Application of similarity analysis in PV sources generation forecasting for energy clusters

Radomir Rogus[1,*], *Maciej* Sołtysik[1], and *Rafał* Czapaj[1]

[1]PSE Innovation Ltd, Al. Jerozolimskie 132, 02-305 Warsaw, Poland

Abstract. The resultant photovoltaic installation powers significantly affect the process of cluster coordination in terms of balancing, which is associated with the need for the most accurate forecast of photovoltaic generation. This article describes the application of similarity analysis in order to use commonly available meteorological data for predicting generation level from photovoltaic sources on the example of several selected installations and their corresponding real production profiles.

1 Introduction

In the light of constantly increasing penetration of the Renewable Energy Sources (RES) in the European Union (Fig. 1), the proper management of the Power Systems (PS) requires development of the prediction tools dedicated for rapidly developing, but instable technologies, such as wind turbines or Photovoltaics (PV).

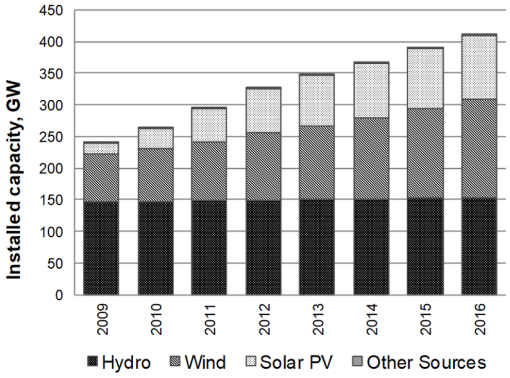

Fig 1. Evolution of installed RES capacity in EU-28 countries (European Commission, 2018).

The improper and inaccurate forecasts may entail numerous detrimental consequences, influencing the financial and technical performance of PS participants [1]. This relation is observed both in the macroscale and microscale, impacting the largest entities in the system, as well as the local individual participants. Considering the upraising concept of the

* Corresponding author: author@email.org

energy clusters in Poland, the integration of particular receiving and generating installations within clusters becomes one of the key activities of the Cluster Coordinator (CC), which are necessary to sustain the power balance and ensure profitability of the entire energy cluster.

Since the PV installations presently constitute one of the most dynamically developing RES sector in Europe and are the most common electricity-generating technology in prosumer-households (often being cumulatively a considerable power source in the energy clusters), the aim of this paper was to construct a predictive model for PV power generation within a hypothetical energy cluster.

The layout of the paper was ordered as follows: firstly, the main concepts of the energy clusters were explained, together with the main responsibilities of the CC from the perspective of Energy Cluster balancing and planning. The next section describes the main distinguishable approaches in modelling the PV generation, as well as the listing of the key meteorological factors influencing the performance of the PV generation units. Afterwards, the methodology applied in forecasting process was presented, along with the data selection process. The final section covers the presentation of the results accompanied by relevant observations and conclusions.

2 Coordination of energy cluster

By the definition, the energy cluster is an agreement of subjects concerning generation and balancing of the energy demand, as well as energy trading and distribution within the Low Voltage network on the terrain no larger than one county or five communes. The energy cluster is represented by the elected CC, whose role is of key importance from the perspective of sustaining adequate and efficient functioning of the energy cluster in the way representing businesses of all its participants [2].

Among others, to the main responsibilities of the CC belong inter alia (i) trading on the energy market in the name of entire cluster, (ii) balancing the power demand and generation capabilities within the cluster, (iii) assuring proper planning of energy carriers delivery from the external suppliers and (iv) supervision on the internal distribution systems [2]. Considering the mentioned duties, the CC has to be equipped with a reliable source of information providing with the expected cumulative power output of the generation sources within the energy cluster. This paper aims particularly at the prediction of the total PV power output from panels located on the terrain of the energy cluster.

