Comparison between different ML approaches for PV and STC production forecasting using real world data

Dea Pujić*a, Marko Jelić*a, Nikola Tomašević a
* School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia
a Institute Mihajlo Pupin, University of Belgrade, Volgina 15, 11060 Belgrade, Serbia
{dea.pujic, marko.jelic, nikola.tomasevic}@pupin.rs

***Abstract*—With the aim of improving ecological interest, the share of renewable energy sources (RES) in the energy production is to be increased. Nonetheless, that growth adversely influences the grid's instability, as a result of the dependency between the RES production and weather conditions. Therefore, with the aim of providing a stable energy system, it is necessary to plan the consumption in advance with respect to the availability of RES production. This paper is focused on comparing current SoA approaches for two different renewable energy sources, photovoltaic panels and solar thermal collector, using real world data from Denmark and Spain.**

I. Introduction

In the 20th century, electrical energy was produced mainly from fossil fuels. However, this created concerns about ecological environment – primary about greenhouse gas emissions, global warming and climate change. Therefore, in recent times, renewable energy sources, such as photo-voltaic (PV) panels, solar thermal collector (STC) and wind turbines (WT), were incorporated in the energy production as well, so as to decrease the use of fossil fuels. Nonetheless, as for the fact that their production highly depends on the weather conditions, this modification destabilizes the grid system. However, when a production forecaster for these sources is available, it is possible to plan and arrange the consumption with the goals of decreasing spending and environment pollution as well as improving other factors also. Consequently, in order to minimize the negative effects of the introduction of intermittent sources in the energy source portfolio, energy production forecasters became a key field of interest among researchers in this domain and are analyzed in this paper.

II. Brief state of the art analysis

Numerous renewable energy sources are available these days, but within this paper PVs and STCs will be covered. Additionally, methodologies differ based on the forecasting horizon. As this paper will focus on one day-ahead forecast, methodologies present in the field of interest will be briefly reviewed.

Different approaches in production forecasting can be found in relevant literature. Chronologically, physical models were first presented. They are defined using physical characteristics and mathematical equations. For STCs, they are mostly the only one currently present [1], [2]. On the other hand, regarding PV production forecasting, physical models are as a rule based on the electrical circuit models, as reviewed in [3]. However, these models require various physical parameters such as panel's surface area, position and angle of the panels, relevant voltages, currents, temperatures from data sheets etc. which often happen to be inaccessible. Therefore, with years, physical approaches were substituted with the statistical ones. Statistical models describe the system using variety of statistical characteristics and accessible data such as Autoregressive moving average methods (such as AR, MA, ARMA, ARIMA) explored in [4] and [5]. Finally, the most precise, especially in cases when huge amounts of data regarding energy production is available, are data driven models, which is why they are currently the most frequently analyzed in the literature. Various approaches can be found such as Neural Networks, supervised learning (linear regression, support vector machines, etc.), clustering and numerous hybrid ones as reviewed in [6].

Taking all of previous into consideration, the main focus of this paper will be developing and comparing different machine learning (ML) approaches for PV and STC production forecasting. The main contribution of this paper is analyzing the same models applied for estimating production of different renewable sources, as in literature comparisons are always given between the estimation performances for the same source. Additionally, for the following simulations, real world data from Denmark and Spain was used.

III. Methodology

To determine the best prediction model, different techniques were employed with each one of them being individually assessed by fine-tuning its respective parameters specific to its implementation. Starting with the optimal linear regression, its model is determined by testing different model orders and regression strength. The support vector regression (SVR) is optimized through the variation of the regularization parameter, the kernel coefficient and its independent term. As for neural networks, the structure of the network is varied through testing different layouts determined by the numbers of hidden layers, their arrangement and different numbers of neurons within them. The optimal random forest regressor is determined by testing different hyper parameters defining the maximum depth of the branching process, polynomial degree of input features and number of estimators. All of these parameters were examined using grid search with the goal of obtaining results which are not on the edge of the considered domain,

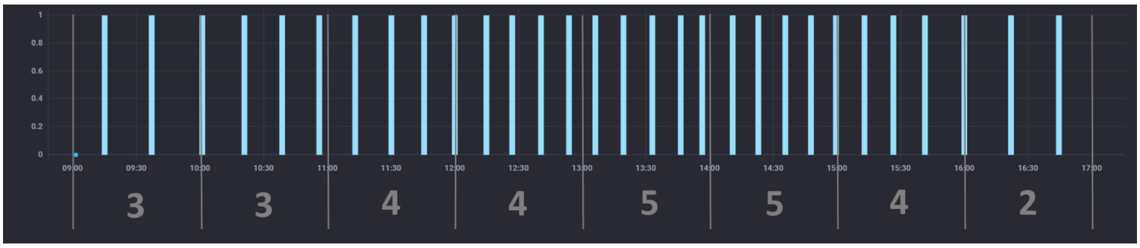

Figure 1. Example of STC production data representation

in order to be as sure as possible that the chosen parameters and architectures are optimal.

