Photovoltaic power generation data filling model based on tensor decomposition

Xiangying Xie¹, Hu Liu², Xiaoguang Ma¹, Dayan Ma¹, and Xin Luo¹
¹ State Grid e-commerce Co., Ltd, Beijing 10053, China
² State Grid Corporation of China, Beijing 100031, China
³ Corresponding author’s e-mail: xiexiangying@sgec.sgcc.com.cn

Abstract. With the development of new energy and the increase in demand, photovoltaic power generation has received more and more attention. However, due to the susceptibility of photovoltaic power generation to the influence of weather factors, especially the influence of solar irradiation, it is intermittent and fluctuating. Understanding the power generation data of a PV plant has a great effect on the subsequent photovoltaic power generation planning and photovoltaic power plant situation assessment. However, these data are often incompletely recorded due to frequent communication problems or equipment failures. Most current studies focus on photovoltaic power generation predictions, it is still a challenge to speculate the missing data around photovoltaic power plants. In this paper, a context-sensitive tensor decomposition model is proposed to intelligently fill the missing power generation entries. The experimental results show that our model can effectively fill missing data and obtain good results.

1. Introduction
In recent years, with the development of all walks of life, the demand for electricity is growing. However, the limitation of conventional energy sources and environmental problems have become increasingly prominent, new energy sources with environmental protection and renewable qualities have entered people's vision. The research and industrial development of new energy is an effective supplementary means of the entire energy supply system and the ultimate option to meet the sustainable development needs of human society. The use of solar energy is one of the important ways. Because solar energy is inexhaustible, safe and reliable, it is the one of the most desirable green sources among many renewable energy sources [1]. Photovoltaic (PV) power generation is a popular solar application method that converts solar energy into electrical energy. China's PV industry continues to expand, and the Chinese market is expected to account for more than 40% of the global market in the next three to five years. For a given photovoltaic system, the output power depends on the weather conditions described by many parameters, such as solar irradiance, ambient temperature, and atmospheric pressure. PV systems are susceptible to random and uncontrollable weather changes [2], and its output power is obviously intermittent and random, which will cause uncertainty in the electricity supply from solar power plants. High uncertainty in supply will increase the possibility of an unexpected power dispatch confusion, which will result into the voltage and frequency of the supply network exceeding safe operating limits, thereby reducing its reliability and safety [3].

Due to the randomness and intermittence of photovoltaic power generation, if the proportion of photovoltaic power generation exceeds 15%, it may cause the grid to be paralyzed [4]. The stable grid-connected operation of the grid requires coordination between different power generation methods. The
adjustment of traditional power generation methods (such as hydropower and thermal power) is very costly. Therefore, mastering the photovoltaic power generation situation and then scientifically correcting the power plan is a reliable way to ensure the safe and economical operation of the power system. In the actual situation, the relevant data records of photovoltaic power generation such as power generation and power generation will be lost due to power station communication failures. Reasonable speculation of these missing entries becomes crucial, and traditional methods require a lot of additional data (such as weather factors) to calculate, increasing the difficulty.

In this paper, we propose a tensor decomposition based model to infer the missing power related data of arbitrary PV stations at different time intervals in the past day. This model only uses historical power generation data of the power station, which reduces the difficulty of obtaining the data set. At the same time, our model obtained excellent results in the speculation experiment on power generation data at the power station level and equipment level.

2. Related work
The estimation of power plant power generation data is conducive to the stable operation of the grid and reduces safety risks. The importance of the issue has boosted the development of many studies worldwide to obtain accurate estimation. Now many studies focused on the prediction of photovoltaic power generation, which can be roughly classified into indirect and direct methods [5]. The indirect forecasting method is generally divided into two steps: firstly, establishes a model for predicting the solar irradiance, and then a model for forecasting the photovoltaic power generation is built based on the predicted value of the solar irradiance and the actual condition of the photovoltaic power plant. And the prediction of solar irradiance is based on numerical weather prediction (NWP). The main disadvantage of this type of model is highly dependent on NWP, which lack of sufficient space-time resolution is one of the main sources of error in this method [6].

