DOI: http://dx.doi.org/10.1109/I2MTC.2013.6555712

Statistical Models approach for Solar Radiation Prediction

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

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

***Abstract*—It is well known that the knowledge of solar radiation represents a key for managing photovoltaic (PV) plants. In a smart grid scenario to predict the energy production can be considered a milestone. However, the unsteadiness of the weather phenomena makes the prediction of the energy produced by the solar radiation conversion process a difficult task. Starting from this considerations, the use of the data collected in the past represents only the first step in order to evaluate the variability both in a daily and seasonal fashion. In order to have a stronger dataset a multi-year observation is mandatory. In this paper, several autoregressive models are challenged on a two-year ground global horizontal radiation dataset measured in Milan, and the results are compared with those of simple predictors.**

I. Introduction

The electric power system is mainly composed by units for energy production i.e. generators, loads and a power grid that connects them. Actual configuration principally includes large central generators which, through the transformers, inject electrical power in the transmission grid. The world energy infrastructure is nowadays subjected to a important transformation such as the growing number of distributed small generation units, based on different technologies, directly connected to the power grid. These small generation units put side by side to the large and traditional ones are defining a grid based on the so called distributed generation. This kind of network architecture implies new problems concerning the management. In fact in traditional network the stability of the power system was achieved by means of the direct control of few large conventional power generators. By introducing distributed generation this approach cannot be followed, since the small generation units are basically not controllable by the network system operator. In particular this scenario is critical when units based on renewable energy resources are used, since they can only provide power as long as the source of energy is available. In many situation the energy production is mainly utilized directly by the producer or for nearby buildings. When energy production exceed the necessity the excess flows into the power grid of the utilities. In order to implement an electric grid allowing a large amount of distributed energy sources, different solutions approach to the problem of the network stability must be followed. It is clear that in this scenario the possibility to predict the capability of the plant to generate power during the day, greatly helps the management of such a power system. Among the renewable energy sources, a very interesting solution is photovoltaic (PV) technology, which allows to obtain electric energy from solar radiation [1]. One of the most important benefits of the electrical energy production based in photovoltaic technology is the low environmental impact. On the other end, the main weakness of this renewable energy source is that its availability cannot be fully controlled. Many aspects need to be considered such as geographic position, local climate, weather and global efficiency of the panel [2], [3], [4]. Among these, the position and the climate influence on the solar radiation can be easily obtained from astronomical and statistical data, but the weather is characterized by a high variability and depends on many physical factors. According to [5], the forecasts required by the activity related to the grid management can be divided in two categories. The first is related to grid stability problem (intra-hour, hour ahead, and day ahead), while the second concerns planning and assets optimization on medium and long-term (monthly and yearly forecasts, respectively). Since the main factor for solar radiation availability is the local weather, approaches based on weather forecast have been widely used in literature. These are based on data obtained from satellite observations and ground stations. The geographic and time availability of data are the main aspects that have to be taken into account. Besides, the sampling rate of the measurement have to be related to the granularity of the forecast.

The solar radiation prediction can be based on data obtained by several data sources, characterized by the type of data they produce, as well the space-time granularity they provide. These data source are, for example: Numerical Weather Predic-

tion (NWP) models, Satellite-base forecast, All-sky imagers, Ground measurements.

Several forecasting approaches have been used in literature. Among these, the most effective in producing hour-ahead predictions are based on empirical regression, neural networks [6] and time-series models (e.g., ARMA, ARIMA) [7][8].

In this paper, a two-year hourly dataset of the global horizontal radiation will be used to feed some autoregressive models to obtain a one-hour forecast. The dataset has been collected in two years by the MeteoLab [9][10]. In previous works [11][12][13], several models have been challenged in the task of predicting the global horizontal illuminance. In the present work, instead, data coming from a new parameter, the global horizontal radiation, will be considered. The performance of the autoregressive models will be compared with those of a naïve predictor, the persistence model, and of a simple predictive model, namely the k-Nearest Neighbor (k-NN) model.

