting that only four PCs are relevant.

[FIGURE 5 NEAR HERE]

Missing R_s data were also estimated according to four alternative methods based on neighbor assignment. This assignment presents in some cases several valid options because the number of available stations is limited and there is not always a single optimum choice. Therefore, only the complete matrix was considered to provide the $\mathbf{t}_1$ and $\mathbf{t}_2$ scores used to establish the neighbor assignment, which was the same for the 3 gap sizes. As observed in Table 4, error indicators are higher if the percentage of missing data increases. After PCA and 2-neighbor-SP, the best results correspond to the 2-neighbor procedure (Table 4) which presents average indicators slightly better than the 1-neighbor and 1-neighbor-SP methods. As could be expected, the approaches based only in the information of one neighboring station are worse than those taking into account data from two neighbors.

4. R_s estimation from geographical data

Score and loading vectors of PC3 and PC4 were extracted from the R_s matrix with 29 stations and 1203 variables. Station s19 was disregarded because it becomes an outlier in both PCs but not for the previous components. Thus, all stations were considered to obtain score and loading vectors of PC1 and PC2.

MLR was applied next to determine if $\mathbf{t}_1$ is correlated with *latitude* (φ), *longitude* (τ), *altitude* (z), and *distance* to the sea. The same study was conducted with $\mathbf{t}_2$, $\mathbf{t}_3$, and $\mathbf{t}_4$. After trying several alternative models, it was decided to consider also two indicator variables and their interactions. $I_{z>400}$ takes the value 1 for the 5 stations with an altitude higher than 400 m and zero otherwise. The indicator variable $I_{\varphi<38.7}$ takes the value 1 for stations that satisfy the condition $\varphi < 38.7$.

$$t_1 = 22403 - 579.1\,\varphi + 363.1\,I_{\varphi<38.7} - 185.6\,\tau + 0.58\,z \qquad (6)$$

$$t_4 = 7932 - 208.0\,\varphi - 333.3\,I_{\varphi<38.7} + 676.9\,\tau - 341.8\,I_{z>400} \qquad (7)$$

$$t_2 = -16846 + 421.4\,\varphi - 540.7\,I_{\varphi<38.7}\,(\varphi-39) + 1.3\,z \qquad (8)$$

$$t_3 = -38125 + 968.1\,\varphi - 1611.9\,I_{\varphi<38.7}\,(\varphi-39.1) - 181\,\tau \qquad (9)$$

All regression coefficients of the best predictive equations are statistically significant ($p \leq 0.002$ for eq. 6, $p \leq 0.0002$ for eq. 7, $p < 0.003$ for eq. 8, and $p < 0.009$ for eq. 9). It was also checked that residuals followed approximately a normal distribution and no outliers were detected. *Longitude* is significantly correlated with *altitude* (r = 0.492, p = 0.006) as well as with *latitude* (r = -0.494, p = 0.0055). Given the correlation among predictive variables, trying to interpret the effect of each parameter in eqs. 6 to 9 might be misleading.

Coefficients of determination are the following: 0.983 (eq. 6), 0.901 (eq. 7), 0.785 (eq. 8) and 0.900 (eq. 9). These high values suggest that a considerable amount of the centered R_s data variability depends on the geographical position of the station. Fig. 6 shows that t_1 and t_2 scores predicted from eqs. 6 and 8 based on geographical information are similar to those originally obtained from the R_s matrix. Predictive MLR equations for $\mathbf{t}_5$ and $\mathbf{t}_6$ were also tried using geographic and climatic variables, but a very poor goodness-of-fit was obtained.

[FIGURE 6 NEAR HERE]

It was found that none of the climatic variables entered in the MLR models. Thus, *latitude*, *longitude* and *altitude*, which are parameters readily available for any location in the region under study, are enough to predict the four relevant scores. Eqs. 6 to 9 are only valid for the R_s estimation in weather stations located on the Mediterranean coast of Spain with similar geographical characteristics as those in Table 1. Nevertheless, the proposed methodology could be applied to any kind of climatic conditions.

