García-Hinde, O., Terrén-Serrano, G., Hombrados-Herrera, M., Gómez-Verdejo, V., Jiménez-Fernández, S., Casanova-Mateo, C., Sanz-Justo, J., Martínez-Ramón, M. & Salcedo-Sanz, S. (2018). Evaluation of dimensionality reduction methods applied to numerical weather models for solar radiation forecasting. *Engineering Applications of Artificial Intelligence, 69*, 157–167.

DOI: 10.1016/j.engappai.2017.12.003

Evaluation of dimensionality reduction methods applied to numerical weather models for solar radiation forecasting

O. García-Hinde[a,*], G. Terrén-Serrano[b], M. Á. Hombrados-Herrera[b], V. Gómez-Verdejo[a], S. Jiménez-Fernández[e], C. Casanova-Mateo[c,d], J. Sanz-Justo[d], M. Martínez-Ramón[b], S. Salcedo-Sanz[e]

[a]*Department of Signal Theory and Communications, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain.*
[b]*Department of Electrical and Computing Engineering, The University of New Mexico, NM, USA.*
[c]*Department of Civil Engineering: Construction, Infrastructure and Transport, UPM, Madrid.*
[d]*LATUV, Universidad de Valladolid, Valladolid, Spain.*
[e]*Department of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain.*

Abstract

The interest in solar radiation prediction has increased greatly in recent times among the scientific community. In this context, Machine Learning techniques have shown their ability to learn accurate prediction models. The aim of this paper is to go one step further and automatically achieve interpretability during the learning process by performing dimensionality reduction on the input variables. To this end, three non standard multivariate feature selection approaches are applied, based on the adaptation of strong learning algorithms to the feature selection task, as well as a battery of classic dimensionality reduction models. The goal is to obtain robust sets of features that not only improve prediction accuracy but also provide more interpretable and consistent results. Real data from the Weather Research and Forecasting model, which produces a very large

[*] Corresponding author.
Email addresses: oghinde@tsc.uc3m.es (O. García-Hinde), guillermoterren@unm.edu (G. Terrén-Serrano), mangelhombradosherre@unm.edu (M. Á. Hombrados-Herrera), vanessa@tsc.uc3m.es (V. Gómez-Verdejo), silvia.jimenez@uah.es (S. Jiménez-Fernández), carlos.casanova@upm.es (C. Casanova-Mateo), julia@latuv.uva.es (J. Sanz-Justo), manel@unm.edu (M. Martínez-Ramón), sancho.salcedo@uah.es (S. Salcedo-Sanz)

January 2, 2018

number of variables, is used as the input. As is to be expected, the results prove that dimensionality reduction in general is a useful tool for improving performance, as well as easing the interpretability of the results. In fact, the proposed non standard methods offer important accuracy improvements and one of them provides with an intuitive and reduced selection of features and mesoscale nodes (around 10% of the initial variables centered on three specific nodes).

Keywords: Dimensionality Reduction; Interpretability; Solar Radiation Forecast; Weather Research and Forecasting Model; Support Vector Regression; Restricted Boltzmann Machine.

1. Introduction

According to the December 13^{th}, 2016 report of Solar Energy Industries Association (SEIA)[1], the total installed solar power capacity in the United States of America reached 35.8 GW in the third quarter of 2016, representing over 60% of the total installed electric capacity. This report states that there are more than 1 million residential solar installations across the country, and their industry growth nearly doubles every year. Regarding European countries, emphasis on solar power is decreasing. A report from Solar Power Europe[2] showed that a total of 1.56 GW in solar capacity were installed from June to September, which was 10% less than in the previous quarter. Nevertheless, during the first quarter of 2016, Europe reached 100GW of installed solar capacity. On the other hand, and according to SEIA, China and Japan lead the solar power market with 50% of new installed capacity. A fundamental tool for this active and growing market is, obviously, solar forecast [1].

Atmospheric behaviour makes solar power highly stochastic. Therefore, an efficient use of solar energy requires intelligent systems, specifically ones that are able to forecast the energy to be produced at different time-horizon scales, ranging from minutes to days. The amount of generated solar power primarily

[1] Source: www.seia.org

[2] Source: www.solarpowereurope.org

depends on cloud coverage, but also on other factors such as the presence of light absorbing particles in the air [2]. Many cloud coverage prediction methods rely on direct measurements, including ground or satellite observations of the clouds. The most widely used approach to address the task of solar radiation prediction consists in the physical modeling of the deterministic part of solar radiation, by computing the relative position of the sun with respect to the facility in order to obtain a clear sky model, and then adding the atmospheric conditions, including rain, wind speed and other variables [3]. However, some causes of radiation attenuation are not easily predicted by direct observation. For this reason, numerical methods to construct weather models, such as the Weather Research and Forecasting (WRF) mesoscale model [4], have also been applied, as they provide significant atmospheric information in the surroundings of the location under study, thus improving predictions [5].

In this work we present a study on feature selection and extraction methods for solar radiation forecast from a WRF model. The WRF model provides forecast of atmospheric variables at different heights for a given area. This model has the drawback of presenting a large number of dimensions (prediction variables). Indeed, the WRF model used in this paper produces around 10^4 prediction variables per time instant, while the number of samples available for the algorithm's training is much lower. This situation makes the use of powerful dimensionality reduction strategies mandatory. We propose and evaluate a series of novel methods that automatically select the most powerful features by adapting strong regression algorithms, such as SVMs or Deep Neural Networks (DNN), to the task of feature selection. We will compare the performance of these methods to other classic selection and extraction strategies and see their effect on interpretability. Our experiments show how our methods can maintain high prediction accuracies, while increasing interpretability by finding relationships and patterns within the data that are opaque to expert human knowledge.

2. Related work

There is a significant amount of work devoted to the prediction of solar radiation. Most approaches tackle the problem from a computational intelligence perspective. These strategies make use of different data sources. Some use meteorological data as inputs for *Machine Learning* (ML) predictors in order to improve their forecasting performance. Indeed, ML models have been used for example in [6], where a Radial Basis Function was used in solar radiation prediction in a power plant using weather data. A comparison of prediction techniques can be found in [7] where the authors take a time series prediction approach where the input data consists of historical solar radiation data. There, Multi-Layer Perceptron (MLP) neural networks, Markov chains, Bayesian inference and ARIMA models are compared. Support Vector Machine models (SVM), and specifically Support Vector Regressors (SVR) [8], have also been widely used in energy production forecast (see e.g. [9]). For example in [10] SVRs are used to predict monthly solar radiation from meteorological data, and the same authors use them in [11] to estimate solar radiation from air temperature. Extreme learning machines (ELM) have also been applied to solar prediction using meteorological variables in [12]; and in [13], a Kernel Extreme Learning Machine (KELM) has been compared to a kernel SVR. Other works introduce neuro-fuzzy approaches [14], or hybrid models combining ARMA and artificial neural networks [15], to cite some.