II. The Prediction Models

A time series is composed of a sequence of observation $\{x_t\}$ sampled by a sequence of random variables $\{X_t\}$. Usually, the ordering value is related to the time and the observation are related to a phenomenon that varies with the time. A practical assumption is that the observations are taken in equally spaced instants.

A. Autoregressive Models

An autoregressive model describes the values of a particular time series in terms of its past values [14]. In particular, the value of X_t is modeled as a combination of a part that is determined by the past values of the series and a part determined by an unpredictable event that happens at the time t (innovation). More formally, given a time series $\{X_t\}$, its autoregressive representation of order p, often denoted by AR(p), is:

$$X_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k X_{t-k} + \varepsilon_t \tag{1}$$

where α_0 is a constant and the innovation ε, is assumed to be white noise ($E(\varepsilon) = 0$, $E(\varepsilon^2) = \sigma^2$) and $\{\varepsilon_t\}$ are supposed to be normal independent and identically distributed (i.i.d.) random variables.

A moving average model describes the time series values in terms of linear combination of (unobserved) innovation values. A moving average representation of order q, often denoted by MA(q), of the time series $\{X_t\}$ is:

$$X_t = \mu + \sum_{h=1}^{q} \beta_h \varepsilon_{t-h} + \varepsilon_t \tag{2}$$

The autoregressive and moving average models can be combined in the autoregressive moving average model (ARMA). An ARMA representation of autoregressive order p and moving average order q, ARMA(p, q) is formally described as:

$$X_t = \alpha_0 + \sum_{k=1}^{p} \alpha_k X_{t-k} + \sum_{h=1}^{q} \beta_h \varepsilon_{t-h} + \varepsilon_t \tag{3}$$

When the time series is sampled from a stationary process, it can be represented by the above mentioned models. However, when the time series shows a trend or a seasonality, a more advanced class of models, namely the autoregressive integrated moving average models (ARIMA), have to be used. The ARIMA model take into consideration also the difference series (i.e., the series resulting by computing the difference of time lagged series). In particular, the notation ARIMA(p, d, q) is commonly used for indicating the ARIMA model with p, d, and q order of respectively autoregression, differencing, and moving average. The formalization of this model is operated through the backward shift operator, B: $X_{t-1} = B\,X_t$. This allows to express X_{t-k} as $B^k X_t$. The ARIMA(p, d, q) representation of the time series $\{X_t\}$ is:

$$\left(1 - \sum_{k=1}^{p} \alpha_k B^k\right)(1-B)^d X_t = \left(1 + \sum_{h=1}^{q} \beta_h B^h\right)\varepsilon_t \tag{4}$$

B. Persistence

In order to assess the performance of model in the short-term prediction of a time series, the persistence model is often used. It is a naïve predictor that assumes that the next value of the time series, x_t will be equal to the last known, x_{t-1}. It is obviously inappropriate for long-term prediction of time-series of interest in real cases, but it can be used as a baseline forecast: it is supposed that any other model will perform better than the persistence model.

C. 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. The k-NN can be used in time series prediction using some previously observed values for composing the input vectors. For instance, when using a two-dimensional feature space, the training dataset will be composed by triples of the form (x_{t-2}, x_{t-1}, x_t), where will be assumed that $x_t = f(x_{t-2}, x_{t-1})$.

III. Experimental Activity

For the experiments here described, a dataset collected by the MeteoLab [9][10] between October 2005 and October 2007 has been used. The MeteoLab station measures and collects with a sampling step of ten minutes the following data: global horizontal irradiance, diffuse horizontal irradiance, global horizontal illuminance, relative humidity, and air temperature. The released dataset provides their hourly average.