The PCR approach was applied to reconstruct the R_s matrix. PCR using four components provides more accurate estimations than the 1- and 2-neighbor-SP approaches, which were also tested (Table 5). Error indicators of 1-neighbor-SP are similar as those of PCR using only one component, which suggests that this method would not be recommended. By contrast, errors of 2-neighbor-SP and PCR with 3 components are similar. Again, these results indicate that it would be better to use data from two neighboring stations instead of just one.

Conclusions

Choosing the right location for a solar energy system is a key factor for maximizing the power generation. Moreover, the estimation of daily R_s values is of interest for the design of photovoltaic systems and energy efficient buildings. Available R_s data are useful for this purpose, particularly if they are recorded from nearby weather stations. PCA was applied to R_s data recorded at 30 stations in the Mediterranean coast of Spain. Four principal components account for the systematic data variation and explain about 66% of the mean-centered R_s variability. By means of MLR, it was found that the latent variables associated to the four relevant PCs can be predicted according to latitude, longitude and altitude. Climatic variables did not increase the predictive goodness-of-fit. Based on the results, a new methodology is proposed to estimate daily R_s values at any location in the region under study when only local geographical parameters are available. The proposed method exhibits a higher accuracy than simpler procedures using data from neighboring stations.

Time series of R_s often present data gaps or discontinuities. In practice, this problem is often solved by adopting the measurements from neighboring stations. The PCA approach characterizes the similarities among weather stations and also allows estimation of present and past R_s values. The proposed method for gap infilling is more accurate than four alternative procedures also tested. The statistical methodology applied here is commonly used in the context of MSPC, particularly in chemical processes, but as far as we know this is the first work that applies such methodology for the estimation of solar radiation.

Mathematical notation

$\mathbf{X}$	matrix (upper case, bold)
$\mathbf{t}$	vector, i.e. column matrix (lower case, bold)
$\mathbf{p}^{\mathrm{T}}$	transposed vector, i.e. row matrix (lower case, bold)
k	scalar (lower case, italicized)

Acknowledgements

We are grateful to the Valencian Institute of Agricultural Research (IVIA) for providing the dataset used in the present work.

References

[1] López G, Rubio MA, Martínez M, Batlles FJ. Estimation of hourly global photosynthetically active radiation using artificial neural network models. Agric Forest Meteorol 2001;107:279-91.

[2] Reddy KS, Ranjan M. Solar resource estimation using artificial neural networks and comparison with other correlation models. Energy Convers Manage 2003;44:2519–30.

[3] Hunt LA, Kuchar L, Swanton CJ. Estimation of solar radiation for use in crop modelling. Agric Forest Meteorol 1998;91:293-300.

[4] Abraha MG, Savage MJ. Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations. Agric Forest Meteorol 2008;148:401-16.

[5] Hontoria L, Aguilera J, Zufiria P. Generation of hourly irradiation synthetic series using the neural network multilayer perceptron. Sol Energy 2002;72(5):441-6.

[6] Zekai Ș. Solar energy fundamentals and modeling techniques: atmosphere, environment, climate change and renewable energy. London: Springer; 2008.

[7] Meza F, Varas E. Estimation of mean monthly solar global radiation as a function of temperature. Agric Forest Meteorol 2000;100:231-41.

[8] Tymvios FS, Jacovides CP, Michaelides SC, Scouteli C. Comparative study of Ångström's and artificial neural networks' methodologies in estimating global solar radiation. Sol Energy 2005;78:752-62.

[9] Muneer T, Younes S, Munawwar S. Discourses on solar radiation modeling. Renew Sustain Energy Rev 2005;11:551-602

[10] Mohandes M, Rehman S, Halawani TO. Estimation of global solar radiation using artificial neural networks. Renew Energy 1998;14(1-4):179-84.

[11] Rehman S, Mohandes M. Artificial neural network estimation of global solar radiation using air temperature and relative humidity. Energy Policy 2008;36:571-6.

[12] Bosch JL, López G, Batlles FJ. Daily solar irradiation estimation over a mountainous area using artificial neural networks. Renew Energy 2008;33:1622-8.

[13] Azadeh A, Maghsoudi A, Sohrabkhani S. An integrated artificial neural networks approach for predicting global radiation. Energy Convers Manage 2009;50:1497-1505.

[14] Mellit A, Kalogirou SA, Hontoria L, Shaari S. Artificial intelligence techniques for sizing photovoltaic systems: A review. Renew Sustain Energy Rev 2009;13:406-19.