In other cases, researchers make use of observations of cloud evolution from satellite data. For example, in [16] SVRs and Multilayer Perceptrons are used to predict the evolution of the clouds seen from satellite images. The data is pre-processed to generate variables related to the size, motion and other factors of the clouds.

In [17], a numerical Weather Forecast System is combined with satellite infrared images to predict several hours ahead. In [18], a model that estimates cloud motion is used to predict the future positions of the clouds from satellite imagery. In [19], a MLP combined with a genetic algorithm is used to perform solar radiation forecast. The authors use satellite images to perform radiation prediction in large areas of Spain. Other works, such as [20, 21, 22, 23], also make use of satellite data in radiation prediction. Forecasting with shorter horizons can be implemented using ground images of the clouds. In [24] a cloud identification model is constructed from RGB images. Cloud monitoring is introduced in [25] for prediction. A short term solar radiance prediction scenario using observations of the whole sky was presented in [26]. Other related works are [27, 28, 29, 30, 31]. While the previous works use RGB images, in [32] a prediction model is developed using a LAPART neural network [33] and infrared images.

A portion of the strategies described above employ ML models for prediction using low or moderate dimensionality databases. However, many of the aforementioned scenarios use sources that produce data of very high dimensionality [34], i.e. satellite images or WRF mesoscale models. In these cases, it is of great importance to use dimensionality reduction methods to manage the structural complexity of the learning algorithm. To this end, there are two basic approaches: feature selection and feature extraction [35, 36, 37, 38]. Regarding specific application of dimensionality reduction methods to solar energy, in [39] a first general review of some works dealing with relevant parameters selection in solar energy prediction problems is offered. In [40] a system for solar irradiance very short-term prediction (minutes time-horizon) is proposed in which a correlation filter is applied to select relevant features. In [41] a study of the main influencing input parameters for solar radiation prediction with neural networks is carried out in different locations of India, by using a Decision Tree variable selection method. In [42] the problem of forecasting the electricity power generation by a solar photo-voltaic system is tackled, comparing multivariate and univariate correlation measurements to select useful features. In [43] an adaptive neuro-fuzzy inference system (ANFIS) has been applied to select the most influential variables in a daily horizontal diffuse solar radiation prediction problem. In [44] two applications of hybrid niching genetic algorithms are presented to solve the problem of variable selection for the estimation of Solar Radiation.

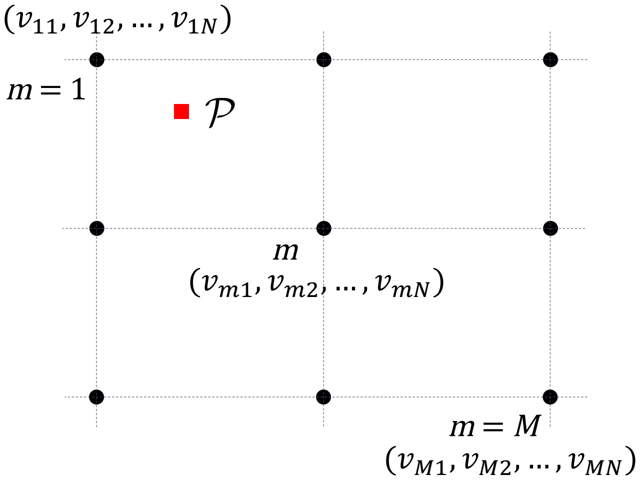

$(v_{11}, v_{12}, \dots, v_{1N})$
$m = 1$
$\mathcal{P}$
m
$(v_{m1}, v_{m2}, \dots, v_{mN})$
$m = M$
$(v_{M1}, v_{M2}, \dots, v_{MN})$

Figure 1: Outline showing a grid of M points and the variables considered at each grid point.

3. Solar radiance problem formulation

The solar radiation prediction problem formulation can be stated in the following way: let $\mathcal{R}_t$ be the global solar radiation registered at a given time t in a location $\mathcal{L}$ of the Earth's surface, and let $\hat{\mathcal{R}}_t$ be the prediction of the global solar radiation under the same considerations. In order to predict $\hat{\mathcal{R}}_t$, let us consider a set of N atmospheric variables $\mathcal{V}$ (outputs of a mesoscale model), some of them referred to different pressure levels (ranging from the pressure level corresponding to the ground level to 50 hPa). Considering each variable at each pressure level as a different predictor, the set of variables can be expressed as $\mathcal{V} = (v_{11}, \dots, v_{1N}, v_{21}, \dots, v_{2N}, \dots, v_{M1}, \dots, v_{MN})$, where M stands for the total number of grid points where the variables were obtained. Figure 1 shows an outline of a generic grid ($m = 1 \dots M$) and the set of variables ($n = 1 \dots N$) considered at each point.

3.1. Model $\mathcal{M}$: the Weather Research and Forecasting model

In this work, the Weather Research and Forecasting (WRF) mesoscale model [4] has been considered to obtain the set of atmospheric variables used as predictive variables. In this study, WRF model version 3.6 has been used and atmospheric and meteorological data have been calculated over a window ranging in latitude from 39° 30' 58"N to 40° 11' 57"N , and in longitude from 4° 42' 14"W to 4° 1' 5"W. In this window, the grid has 5 elements or nodes from West to East and 5 from North to South, summing up a total of 25 nodes, covering roughly, 15×30 km^2 each. Atmospheric values are calculated, in the vertical dimension, at 37 levels above the ground, at ground-level, and at four additional levels beneath the surface. The grid type is Arakawa, that is to say that the data is calculated at the center of each element, with a 72 second time step.

The WRF model used in this study provides 423 variables at each of the 25 nodes, including wind speed components, temperature values, upper atmosphere outgoing long wave radiation, cloud coverage per cell, etc. This results in 25×423 variables to determine the predicted global solar radiation $\hat{\mathcal{R}}_t$ at each time instant. As objective variable data to train and test the algorithms, the global solar radiation measured at Toledo's radiometric station (Spain), located at (39° 53' 5"N, 4° 02' 43"W) and at an altitude of 515 m, is considered. One complete year of hourly data (from May 1st, 2013 to April 30th, 2014) was used. Figure 2 shows the geographical locations of the 25 mesoscale nodes as well as the Toledo measuring station.