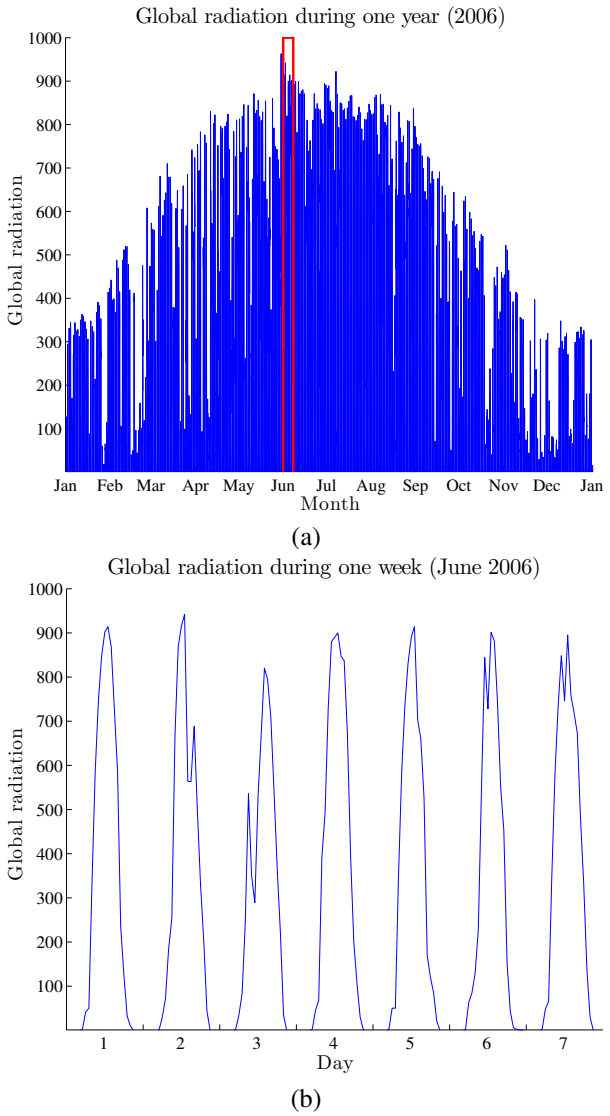


Fig. 1. One year (a) and one week (b) of the measured global horizontal radiation. Note the trend in the year and in the day, but also the strong variability in the intraday values.

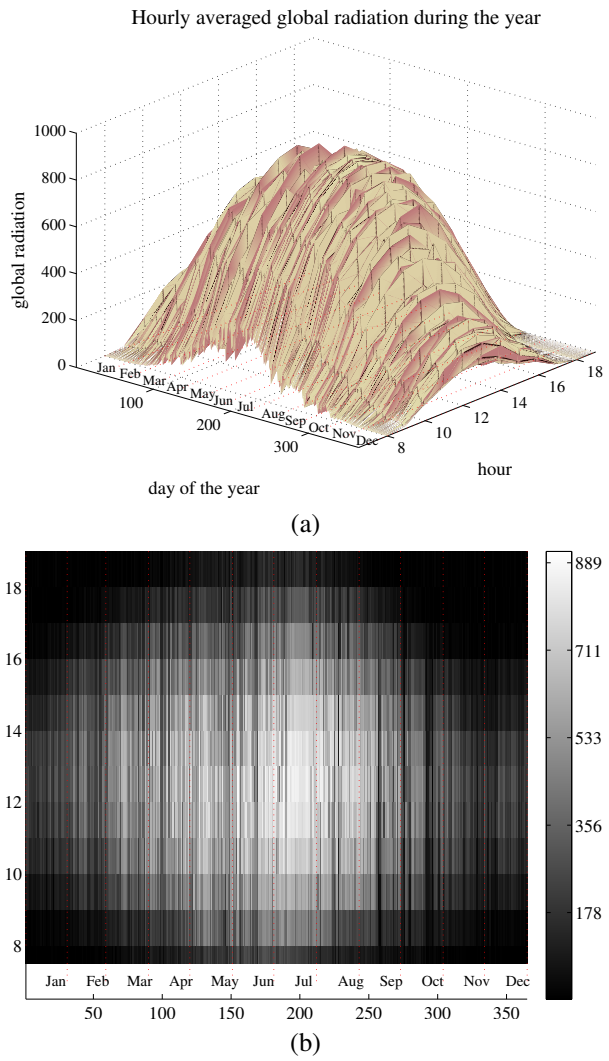


Fig. 2. The average global radiation for each day of the year and hour have been is plotted as a surface. The roughness of the surface is due to variability, although a clear trend of the phenomenon can be acknowledged.

