DOI: http://dx.doi.org/10.1109/EESMS.2013.6661694

A Data Approximation based Approach to Photovoltaic Systems Maintenance

Stefano Ferrari, Massimo Lazzaroni, and Vincenzo Piuri
Università degli Studi di Milano
Milan, Italy
Email: {stefano.ferrari,massimo.lazzaroni, vincenzo.piuri}@unimi.it

Ayse Salman
Doğuş University
Istanbul, Turkey
Email: asalman@dogus.edu.tr

Loredana Cristaldi and Marco Faifer
Politecnico di Milano
Milan, Italy
Email: {loredana.cristaldi, marco.faifer}@polimi.it

***Abstract*—The solar panel, which transforms the energy carried by the light in electricity, is a reliable component of a photovoltaic (PV) system, but its efficiency depends on several factors, such as its orientation, its working temperature, and its tidiness. Since maintenance is an expensive activity, a careful evaluation of the degradation of the panel and the resulting production loss has to be carried out. Besides, an accurate estimation of the potential production with respect to the weather condition requires expensive instruments and skilled operators. In this paper, we propose an alternative approach based on the prediction of the potential production based on a public weather station in the nearby of the considered plant. Several computational intelligence paradigms as well as several prediction setups are here challenged and compared.**

I. Introduction

Renewable energy sources play an important role in the nowadays economy and their relative importance is ever increasing. Among these, photovoltaic (PV) technology, which allows to obtain electric energy from solar radiation [1], represents a very interesting solution. In fact, besides large industrial plants, also small plants for fulfilling local needs are now quite common.

Among the several parameters that affect the efficiency of the solar panels (such as latitude, orientation, and inclination) the environmental conditions (such as, temperature [2], dust [3][4], snow [5], soiling [6]) are the most subtle to model and to control [2][7][8]. Due to the exposure in outdoor field, in fact, the performance of the panels can degrade. A simple approach for detecting and estimating the effects of obstructive phenomena (dirt, dust, or soiling) is to compare the power yielded by the solar panel with the theoretical output estimated using on-field measurement of the solar radiation and the other physical quantities (e.g., the air and panel temperature). However, this approach is not feasible in real cases. In fact, in order to provide an effective estimation, reliable measurements are needed, which in turn require an accurate policy of maintenance of the instrumentation operated by trained staff. This, especially for small plants, can make the economical loss due to the degradation of the performances preferable to the maintenance costs.

Recently, in [6] a direct approach for estimating the loss in performance has been proposed. It is based on the calibration of a panel in outdoor conditions with respect to an identical panel maintained in perfect conditions. Using some knowledge on the meteorological conditions and on the context (e.g., desert or agricultural region), the historical data collected by the two panels allows to estimate the loss in the generated power from the output of the unmaintained panel.

In previous work, we studied the effectiveness of autoregressive models and computational intelligence models for short term prediction of illuminance [9][10] and global radiation [11][12]. In particular, in [12] the data from a public weather station has been employed to predict, through the use of computational intelligence paradigms, the solar radiation on a site ten kilometers away. In this paper we experimented the use of a similar setup for estimating the power production loss of a panel due to the maintenance absence. The advantage of this approach is that the well maintained public weather station can provide accurate measurements, which can be exploited for estimating the performance of small solar plants in the neighborhood. A stable deviance of the real production from the estimated can be considered for scheduling the maintenance intervention.

II. The Prediction Models

Several models have been challenged for approximating the power produced by an observed solar panel under several working conditions (panel's temperature, irradiance, and length of the drought period). In particular, we used the Support Vector Regression (SVR) [13], the Feed-forward Neural network (FFN) [14], and the k-Nearest Neighbor (k-NN) predictor [15]. All these models allow to approximate a mapping between an input and an output space, automatically setting up their parameters, through suitable training algorithms that make use of only a finite set of input-output pairs (possibly affected by error), called training set.

A. k-Nearest Neighbor Interpolator

The k-Nearest Neighbor (k-NN) model is a instance-based or lazy learning paradigm used both for function approximation and classification [15]. It is used to predict the value of a function, f, in unknown points, given a sampling of the function itself (training data), $\{(x_i,\, y_i)\,|\,y_i = f(x_i)\}$. For an unknown point, x, the value of $f(x)$ is estimated from the value of its k nearest neighbors, for a given k, using a suitable voting scheme or an average. The most simple scheme, often used in classification, estimates $f(x)$ as the most common output value among its neighbors, while in function approximation the average output value is often used. More complex schemes, such as the use of weighted averaging, or a sophisticated norm for computing the distance can be used as well.