The performance of each of the models is evaluated using two standard metrics: the mean absolute error (MAE) and the mean squared error (MSE) between the values that are supposed to be estimated and the ones that are outputs of the model. The results table that will be presented later in the text only depicts models with the best performance for each of the techniques with their corresponding MAE and MSE values.

IV. Results

A. *Weather data*

After the models that were to be used have been chosen, in the first part of this section, selected input features will be discussed. Having in mind the nature of the considered renewable sources, it is undoubtable that weather conditions are highly correlated with the production that is to be estimated, which is why they are inevitable when picking the inputs. Expectedly, solar irradiation is by far the most correlated weather parameter with the desired output, which is why it was decided to include it as an input predictor. Consequently, the list of acceptable forecast weather services was limited, as this decision required providing 24-hour ahead irradiation forecast with hourly resolution. As an appropriate one, Weatherbit [7] online weather forecasting service has been chosen. Additionally, as the inputs apart from the aforementioned irradiation, information about humidity are considered, wind speed and direction, UV, outdoor temperature, cloud coverage and pressure were chosen and are also reported by Weatherbit. Finally, for STC, previous production has also been used.

B. *Production data*

Apart from the weather data obtained using the Weatherbit service, historical production data has is also required for models training purposes. As part of H2020 RESPOND project, this data was collected. Namely, hourly historical production data from the Danish pilot has been collected through the Evishine platform [8] which contains production data for different buildings from a couple of previous years. This data is scaled down to represent the amount of energy produced for users who are taking part in the project and an example of the one-day production measurements data. On the other hand, STC production data has been obtained directly from the RESPOND platform. Namely, as a part of the RESPOND project, STC production measurements are being collected through the sensor which sends a pulse after each kilowatt-hour of energy is produced, representing the amount of produced energy through the frequency of pulses, as shown in Figure 1.**Error! Reference source not found.** Unlike with the PV panels, whose production is influenced just by weather parameters and panel's physical characteristic, STC production could also vary depending on the user demand. In the simple case where the collector is directly connected or closely coupled with the user loop, the user's demand habits influence the production data. In other words for direct connections specifically, when the users require high volumes of thermal energy, the temperature of the return circuit to the collector is significantly lower that what is the case when little to no energy is required. This, in turn, influences the total energy leaving the collector through the direct circuit. On the other hand, if the collector is not directly connected with the user and there are one or more how water tanks for heat storage, demand is much less influential on the production since the tanks act as filters. Therefore, STC production data can vary drastically depending on the data source with topology playing a key role. In this paper, influence of the demand is negligible and the influence irradiation is significant, allowing for a precise estimator to be implemented.

C. *Hyperparameter selection*

After collecting the historical data for both weather and production outputs, the training process has been carried out in Python. All the data has been separated into three groups – training, validation and testing sets. Moreover, it was normalized by the standard deviation and centered using mean value of training data. Thereafter, as it has already been mentioned, the training process has been performed on various combination of model hyper parameters which are shown in Table I and Table II. For support vector machine models, three different kernel functions have been used – radial basis function (RBF), linear and sigmoid. For each of them, the regularization parameter has been varied in the interval from 1e-6 to 1 following a logarithmic law. Additionally, for RBF and the sigmoid one, kernel coefficients were taken from the interval between 1e-3 and 1e3, again following a logarithmic law, whilst the independent term for sigmoid kernel took values from the {0, 1, 10, 100} set. For linear regression, two hyper parameters were optimized – the regularization parameter (with values from 1e-4 to 1e3 following a logarithmic law) and the polynomial degree of input values taking integer values from 1 to 10. Following that, neural network models have been tested with 4 groups of hyper parameter being optimized – the regularization parameter, taking values from the {0, 1e-4, 1e-3, 1e-2, 1e-1} set, the polynomial degree of input being one, two or three, the number of hidden layers and number of neuron in each hidden layer. Architectures of hidden layers that have been tested were {30}, {40}, {40, 5}, {40, 10}, {40, 20}, {40, 40}, {50}. For KNN and KNN weighted models, the number of estimators was optimized taking integer values from the interval from 3 to 50. Finally, for the random forest model, the following parameters were tuned -