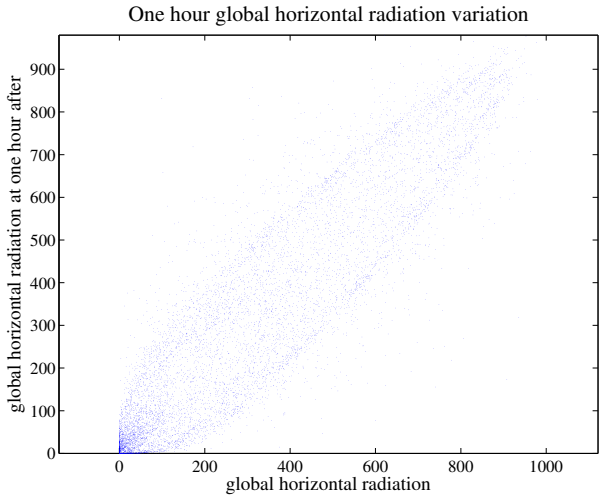
For this work, only the global horizontal radiation has been considered. Subsets of the available samples are reported in Fig. 1. In particular, Fig. 1a describes the global horizontal radiation measured in the year 2006, in Fig. 1b only one week is reported (the first week of June). It can be noticed that regularities are apparent both in the yearly and in the daily scale, but also that large deviations from the average behavior are possible, due to meteorological variability.

As shown by surface depicted in Fig. 2, the global horizontal radiation varies both on daily and seasonal basis. The surface has been obtained by averaging the samples acquired in the same hour of the same day of the year. Although a trend is clearly recognizable, the variability of the global horizontal radiation (which depends also by fast changing meteorological phenomena) makes the surface very wrinkled.

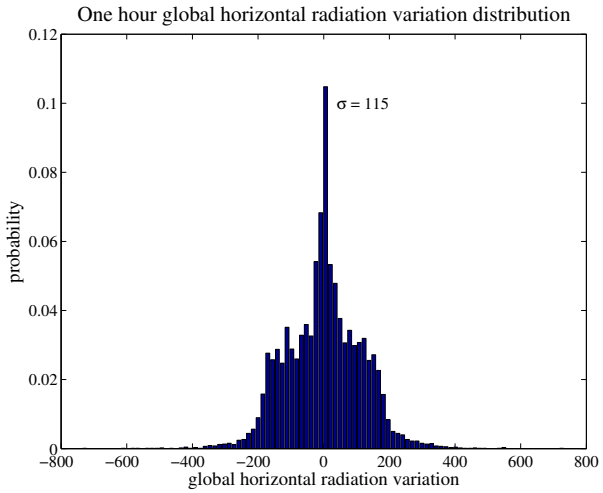
Figure 3, instead shows the relation between the global horizontal radiation acquired at two consecutive hours. In particular, in Fig. 3a the distribution of the points along the identity line supports the use of the persistence predictor. However, the maximum of the prediction error of the persistence can be considerably high since the length of the vertical section of the cloud of points is at least 300, where the maximum value of the radiation is about 900. The histogram in Fig. 3b resembles a mixture of two normal distributions with the same mean. This is due to the fact that in the early and the late daylight hours the global radiation does not change very much (especially in the winter). Hence, the consecutive samples acquired in those period of time are quite similar, while the other moments of the day show a larger variability.

A. *Dataset Pre-Processing*

Since time series models requires that all the values are equally time spaced, the few values that are missing are interpolated using a simple rule that exploits the daily sea-



(a)



(b)

Fig. 3. The persistence predictor uses the global radiation value measured one hour before as predicted value. Panel (a) shows the relationship between the two measurements of the global horizontal radiation performed at the distance of one hour. Evidently, the samples distribute along the identity line. In panel (b), the estimated probability density function of the difference between subsequent samples (with a standard deviation of 115).

sonality of the solar radiation. For each missing value, x_t, the set $\{x_{t-1}, x_{t+1}, x_{t-24}, x_{t+24}\}$, i.e., the set composed of the global radiation one hour before and ahead, and one day before and ahead are considered. The missing value is then replaced with the average of the collected values. Since the missing data are few, the selected set has a meaningful number of elements even though some of the selected elements are missing too.