B. Support Vector Regression

Support Vector Machines (SVM) is a powerful method for classification [16][17] and regression [13]. In the latter domain, the method is usually named Support Vector Regression (SVR). In its original formulation, the regression function is obtained as the linear combination of some samples, called Support Vectors (SV), but it can be extended to non-linear mapping through the use of suitable functions called kernels. The solution to the regression problem is obtained as the minimization of a suitable loss function, which can be chosen such that the optimization problem results to be convex. The loss function is ruled by three hyperparameters: the accuracy, ε, that represents the accepted distance between the training data and the solution; the trade-off, C, that balance the closeness of the solution to the training data and the robustness of the solution; and the width of the Gaussians used as kernels, σ, which in the basic SVR algorithm are constrained to have the same width. The convexity of the problem guarantees that the optimal solution (which identifies the SVs, $\{\mu_i\}$, and the corresponding coefficients, $\{\beta_i\}$) is unique. The mapping is modeled as a linear combination of kernel functions (usually Gaussians):

$$f_{\text{SVR}}(x) = \sum_{i=1}^{L} \beta_i\, G(x;\, \mu_i,\, \sigma) + b \tag{1}$$

C. Feed-forward Neural Networks

The feed-forward neural networks (FFN) [14][18][19] are composed of processing units (called neurons) organized in layers. Each neuron computes its output as a function of a linear combination of the output of the neurons of the previous layer (this function is often called transfer function or activation function). The information, hence, flows only from the input layer to the output layer. It can be proved that a network with one hidden layer (i.e., a layer between the input and the output layers) has the universal approximation property. A FFN is characterized by the number of neurons of the hidden layer, the activation function, Ψ, (usually sigmoidal), and by the learning algorithm used (usually gradient descent based, such as Marquardt algorithm [19]).

More formally, the output, $f_{\text{FNN}}(\cdot)$, of a single layer FNN is

$$f_{\text{FNN}}(x) = \beta_0 + \sum_{j=1}^{L} \beta_j \Psi(\gamma_j^T \cdot x) \tag{2}$$

where L is the number of units of the hidden layer, the β_j is the weight of each neuron (β_0 is a bias term), the γ_j represents the weight vector of the linear combination of input for the j-th neuron.

The function Ψ, which can be chosen among different functions, is often the hyperbolic tangent

$$\Psi(z) = \frac{1 - \exp(-2z)}{1 + \exp(-2z)}. \tag{3}$$

When more than one hidden layer is used, the approximation capabilities of the neural model family (i.e., the class of functions that can be represented by any of the FNN with the given architecture) do not increase. However, when only a limited budget of neurons is taken into account, the approximation performance can change, although in practical cases it rarely improves considerably. When the FNN has two hidden layers (with L_k neurons in the k-th layer), its output, $f_{\text{FNN2}}(\cdot)$ is

$$f_{\text{FNN2}}(x) = \beta_0 + \sum_{i=1}^{L_2} \delta_i \Psi \left(\sum_{j=1}^{L_1} \beta_{ij} \Psi(\gamma_j^T \cdot x) \right) \tag{4}$$

where the input, x, is weighted by γ_j before being input to the j-th neuron of the first layer, which feed the i-th neuron of the second layer with its output weighted by β_{ij}; the output of the FNN is then computed weighting by δ the output of the second layer neurons.

III. The Datasets

For this work, we used as reference a dataset collected by ARPA Lombardia [20]. This dataset contains the hourly measurement of the global radiance in the site of Lambrate (Italy) which is separated by about 2 km from the site where the plant is situated. The ARPA dataset includes the global radiation, the air temperature and rainfall, measured hourly.

The plant dataset includes a set of measured voltage-current (V-I) characteristic curves (the produced power depends also on the applied load), the working temperature and solar radiation which cover most of the possible working condition of the photovoltaic panel. For this reason a measurement campaign has been performed and starting from the V-I curves the Maximum Power Point (MPP) values have been estimated (with a sampling period of about one minute). The rated parameters of the panel are the following: the maximum power, $P_{\text{MAX}} = 5$ W; the voltage and the current at which maximum power is produced, $V_{\text{PM}} = 17.5$ V and $I_{\text{PM}} = 0.285$ A; the open circuit voltage, $V_{\text{OC}} = 21.3$ V; and the short circuit current, $I_{\text{SC}} = 0.31$ A.