4. Methods for Feature Selection

In this section, the three non standard feature selection methods included in this study are described. Two of them are based on a bootstrapping technique [45] to approximate the sample distribution of the weights of two ML models: a linear SVR and a Restricted Boltzmann Machine (RBM). In the third method we implement a selection method based on the reconstruction error during the

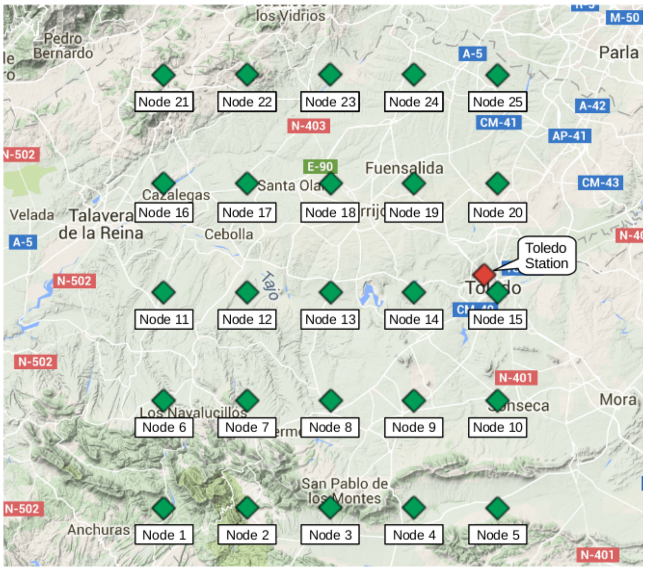


Figure 2: Map showing the geographical locations of the 25 mesoscale nodes as well as the Toledo's measuring station.

unsupervised RBM training phase. All three methods then feed the selected set of features to an SVR for prediction.

Before describing the proposed feature selection methods, let us briefly establish the concept of the feature selection problem and the notation used throughout this work. In its more general form, feature selection for a machine learning scenario can be defined as follows: given a set of labeled data samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ (or $y_i \in \{\pm 1\}$ in the case of classification problems), choose a subset of m features ($m < n$), that achieves the lowest error in the prediction of the variable y_i.

4.1. Feature selection through Bootstrapped SVRs

The first of the non standard methods employed in the experiments is the Bootstrapped SVM feature selection method (BSFS) presented in [46]. This algorithm trains a large number of SVMs, each of them using different subsets sampled at random and with replacement from the training data. Thus, by training K SVRs we will obtain a set of K corresponding weight vectors,

$\langle \mathbf{w}^{(k)} \rangle_{k=1}^{K}$. This is akin to the idea of using an ensemble paradigm for the purpose of feature selection [47, 48].

One of the key ingredients of this algorithm is the use of a linear SVM with an l_2 regularisation term. This exploits two basic theoretical properties of this regression scenario:

- Features that are *irrelevant* will present a weight distribution centered in zero.
- Isolated features that are *relevant* will present non-zero mean weight distributions.

Determining which are the irrelevant features is performed by applying a one sample location Student's t-test [49] where the null hypothesis is that a feature-weight's sample distribution presents a mean of zero. After applying the test to each component of the feature-weight space, the p-values are used to determine the probability that the weight distribution for each component has a mean of zero (a lower p-value rejects the null hypothesis and indicates a non-zero mean). Algorithm 1 shows the pseudo-code for the BS-FS method.

Algorithm 1: The BS-FS algorithm.

```
Data:
$\mathbf{X}_{tr}, \mathbf{y}_{tr}$
Input:
$K$ = number of bootstrap iterations
$S$ = number of samples per subset
for $k \in K$ do
    $\mathbf{X}_{tr}^{(k)}, \mathbf{y}_{tr}^{(k)}$ = random sample of size $S$ from $\mathbf{X}_{tr}$ & $\mathbf{y}_{tr}$
    Train SVR using $\mathbf{X}_{tr}^{(k)} \rightarrow$ obtain $\mathbf{w}^{(k)}$
    $\mathbf{W}(k,:) := \mathbf{w}^{(k)}$
for $d \in D$ do
    Perform Student's t-test for $\mathbf{W}(:,d)$ with null hypothesis: $\bar{w}_d = 0$
    Store p-value for feature $d$
Rank features according to their corresponding p-values.
```

Two different approaches can be used to determine the optimal feature set. One option is to establish a confidence level threshold and reject all feature-weights that present a p-value that surpasses it. This results in a very fast implementation, but has the disadvantage of an arbitrarily set threshold that can lead to weak selections. A better approach is to cross-validate the optimal set of features. Although this method is slower, it will yield much more significant feature sets.

The computational cost of this algorithm during training depends on the method used to compute the weights of the SVM. The most widespread method, and indeed the one used in this work, is the Sequential Minimal Optimization algorithm (SMO) [50]. In the worst case, the cost of SMO is $\mathcal{O}(N^3)$, while on average it is of $\mathcal{O}(N^2)$. Performing the Student's t-test has linear cost. Therefore, the worst case scenario cost of the BS-FS algorithm can be approximated by:

$$\mathcal{O}(KS^3 + DN) \tag{1}$$

where K is the number of bootstrap iterations, S is the subsample size, D is the number of input dimensions and N is the total number of training samples. Note that S will always be smaller than N. Indeed, both S and K can be tuned to control computational complexity, while reasonably maintaining the model's functionality. Therefore, this model compares very favorably to other approaches like Recursive Feature Elimination, where the complexity is a direct function of the number of training data, N, times the dimensionality, D.

4.2. Feature selection with RBMs

Another strategy considered in this work is to use a restricted Boltzmann Machine (RBM) or a stack of them (deep Boltzmann Machine or DBM) as a feature selector whose output is fed to an SVR. A DBM can be used as a pre-processing unit in which each layer of the machine produces a higher level of abstraction. An RBM can be viewed as a type of Markov Random Field [51] with hidden and visible nodes where the relationships between them are restricted to connections between a hidden node and all visible nodes and vice-versa. This structure allows to approximate a joint probability distribution model for the hidden and visible nodes [52, 53] that can be factorized [54]). In this application, we treat the visible nodes as the data input and the hidden nodes are fed into a SVR. The training of the RBM is fully unsupervised and based on the Contrastive Divergence strategy introduced in [52].

The RBM and deep or multilayer versions of the RBM have been used for feature selection in [55]. The main idea is that during the reconstruction phase, this is, when visible states are synthesized from hidden states, those features that contain information will be better reconstructed, but those that contain only noise will have a significant reconstruction error. Therefore, once the training phase has finished, the expectation of the reconstruction error of each input feature can be computed, i.e.

$$\mathbb{E}(e_j) = \frac{1}{N}\sum_n x_{j,n} \tag{2}$$

where $x_{j,n}$ represents feature j of data vector $\mathbf{x}_n$. All features are simply ranked using this measure. We call this approximation RE-RBM.

Another alternative for feature selection, and the third model proposed in this work, is to use the BS-FS Algorithm 1, but where the SVR is substituted with the hybrid RBM SVR structure. The RBM weights are selected instead of the SVM weights. The features are ranked according to the p-values of their associated weights. This model is abbreviated as the BS-RBM.

The complexity of the RBM section is dominated by matrix multiplication. The RE-RBM uses a Contrastive Divergence forward step, and a backpropagation step. If the total number of parameters is P, the total number of samples is N and the number if iterations until convergence is L, then computational burden of the training is approximately [56]:

$$\mathcal{O}(LNP + N^3) \tag{3}$$

where the N^3 term is due to tthe SVM training with the whole training dataset. The BS-RBM has a training computational burden that depends on the matrix parameter manipulation of the RBM and SVM trainings plus the Student's test, i. e, its computational burden is:

$$\mathcal{O}(LKSP + N^3 + DN) \tag{4}$$

5. Experiments and results

This work evaluates the ability of the methods presented in Section 4, along with a battery of classical dimensionality reduction models, to provide a good selection of variables in terms of both predictive performance and interpretability. The experimental results will now be presented and analyzed.