[15] Mellit A, Benghanem M, Hadj Arab A, Guessoum A. An adaptive artificial neural network model for sizing stand-alone photovoltaic systems: application for isolated sited in Algeria. Renew Energy 2005;30:1501-24.

[16] Bakirci K. Models of solar radiation with hours of bright sunshine: A review. Renew Sustain Energy Rev 2009;13:2580-8.

[17] Liu X, Mei X, Li Y, Wang Q, Jensen JR, Zhang Y, Porter JR. Evaluation of temperature-based global solar radiation models in China. Agric Forest Meteorol 2009;149:1433–46.

[18] Hargreaves GH, Samani ZA. Estimating potential evapotranspiration. J Irrig Drain Eng 1982;108(3):225-30.

[19] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric Forest Meteorol 1984;31:159-66.

[20] Allen RG. Self-calibrating method for estimating solar radiation from air temperature. J Hydrol Eng 1997;2(2):56-67.

[21] Fadare DA. Modelling of solar energy potential in Nigeria using an artificial neural network model. Appl Energy 2009;86:1410–22.

[22] Al-Alawi SM, Al-Hinai HA. An ANN-based approach for predicting global radiation in locations with no direct measurement instrumentation. Renew Energy 1998;14(1-4):199-204.

[23] Mohandes M, Balghonaim A, Kassas M, Rehman S, Halawani TO. Use of radial basis functions for estimating monthly mean daily solar radiation. Sol Energy 2000;68(2):161-8.

[24] Sfetsos A, Coonick AH. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Sol Energy 2000;68(2):169-78.

[25] Dorvlo ASS, Jervase JA, Al-Lawati A. Solar radiation estimation using artificial neural networks. Appl Energy 2002;71:307–19.

[26] Sözen A, Arcaklioğlu E, Özalp M. Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Convers Manage 2004;45:3033–52.

[27] Lam JC, Wan KKW, Yang L. Solar radiation modelling using ANNs for different climates in China. Energy Convers Manage 2008;49:1080–90.

[28] Benghanem M, Mellit A, Alamri SN. ANN-based modelling and estimation of daily global solar radiation data: A case study. Energy Convers Manage 2009;50:1644-55.

[29] Kalogirou S, Michaelides S, Tymvios F. Prediction of maximum solar radiation using artificial neural networks. Proceedings of World Renewable Energy Congress VII, Cologne, Germany; 2002.

[30] López G, Batlles FJ, Tovar-Pescador J. Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 2005;30:1675-84.

[31] Hontoria L, Aguilera J, Zufiria P. An application of the multilayer perceptron: Solar radiation maps in Spain. Sol Energy 2005;79:523-30.

[32] Dunia R, Qin SJ, Edgar TF, McAvoy TJ. Identification of faulty sensors using principal component analysis. AIChE J 1996;42:2797-2812.

[33] Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer-Verlag; 2002.

[34] García-Diego FJ, Zarzo M. Microclimate monitoring by multivariate statistical control: The renaissance frescoes of the Cathedral of Valencia (Spain). J Cult Heritage 2010;11:339-44.

[35] Wold H. Estimation of principal components and related models by iterative least squares. In: Krishnaiah PR, editor. Multivariate analysis, New York: Academic Press; 1966, p. 391-420.

[36] Nomikos P, MacGregor JF. Monitoring of batch processes using multiway principal component analysis. AIChE J 1994;40:1361-75.

[37] Rännar S, MacGregor JF, Wold S. Adaptive batch monitoring using hierarchical PCA. Chemom Intell Lab Syst 1998;41:73-81.

[38] Zarzo M, Ferrer A. Batch process diagnosis: PLS with variable selection versus block-wise PCR. Chemom Intell Lab Syst 2004;73:15-27.

[39] Diana G, Tommasi C. Cross-validation methods in principal components: a comparison. Statist Methods Applicat 2002;11(1):71-82.

[40] Nelson PRC, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS: Score calculations with incomplete observations. Chemom Intell Lab Syst 1996;35:45-65.

[41] Arteaga F, Ferrer A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. J Chemometrics 2002;16:408-18.

[42] Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS). Umea, Sweden: Umetrics AB; 1999.