In order to match with the ARPA dataset, the data has been resampled with a sampling period of one hour. The data has been collected in two different periods, from July to October 2012 (66 samples) and from May to June 2013 (307 samples) for a total of 373 hourly samples.

For the experiment here described, a dataset composed of:

- the working temperature (in K),
- the global irradiance (in W/m^2),

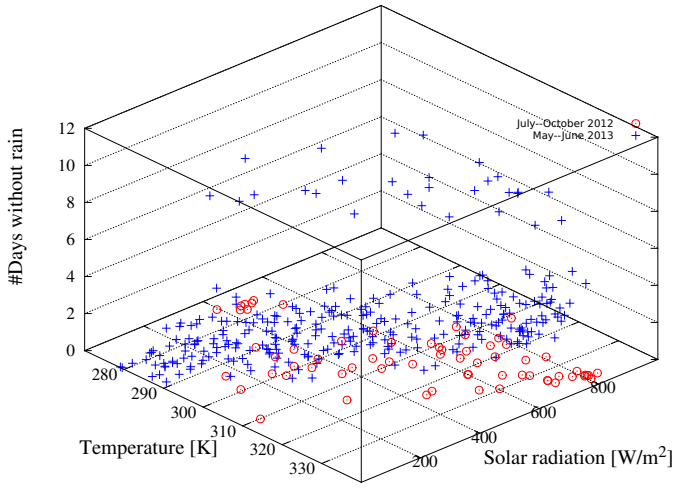

Fig. 1. The distribution of the samples of the dataset projected onto the Temperature-Irradiance-Drought subspace. The dataset is composed of samples measured in two different periods of time (July–October 2012 and May–June 2013). The two subsets are reported using the circle and the cross markers, respectively. Although there are two clusters characterized for the number of days without meaningful rainfall, it can be noticed that they belong to the same subset and can be ascribed to the normal variability of the meteorological phenomena. A weak correlation between the temperature and the solar radiation can also be noticed.

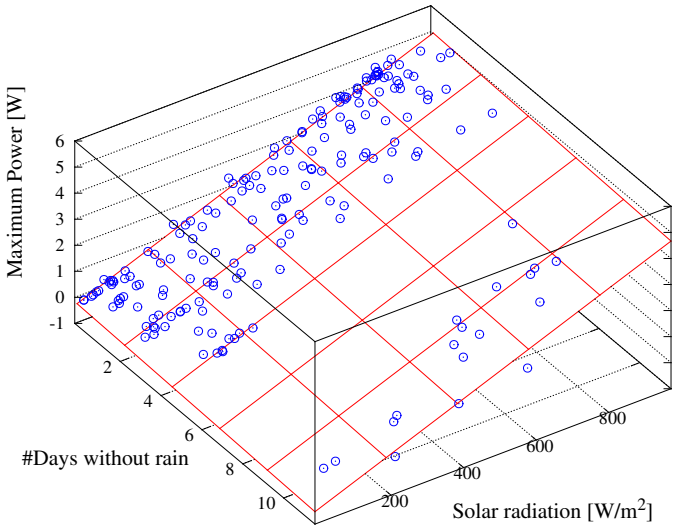

Fig. 2. The distribution of the samples of the dataset projected onto the MPP-Irradiance-Drought subspace. For making easier to understand the relation between the variables, the best fitting plane is also plotted. It is apparent that the MPP is directly proportional to the irradiance, but also a slight influence of the length of the drought period can be appreciated.

- the drought period length (in days),
- the MPP, (in W)

has been used. The drought period has been defined as the number of days (or fraction of day) since it rained less than 1 mm/m^2 of precipitation per day. Hence, rainfalls for less than 1 mm has been considered, somehow arbitrarily, too light for cleaning the panel.

In order to appreciate the distribution of the samples values, a plot of the temperature, solar radiation, and drought period length is reported in Fig. 1. Since two clusters are apparent along the drought period axis, the homogeneity of the two measurements periods can be questioned. However, in Fig. 1 the samples belonging to the two periods are reported using different markers (circles for July-October 2012 and crosses for May–June 2013) and it is evident that the long drought period cluster belongs to the same subset, while the other samples are quite mixed: this can be considered as evidence that the two subsets are homogeneous and the two clusters can be accounted to the normal variability of the meteorological phenomena.