5.1. Experimental setup

The first step in the experimental process of this study was to perform a preliminary processing of the data described in Section 3, revealing that many features have a standard deviation smaller than 0.01. These predictive variables are deemed to be irrelevant and were eliminated from the study. As a consequence, 301 variables per mesoscale point remained, resulting in a dimensionality reduction of close to 40%, and an input matrix of dimension 25×301. Table 1 shows the remaining variables considered for each grid point. Note that a variable (i.e. CLDFRA) obtained at two different pressure levels is analyzed in this work as two separate predictive variables.

Table 1: Outputs of the WRF model used in the experiments as predictive variables (301 variables per WRF model node).

Variable	Description
OLR	Top of the atmosphere outgoing long-wave radiation
GLW	Downward long-wave flux at ground level
SWDOWN	Downward short-wave flux at ground level
TSK	Surface skin temperature at ground level
v10	y-wind component at 10 meters above the ground
u10	x-wind component at 10 meters above the ground
PSFC	Surface atmospheric pressure
TH2	Potential temperature at 2 meters above the ground
u	x-wind component at the 37 vertical levels of the WRF model
v	y-wind component at the 37 vertical levels of the WRF model
w	z-wind component at the 37 vertical levels of the WRF model
CLDFRA	Fraction of clouds in each cell at 30 vertical levels of the WRF model
QVAPOR	Water vapor mixing ratio at the 37 vertical levels of the WRF model
T'	Perturbation potential temperature at the 37 vertical levels of the WRF model
P	Perturbation pressure at the 37 vertical levels of the WRF model
P_HYD	Hydrostatic pressure at the 37 vertical levels of the WRF model
SMCREL	Relative soil moisture at 4 depth levels

From the aforementioned model, a data set consisting of 5840 samples with 7525 input features is obtained (resulting from the 25 × 301 node-variable matrix).

For the experiments, the dataset is randomly split into ten partitions that are used to calculate independent performance measurements. The measurements for the ten partitions are then averaged to obtain a single stable performance measurement for each of the methods under study.

The feature selection methods described in Section 4 are tested alongside the following well known dimensionality reduction strategies, ranging from classic statistical analysis tools to modern ML models:

- A principal component analysis (PCA) feature extraction algorithm [57]. This is a non-supervised, multivariate model.

- A partial least squares approximation (PLS) feature extraction algorithm [58]. This is a supervised, multivariate model.

- A filter that ranks features in terms of their variance (variance filter). This is a non-supervised, univariate method. Features with higher variance are considered to be more relevant.

- A filter that ranks features in terms of their individual correlation with the target vector (correlation filter). This is a supervised, univariate method. Features with higher correlation are considered to be more relevant.

- A Lasso regression model [59]. This is a supervised, multivariate method that pairs a least mean square error regressor with an L_1 regularisation term that provides a sparse selection of features.

- An Recursive Feature Elimination (RFE) algorithm [60]. This is a supervised, multivariate algorithm.

A baseline method without any feature selection or extraction is also implemented to give a reference performance value. PCA, PLS, the variance and correlation filters, the BS-FS algorithm and the RFE algorithm were all paired with an SVR trained with the resulting feature sets. The Lasso algorithm in turn uses its own embedded least mean square error regressor. The algorithms labeled as BS-RBM and RE-RBM make use of a Restricted Boltzmann Machine with sigmoidal activations and 500 output nodes. The output of these models is again fed into a linear ν-SVR.

5.2. Parameter validation

All parameters across the board were validated using a 10-fold cross-validation strategy applied to each of the ten test experimental partitions. This includes the determination of the optimal number of dimensions for each model and test partition.

After validation, the optimal values for the parameters of the baseline SVR, namely the regularization coefficient C and the epsilon-tube coefficient ϵ, were found to be: $C = 1.0$, after sweeping a range of 7 values logarithmically spaced between 10^{-3} and 10^{3}; and $\epsilon = 0.2$, swept linearly within the range of 0.1 and 1.0 with increments of 0.1.

Having realized that, for this database, there is a very flat region around these values where the SVR performs well, it was decided that the same parameters would be used for the final SVR regression phase of all the dimensionality reduction models. Similarly, these same values are used for the internal SVR parameters of the BS-FS and RFE algorithms. In the case of the Lasso model, the regularization term is what dictates the sparsity of the solution, therefore, it's C parameter is what determines the final number of selected features and needs to be independently validated. Thus, the L_1 regularization coefficient for the Lasso was validated using the same 10-fold cross-validation strategy, sweeping 50 equally spaced values in the range between 10^{-5} and 10^{-2}.

In the case of the RBM methods, cross-validation included the discovery of the optimal number of output nodes [61]. The best validation error was achieved with one layer, after sweeping values between 1 & 3. In order to control the computational burden, only 50, 100, 100, 500 and 1000 nodes were swept during the validation stage.

5.3. Performance measurements

To evaluate the performance of these algorithms, the R^2 metric is used, which is defined as:

$$R^2 = 1 - \frac{\sum_{l \in C_{tst}} (y_l - \hat{y}_l)^2}{\sum_{l \in C_{tst}} (y_l - \bar{y})^2}, \tag{5}$$

where y_l is the true target value for the l^{th} test sample, $\hat{y}_l$ is the predicted target value for the l^{th} test sample, $\bar{y}$ is the true mean of the test target values and C_{tst} is the set of test samples.

Table 2 shows the validated performances for the feature extraction and selection methods, both in terms of the R^2 score and the number of features or components. Note that in the case of the two extraction models, the reduced dimensionality results not from a selection of a subset of original features, but from their projection onto the space of the optimal components selected by cross validation. This is an important distinction in terms of interpretability, as will be discussed in the following subsection.

The results indicate that, as a rule, dimensionality reduction, be it by means of extraction or selection, leads to an improvement not only in terms of complexity, as there is a considerable dimensionality reduction in all cases, but also, albeit slightly, in terms of performance. All the multivariant methods (the PCA and PLS feature extraction models as well as the BS-FS, BS-RBM, RE-RBM, BS-RBM, Lasso and RFE feature selection models) outperform the univariant algorithms (the correlation and variance filters).

Table 2: Feature selection & extraction performance analysis. All values are obtained by crossvalidation of the working point. All results are averaged over the 10 test partitions and their standard deviations are included.