The direct prediction methods is generally based on the historical data of photovoltaic power generation as the input to the model which uses the relevant prediction theory and prediction algorithm to make output prediction. Related prediction theories mainly include statistical methods, machine learning methods and hybrid methods. The statistical methods include Autoregressive moving average (ARMA), Autoregressive Integrated Moving Average model (ARIMA), and Markov chain. Optimization of the set of inputs can be performed by the application of GA[7], Genetical Swarm Optimization (GSO) [8], fuzzy logic [9], Principal Component Analysis (PCA)[10] or firefly optimization[11]. The main shortcoming of the ARMA model is that the used time series data must be stationary. [12] uses ARMA model which consider both lagged past values and errors;[13] designed ARMAX model based on ARMA model which allows the model to receive external weather data and improve the accuracy of the prediction model;[14] use Markov chain to predict PV power generation which get acceptable performance.

The machine learning methods mainly include Artificial Neural Networks (ANNs) based methods and Support Vector Machines (SVM) models etc. ANN learns the nonlinear mapping between the input and output of the data samples through the relevant training algorithm, and then uses the learned nonlinear mapping to achieve prediction. However, the main drawbacks of ANN are requiring a large number of data during the training process and overfitting. [15] uses ANN to predict insolation by computer simulations by taking the insolation of each month into consideration. SVM stands out for their strong generalization capacity and for their ability to deal with non-linear problems. Three parameters dominate the performance of the model [16], which are penalty (C), tube radius (ε) and the kernel function's parameter. They jointly determine the performance of the model, which is also the limitation of SVM based model [17] and [18] uses SVM based model which has shown great potential in several studies.

Hybrid models combine technologies to better utilize each individual topology and enhance its advantages and ensure the better forecasting performance. [19] used ARIMA with ANN model which utilizes daily weather forecast to predict the 24 hour advance random energy output of the photovoltaic system and adapts to less weather parameter data.[20] used fuzzy inference model with RNN which
fuzzy inference model has been utilized to smoothen the meteorological data to forecasting PV power generation. However, due to the use of multiple technologies, the computational complexity of the hybrid model increases, while its performance is affected by the choice of a single technology that may perform poorly.

As far as we know, there are still few studies on the filling of photovoltaic power generation data, and the tensor decomposition method has not been widely applied in the prediction of it, but it performs well in missing data inference and can adapt to sparse data [21]. This paper proposes a tensor decomposition based method to complete the data at the missing moment.

3. Context-sensitive Tensor Decomposition

3.1. Tensor construction

We model the photovoltaic power generation data for each site using a tensor, \( \mathcal{X} \in \mathbb{R}^{N \times M \times L} \), with three dimensions denoting \( N \) sites, \( M \) time slots, and \( L \) days, respectively:

- **An Entry:** An entry \( \mathcal{X}(i, j, k) \) stores the PV power generation data of site \( s_i \) in time slot \( t_j \) and day \( d_k \) over the given period of time (e.g., a month). For the missing entries we will filled with an inferred value in the experiment. The value of each entry in tensor \( \mathcal{X} \) is then normalized to [0, 1] for decomposition.

- **Site dimension:** In the station-level power generation data speculation experiment, we regard each photovoltaic power station as an experimental site, and in the device-level power generation data speculation experiment, each power station has multiple devices and each device is regarded as an experiment site, \( s = \{ s_1, s_2, ..., s_i, ..., s_N \} \).

- **Time span dimension:** We divide a day into equal slots, \( t = \{ t_1, t_2, ..., t_j, ..., t_M \} \). Each time slot lasts for a fixed period of time, such as an hour or five minutes. Because the solar irradiance can only be collected during the day, which affect PV power generation, so we only take the time slot of 6:00am -7:00pm for the experiment.