3 PV generation forecasting models

The multiplicity of factors impacting the performance of the PV plant translates into large number of distinctive approaches in the construction of pertinent forecasting models [3]. Thus, the choice of a model characterized by the highest achievable accuracy, but burdened with reasonable computational complexity becomes a demanding challenge for the scientists involved in the topic [4]. As presented in [4] there are two main types of the reviewed PV generation forecasting models: direct and indirect. The former one encompasses models predicting the power generation sourcing from historical production and weather data, while the latter approach is based on the single input expressed in solar radiance, which in turn is an independent subject of forecasting processes (e.g. Numerical Weather Forecasts). The studies comparing the performance of the direct and indirect forecasting methods for PV generation reveal the advantage of the direct methods as the more accurate and reliable techniques [4]. Although, the applicability of a given forecasting method is limited by the access to the potential input data and horizon of the forecast [3].

Besides the mentioned time horizon of the forecast (short, medium and long-term), the further sub-classification of the PV generation predictive models concerns the methodology used, from which one can mention, for example (i) persistence model, (ii) statistical models, (iii) Machine Learning models and (iv) and hybrid models, which internally combine and capture the most advantageous features of individual models (i-iii) [3].

Among the weather parameters influencing the performance of the PV panels, the solar irradiance is accordingly recognized as the most important one, revealing a linear relation with the PV power output [4]. Further, the measures such as cloud cover, air temperature, air humidity, wind speed and direction are also considered as determinants for predicting the energy generation of the PV plant. However, the influence of a particular parameter on the PV performance is not unequivocal; it is emphasized that the effect of an individual parameter may differ with respect to the overall climate conditions, geographical location or design characteristics of applied technology [3,4].

4 Data used

The dataset used for prediction of the generated power comprised of historical generation profiles of three PV installations located in Katowice, in the South of Poland in hourly fragmentation, recorded in 2017 calendar year (Fig. 2).

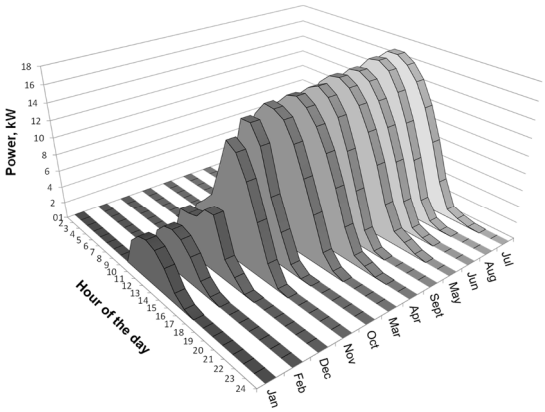

Fig. 2. Selected exemplary generation profiles of PV plant in 2017.

For the purposes of forecasting, the PV generation profiles data were supplemented by the meteorological records sourced from the Institute of Meteorology and Water Management [5] from the measurement station located in Katowice, for the calendar year of 2017 in hourly granulation as well (Fig. 3).

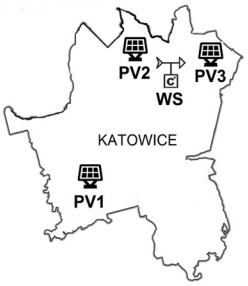

Fig. 3. Locations of analyzed PV plants and Weather Station.

The meteorological data consists of the following parameters: (i) visibility (discrete scale from 0 to 8), (ii) overcast (discrete scale from 0 to 8), (iii) wind speed (km/h), (iv) wind direction (degrees), (v) air temperature (°C), (vi) relative humidity (%), and (vii) pressure on the level of the station. As the completion of the data, the information about the sunrise and sunset time was collected in order to preliminarily point out the instances of data when the overall production will be certainly equal to zero. The '*hour of the day*' attribute was replaced by the '*time to noon*' attribute, which in better way reflects the maximum available peak power at a given time of the day.

The analyzed installations differ from each other by the set-up date, type of the panels, achievable power and the inverter used, what is presented in the table 1.

Table 1. Characteristics of analyzed PV plants.

Parameter	PV1	PV2	PV3
Start date	07-30-2015	04-04-2014	04-14-2015
Installed Power, [kW]	20.0	100.5	28.0
Modules	Kingdom Solar KD-M250	Canadian Solar Inc. CS6P-240P	Canadian Solar Inc. CS6P-250P
Inverter	Sunny Tripower 17000TL-10	Sunny Tripower 17000TL-10	Sunny Tripower 17000TL-10

As it can be read from the table 1, the representative PV plants vary significantly in terms of the scale and the power, since they serve in both commercial and private usage manner, what reflects the diversity of energy suppliers in the energy clusters.

