Solar Energy 150 (2017) 423–436

Contents lists available at ScienceDirect

Solar Energy

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

Multi-site solar power forecasting using gradient boosted regression trees

Caroline Persson [a,*], Peder Bacher [a], Takahiro Shiga [b], Henrik Madsen [a]

[a] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark
[b] Social Systems Research Domain, Toyota Central R&D Labs., Inc., Nagakute, Aichi, Japan

ARTICLE INFO

Article history:
Received 31 January 2017
Received in revised form 25 April 2017
Accepted 26 April 2017

Keywords:
Solar power forecasting
Multi-site forecasting
Spatio-temporal forecasting
Regression trees
Gradient boosting
Machine learning

ABSTRACT

The challenges to optimally utilize weather dependent renewable energy sources call for powerful tools for forecasting. This paper presents a non-parametric machine learning approach used for multi-site prediction of solar power generation on a forecast horizon of one to six hours. Historical power generation and relevant meteorological variables related to 42 individual PV rooftop installations are used to train a gradient boosted regression tree (GBRT) model. When compared to single-site linear autoregressive and variations of GBRT models the multi-site model shows competitive results in terms of root mean squared error on all forecast horizons. The predictive performance and the simplicity of the model setup make the boosted tree model a simple and attractive compliment to conventional forecasting techniques.

1. Introduction

Today's society is facing huge challenges in form of climate changes caused by increased emissions of greenhouse gasses into the atmosphere. Different initiatives in progress are aiming to lower human caused CO_2-emissions, e.g. the European goals of a CO_2-neutral Europe by 2050 (European Commision, 2011). The urge for a green transition has initiated research and development focusing on sustainable and green technologies. Investments are constantly made in the search for less expensive and more efficient technologies and strategies to promote the transition from the current dependence on fossil fuels to a future society relying on energy from renewable resources. For non-dispatchable renewables such as wind and solar power new challenges arise, since power generation from these sources depends heavily on climatological and meteorological conditions. This complicates the task of forecasting their power generation, which is important in order to schedule power generation and manage the allocation of reserve capacities. Thus, good forecasts of renewable power generation are necessary when striving for an efficient and cost beneficial incorporation of renewable power into the electricity market.

As the scale of energy systems influenced by variable energy sources is growing the need for efficient forecasting tools is increasing. It is no longer sufficient to only have small scale forecasting models for a single wind farm or photovoltaic (PV) installation. To optimize integration of renewable energy, forecasting models need to be able to deal with production spanning larger geographical areas, increasing amounts of historical data and computational time limits. Non-parametric machine learning algorithms have shown to be efficient in statistical modelling involving large amounts of data and highly non-linear patterns. Thus, it is relevant to explore the application of these methodologies in the field of solar power forecasting.

1.1. Bibliographic review

Several examples of non-parametric modelling techniques applied in the field of solar power forecasting exist in the literature. Sharma et al. (2011) use weather forecasts to predict hourly solar intensity as a proxy for solar power generation. The study compares multiple regression techniques for generating prediction models, including linear least squares and Support Vector Machines (SVM) using multiple kernel functions. Furthermore, dimensionality reduction is explored using Principal Component Analysis (PCA). The best performance is obtained by an SVM model with a Radial Basis Function (RBF) kernel being 51% more accurate than a naive persistence model (Antonanzas et al., 2016) in terms of root mean squared error (RMSE). The choice of an RBF kernel and the dimensionality reduction using PCA improved the SVM

* Corresponding author.
E-mail address: caroline.st.persson@gmail.com (C. Persson).

http://dx.doi.org/10.1016/j.solener.2017.04.066
0038-092X/

model significantly. Similarly, Ragnacci et al. (2012) propose a model based on SVM for forecasting of PV power, where multiple techniques for dimensionality reduction of the input variables are exploited. The best results are obtained with the supervised dimensionality reduction method Covariance Operator Inverse Regression (COIR) (Kim and Pavlovic, 2008). These results clearly support the importance of proper feature and model selection. This paper takes advantage of the automated feature selection of the Gradient Boosted Regression Trees (GBRT). By introducing a regularization term to the fitting algorithm of the GBRT, the feature selection is performed automatically (Hastie et al., 2011) and no preliminary dimensionality reduction step is needed. Moreover, in contrast to PCA and its variants, GBRT do not suffer from any loss of interpretation due to transformations of the input variables.

Marquez and Coimbra (2011) propose the use of Artificial Neural Networks (ANN) to forecast global horizontal irradiance (GHI) and direct normal irradiance (DNI) using weather forecasts as predictors. A preliminary feature selection is performed using a Genetic Algorithm (GA) and a Gamma Test. All ANN models proposed show major improvements over the reference model. The improvements are particularly accentuated for models forecasting DNI. Pedro and Coimbra (2012) fit several forecasting models, which predict the hourly PV power generation for one and two hours ahead only using endogenous variables. The methods studied in the paper are among others Autoregressive Integrated Moving Average (ARIMA), k-Nearest-Neighbors (kNN), ANN and ANN optimized by Genetic Algorithms (ANN/GA). The ANN based methods outperform the other models on several error metrics. Only in terms of mean bias error (MBE) the ARIMA model is slightly better for certain time intervals. The ANN/GA model performs better than the ANN in all tests, emphasizing the positive effect of optimizing the parameters and input variables of the ANN model. Finally, Pedro and Coimbra (2012) reports an improvement of 32.2% over a persistence model for the ANN/GA in terms of RMSE for forecasts one hour ahead.