- **Day dimension:** The third dimension denotes days, \( d = \{ d_1, d_2, ..., d_k, ..., d_L \} \).

3.2. Tensor decomposition

A common way to fill the missing entries of the tensor is to decompose it into the product of several (low-rank) matrices and a core tensor based on its non-zero entries using a tucker decomposition model.

\[
\mathcal{X} \approx \mathcal{G} \times_1 S \times_2 T \times_3 D = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr} s_p t_q d_r = \left\langle \mathcal{G} ; S, T, D \right\rangle
\]

\[
S \in \mathbb{R}^{I \times P}, T \in \mathbb{R}^{J \times Q}, \text{ and } D \in \mathbb{R}^{K \times R} \text{ are the factor matrices and can be seen as the principal}
\]
components in the respective tensor mode. The tensor $\mathcal{G} \in \mathbb{R}^{P \times Q \times R}$ is the core tensor and its entries indicates how different components interact with each other. The Symbol “$\times$” denotes the matrix multiplication, $x_1$ stands for the tensor-matrix multiplication, where the subscript I stands for the mode of a tensor, e.g., $H = \mathcal{G} \times_1 S$ is $H_{ijk} = \sum_{i=1}^{P} \mathcal{G}_{ijk} \times S_{ij}$. The symbol “$\circ$” represents the vector outer product. This means that each element of the tensor is the product of the corresponding vector elements.

$P$, $Q$, and $R$ are referred to as the number of components (i.e., columns) in the factor matrices $A$, $B$, and $C$, respectively. If $P$, $Q$, $R$ are smaller than $I$, $J$, $K$, this will result in a compression of $\mathcal{X}$, which means the core tensor $\mathcal{G}$ will be thought of as a compressed version of $\mathcal{X}$. In our cases, $P$, $Q$, and $R$ are very small.

The optimization problem that we wish to solve is

$$\min_{\mathcal{G}, S, T, D} \| \mathcal{X} - [\mathcal{G}; S, T, D] \|$$ (2)

According to Equation (1), we can rewrite the objective function of the optimization tensor decomposition as:

$$Loss_i(\mathcal{X}) = \| \mathcal{X} - \mathcal{G} \times_1 S \times_2 T \times_3 D \|_2^2 + \frac{\lambda}{2} (\|\mathcal{G}\|_2^2 + \|S\|_2^2 + \|T\|_2^2 + \|D\|_2^2)$$ (3)

Where $\|\cdot\|_2$ denotes the $l_2$ norm; the first part is to control the decomposition error and $\frac{\lambda}{2} (\|\mathcal{G}\|_2^2 + \|S\|_2^2 + \|T\|_2^2 + \|D\|_2^2)$ is a regularization penalty to avoid over-fitting; $\lambda$ is a parameter controlling the contribution of the regularization penalty; $i$ represents the i-th optimization. By minimizing the objective function, we can get optimized $S$, $T$, and $D$. Afterwards, we can recover the missing values in $\mathcal{X}$ by Equation (1).

We use numerical methods of gradient descent to find local optimal solutions when $Loss_i(\mathcal{X}) - Loss_{i-1}(\mathcal{X}) < \varepsilon$, i.e., when the objective function begins to converge. We use an element-wise optimization algorithm [23] to update each element of the tensor individually.