4 Methods

For the purpose of the study, it was decided to apply the Decision Tree (DT) algorithm, possible to use within MS SQL Server Analysis Services. Before model estimation, the selection of input variables was conducted in order to include only the information-contained variables. The selection of input attributes was based on correlation coefficient criterion, Naive Bayes (NB) posterior probability criterion, expertise of the authors and naive approach, saying that all available data should be included in the model.

Naive Bayes algorithm (classifier) is a classification technique which assumes the lack of relation between the predictors. In other words, the information contribution of each feature to the model is independent. The NB algorithm seeks for the classification of the predicted feature with maximum a posteriori probability (eq. 1,2) [6].

$$PP(C|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C)P(C)}{P(\boldsymbol{x})} \tag{1}$$

$$P(C|\boldsymbol{x}) = \prod_{i=1}^{n} P(x_i|C) \tag{2}$$

where: $P(C|\boldsymbol{x})$ – posterior probability of class C given predictor $\boldsymbol{x}$, $P(C)$ – prior probability of class C, $P(\boldsymbol{x}|C)$ – probability of predictor given class, $P(\boldsymbol{x})$ – prior probability of predictor. As the criterion for comparison of relevance of the predictors, the confidence score is used, which takes the value from 0 to 1 and reflects the importance of a given feature in forecasting the output [7].

The scheme of DT represents the breakdown of the values of a feature in classes. The nodes of a DT represent the way the breakdown is made, while the leafs correspond to the classes for which the particular instances of attributes were assigned. The features can be sub-classified in terms of the datatype, which can be qualitative or quantitative, what enforces the usage of different approaches in training process of the model. The training (learning) process aims to reflect the model output most accurately by classifying the historical data. After the learning process the new instances of data are subjected to the constructed classifier (validation phase), what enables to measure the model's effectiveness. Despite the variability of developed DT algorithms, all of them generally follow the four key steps in construction of the model, which are [8]: (i) selection of the most information-valuable attribute (root), (ii) expansion of the DT by adding the to the root branches representing the breakdown of the attribute's values, (iii) assignment the instances of data with respect to defined classes, (iv) if all the instances belong to a single class, termination of the breakdown and ending the branch with a leaf with assigned class; otherwise, recurrent expansion of the tree until all the instances have an assignment. Although, on the way of set-up of the DT following this top-down approach, several issues have to be resolved, which include the recognition of the most valuable attribute, the overfitting and definition of splitting point in the case of continuous numeric features [8]. One of the most widely applied methods in DTs is the ID3 algorithm developed by J.R. Quinlan, which incorporates the concepts Information Gain (IG) and entropy in order to construct the tree. The IG is a parameter allowing to select the most information-contained attributes and is expressed as the difference of entropy between the parent node and child nodes [8]. In the case of selecting the root in a DT, the IG is calculated for each considered attribute as it is presented in eq. 3 [8].

$$IG(A) = E(S) - E(A) \tag{3}$$

where: A – attribute, E – entropy, S – learning dataset. The entropy of the entire set $E(S)$ is obtained according to equation 4.

$$E(S) = \sum_{i=1}^{k} -\frac{n_i}{n} log_2(\frac{n_i}{n}) \tag{4}$$

where: k – number of subsets of the predicted attribute, n_i – number of instances belonging to a i-th subset. The entropy of a single attribute, which divides the S set into subsets S_v, each having m_j elements is then obtained by applying the equation 5.

$$E(A) = \sum_{j=1}^{v} -\frac{m_j}{n} E(S_j) \tag{5}$$

The entropy measures the homogeneity of a sample. Its value can take the value from 0 to 1, depending on the uncertainty probability of a given statistical event [8].