A plot of MPP value vs. drought period and solar radiation is reported in Fig. 2. The depicted relationship is a projection of the one that is object of the study in this work, obtained not considering the influence of the working temperature on the power produced by the solar panel. For making easier to understand the relation between the variables, the best fitting plane has also been plotted. It is apparent that the MPP is directly proportional to the irradiance, but also that the length of the drought period can affect it negatively.

IV. Experiments

The prediction models described in Section II have been applied to the dataset described in the previous Section for predicting the produced power (in terms of MPP) given the working temperature of the panel, the irradiance, and the length of the drought period. Since the power produced cannot be negative, all the predictors have been enriched with a post-processing module that remaps to zero the negative values eventually output by the original models.

A. Data preprocessing

The dataset has been randomly partitioned in three sets: the training set, for computing the model parameters; the validation set, for selecting the best one among the challenged models; and the testing set, for assessing the final performance. The available samples have been distributed in the three set in the proportion of 80-10-10%, respectively. It results in 298, 37, and 38 samples for, respectively, the training, the validation and the testing set.

Since the scarcity of the available dataset, 30 trials (with different randomization of the data) have been carried out with each configuration of each model, and the average performance have been considered.

In order to balance the relative importance of the input features, the data has been normalized by dividing each feature by its standard deviation before to feed the predictors.

B. Performance Evaluation

For the evaluation of the performance, the prediction error has been measured by means of the average of the absolute error achieved on the testing set data:

$$\mathrm{Err}(f) = E(|y - f(x)|) \tag{5}$$

where $f(x)$ is the value predicted by the model for the sample (x, y), where x is the vector composed of the temperature, the irradiance, and the length of the drought period, while y is the corresponding measured MPP.

C. Prediction through k-NN Models

The performance of a k-NN predictor depends on several hyperparameters. Since it does not requires other training process than just storing the training values, all the hyperparameters of a k-NN predictor operate in the prediction stage. In particular, the behavior of the k-NN predictor is ruled by:

- k: the number of neighbors;
- the weighting scheme: the law to assign the weights for the weighted averaging prediction;
- the norm of the input space.

The following values for the hyperparameter k have been challenged:

$$k \in [1, 15] \tag{6}$$

Three weighting schemes have been tried: equal weight, weight proportional to the inverse of the neighborhood rank, and weight proportional to the inverse of the distance. Only the Euclidean norm has been used to compute the distance in the input space.

The k-NN predictors have been coded in GNU Octave, version 3.2.4.

D. Prediction through SVR

In order to train a SVR predictor, the hyperparameters that regulate the optimization procedure, have to be set to the proper value. Since the optimal values cannot be estimated a-priori, several combinations have to be tried and their effectiveness have to be assessed by cross validation.

The challenged hyperparameter values have been:

- for the accuracy, ε: $\{0.01,\ 0.05,\ 0.1,\ 0.5,\ 1\}$;
- for the regularization trade-off, C: $\{0.1,\ 1,\ 10,\ 100,\ 500,\ 1000\}$.

Besides, since the Gaussian kernel has been used, the width of the Gaussian, σ, which regulates the extent of the influence region of the corresponding SV (in regions further then 3σ from μ, the output is negligible), has to be set before the optimization is carried on. The values challenged for σ have been:

$$\sigma \in \{0.01,\ 0.05,\ 0.1,\ 0.5,\ 1,\ 2,\ 5\}. \tag{7}$$

For the simulations, the SVR implementation provided by SVM*light* [21] has been used.

E. Prediction through FFN

Since the FFN are a very variegated class of models, with different architectures and different learning algorithms, some a-priori choices are required in order to limit the number of simulations. For this experiment, two different architectures have been considered: single hidden layer (as described in (2)) and two hidden layers, (4). The only activation function considered for the hidden layer neurons has been the hyperbolic tangent function, (3), while the linear function has been chosen for the output layer.

TABLE I. TEST ERROR ACHIEVED BY THE PREDICTORS.

Predictor	Err(f) (std)
k-NN	0.274 (0.248)
SVR	0.247 (0.253)
FFN	0.260 (0.235)
FFN2	0.267 (0.237)

Several values for the number of neurons, L, have been tried:

$$L \in \{10,\ 20,\ 40,\ 100\}. \tag{8}$$

For the models with two hidden layers, the neurons have been equally distributed in the layers.

The networks have been trained using the Levenberg-Marquardt backpropagation algorithm.