[43] Martí P, M. Gasque M. Improvement of temperature-based ANN models for solar radiation estimation through exogenous data assistance. Energy Convers Manage. doi:10.1016/j.enconman.2010.08.027

Figure captions

Fig. 1. R_s values averaged for the 30 stations. Gaps correspond to days with missing data for at least one station.

Fig. 2. *Left*: Map of the eastern coast of Spain (provinces of Alicante, Valencia and Castellón) indicating the location of the 30 weather stations (codes as in Table 1). *Right*: rotated score plot of PC1/PC2 obtained from the initial R_s matrix (30 stations by 1203 variables).

Fig. 3. Score plot of PC3/PC4 from the R_s>200 model (filled triangles) and R_s<200 (empty triangles). Station s19 was disregarded. Both plots were slightly rotated to achieve a better fit between scores corresponding to the same station

Fig. 4. R_s centered trajectories (i.e. difference with respect to the mean trajectory) averaged for stations with a similar performance (A: s1-s4-s6-s15, B: s2-s5-s9-s10, C: s26-s29-s30, D: s11-s12-s14).

Fig. 5. Error parameters showing the gap infilling performance of PCA for different gap sizes with an increasing number of PCs.

Fig. 6. Comparison between t_1 and t_2 scores from the R_s matrix with those obtained by applying eqs. 6 and 8 taking into account the geographical data of stations.

Table 1

Geographical parameters of the 30 weather stations. z : altitude (m) with respect to sea level; φ : latitude (degrees); τ : longitude (degrees).

Station	c[a]	z	φ	τ [b]		c[a]	z	φ	τ [b]
Benavites	1	8	39.7333	0.2150	Dénia-Gata	16	102	38.7939	-0.0836
Tavernes de Valldigna	2	15	39.0964	0.2367	Vila Joiosa	17	138	38.5294	0.2553
Catral	3	27	38.1544	0.8042	Pedralba	18	200	39.5678	0.7164
Sagunt	4	33	39.6492	0.2925	San Rafel del Riu	19	205	40.5956	-0.3703
Carcaixent	5	35	39.1167	0.5047	Altea	20	210	38.6056	0.0775
Vila Real	6	42	39.9333	0.1000	Monforte del Cid	21	244	38.3997	0.7289
Ondara	7	49	38.8197	-0.0075	Llíria	22	250	39.6919	0.6253
Moncada	8	58	39.5877	0.3992	Turís	23	299	39.4006	0.6836
Vilanova de Castelló	9	58	39.0667	0.5228	Cheste	24	323	39.5217	0.7417
Carlet	10	66	39.2264	0.5459	Agost	25	345	38.4278	0.6433
Almoradí	11	74	38.0908	0.7714	Villena	26	495	38.5967	0.8733
Pilar de la Horadada	12	77	37.8700	0.8103	Campo Arcís	27	584	39.4344	1.1608
Elx	13	86	38.2667	0.7000	El Pinós	28	606	38.4286	1.0594
Orihuela	14	99	38.1828	0.9536	Camp de Mirra	29	627	38.6803	0.7717
Vall d'Uixó	15	100	39.7975	0.2272	Castalla	30	708	38.6053	0.6728

[a]Station code, which was assigned according to an increasing altitude.

[b]Positive values: West; negative values: East, with respect to the Greenwich meridian.

Table 2

Summary overview of 3 PCA models: (i) R_s matrix with 1203 variables (‘all’), (ii) subset of 604 variables with an average R_s <200, and (iii) subset of 599 variables with an average R_s >200. These models were repeated after discarding station s19. Goodness-of-fit (R^2_X), eigenvalue (λ), goodness-of-fit by cross-validation (Q^2), and threshold value (Q^2_{limit}).