polynomial degree (integers between 1 and 9), number of estimators (values from {5, 10, 20, 50, 100, 200, 300, 500, 1000} set and maximum depth of each estimator (values from set {None, 5, 10, 20, 100}).

D. Model selection

Finally, for each methodology and source, the smallest MSE in percent, i.e. the performance of the model with the optimal hyperparameters is shown in Table I. It can be noticed that optimal methodology differs between the sources – for PV production forecasting NNs are found to be optimal, whilst for STC the RF model is the most suitable. Neural network that performed the best for PV production forecaster had **2 hidden layers** with **40** and **5 neurons** in each of them, respectively. It was trained using **regularization factor 0.001** and **polynomial degree 2**. On the other hand, optimal STC production forecaster had **50 estimators**, with **linear inputs** and **no** max depth.

TABLE I.
VALUES OF CONSIDERED HYPER PARAMETERS FOR EACH APPROACH

		Parameters							
		Regularization parameter	Kernel coefficient	Independent term in kernel	Polynomial degree	No. neurons in 1st hidden layer	No. neurons in 2nd hidden layer	Number of neighbors/estimators	Max depth
Methodology	SVR - RBF	1e-6 - 1	1e-3 – 1e3						
	SVR - linear	1e-6 - 1							
	SVR - sigmoid	1e-6 - 1	1e-3 – 1e3	{0,1,10, 100}					
	Linear regression	1e-4 – 1e3			1 - 10				
	Neural network	0 – 0.1			1 - 3	{30, 40, 50}	{5, 10, 20, 40}		
	KNN (weighted)							3 - 50	
	Random forest				1 - 9			5 – 1e3	{None, 5, 10, 20, 100}

TABLE II.
NUMBER OF DIFFERENT HYPER PARAMETERS AND MODELS THAT HAVE BEEN TESTED

		Parameters								**Number of models**
		Regularization parameter	Kernel coefficient	Independent term in kernel	Polynomial degree	No. neurons in 1st hidden layer	No. neurons in 2nd hidden layer	Number of neighbors/estimators	Max depth	
Methodology	SVR - RBF	7	7							49
	SVR - linear	7								7
	SVR - sigmoid	7	7	4						196
	Linear regression	8			10					80
	Neural network	5			3	7				105
	KNN (weighted)							48		48
	Random forest				9			10	5	450
										935