The selection of the node attribute is made by comparing the IG for each of the attributes and typing the one with the highest IG value. This procedure carries on recursively as long as the addition of nodes brings the increment of the IG value [8]. Although, it should be pointed that the abovementioned procedures find and application only for discretized data. In the case of continuous, numeric data, such as e.g. temperature, one of the solutions proposed is Least Squares Regression. Basically, the values of an attribute are divided into partitions, for which an individual function is estimated based on Least Square Error (LSE) criterion [9].

$$\sum_{y_i \in p}(y_i - y_i^*(p))^2 \rightarrow min \quad (6)$$

where: y_i – instance value, y_i^*– estimated value of instance y_i in partition p. The splitting point is sought iteratively until obtaining the attribute breakdown characterized by the lowest aggregated LSE in the node [9].

Another aspect of the DT which must be taken into account when constructing the forecasting model is the overfitting, which is an excessive adjustment of the model to the learning dataset. Too detailed decomposition of the tree may lead to relatively small errors of the model with respect to the learning data, but may reveal significant inaccuracies when subjected to the validation data. For avoiding this adverse effect, a common method constitutes the pruning [10]. The overfitting of the DT can be handles in two ways of the pruning: (i) pre-pruning, which prevents the excessive growth of the DT by means of pre-defined stopping criteria, or (ii) post-pruning, which reduces the number of nodes of the DT after overfitting took place [11].

In the range of available post-pruning methods, the Minimum Error Pruning approach which looks for a DT structure showing the lowest value of expected error when subjected to the testing dataset. The error estimate is derived by applying formula shown in the equation 7 [10].

$$e(t) = \frac{n_t - n_{t,c} + k - 1}{n_t + k} \quad (7)$$

where: k –number of class, n_t – number of instances in node t, $n_{t,c}$ – number of instances assigned to class c in node t. The comparison of the expected error values before and after pruning allows to decide whether the pruning is justified or not [10].

The sequence of steps leading to forecast the PV power within a cluster is presented in the figure 4.

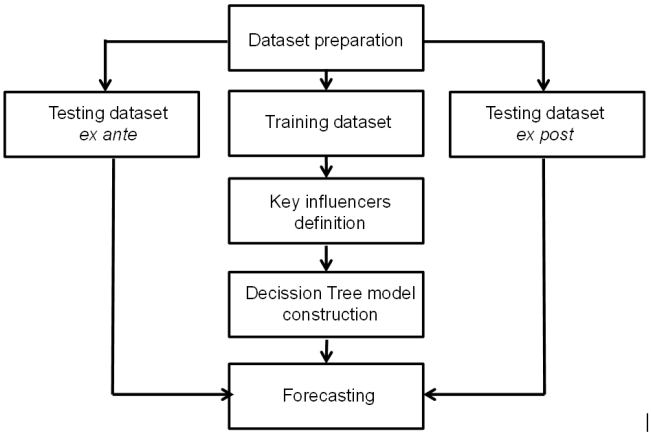

Fig. 4. PV power forecasting process flow chart.

With use of the sunrise/sunset information for each day, the instances for which the PV production are certainly zero were distinguished from the productive hours by defining a correspondent binary variable. The data is then divided into learning set (70% of total records) and testing set (30% of total samples). The testing set that is consisted of randomly selected records from the 8760 instances in the entire year 2017, for which the ex post error was derived. Based on data selection criteria indicating the most information-contained

attributes, five individual cases were examined, what was followed by constructing and estimating the DT model. In order to test the reliability of the models, they were additionally queried with the use of 744 sequential data records from January 2018 to estimate the ex ante error. Evaluation of the constructed DT were performed by calculating the Mean Absolute Error (MAE) and normalized Root Mean Square Error (nRMSE) for the testing samples, according to equation 8 and 9 [3].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i^* - y_i| \tag{8}$$

$$nRSME = \frac{1}{y_{max}-y_{min}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^* - y_i)^2} \tag{9}$$

where: N – number of records in testing sample, y_i^* forecasted power, y_i – actual value of power.

5 Forecasting results and conclusion

The table 2 below shows the data selected to the DT model in each analyzed case, for which the chosen variables were further processed by the algorithm. The forecasting accuracy measures results were presented as well.

Table 2. Description of considered cases in DT modelling and corresponding forecasting results.