The resulting dataset is composed of 18096 samples. Since the dataset covers a period of time of two years, and the autoregressive models require a training set composed of consecutive data, the first year has been used as training set. In this way, the yearly variability have a chance of being captured by the models. The data belonging to the second year has been randomly partitioned in the validation and training set. Hence, training, validation, and testing set are composed of, respectively, 9048, 4524, and 4524 samples.

B. Performance Evaluation

For the evaluation of the performances, only the daylight hours data ($[8, 19]$) has been considered. Besides, since the solar radiation cannot be negative, all the negative values predicted by the models are set to zero.

The prediction error has been evaluated as the average of the absolute error achieved on the testing data:

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

where $f(x_t)$ is the value for x_t predicted by the model f.

C. k-NN Models Prediction

The performance of a k-NN predictor depends on several hyperparameters, which operate only in the prediction stage, since the k-NN predictor does not requires other training process than just storing the training values. In particular, the behavior of the k-NN predictor is ruled by the number of the considered neighbors, k; the number of dimension of the input space, D, which corresponds to the number of previous values used for the prediction; the weighting scheme, i.e., the law to assign the weights for the weighted averaging prediction.

The following values for the hyperparameters has been challenged:

$$k \in [1, 30] \quad \text{and} \quad D \in [1, 10] \tag{6}$$

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.

For the sake of comparison, the rules for generating the training, validation and test set will be the same used for the autoregressive models, described in III-A.

D. Autoregressive Models Prediction

In order to train an autoregressive predictor, a suitable value for the hyperparameters that rule the optimization procedure (i.e., the autoregression order, p, the moving average order, q, and the differencing order, d), have to be chosen. Several combination of the hyperparameters values have been tried and their effectiveness have been estimated through cross validation. In particular, the AR models have been challenged with $p \in \{1, \ldots, 100\}$; the ARMA models have been challenged with the combination of p and q for $p \in \{1, \ldots, 50\}$ and $q \in \{1, \ldots, 50\}$; and the ARIMA model have been challenged with the combination of the following values of p, d, and q:

$$p \in \{1, \ldots, 30\}, \quad d \in \{1, \ldots, 3\}, \quad q \in \{1, \ldots, 30\} \tag{7}$$

Since the training of the ARMA and ARIMA models requires consecutive training data, for avoiding of considering two separated periods of time for evaluating the validation and training error (which involves the risk of biased estimation due to the seasonality of the phenomenon under study), the prediction on the data not used to train the predictor has been carried out first and then the predicted period has been sampled for obtaining the validation and testing data.

TABLE I
TEST ERROR ACHIEVED BY THE PREDICTORS.

Predictor	Err(f) (std)
Persistence	88.3 (74.2)
k-NN	47.7 (59.7)
AR	43.5 (56.9)
ARMA	42.7 (56.5)
ARIMA	43.3 (56.5)

IV. RESULTS AND DISCUSSION

The persistence and k-NN predictors, described in Section II, have been coded in Matlab, while for the autoregressive models (AR, ARMA, and ARIMA) their implementation in R have been used. Their performances have been evaluated using the prediction error, Err(f), described in (5). Since the persistence predictor configuration does not need any hyperparameters, the whole dataset described in Section III-A has been used to assess its performances. Instead, the training of the k-NN and the autoregressive models are regulated by a pool of hyperparameters. Hence, the training set has been used to estimate the model's parameters for each combination of the hyperparameters, then the validation dataset has been used to identify the best model (i.e., the one that achieved the lowest prediction error on the validation dataset) and the prediction error of that model on the testing set has been used to measure the performance of the class of the predictors.

As reported in Table I, the persistence predictor has achieved an error $\text{Err}(f_{\text{P}}) = 88.3$, while the k-NN achieved an error $\text{Err}(f_{k\text{-NN}}) = 47.7$, for $D = 4$, $k = 17$, and using the inverted distance weighting scheme.