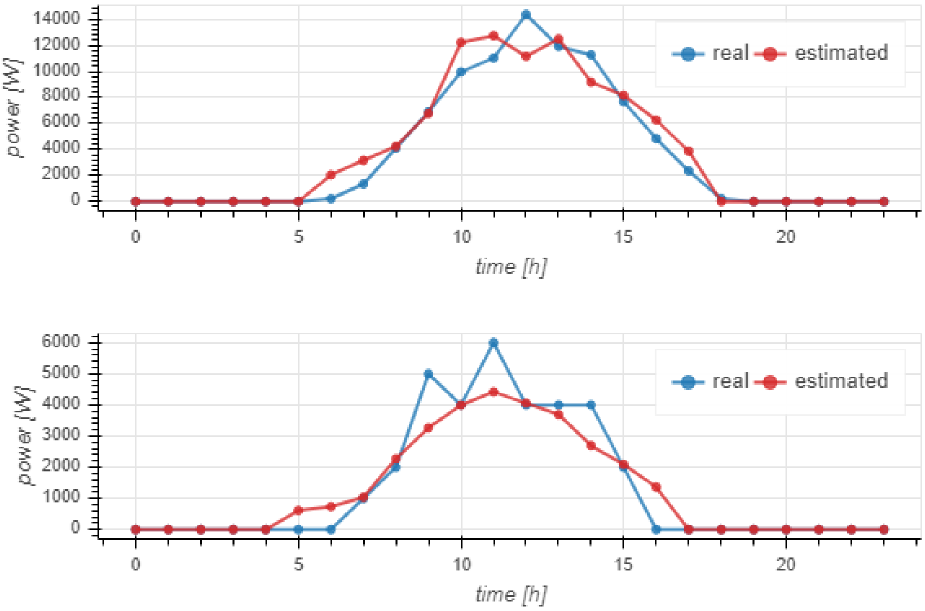

Figure 2. Examples of PV (top) and STC (bottom) production forecast for one day

TABLE III.
COMPARISON BETWEEN THE OBTAINED PERFORMANCES FOR PV AND STC FORECASTERS

		MAE [%]	
		PV	STC
Methodology	SVR - RBF	8.99	8.52
	SVR - linear	8.87	9.17
	SVR - sigmoid	8.79	9.01
	Linear regression	9.63	8.60
	Neural network	**8.5**	7.63
	KNN	8.75	7.90
	KNN weighted	8.45	7.85
	Random forest	8.61	**6.2**

Suitability of the chosen models for RES production forecasters is additionally illustrated by the fact that the second best performing model for PV is the RF whilst for STC it is the RF, distinguishing them as the most appropriate in context of day-ahead renewable energy production forecasting. Additionally, to further illustrate the obtained results, estimations for one day outputs for each resources are given in Figure 2, from where it can be confirmed that model estimations indeed follow the real production, and so, that the presented methodology can be exploited in real world practice. It can be noticed that in times in which there is no irradiation, the production is zero, which is the heuristic used to improve the performance of all the models. In other words, in the periods of the day when the global horizontal irradiation, obtained from Weatherbit, service is zero, the output of the estimator is set to be zero.

Finally, these models have been deployed and integrated as a part of a cloud platform which is intended to influence the users to increase energy savings and improve grid stability by responding to the demand response events. Therefore, its primary role was to provide necessary input for an energy dispatch optimization service which is supposed to provide an optimal curve, thus giving users guidance towards best possible behavior depending on the load curve, future RES production and pricing tariffs. Nonetheless, it turned out that end users are interested in more than the optimal load curve but the production forecast values themselves since these values are essentially giving useful feedback for them in order to adapt their habits and demand depending on the availability of renewable production.

Taking all of previous into consideration, it can be concluded that with machine learning algorithms, especially random forest and neural network models, high performances could be achieved for day-ahead RES production forecasting, in turn providing the end user with valuable information in order to motivate them to adapt the consumption and decrease the necessity of burning fossil through optimal utilization of renewable sources.

ACKNOWLEDGMENT

The research presented in this paper is partly financed by the European Union (H2020 LAMBDA project GA. No: 809965, H2020 RESPOND project GA. No: 768619, H2020 REACT project GA. No: 824395), and partly by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

REFERENCES

[1] P. Bacher, H. Madsen, and B. Perers, “Short-Term Solar Collector Power Forecasting,” in *Proceedings of ISES Solar World Conference 2011*, 2011, Accessed: Jan. 15, 2020. [Online]. Available: https://orbit.dtu.dk/en/publications/short-term-solar-collector-power-forecasting.

[2] E. W. Law, A. A. Prasad, M. Kay, and R. A. Taylor, “Direct normal irradiance forecasting and its

application to concentrated solar thermal output forecasting – A review," *Solar Energy*, vol. 108, pp. 287–307, Oct. 2014, doi: 10.1016/j.solener.2014.07.008.

[3] A. Dolara, S. Leva, and G. Manzolini, "Comparison of different physical models for PV power output prediction," *Solar Energy*, vol. 119, pp. 83–99, Sep. 2015, doi: 10.1016/j.solener.2015.06.017.

[4] Rui Huang, T. Huang, R. Gadh, and Na Li, "Solar generation prediction using the ARMA model in a laboratory-level micro-grid," in *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*, Nov. 2012, pp. 528–533, doi: 10.1109/SmartGridComm.2012.6486039.

[5] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, Nov. 2013, doi: 10.1016/j.rser.2013.06.042.

[6] C. Voyant *et al.*, "Machine learning methods for solar radiation forecasting: A review," *Renewable Energy*, vol. 105, pp. 569–582, May 2017, doi: 10.1016/j.renene.2016.12.095.

[7] "Weatherbit | Weather API - Historical Weather API." https://www.weatherbit.io/ (accessed Jan. 15, 2020).

[8] "eviShine." https://evishine.dk/ALBOA (accessed May 28, 2020).