		PCA with 30 stations				PCA with 29 stations			
model	PC	R^2_X	λ	Q^2	Q^2_{limit}	R^2_X	λ	Q^2	Q^2_{limit}
All	1	0.366	11.0	0.300	0.034	0.376	10.9	0.312	0.035
All	2	0.129	3.86	0.097	0.035	0.138	4.01	0.120	0.036
All	3	0.104	3.12	0.111	0.036	0.100	2.90	0.140	0.038
All	4	0.061	1.83	0.041	0.038	0.057	1.67	0.070	0.039
All	5	0.045	1.34	0.011	0.039	0.043	1.25	-0.038	0.041
All	6	0.039	1.18	0.012	0.041	0.038	1.10	0.004	0.042
R_s<200	1	0.380	11.4	0.322	0.035	0.389	11.3	0.334	0.036
R_s<200	2	0.101	3.04	0.061	0.036	0.107	3.11	0.038	0.037
R_s<200	3	0.096	2.87	0.076	0.037	0.098	2.84	0.125	0.039
R_s<200	4	0.072	2.15	0.061	0.039	0.065	1.88	0.044	0.040
R_s<200	5	0.047	1.41	-0.020	0.040	0.049	1.42	-0.059	0.042
R_s<200	6	0.043	1.29	-0.007	0.042	0.045	1.30	0.028	0.043
R_s>200	1	0.363	10.9	0.281	0.035	0.375	10.9	0.294	0.036
R_s>200	2	0.156	4.69	0.153	0.036	0.165	4.79	0.192	0.037
R_s>200	3	0.107	3.22	0.130	0.037	0.105	3.04	0.156	0.039
R_s>200	4	0.057	1.70	-0.003	0.039	0.049	1.41	0.047	0.040
R_s>200	5	0.042	1.26	0.026	0.040	0.042	1.21	-0.021	0.042
R_s>200	6	0.038	1.14	-0.017	0.042	0.037	1.08	0.035	0.043

Table 3

PCA of the R_s matrix (29 stations, 1203 variables): goodness-of-fit (R^2_X) and goodness-of-fit by cross-validation (Q^2) considering different gap sizes.

	0% gaps		5% gaps		10% gaps		15% gaps	
PC	R^2_X	Q^2	R^2_X	Q^2	R^2_X	Q^2	R^2_X	Q^2
1	0.376	0.312	0.375	0.324	0.376	0.321	0.385	0.327
2	0.138	0.120	0.139	0.123	0.142	0.126	0.140	0.120
3	0.100	0.140	0.100	0.133	0.101	0.136	0.099	0.129
4	0.057	0.070	0.058	0.042	0.058	0.043	0.059	0.048
5	0.043	-0.038	0.044	0.005	0.044	-0.017	0.043	-0.013
6	0.038	0.004	0.037	0.004	0.037	0.011	0.036	0.006

Table 4

Error parameters (eqs. 3 to 5) as performance indicators for several infilling methods and gap sizes: PCA approach based on 4 components (A), 1-neighbor (B), 2-neighbor (C), 1-neighbor-SP (D), and 2-neighbor-SP (E).

gap size	method	MSE	MAE	AARE
5%	A	297.54	11.840	0.0814
10%	A	343.31	12.505	0.0931
15%	A	391.21	13.306	0.0949
5%	B	373.64	13.736	0.0857
10%	B	435.44	14.388	0.0959
15%	B	459.40	14.775	0.0944
5%	C	340.14	12.965	0.0834
10%	C	393.27	13.290	0.0910
15%	C	414.81	13.858	0.0912
5%	D	457.63	13.941	0.1070
10%	D	435.51	14.480	0.0941
15%	D	461.63	14.342	0.0994
5%	E	309.96	12.459	0.0861
10%	E	365.77	12.983	0.0969
15%	E	373.31	13.119	0.0925

Table 5

Error parameters (eqs. 3 to 5) as indicators of the goodness-of-fit for the reconstruction of the R_s matrix according to four different methods.

method	N	MSE	MAE	AARE
PCA [a]	1	422.76	14.283	0.1135
	2	336.16	12.425	0.1015
	3	266.71	11.170	0.0870
	4	225.72	10.295	0.0767
PCR [b]	1	427.03	14.409	0.1142
	2	359.78	13.104	0.1049
	3	311.93	12.252	0.0923
	4	286.04	11.784	0.0853
1-neighbor-SP	-	425.19	14.044	0.0990
2-neighbor-SP	-	309.45	11.853	0.0842

[a] R_s matrix (centered values) reconstructed according to eq. 1 with an increasing number of components (N).

[b] Same as the PCA method but t_i scores were obtained using eqs. 6 to 9.