	$\mathbf{R^2}$	**Features**	**Components**	$\mathcal{O} \sim$
Baseline	0.922 ± 0.006	All	-	N^3
PCA	0.931 ± 0.003	-	217.0 ± 42.2	$D^2N + D^3 + N^3$
PLS	$\mathbf{0.933 \pm 0.004}$	-	$\mathbf{12.7 \pm 1.0}$	$D^2N + D^3 + N^3$
Correlation	0.928 ± 0.004	1030 ± 603.41	-	$D^2N + N^3$
Variance	0.923 ± 0.004	270 ± 161.55	-	$DN + N^3$
Lasso	$\mathbf{0.933 \pm 0.004}$	$\mathbf{53 \pm 2.25}$	-	N^3
RFE	0.926 ± 0.005	1355 ± 140.46	-	$(D + 1)N^3$
BS-FS	0.931 ± 0.004	450 ± 143.37	-	$KS^3 + DN + N^3$
RE-RBM	0.928 ± 0.005	3140 ± 337.31	-	$LNP + DN + N^3$
BS-RBM	0.930 ± 0.004	3327 ± 376.00	-	$LKSP + DN + N^3$

5.4. Interpretability analysis

Figures 4 to 12 have been designed to summarize the consistency with which each dimensionality reduction method has deemed each feature to be important. Figure 3 explains the construction process: the selection maps (or weight maps in the case of the PLS and PCA models) resulting from the ten experiments are averaged and normalized to give a picture of how consistent each method is when selecting variables. This is shown in a 25×301 matrix representing the 25 mesoscale WRF nodes and their corresponding 301 variables. We have called this matrix the *aggregated map*. Dark red colors indicate that the variable was consistently considered important over the ten experiments, whereas deep blue colors indicate that a variable was consistently considered irrelevant. Lighter blues, greens and yellows indicate inconsistency in the selection, and therefore, a lack of informativeness for a particular variable.

The *high relevance map* applies a hard threshold of 0.75 to the aggregated map. This means that any variable with a high relevance and consistency will appear in black, giving us a visual clue of which variables are consistently more important for a particular method. This map also includes two information bars: the bar on the right hand side represents the row-average of the high relevance map, therefore giving us a visual indication of what mesoscale nodes are contributing more towards the final prediction process; likewise, the bar on the bottom represents the column average of the high relevance map, giving us an idea of what variables are more contributive over all the nodes.

From these maps we can establish how consistently selective and visually informative each method is in terms of both the node space and the variable space. This information, coupled with the results from Table 2, can be used to evaluate each method in terms of effectiveness and interpretability.

The worst performers seem to be the RFE algorithm (Figure 4) and the variance filter (Figure 5). Their selections are all over the place in both the mesoscale node and variable spaces. They both appear to be extremely noisy and inconsistent, giving us almost no information as to what variables or nodes may be the most helpful for the prediction of solar energy. RFE fairs only slightly better than the variance filter in the mesoscale node space, yet it still fails to provide useful information in the variable space. These two methods are also the worst performers in terms of R^2, so it is obvious that they are inadequate in this particular scenario. It is also worthy of note that the RFE algorithm has a very high computational cost. The correlation filter is also a weak performer in terms of R^2, although it seems to be very consistent in it's selection over the variable space.

The BS-RBM model (Figure 6) also provides a very noisy selection of variables, although with much greater consistency. Interestingly, this model seems to very consistently ignore a specific band of variables related to cloud cover, which is a counterintuitive result. In terms of R^2 it fares better than the RFE model and the variance filter. The RE-RBM algorithm (Figure 7) is able to locate a number of variables related to pressure and cloud cover which are significant for the prediction in a consistent way. It is not particularly strong in terms of R^2 performance, but it still manages to produce visual results that are somewhat informative.

The correlation filter (Figure 8) seems to be quite selective in the variable space, specifically with variables related to wind speed and temperature, whilst completely loosing focus on the node space. It also offers only average performance in terms of R^2.

The first component of the PLS algorithm (Figure 9) seems to place most of the importance in a few of the variables related to long-wave radiation. At the same time it makes little distinction over the mesoscale nodes [3]. It is important to note that this method is among the most powerful in terms of both R^2 and dimensionality reduction, although it seems to be very sparse in its informativeness. The first component of PCA (Figure 10) has very strong preference for the variables related to pressure. Considering its good performance in terms of R^2 and the variables it gives preference to, it seems to be a useful model. It is important to remember that both PLS and PCA are feature extraction methods that do not select specific variables, but rather define linear combinations of all the variables. This implies that the results are not as directly interpretable.

Both the Lasso and BS-FS approaches have a good performance in terms of R^2, and they also seem to provide a narrow and consistent set of variables and nodes, specially in the case of the BS-FS. This can be attested by the fact that the aggregated map shows little inconsistencies, with high percentages of strongly rejected variables with a few confidently selected ones. Lasso tends to favor cloud cover related variables at different levels over most of the nodes, although it's selection is more scattered. BS-FS seems to favor the downward shortwave flux over most nodes as well as atmospheric pressure variables for nodes 1, 15 and 20. It is interesting to note that the Toledo measuring station is very close to nodes 15 and 20, whereas node 1 is very far away. Given their good performance and strong selectiveness, these seem to be among the most useful and informative models in this particular scenario. It is also important to consider that they both present good performance in terms of computational

[3] The first component in the PLS model is the most important in terms of the covariance of the input with the target values. Other components (not shown) behave in a similar fashion.

load, considering the BS-FS has good performance even with small values for S and K (see Section 4).

We can now perform an analysis of the variables selected by the best performers for this problem. One interesting fact is that the more informative algorithms (Lasso, BS-FS, PCA and RE-RBM) tend to give importance to similar bands of variables, although they vary on the nodes they prefer. It is also interesting that these models tend to select large groups of leveled variables such as pressure, temperature and fraction of cloud cover (see Table 1).

Lasso and RE-RBM lean towards the selection of the cloud fraction $CLDFRA$ at different altitude levels. This seems like a good indicator, since the more clouds at different levels of the atmosphere, the less solar radiation will reach the ground. Pressure and temperature are key variables in the cloud formation, so it seems that PCA, BS-FS, and RE-RBM all focus on the selection of variables related to cloud and cloud-formation in the area under study, which seems quite intuitive.

Note that the algorithms have automatically selected these variables based only on a statistical processing of the information, and not on the physical meaning of the variables, as a human guided by expert knowledge would do. Nevertheless, the physics underneath the phenomenon arises in the analysis of the best solutions found by the algorithms.