The AR model that scores the lower validation error has been trained using $p = 97$ and achieved $\text{Err}(f_{\text{AR}}) = 43.5$; the best ARMA model has been trained using $p = 28$ and $q = 22$, achieving $\text{Err}(f_{\text{ARMA}}) = 42.7$; the best ARIMA model, trained using $p = 23$, $d = 1$, and $q = 16$, achieved a testing error $\text{Err}(f_{\text{ARIMA}}) = 43.3$.

Figure 4 shows the distribution of the prediction error of AR, ARMA and ARIMA models. Hardly some difference can be spotted in Figs. 4a–c, although Figs 4d–f reveal a slightly compact histogram for ARMA and ARIMA. In particular, the error peaks in Figs 4a–c are in the same position, probably due to some fast changing meteorological events happened in that period that modified the usual global radiation pattern.

In Figs. 5 and 6, the performance of the autoregressive models with respect to their hyperparameters are represented. The continuous line represents the projection of the error onto the considered parameter when the other parameters are fixed to the values that achieved the best prediction error. Except for the AR model, the line rarely touch the lowest error point. A different randomization of the data could change the hyperparameters combination that gives the best setup.

V. CONCLUSIONS

Although all the models have achieved a similar testing error, with also a similar distribution, the training time for the AR model has been larger than the time required by the ARMA and ARIMA models (respectively 45 and 49 times larger). A deeper analysis of the results unveil that a comparable error can be achieved by the AR model using a smaller number of coefficients, but with a required time just 2 times larger than the other models (e.g, for $p = 51$). The analysis of the behavior of the error with respect to the hyperparameters of the models shows the tendency of increasing the performance as the numbers of previous values of the global radiation considered in the prediction approach the seasonality of the time series or a multiple of it. However, since the global radiation has a daily seasonality of 24 hours, the large number of parameters of the model can give rise to numerical error in the estimation procedure and in any case, can slow down the convergence of the estimation procedure.

Since a simpler model is preferable, the ARIMA model, which requires less parameters than AR and ARMA, can be considered to best fit the time series here considered.

Future works can consider the exploitation of other information (both temporal and meteorological) in order to improve the robustness of the estimation of the model parameters.

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

[3] 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.

[4] L. Cristaldi, M. Faifer, S. Ierace, M. Lazzaroni, and M. Rossi, "An approach based on electric signature analysis for photovoltaic maintenance," in *IEEE International Energy Conference and Exhibition (ENERGYCON 2012)*, Sep. 2012, pp. 1–8.

[5] V. Kostylev and A. Pavlovski, "Solar power forecasting performance — towards industry standards," in *1st Int. Workshop on the Integration of Solar Power into Power Systems*, Aarhus, Denmark, Oct. 2011.

[6] A. Mellit, M. Benghanem, and S. Kalogirou, "An adaptive wavelet-network model for forecasting daily total solar-radiation," *Applied Energy*, vol. 83, no. 7, pp. 705–722, 2006.

[7] R. Perdomo, E. Banguero, and G. Gordillo, "Statistical modeling for global solar radiation forecasting in Bogotà," in *Photovoltaic Specialists Conference (PVSC), 2010 35th IEEE*, jun 2010, pp. 002 374–002 379.

[8] M. H. T.A. Raji, A.O. Boyo, "Analysis of global solar radiation data as time series data for some selected cities of western part of Nigeria," *International Journal of Advanced Renewable Energy Research*, vol. 1, no. 1, pp. 14–19, 2012.

[9] T. Poli, L. P. Gattoni, D. Zappalà, and R. Gottardi, "Daylight measurement in Milan," in *Proc. of PLEA2006, Conf. on Passive and Low Energy Architecture*, 2006.

[10] ——, "Daylight measurement in Milan," in *Clever Design, Affordable Comforta Challenge for Low Energy Architecture and Urban Planning*. Geneve - CH: Raphael Compagnon & Peter Haefeli and Willi Weber, 6 2006, pp. 429–433.