4. Experiments and results

4.1. Dataset

We used data about power generation from several PV plants in July 2019. We have three categories of power generation data: hourly power generation data, power generation efficiency data, and cumulative power generation data. Each type of data is divided into station level and device level. The time interval of hourly power generation data is 1 hour, and the time interval of power generation efficiency data and cumulative power generation data is 5 minutes. Also, due to the influence of solar irradiation, the generation and power after sunset is almost zero and the cumulative generation variation is small. In order to reduce interference, only the data from 08:00am-7:00pm were selected. The number of sites in the different experiments also varies depending on the data recorded by the power plant, and Table 1 records the scale and tensor’s dimension of the data set for each experiment.

| Dataset                     | Scale         | dimension          |
|-----------------------------|---------------|--------------------|
| Hourly power generation     | 1122          | 4484               |
| power generation efficiency | 24294         | 101252             |
| cumulative power generation| 23791         | 60523              | 12 × 14 × 30         |
|                            | 12 × 14 × 30  | 6 × 168 × 30       |
|                            | 15 × 14 × 30  | 6 × 168 × 30       |

4.2. Evaluation metric

In this paper, two metrics are used to evaluate the forecasting accuracy of the model. One is the Root Mean Square Error (RMSE) (Equation (5)), and the other is the Mean Squared Error (MSE) (Equation (6)). Smaller values of the error metrics indicate higher forecasting accuracy.
\[ MAE = \frac{\sum |y_i - \hat{y}_i|}{n} \]  
\[ RMSE = \left( \frac{\sum (y_i - \hat{y}_i)^2}{n} \right)^{1/2} \]

Where \( n \) is the total number of forecast values, \( y_i \) and \( \hat{y}_i \) denote the \( i \)-th entry’s actual and forecast value, respectively.

4.3. Photovoltaic power generation data fill experiment

We evaluate the tensor decomposition based model by randomly removing 30% non-zero entries from the tensor and filling in these entries using our model. We then use the original values of these entries as the ground truth to measure the inferred values. The average method means the value in empty entry will be averaged from the power generation data at the same time slot on the previous day and the following day of this PV station or averaged from the power generation data at the same time slot on the previous two day for the same PV station.

Our final experimental result is calculated by the above method and corresponds to the following parameters: \( P=Q=R=10, \epsilon=0.001 \). As the results demonstrate in Table 2, our model can fill the missing entries in the data, which achieved good result and brings great improvement comparing with the average method.

| experiment                                      | MAE       | RMSE      |
|------------------------------------------------|-----------|-----------|
| Hourly power generation                        | 0.1098    | 0.1013    |
| Average method on hourly data                  | 0.4647    | 0.4699    |
| power generation efficiency                    | 0.0661    | 0.0916    |
| Average method on efficiency data              | 0.2350    | 0.3814    |
| cumulative power generation                    | 0.0080    | 0.0038    |
| Average method on cumulative data              | 0.4473    | 0.7620    |

The histograms of MAE and RMSE for station-level cumulative power generation experiment are illustrated in the figure 2 and figure 3 below.

Figure 2. Histogram of MAE for station-level cumulative power generation experiment

Figure 3. Histogram of RMSE for station-level cumulative power generation experiment

5. Conclusion

In this paper, we propose a tensor decomposition based model, which has the advantages of simplicity
and can accurately infer the missing PV power generation data. We use the power generation data recorded by sensors of several PV stations and model the data with a three dimension tensor, filling the missing entries of the tensor using our model. We did filling experiment on three types of power generation data, which will provide data basis for the subsequent photovoltaic power dispatch decision and photovoltaic power station performance evaluation based on power generation situation. Experiments prove that our model is robust and achieve good result.

Acknowledgments
The authors gratefully acknowledge the financial support from National Key R&D Program of China (2018YFB1500800), Science and Technology project of State Grid Corporation of China(SGTJDK00DYJS2000148),and Self-built project of State Grid E-Commerce Co., Ltd. In 2010(1700/2020-72001B)