The simulations have been run in GNU Octave 3.2.4, using the octave-nnet package (version 0.1.13).

V. RESULTS AND DISCUSSION

The test error of the challenged models are reported in Table I. The best performance for the k-NN model has been obtained using $k = 5$ and the inverted distance weighting scheme. The best performing SVR used $\sigma = 2$, $\varepsilon = 0.005$, and $C = 500$. The FFN models that achieved the lowest average validation error used $L = 20$ neurons for the single hidden layer case, and 50 neurons for each layer in the two hidden layer case (hence a total of $L = 100$ neurons).

From the data reported, the lowest test error has been achieved by the SVR (0.247 W), followed by the two neural models (0.260 and 0.267 W for respectively, the single and the two hidden layer networks), and then by the k-NN predictor (0.274 W). Although the SVR error is 5% relatively smaller than the error of the FFN model, their difference is not very meaningful, since the standard deviation of each test error is abundantly larger than their difference. The same consideration also applies to the comparison of errors of the SVR with respect to those of FFN2 and k-NN.

In Fig. 3, the approximation of the function that relates MPP to the considered input features (working temperature, irradiance, and length of the drought period) is reported for the four prediction model considered. For each of the models, the surface has been obtained by sampling regularly the input space and averaging the output of the best models (i.e., those obtained with the parameter above reported) for all the trials. In order to plot them, the output have been averaged also along the temperature. All of the surfaces resembles the orientation and the shape of the best fitting plane reported in Fig. 2. Subtle differences among the resulting surface can be observed: the k-NN surface is less smooth than the others and the SVR surface reaches higher values than the others for high irradiance and short drought period. It should be remarked that the surfaces has been obtained under the simplifying hypothesis of uniformly distributed temperatures values. Instead, as can be noticed from Fig. 1, the temperature is partially correlated to the irradiance. Hence, the relative importance of the contribution of a given value in the temperature dimension could depend also on the corresponding irradiance value.

In order to evaluate the production loss due to the absence of maintenance, the output of the best models (averaged over