[11] F. Bellocchio, S. Ferrari, M. Lazzaroni, L. Cristaldi, M. Rossi, T. Poli, and R. Paolini, "Illuminance prediction through SVM regression," in *Environmental Energy and Structural Monitoring Systems (EESMS), 2011 IEEE Workshop on*, Sep. 2011, pp. 1–5.

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

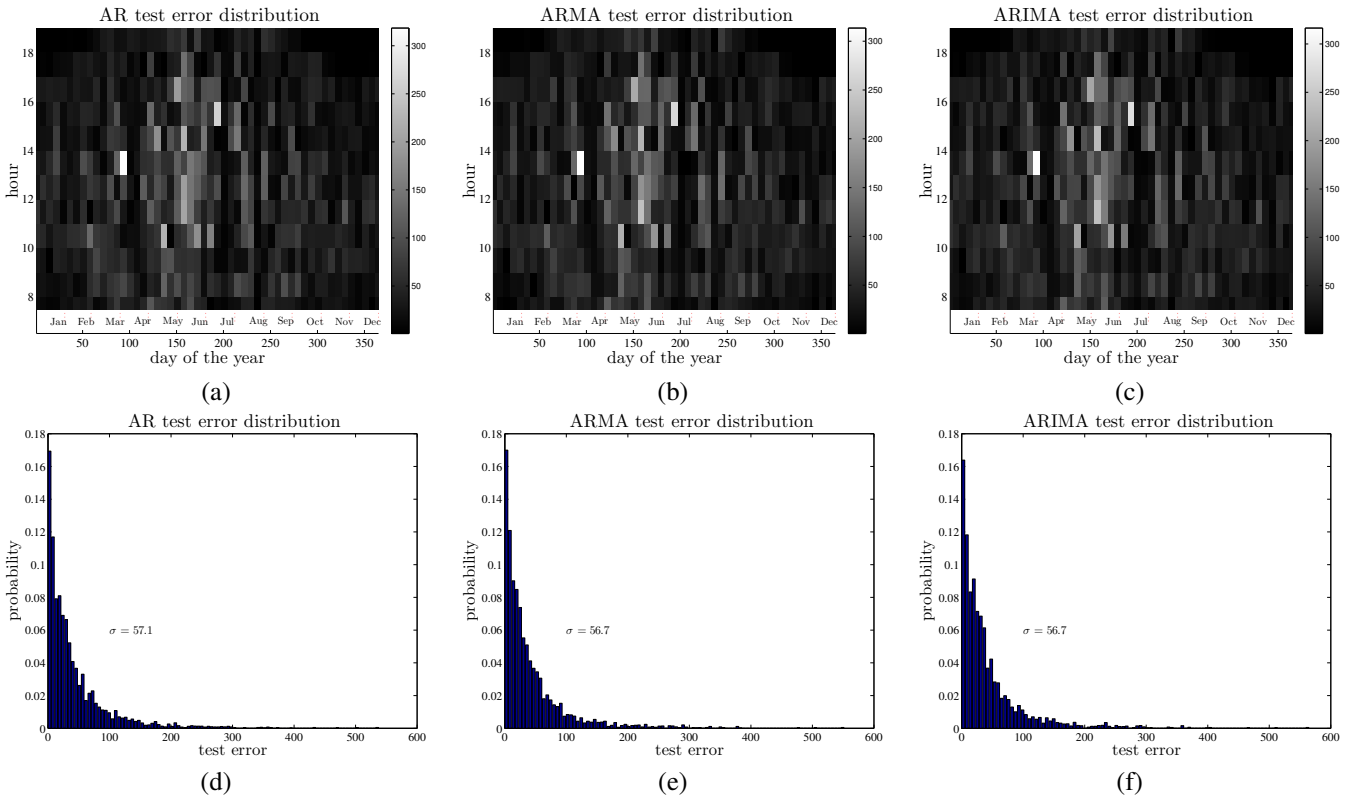


Fig. 4. Test error distribution. In panels (a)–(c), the test error produced by respectively the AR, the ARMA, and the ARIMA predictor are reported with respect to the day of the year and the hour.In panel (d)–(f), the estimated probability density function of the test error of the autoregressive models.