Attribute	Data selection cases				
	R^2>0.5 (C1)	R^2>0.3 (C2)	Naive Bayes (C3)	Expert (C4)	Naive (C5)
Sunrise		x		x	x
Sunset		x			x
Visibility		x	x	x	x
Overcast		x	x	x	x
Wind direction					x
Wind speed				x	x
Air temperature	x	x	x	x	x
Humidity	x		x	x	x
Pressure					x
Time to noon	x	x	x	x	x
nRMSE *ex post* [%]	9.24	7.64	8.13	8.09	7.13
MAE *ex post* [kW]	5.61	4.38	4.87	4.80	4.31
nRMSE *ex ante* [%]	9.94	16.96	12.38	13.52	13.40
MAE *ex ante* [kW]	3.00	4.86	3.59	3.91	4.06

Due to the complexity of the constructed DT model, it could not be presented illustratively in this paper; the number of created levels was varying between 10 and 13 layers. In each of the considered cases the root node of the DT splitted with regard to the binary variable distinguishing the productive/non-productive time of the day. Secondarily,

the most important parameters were 'sunset/sunrise', which in a good way reflected the season of the year, followed by relative humidity and air temperature. The overcast, which intuitively was considered as the key meteo-input to the model occurred to be not reliant variable. This could come from the fact that the cloud cover is a very dynamic parameter, which expressed in the hourly fragmentation can be burdened by high error.

Additional source of uncertainty could result from the dissimilarities in production profiles of the analyzed PV plants – the correlation coefficient for generation profiles of two chosen PV plants has not exceeded the value of 0.97 in any case, what indicates the ambiguities in the performance of analyzed installations. The obtained errors of PV forecasting demonstrate relatively high values comparing to related works presented in the literature. Nevertheless, these works were concerning the performance of a single farm, usually with adjoined weather station, capable to measure the irradiation at the spot. In this case, the PV stations were dissipated on the 160 km^2 city area and the model was basing on a single weather station.

Analyzing the errors of the DT model for each of the cases, reverse tendencies can be observed in models' accuracy when comparing the *ex post* and *ex ante* results. For the *ex post* case, the lowest values of errors are observed for model supplied with maximum number of attributes. On the other hand, the three-inputs DT models demonstrated the most accurate results for the *ex ante* testing dataset. Nonetheless, it should be underlined that in the first case, the data consisted of random results from the entire year, while the results of *ex ante* simulation were obtained for time-sequential records of the month of January 2018, when the overall production is relatively low. This leads to the conclusion that the weather conditions play more significant role when predicting the relatively low values of power in winter time.

Further development of the forecasting concept presented in this paper will include the expansion of the single analyzed area of cluster and/or creation of model for predicting the PV output from a group of clusters.

References

1. C. Monteiro, T. Santos et. al., Energies **6**, 2624-2643 (2013)
2. Consortium: KAPE, WiseEuropa, Atmoterm, KIER, *Koncepcja Funkcjonowania Klastrów Energii w Polsce (*eng. *The concept of functioning of Eneregy Clusters in Poland)*, (2017)
3. C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, Z. Hu, , CSEE JPES **4**, 38-46 (2015)
4. U. K. Das, K.S. Tey et al., Renew. Sust. Energ. Rev. **81,** 912-928 (2018)
5. Publicly available meteorological measurement and observation data, *Instytut Meteorologii i Gospodarki Wodnej,* https://danepubliczne.imgw.pl/
6. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Second Edition* (Springer, 2008)
7. A. Mandelbaum, D. Weinshall, Computing Research Repository, abs/1709.09844 (2017)
8. J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się, wydanie drugie*, (eng. *Statistical learning systems, second edition),* (Exit, Warsaw, 2008)
9. W. Rea, M. Reale, C. Cappelli, J. A. Brown, Economic Reviews **29**, 754-777 (2010)
10. E. Gatnar, *Symboliczne metody klasyfikacji danych (*eng. *Symbolic methods of data classification),* (PWN, 1998)
11. N. Patel, S. Upadhyay, IJCA **12,** 20-25 (2012)