Applied Energy 205 (2017) 116–129

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

# Comparative analysis of data-driven methods online and offline trained to the forecasting of grid-connected photovoltaic plant production

S. Ferlito, G. Adinolfi, G. Graditi*

*ENEA – Italian National Agency for New Technologies, Energy and Sustainable Economic Development Research Center, Piazza E.Fermi, 1, 80055 Portici (Napoli), Italy*

## HIGHLIGHTS

- Eleven forecasting data-driven methods are compared to 12 h ahead prediction of PV power.
- Linear, nonlinear and ensemble forecasting models are evaluated.
- An "online" and "offline" methods are applied to train forecasting methods.
- Optima model main parameters and optimum training dataset length are identified for each model.
- Performances are compared by accuracy metrics, mean RMSE confidence intervals and Box Plots describing RMSE distribution.

## ARTICLE INFO

*Keywords:*
Photovoltaic generation
Machine learning
Performance evaluation
Prediction error
Training dataset optimization

## ABSTRACT

Actual technology improvements are contributing to DC and AC micro grid diffusion characterized by renewables photovoltaic and storage systems. Photovoltaic technologies have the advantage of a capillary distribution, but they are characterized by an intrinsic variable behavior due to continuously changing weather conditions. This drawback can be overcome by an appropriate temporal and energetic match among photovoltaic generation and storage capacity, so increasing micro grids reliability and efficiency levels. An accurate forecast of photovoltaic production can contribute to smooth photovoltaic systems intermittency problems so supporting generation and storage balance.

Many different forecasting algorithms have been proposed in literature to provide long, medium and short-term predictions of photovoltaic production. A criterion for selection among them is not a priori and it is not uniquely identifiable. In this paper, the attention is focused on eleven data-driven models to obtain 12 h ahead forecast, also including some models not usually used in solar field but successfully employed in other expertise fields. This study addresses simple linear models, as Multiple Linear Regression, nonlinear models, such as Classification And Regression Tree, Model Tree M5, Extreme Learning Machines, weighted k-Nearest Neighbors, Multivariate Adaptive Regression Spline, Support Vector Machines, Bayesian Regularized Neural Networks and ensemble methods, as Random Forests, Cubist and Extreme Gradient Boosting. The goal is to compare methods characterized by different complexity levels to understand if a higher complexity model can provide better performances. Furthermore, the considered forecasting methodologies are compared applying two different training methodologies (online and offline) to identify the most performing training mode. The application of optimization algorithms permits to identify optima parameters and the optimum training dataset length for each model. After the optimization step, a statistical analysis is carried out to compare methods forecasting performances for the production of a 1 $kW_p$ photovoltaic plant installed at the ENEA Research Center of Portici. The case study results demonstrate promising forecasting performances applying the online training mode. Among all studied methods, Support Vector Machines, M5 and the Cubist can assure minima prediction errors and satisfying accuracy with optima dataset lengths. Cubist and M5 represent the best performing models since they are able to minimize prediction errors both in case of optima and minima training datasets.

* Corresponding author.
*E-mail addresses:* sergio.ferlito@enea.it (S. Ferlito), giovanna.adinolfi@enea.it (G. Adinolfi), giorgio.graditi@enea.it (G. Graditi).

http://dx.doi.org/10.1016/j.apenergy.2017.07.124
Received 25 March 2017; Received in revised form 22 May 2017; Accepted 27 July 2017

**Nomenclature**

| | |
|---|---|
| *ANN* | Artificial Neural Network |
| *AZ* | Solar Azimuth |
| *BRNN* | Bayesian Regularized Neural Network |
| *CART* | Classification and Regression Tree |
| *CARET* | Classification And Regression Training |
| *CC* | Cloud Cover |
| *CV* | Cross Validation |
| *DEoptim* | differential evolution optimization algorithm |
| *EL* | solar elevation |
| *ELM* | Extreme Learning Machine |
| *ES* | Exponential Smoothing |
| *FFNN* | Feed Forward Neural Network |
| *GHI* | Global Horizontal Irradiance |
| *GRNN* | Generalized Regression Neural Network |
| $I_{poa\_cs}$ | plane of array irradiance by clear sky model |
| *IQR* | interquartile range |
| *KNN* | weighted k-nearest neighbors |
| *MA* | Moving Average |
| *MLR* | multiple linear regression |
| *Pac* | photovoltaic power production |
| *PV* | photovoltaic power |
| *RF* | Random Forest |
| *MARS* | Multivariate Adaptive Regression Splines |
| *M5* | model trees M5 |
| *NWP* | Numerical Weather Prediction |
| *SARIMA* | Seasonal AutoRegressive Integrated Moving Average |
| *SARIMAX* | Seasonal AutoRegressive Integrated Moving Average with eXogenous variables |
| *SR* | stepwise regression |
| *T2m* | temperature at 2 m |
| *Tcc* | Total Cloud Cover |
| *Temp* | ambient temperature |
| *Wcond* | weather conditions |
| *XGBoost* | Extreme Gradient Boosting |