As exemplified in Marquez and Coimbra (2011) and Pedro and Coimbra (2012) the use of ANN has resulted in good predictive performance. Some of the advantages of ANN models are their ability to model highly non-linear relations, moreover the models do not require any assumptions about the underlying process relating input and output variables. On the other hand, it is not always enough to simply produce good predictions. Often it is also desirable to have information providing qualitative understanding of the relationship between joint values of the input variables and the resulting predicted response value. Thus, black box methods such as ANN models (Hastie et al., 2011), which can be quite useful in purely predictive settings, do not represent the ideal methodology if a transparent interpretation of the models is desired. As shown later, it is possible to compute the importance of the input variables of the GBRT models. Visualization of the importance of input variables is a time efficient way to identify and compare dominant input variables across models, in this case models of different forecast horizons.

Zamo et al. (2014a) evaluate the performance of several different statistical model approaches including SVM, Binary Regression Trees, Random Forest (RF), GBRT and Generalized Additive Models (GMA) for forecasting of hourly PV power generation for lead times between 28 and 45 h, i.e. one day ahead with an hourly resolution. Predictors consist of 31 different outputs of a NWP model. The study evaluates the predictive performance of 28 individual PV plants, as well as the performance of predicting the aggregated power generation over all plants. The benchmarking designates RF as the best forecast model. For individual power plants, the median RMSE over 30-fold cross-validation for the RF model is between 10% and 12% of the maximum measured production. In general, this paper supports the application of decision tree models within the field of solar power forecasting.

Bessa et al. (2015) investigate the influence of including spatial and temporal information when forecasting solar power generation one to six hours ahead. Data consists of power series from 44 micro-generation units. The paper uses a vector autoregressive (VAR) framework, where lagged terms from the target installation and the neighboring installations are included as predictors resulting in a multi-output linear regression model. The VAR model is compared to a recursive autoregressive (AR) model with no spatial information. The results show that for point-forecasts the spatio-temporal VAR model on average outperforms the AR model with 12.5% for lead time one and 0.1% for lead time six in terms of normalized RMSE. Moreover, Bessa et al. (2015) apply gradient boosting to estimate the coefficients of a probabilistic VAR model and a VAR model with exogenous input (VARX). The boosting uses linear base learners and a quantile loss function to fit quantiles ranging between 5% and 95% with 5% increments.

Tree based models have also been applied for probabilistic forecasting of solar power generation, e.g. Zamo et al. (2014b) and Almeida et al. (2015) successfully apply Quantile Regression Forests to estimate quantiles of the PV power generation in order to produce probabilistic forecasts.

In this paper the application of regression trees, or more precisely, Gradient Boosted Regression Trees (GBRT) for prediction of future power generation from PV rooftop installations are thoroughly analyzed. Data consists of hourly power measurements from 42 individual PV installations and associated weather forecasts for their respective locations. The objective is to fit a model which is capable of predicting future power generation for a PV installation given its location, historical power generation and available weather forecasts. When using the location of the PV installations as inputs to the GBRT it is possible to make predictions in a multi-site framework. Instead of fitting an individual model for each PV installation the data from all 42 PV installations are used to fit one multi-site model. An individual multi-site GBRT model is trained for each forecast horizon resulting in six individual models. Together the combined predictions from these six models constitute the multi-site forecasts one to six hours ahead.

The remainder of this paper is structured as follows. In Section 2 the data and techniques for normalization are presented. In Section 3 the different models including benchmark models and the GBRT models are described. In Section 4 the intermediate and final results are presented. Furthermore, model selection, parameter tuning and feature exploration are thoroughly described. In Section 5 shortcomings and potential improvements are discussed. Finally, in Section 6 the paper is concluded and the major findings are summarized.

2. Data analysis and preprocessing

The data consists of historical power observations, information about the location of the PV installations and Numerical Weather Predictions (NWP). The data spans the period from April 19, 2014 to February 28, 2015. Power observations and weather forecasts are available in an hourly resolution. The data is divided into a training and a test set. The training set consists of the first 75% of the data (April 19, 2014 to December 12, 2014) and is further divided into K equally sized subsets without shuffling, referred to as the validation sets. The validation sets are used for K-fold cross-validation in order to optimize parameters for model selection as described in Section 4. The test set consists of the last 25% of the data (December 13, 2014–February 28, 2015) and is exclusively used for evaluation and comparison of the final models.