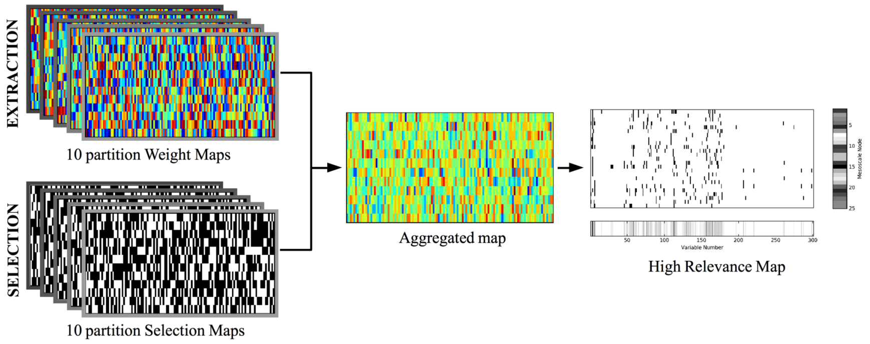
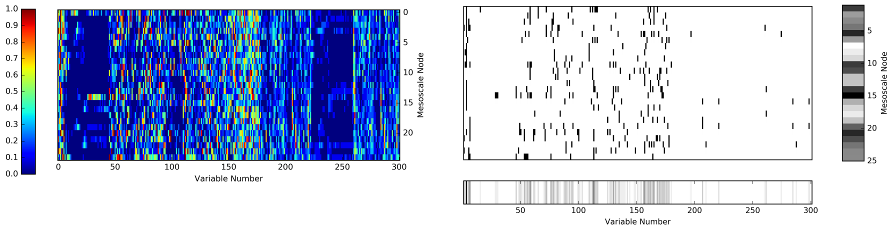


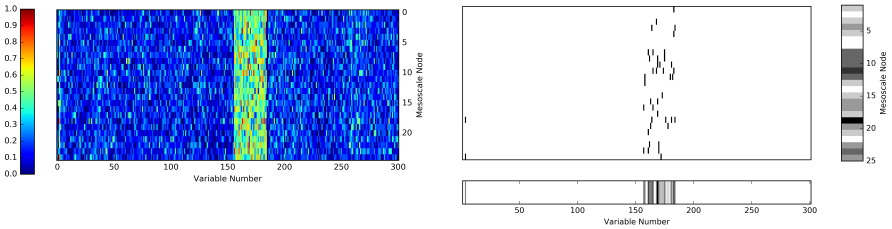
Figure 3: Visual explanation for the construction of the descriptive figures. A binary selection map (or weight map in the case of the PCA and PLS models) is generated for each of the ten data partitions. These maps are averaged and normalized to 1 to generate the Aggregated Map. The Aggregated Map is in turn thresholded at 0.75 to generate the High Relevance Map. The node and variable information bars to the right and to the bottom of the High Relevance Map are generated by computing its row and column averages respectively.



(a) Aggregated map.

(b) High relevance map.

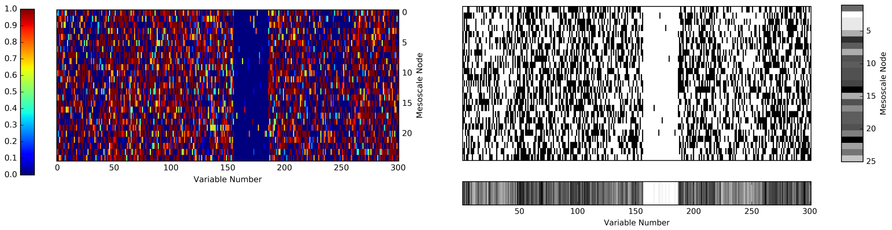
Figure 4: RFE model relevance maps.



(a) Aggregated map.

(b) High relevance map.

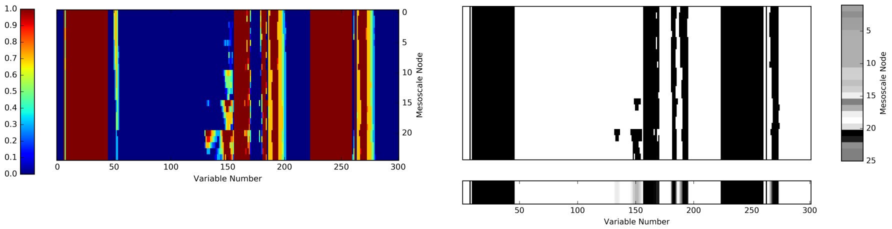
Figure 5: Variance Filter relevance maps.



(a) Aggregated map.

(b) High relevance map.

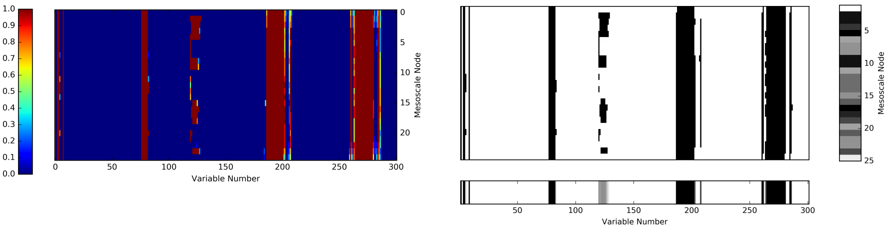
Figure 6: BS-RBM relevance maps.



(a) Aggregated map.

(b) High relevance map.

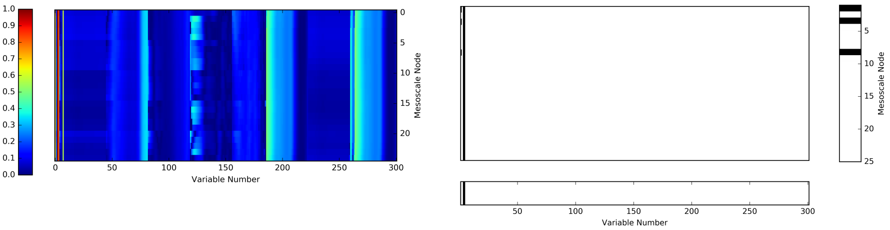
Figure 7: RE-RBM relevance maps.



(a) Aggregated map.

(b) High relevance map.

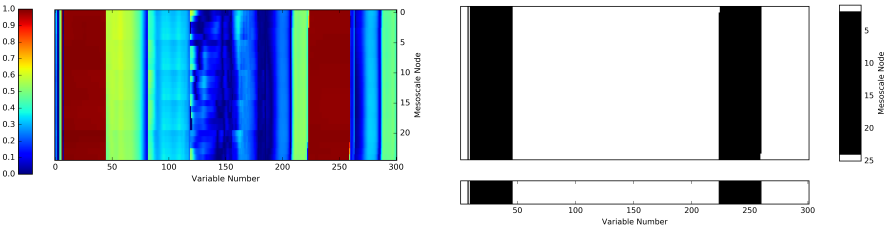
Figure 8: Correlation Filter relevance maps.



(a) Aggregated map.

(b) High relevance map.

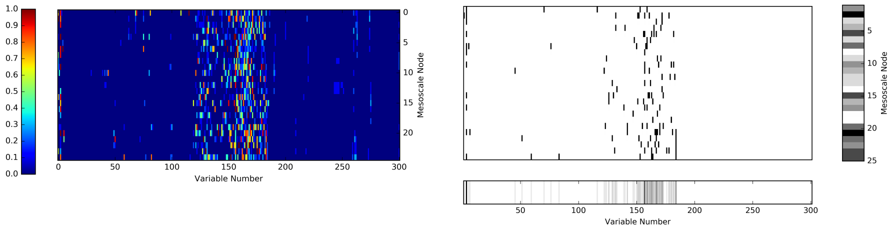
Figure 9: PLS relevance maps.



(a) Aggregated map.

(b) High relevance map.