## 1. Introduction

### 1.1. Motivation and approach

During last decade energy production from renewable resources has considerably increased due to PhotoVoltaic (PV) technologies and BOS (Balance Of System) price reduction. The capillary installation of non-programmable renewables generators has permitted to cover about 14% [1,2] of Italian total energy demand. In spite of the worldwide reached results, PV plant production is adversely affected by intrinsic unpredictability due to continuous weather variability. Ambient temperature [3,4], total cloud cover, global horizontal irradiance represent factors deeply influencing PV generators efficiency. In addition, also environmental and mechanical factors, such as hot spots, dust [5] and shadow, can negatively contribute to the decrease of PV devices performances. In this scenario, researchers attention is focused on strategies able to overcome or adequately manage PV plants continuous changing behavior and efficiencies. In fact, the application of algorithms to accurately forecast PV production have become an important task to real time match among renewable, traditional generation, consumptions and storage.

This action can provide a significant contribution to the smart management of the energy networks based on poly-generation and storage systems, actively contributing to the adequate schedule of storage charge and discharge operations. In this way, economic and energy savings strategies are fostered in AC and DC micro-grids [6]. An accurate forecast characterized by a minimum prediction error can represent a useful tool for energetic and purchasing scopes.

### 1.2. Literature review

In literature, a great amount of papers has been dedicated to the implementation of forecasting algorithms, employed in different expertise areas such as financial, medical, biological and engineering ones. Different classifications have been proposed on the base of their time horizon, specific variables or complexity level.

In this paper the considered forecasting methods are classified as linear, non linear and ensemble ones. In detail, in [7] the Multiple Linear Regression (MLR) technique is applied to obtain electric load forecast. This method is used to improve the purchase process of an American utility company.

The Classification And Regression Tree (CART) is a non linear model employed in [8] to identify explanatory variables in a global solar radiation forecasting application. This technique contributes to elaborate a Variable Importance classification attributing priority levels among the whole considered variables. A case study referring to real Tokyo data is carried out.

In [9] the weighted k-Nearest Neighbors (KNN) method is applied in an economic scenario. The aim is to forecast the Chinese Shanghai and Shenzhen stock market indices. On the base of these indices, suitable exchange strategies can be developed to obtain low risk profits.

A Multivariate Adaptive Regression Spline (MARS) model is examined in [10] to forecast one step ahead daily PV power production. All reported data-driven models need a training and a test dataset. The subdivision of the whole dataset in training and test subsets is randomly done with a quite large proportion dedicated to the training phase. An hybrid model composed by Support Vector Machines (SVM) method and a Seasonal Auto Regressive Integrated Moving Average (SARIMA) one is proposed in [11] for short-term forecasting of PV production; results are validated using a dataset of about five months.

Furthermore, in [12], global incident solar radiation is forecasted basing on the sunshine hours by means of the SVM and the wavelet-coupled SVM models. This study is carried out for three Australian cities (Townsville, Brisbane and Cairns).

Many literature papers are available on Artificial Neural Network (ANN) applied to forecasting issues. In detail, three types of ANN are compared in [13] to forecast power production one step-ahead P(t + 1), using future values of solar irradiance on plane of array Gt(t + 1), solar cell temperature Tc(t + 1) and the present value of production P(t) as inputs. Four data-driven approaches (ANN, SVM, KNN and MLR) are used in [14] to predict daily solar power values three steps-ahead. These four models were evaluated considering two base scenarios: one including meteorological parameters as inputs and the other using only time-series parameters.

Furthermore, the Extreme Learning Machines (ELM) methods is employed in [15] to forecast solar irradiance. The goal is the optimization of power fluxes in off-grid applications.

The PV power production issue is debated in [16] with the use of data-driven methods and real data acquired at the Savona Campus, located in Italy, by the Energy Management Systems (EMS). Authors compare the Extreme Learning Machines, Random Forests and Regularized Least Squares methods to improve the EMS forecasting system.

An “analog ensemble” method is proposed in [17] and its forecasting results are compared with a quantile regression and an ANN ones. This model is based only on forecasted variables provided by a Numerical Weather Prediction (NWP) system. Results are validated using three datasets relative to three different sites. The model proposed in [17] is able to give a probabilistic forecast with a quantification of uncertainty associated to each prediction (prediction interval).