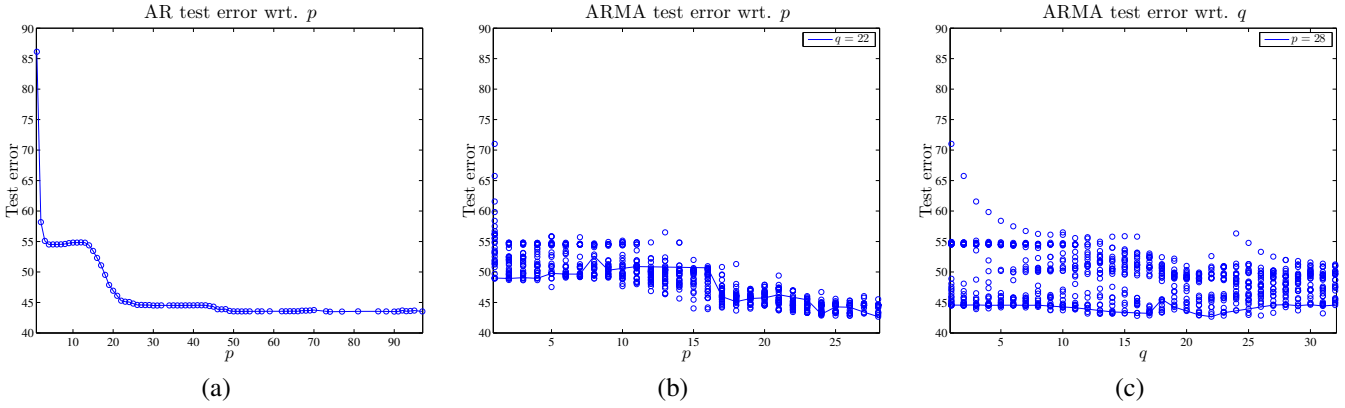


Fig. 5. AR, (a), and ARMA, (b)–(c), test error wrt. the hyperparameters.

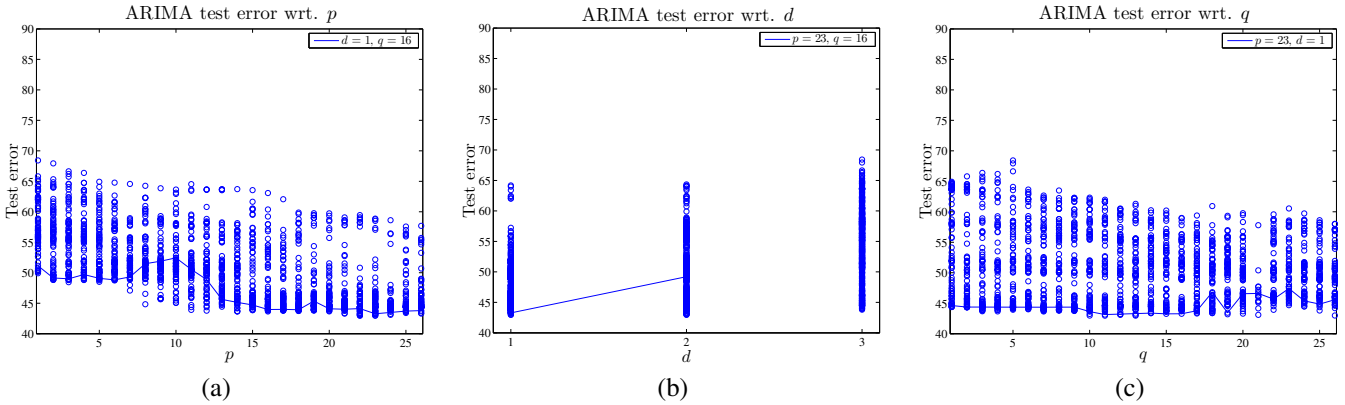


Fig. 6. ARIMA test error wrt. the hyperparameters.

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

[14] G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.

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