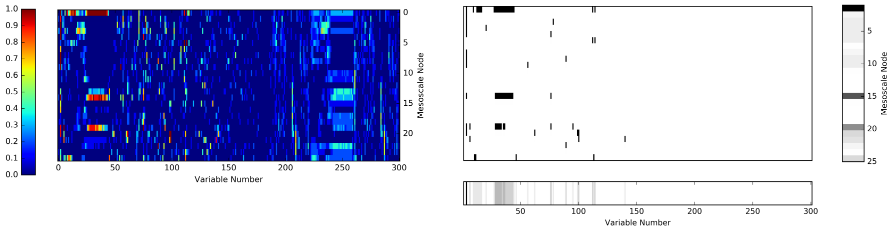
Figure 10: PCA relevance maps.



(a) Aggregated map.

(b) High relevance map.

Figure 11: Lasso relevance maps.



(a) Aggregated map.

(b) High relevance map.

Figure 12: BS-FS model relevance maps.

6. Conclusions

In this work we have presented a detailed study of the effectiveness of several automatic dimensionality reduction algorithms when applied to the context of solar radiation prediction. An important aspect of automatic dimensionality reduction is its ability to improve our understanding of the input-variable space by pointing out those sets of variables that increase the power of the prediction task, thus increasing the problem's interpretability. Therefore, in this work we have not only analyzed the increase in accuracy provided by dimensionality reduction algorithms, but also the resulting variable subsets and their relationship with the radiation problem.

To achieve this goal, three innovative dimensionality reduction algorithms, the BS-FS, RE-RBM and BS-RBM models described in Section 4, have been tested alongside a battery of classic dimensionality reduction strategies, including feature selection as well as feature extraction algorithms: PCA and PLS models, a Lasso model, an RFE algorithm and a Variance and Correlation filter. These models have been applied to a Weather Research and Forecasting model consisting of 25 nodes, each of them containing a set of 423 variables related to weather conditions, along one year (from May 1st, 2013 to April 30th, 2014) of hourly sampled data. This results in a high dimensionality problem that is well suited to the application of feature selection and extraction strategies.

We have found that dimensionality reduction, be it through feature selection or extraction, proves to be a useful tool in this context since all the applied algorithms have produced an increase in predictive accuracy (see Section 5). Specifically, the Lasso feature selection algorithm and the PLS extraction method achieve the highest increases in performance.

However, if we pay attention to the gain in interpretability, approaches such as PCA, the BS-FS algorithm have lead to feature selections of great relevance, leading to the automatic discovery of strong and intuitive selections of input variables. This shows that automatic variable selection is an effective way of improving a problem's interpretability, as it can lead to feature sets that agree with an intuitive understanding of the problem.

If we focus on the non standard methods presented in this work, the BS-FS offers the most intuitive and interpretable results both in the node and variable domains (selecting approximately 10% of the variables centered around three nodes), as well as achieving high performance levels in terms of R^2. Of the RBM based methods, the RE-RBM model achieves somewhat interpretable results, yet its performance is unremarkable; the BS-RBM model on the other hand offers slightly better performance, but it's selection capabilities are noisy and uninformative.

Future work proposals include the use of mesoscale models at a larger scale in order to produce solar forecast in the range of 24 hours ahead. To this end, we will apply Multi-Input Multi-Output (MIMO) structures, that allow a block prediction of all set points of the forecast. Also of interest is the estimation of confidence intervals of the forecast solar radiation values, that can be computed using MIMO versions of Gaussian Process regressors, in combination with the above introduced feature selection methods.

Acknowledgments

This work has been partially supported by the projects TIN2014-54583-C2-2-R, TEC2014-52289-R and TEC2016-81900-REDT of the Spanish Ministerial Commission of Science and Technology (MICYT), and by Comunidad Autónoma de Madrid, under project PRICAM P2013ICE-2933.

References

[1] C. Voyant, N. G., S. Kalogirou, M. L. Nivet, C. Paoli, F. Motte, and A. Fouilloy, “Machine learning methods for solar radiation forecasting: A review,” *Renewable Energy*, vol. 105, pp. 569 – 582, 2017.

[2] T. Khatib, A. Mohamed, and K. Sopian, “A review of solar energy modeling techniques,” *Renewable and Sustainable Energy Reviews*, vol. 16, no. 5, pp. 2864–2869, 2012.

[3] M. Bilgili and M. Ozoren, “Daily total global solar radiation modeling from several meteorological data,” *Meteorology and Atmospheric Physics*, vol. 112, no. 3-4, pp. 125–138, 2011.

[4] W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, and J. Powers, “A description of the advanced research wrf version 2,” tech. rep., National Center for Atmospheric Reserach, Mesoscale and Microscale Meteorology Division, 2005. Technical Note.

[5] B. V. Dasarathy, “Information fusion as a tool for forecasting/prediction–an overview,” *Information Fusion*, vol. 12, no. 2, pp. 71–73, 2011.

[6] M. Benghanem and A. Mellit, “Radial basis function network-based prediction of global solar radiation data: Application for sizing of a stand-alone

photovoltaic system at al-madinah, saudi arabia," *Energy*, vol. 35, no. 9, pp. 3751–3762, 2010.

[7] C. Paoli, C. Voyant, M. Muselli, and M. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, no. 12, pp. 2146 – 2160, 2010.

[8] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 1–32, 1998.

[9] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, "Support vector machines in engineering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 3, pp. 234–267, 2014.

[10] J. Chen, H. Liu, W. Wu, and D. Xie, "Estimation of monthly solar radiation from measured temperatures using support vector machines – a case study," *Renewable Energy*, vol. 36, no. 1, pp. 413 – 420, 2011.

[11] J. Chen, G. Li, B. Xiao, Z. Wen, M. Lv, C. Chen, Y. Jiang, X. Wang, , and S. Wu, "Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature," *Energy Conversion and Management*, vol. 89, pp. 318 – 329, 2015.

[12] S. Salcedo-Sanz, C. Casanova-Mateo, A. Pastor-Sánchez, and M. Sánchez-Girón, "Daily global solar radiation prediction based on a hybrid coral reefs optimization – extreme learning machine approach," *Solar Energy*, vol. 105, pp. 91 – 98, 2014.

[13] S. Shamshirband, K. Mohammadi, H. Chen, G. Samy, D. Petković, and C. Ma, "Daily global solar radiation prediction from air temperatures using kernel extreme learning machine: A case study for iran," *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 134, pp. 109 – 117, 2015.

[14] L. Olatomiwa, S. Mekhilef, S. Shamshirband, and D. Petković, “Adaptive neuro-fuzzy approach for solar radiation prediction in nigeria,” *Renewable and Sustainable Energy Reviews*, vol. 51, pp. 1784 – 1791, 2015.

[15] C. Voyant, M. Muselli, C. Paoli, and M. Nivet, “Hybrid methodology for hourly global radiation forecasting in mediterranean area,” *Renewable Energy*, vol. 53, pp. 1 – 11, 2013.

[16] H. S. Jang, K. Y. Bae, H. S. Park, and D. K. Sung, “Solar power prediction based on satellite images and support vector machine,” *IEEE Transactions on Sustainable Energy*, 2016.