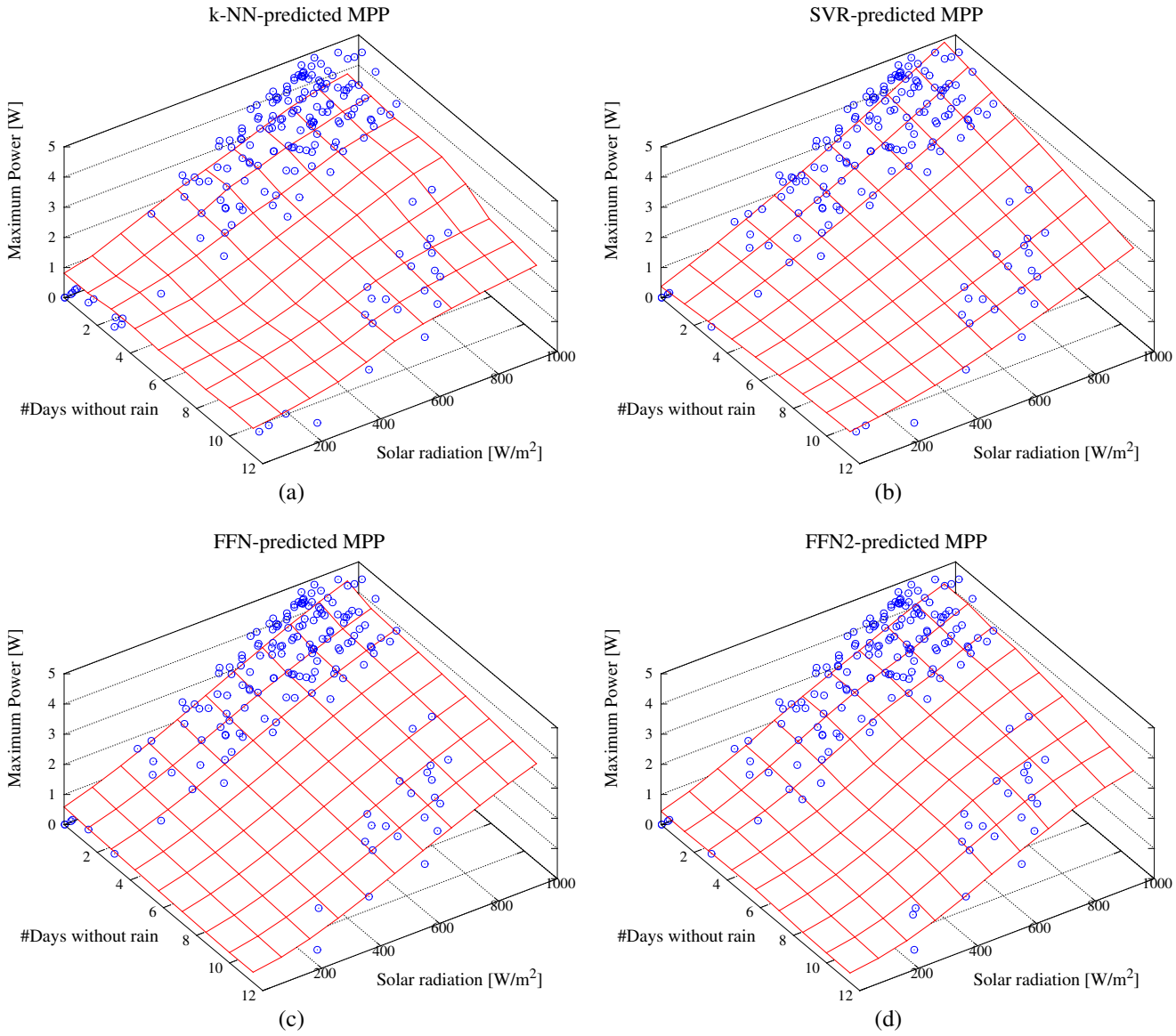

Fig. 3. The approximation of the MPP as a function of the irradiance and drought period length operated by (a) k-NN, (b) SVR, (c) FNN, and (d) FNN2 predictors.

the trials) has been evaluated for several values of length of the drought period. The simulation has been computed using as input those data that has an irradiance value higher than the average. This limitation is motivated by both the intention of excluding the data with a potentially high relative error, and considering only the situation where the productivity of the panel is potentially high. The resulting curves have been then normalized with respect to the maximum value of the MPP and plotted in Fig. 4. It can be noticed that the SVR curve is the smoothest, followed by the FNN curve. Moreover, the SVR curve is closest to the average curve, and hence is a good candidate for modeling the average production lost curve.

Some final remarks should be dedicated to the data. Since the models are generated from the data, the quality of the dataset (i.e., its ability to correctly represent the studied phenomenon) both in terms of accuracy and the coverage of the possible cases can affect the results.

In particular, it should be noticed that the SVR model uses a number of SVs close to the number of training examples (292 vs. 298). This means that almost all the training data are required to describe the relationship and the generalization is due to the interference between SVs (σ is equal to the double of the standard deviation of the data). This consideration is also supported by the value of ε and C (0.005 and 500, respectively) that are indicators of a solution close to the data.

VI. Conclusion

In this paper, the productivity loss of a solar panel due to the lack of maintenance has been studied. The energy production of the panel has been modeled as a function of its working temperature, the irradiance, and the length of the drought period.

Several models have been challenged in the task of approximating the relationship on the above mentioned parameters

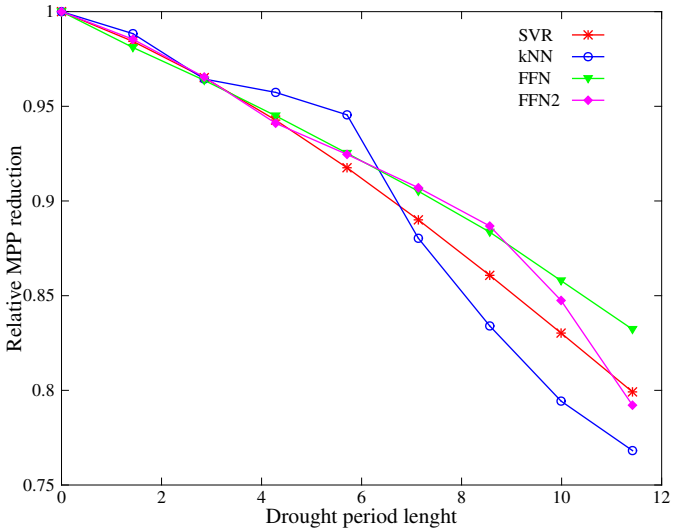

Fig. 4. Relative reduction curves estimated on the data that has an irradiance value higher than the average.

from a dataset of samples collected between July 2012 and June 2013.