References
[1] Wang, F., Zhen, Z., Mi, Z., Sun, H., Su, S., Yang, G. (2015). Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. Energy and Buildings, 86: 427–438.
[2] Shi, J., Lee, W.-J., Liu, Y., Yang, Y., Wang, P. (2012). Forecasting Power Output of Photovoltaic Systems Based on Weather Classification and Support Vector Machines. IEEE Transactions on Industry Applications, 48(3): 1064–1069.
[3] Law, E. W., Prasad, A. A., Kay, M., Taylor, R. A. (2014). Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting – A review. Solar Energy, 108: 287–307.
[4] Eltawil, M. A., Zhao, Z. (2010). Grid-connected photovoltaic power systems: Technical and potential problems—A review. Renewable and sustainable energy reviews, 14(1):112-129.
[5] Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. Renewable and Sustainable Energy Reviews, 81: 912-928.
[6] Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., Martinez-de-Pison, F. J., Antonanzas-Torres, F. (2016). Review of photovoltaic power forecasting. Solar Energy, 136: 78-111.
[7] Pedro, H. T., Coimbra, C. F. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy, 86(7): 2017-2028.
[8] Ogliari, E., Grimaccia, F., Leva, S., Mussetta, M. (2013). Hybrid predictive models for accurate forecasting in PV systems. Energies, 6(4): 1918-1929.
[9] Junior, J. G. D. S. F., Oozeki, T., Ohtake, H., Shimose, K. I., Takashima, T., Ogimoto, K. (2014). Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis. Renewable Energy, 68: 403-413.
[10] Simonov, M., Mussetta, M., Grimaccia, F., Leva, S., Zich, R. (2012). Artificial intelligence forecast of PV plant production for integration in smart energy systems, 3454-3460.
[11] Haque, A. U., Nehrir, M. H., Mandal, P. (2013, July). Solar PV power generation forecast using a hybrid intelligent approach. In 2013 IEEE Power & Energy Society General Meeting, pp: 1-5.
[12] Chu, Y., Urquhart, B., Gohari, S. M., Pedro, H. T., Kleissl, J., Coimbra, C. F. (2015). Short-term reforecasting of power output from a 48 MWe solar PV plant. Solar Energy, 112: 68-77.
[13] Li, Y., Su, Y., Shu, L. (2014). An ARMAX model for forecasting the power output of a grid connected photovoltaic system. Renewable Energy, 66: 78-89.
[14] Li, Y. Z., Luan, R., Niu, J. C. (2008). Forecast of power generation for grid-connected photovoltaic system based on grey model and Markov chain. In 2008 3rd IEEE Conference on Industrial Electronics and applications, pp. 1729-1733.
[15] Yona, A., Senjyu, T., Saber, A. Y., Funabashi, T., Sekine, H., Kim, C. H. (2008). Application of neural network to 24-hour-ahead generating power forecasting for PV system. In 2008 IEEE Energy and Society General Meeting-Conversion and Delivery of Electrical Energy in
the 21st Century, pp: 1-6.

[16] Cherkassky V, Ma Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw, 17:113–26.

[17] da Silva Fonseca, J. G., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y., Ogimoto, K. (2011). Photovoltaic power production forecasts with support vector regression: A study on the forecast horizon. In 2011 37th IEEE Photovoltaic Specialists Conference, pp: 002579-002583.

[18] Rana, M., Koprinska, I., Agelidis, V. G. (2015). 2D-interval forecasts for solar power production. Solar Energy, 122, 191-203.

[19] Ramsami, P., Oree, V. (2015). A hybrid method for forecasting the energy output of photovoltaic systems. Energy Conversion and Management, 95, 406-413.

[20] Yona, A., Senju, T., Funabashi, T., Kim, C. H. (2013). Determination method of insolation prediction with fuzzy and applying neural network for long-term ahead PV power output correction. IEEE Transactions on Sustainable Energy, 4(2), 527-533.

[21] Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., Chang, E. (2014) Diagnosing New York city's noises with ubiquitous data. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp: 715-725.

[22] Kolda, Tamara G., and Brett W. Bader. (2009) Tensor decompositions and applications. SIAM review 51(3): 455-500.

[23] Karatzoglou, A., Amatriain, X., Baltrunas, L., Oliver, N. (2010). Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In Proceedings of the fourth ACM conference on Recommender systems, pp: 79-86.