[17] T. Kato, Y. Manabe, T. Funabashi, K. Yoshiura, M. Kurimoto, and Y. Suzuoki, “A study on several hours ahead forecasting of spatial average irradiance using nwp model and satellite infrared image,” in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, 2016.

[18] S. Cros, O. Liandrat, N. Sébastien, and N. Schmutz, “Extracting cloud motion vectors from satellite images for solar power forecasting,” in *2014 IEEE Geoscience and Remote Sensing Symposium*, 2014.

[19] A. Linares-Rodriguez, J. Ruiz-Arias, D. Pozo-Vazquez, and J. Tovar Pescador, “An artificial neural network ensemble model for estimating global solar radiation from meteosat satellite images,” *Energy*, vol. 61, pp. 636 – 645, 2013.

[20] R. Perez, S. Kivalov, J. Schlemmer, K. H. Jr., D. Renné, and T. E. Hoff, “Validation of short and medium term operational solar radiation forecasts in the us,” *Solar Energy*, 2010.

[21] C. Schillings, H. Mannstein, and R. Meyer, “Operational method for deriving high resolution direct normal irradiance from satellite data,” *Solar Energy*, 2004.

[22] A. Hammer, D. Heinemann, E. Lorenz, and B. Lückehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, 1999.

[23] R. Ineichen, P.and Perez, "Derivation of cloud index from geostationary satellites and application to the production of solar irradiance and daylight illuminance data," *Theoretical and Applied Climatology*, 1999.

[24] H. Li, F. Wang, H. Ren, H. Sun, C. Liu, B. Wang, J. Lu, Z. Zhen, and X. Liu, "Cloud identification model for sky images based on otsu," in *International Conference on Renewable Power Generation (RPG 2015)*, 2015.

[25] R. Tapakis and A. G. Charalambides, *Monitoring Cloud Coverage in Cyprus for Solar Irradiance Prediction.* 2013.

[26] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, and B. Washom, "Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed," *Solar Energy*, 2011.

[27] H. Huang, J. Xu, Z. Peng, S. Yoo, D. Yu, D. Huang, and H. Qin, "Cloud motion estimation for short term solar irradiation prediction," in *2013 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct 2013.

[28] M. Cervantes, H. Krishnaswami, W. Richardson, and R. Vega, "Utilization of low cost, sky-imaging technology for irradiance forecasting of distributed solar generation," in *2016 IEEE Green Technologies Conference (GreenTech)*, April 2016.

[29] R. Marquez and C. F. Coimbra, "Intra-hour dni forecasting based on cloud tracking image analysis," *Solar Energy*, 2013.

[30] H. Y. Cheng and C. C. Yu, "Solar irradiance now-casting with ramp-down event prediction via enhanced cloud detection and tracking," in *2016 IEEE International Conference on Multimedia and Expo (ICME)*, July 2016.

[31] S. Sun, J. Ernst, A. Sapkota, E. Ritzhaupt-Kleissl, J. Wiles, J. Bamberger, and T. Chen, "Short term cloud coverage prediction using ground based all sky imager," in *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Nov 2014.

[32] A. Mammoli, A. Ellis, A. Menicucci, S. Willard, T. Caudell, and J. Simmins, "Low-cost solar micro-forecasts for pv smoothing," in *2013 1st IEEE Conference on Technologies for Sustainability (SusTech)*, Aug 2013.

[33] T. P. Caudell and M. J. Healy, "Studies of inference rule creation using lapart," in *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems*, vol. 3, pp. ICNN1–ICNN6, 1996.

[34] V. Bolon-Canedo, N. S.-M. no, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowledge-Based Systems*, vol. 86, pp. 33 – 45, 2015.

[35] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[36] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*, vol. 207. Springer, 2008.

[37] I. K. Fodor, "A survey of dimension reduction techniques," *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, vol. 9, pp. 1–18, 2002.

[38] N. Martin and H. Maes, *Multivariate analysis.* Academic press, 1979.

[39] A. K. Yadav and S. Chandel, "Solar radiation prediction using artificial neural network techniques: A review," *Renewable and Sustainable Energy Reviews*, vol. 33, pp. 772–781, 2014.

[40] C.-L. Fu and H.-Y. Cheng, "Predicting solar irradiance with all-sky image features via regression," *Solar Energy*, vol. 97, pp. 537–550, 2013.

[41] A. K. Yadav, H. Malik, and S. Chandel, "Selection of most relevant input parameters using weka for artificial neural network based solar radiation prediction models," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 509–519, 2014.

[42] M. Rana, I. Koprinska, and V. G. Agelidis, "Univariate and multivariate methods for very short-term solar photovoltaic power forecasting," *Energy Conversion and Management*, vol. 121, pp. 380–390, 2016.

[43] K. Mohammadi, S. Shamshirband, D. Petković, and H. Khorasanizadeh, "Determining the most important variables for diffuse solar radiation prediction using adaptive neuro-fuzzy methodology; case study: City of kerman, iran," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 1570–1579, 2016.

[44] A. Will, J. Bustos, M. Bocco, J. Gotay, and C. Lamelas, "On the use of niching genetic algorithms for variable selection in solar radiation estimation," *Renewable energy*, vol. 50, pp. 168–176, 2013.

[45] B. Efron, "Bootstrap methods: another look at the jackknife," *The annals of Statistics*, pp. 1–26, 1979.

[46] O. García-Hinde, V. Gómez-Verdejo, M. Martínez-Ramón, C. Casanova-Mateo, J. Sanz-Justo, S. Jiménez-Fernández, and S. Salcedo-Sanz, "Feature selection in solar radiation prediction using bootstrapped svrs," in *Evolutionary Computation (CEC), 2016 IEEE Congress on Computational Intelligence*, pp. 3638–3645, IEEE, 2016.

[47] B. Pes, N. Dessì, and M. Angioni, "Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data," *Information Fusion*, vol. 35, pp. 132–147, 2017.

[48] A. Tsymbal, M. Pechenizkiy, and P. Cunningham, "Diversity in search strategies for ensemble feature selection," *Information fusion*, vol. 6, no. 1, pp. 83–98, 2005.

[49] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[50] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.

[51] H. Ackley, E. Hinton, and J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive Science*, pp. 147–169, 1985.

[52] G. E. Hinton, "Training products of experts by minimizing contrastive divergence.," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[53] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*, pp. 791–798, ACM, 2007.

[54] K. P. Murphy, *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: The MIT Press, 2012.

[55] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," in *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*, pp. 1–6, Dec 2014.

[56] E. Mizutani and S. E. Dreyfus, "On complexity analysis of supervised mlp-learning for algorithmic comparisons," in *Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference on*, vol. 1, pp. 347–352 vol.1, 2001.

[57] I. Jolliffe, *Principal component analysis.* Wiley Online Library, 2002.

[58] H. Wold, "Partial least squares," *Encyclopedia of statistical sciences*, 1985.

[59] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[60] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[61] A. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.