The resulting models provide an approximation that well describes the dataset, although with slightly differences and a different degree of smoothness. Among them, the SVR model seems to be the best candidate, although the average model can also be considered.

This research work can be improved considering more data which can enrich the working situation considered. For instance, adding samples from a winter period can add information on the behavior of the panel for high levels of irradiance, but with low working temperatures. Besides, the availability of ground truth data can help in establishing the best model for the phenomenon.

REFERENCES

[1] M. Catelani, L. Ciani, L. Cristaldi, M. Faifer, M. Lazzaroni, and P. Rinaldi, "FMECA technique on photovoltaic module," in *IEEE International Instrumentation and Measurement Technology Conference (I2MTC 2011)*, May 2011, pp. 1717–1722.

[2] E. Meyer and E. Ernest van Dyk, "Assessing the reliability and degradation of photovoltaic module performance parameters," *Reliability, IEEE Transactions on*, vol. 53, no. 1, pp. 83–92, 2004.

[3] H. K. Elminir, A. E. Ghitas, R. Hamid, F. El-Hussainy, M. Beheary, and K. M. Abdel-Moneim, "Effect of dust on the transparent cover of solar collectors," *Energy Conversion and Management*, vol. 47, no. 18—19, pp. 3192–3203, 2006.

[4] M. Mani and R. Pillai, "Impact of dust on solar photovoltaic (pv) performance: Research status, challenges and recommendations," *Renewable and Sustainable Energy Reviews*, vol. 14, no. 9, pp. 3124–3131, 2010.

[5] L. Powers, J. Newmiller, and T. Townsend, "Measuring and modeling the effect of snow on photovoltaic system performance," in *Photovoltaic Specialists Conference (PVSC), 2010 35th IEEE*, 2010, pp. 000 973–000 978.

[6] J. Caron and B. Littmann, "Direct monitoring of energy lost due to soiling on first solar modules in California," *Photovoltaics, IEEE Journal of*, vol. 3, no. 1, pp. 336–340, 2013.

[7] M. Catelani, L. Cristaldi, L. Ciani, M. Faifer, M. Lazzaroni, and M. Rossi, "Characterization of photovoltaic panels: the effects of dust," in *IEEE Int. Energy Conference and Exhibition (ENERGYCON 2012)*, Sep. 2012, pp. 49–54.

[8] L. Cristaldi, M. Faifer, M. Rossi, and F. Ponci, "A simple photovoltaic panel model: Characterization procedure and evaluation of the role of environmental measurements," *IEEE Trans. on Instrumentation and Measurement*, vol. 61, no. 10, pp. 2632–2641, 2012.

[9] S. Ferrari, A. Fina, M. Lazzaroni, V. Piuri, L. Cristaldi, M. Faifer, and T. Poli, "Illuminance prediction through statistical models," in *Environmental Energy and Structural Monitoring Systems (EESMS), 2012 IEEE Workshop on*, Sep. 2012, pp. 90–96.

[10] S. Ferrari, M. Lazzaroni, V. Piuri, A. Salman, L. Cristaldi, M. Rossi, and T. Poli, "Illuminance prediction through extreme learning machines," in *Environmental Energy and Structural Monitoring Systems (EESMS), 2012 IEEE Workshop on*, Sep. 2012, pp. 97–103.

[11] S. Ferrari, M. Lazzaroni, V. Piuri, L. Cristaldi, and M. Faifer, "Statistical models approach for solar radiation prediction," in *Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International*, May 2013, pp. 1734–1739.

[12] S. Ferrari, M. Lazzaroni, V. Piuri, A. Salman, L. Cristaldi, and M. Faifer, "Computational intelligence models for solar radiation prediction," in *Instrumentation and Measurement Technology Conference (I2MTC), 2013 IEEE International*, May 2013, pp. 757–762.

[13] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[14] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Trans. on*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[16] V. N. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[17] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[18] R. Lippmann, "An introduction to computing with neural nets," *ASSP Magazine, IEEE*, vol. 4, no. 2, pp. 4–22, 1987.

[19] M. Hagan and M.-B. Menhaj, "Training feedforward networks with the marquardt algorithm," *Neural Networks, IEEE Transactions on*, vol. 5, no. 6, pp. 989–993, 1994.

[20] ARPA Lombardia. [Online]. Available: http://ita.arpalombardia.it/

[21] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, ch. 11, pp. 169–184.