This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2020.3029731, IEEE Transactions on Sustainable Energy

PV Generation Forecasting with Missing Input Data: A Super-Resolution Perception Approach

Wei Liu, *Student Member, IEEE*, Chao Ren, *Student Member, IEEE* and Yan Xu, *Senior Member, IEEE*

***Abstract*—For practical PV generation forecasting, it is sometimes the case that the input data is missing or incomplete due to measurement and recording errors, which makes the application of a machine learning-based PV forecasting model difficult or impossible. This letter presents a new forecasting framework to address this practical and important issue. The super-resolution perception convolutional neural network (SRPCNN) is used to recover the missing data and stochastic configuration network (SCN) is used to forecast PV generation with the recovered data. The proposed method is tested on an open dataset, and two existing methods are compared as benchmarks. The testing results verify its excellent missing-data recovery ability for accurately forecasting PV generation even under severe missing-data conditions.**

***Index Terms*—PV generation forecasting, missing data, stochastic configuration network, super-resolution perception convolutional neural network.**

I. Introduction

UNDER the pressures of carbon emission and fossil fuel depletion, PV power units have been extensively installed in power grids [1]. The inherent intermittency of PV power brings several technical challenges to the power system, making it essential to predict PV generation to ensure reliable operation and economic dispatch of the power grids [2].

In the literature, various methods including higher-order Markov chains [1] and artificial neural networks (ANN) [3] have been proposed to forecast PV generation. These traditional methods rely on the completeness of the PV generation dataset, while in practice, the measured data may be incomplete. For example, the ratio of missing data was approximately 19.0% in 2017 according to the Korea Meteorological Administration [4]. A flawed dataset will make the use of a machine learning-based PV forecasting model impossible or significantly decrease the accuracy of forecasting models. However, very few research works have investigated this missing data issue in the literature.

To counteract the missing data and enable more accurate PV generation forecasting for practical applications, this paper presents a new forecasting framework based on super-resolution perception convolutional neural network (SRPCNN) for missing data imputation. With the data recovered by SRPCNN, a novel randomized learning algorithm named SCN [5] is used to train a regression model for PV generation forecasting.

II. Proposed Methodology

A. Proposed Forecasting Framework against Missing Data

The proposed framework is illustrated in Fig.1, comprising the offline training stage and the online prediction stage. At the offline training stage, two models are separately trained: 1) an SRPCNN for data imputation and 2) an SCN for PV forecasting. The inputs of the SRPCNN model are the flawed data due to missing data and the outputs are the reconstructed full data. For the SCN model, only the full training dataset is used. The inputs and outputs are the PV generation of the last N time steps and k time steps after, respectively.

At the online prediction stage, if the measurement input is complete, then it will be directly imported to the trained SCN model. Otherwise, the incomplete measurement input is firstly imported to the trained SRPCNN model, which will recover the incomplete data. Then, the estimated complete data is used as the input of trained SCN and the final predicted PV generation will be exported.

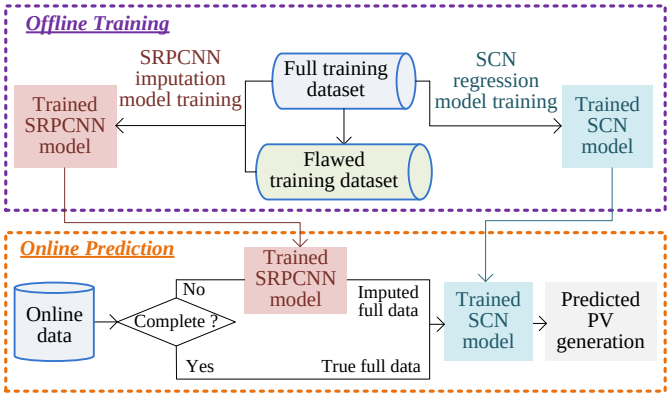

Fig. 1. The framework of the proposed method

B. Super-Resolution Perception for Missing Data Imputation

In [6], the super-resolution perception convolutional neural network (SRPCNN) is proposed and has shown its effectiveness to reconstruct high-frequency data from low-frequency data of industrial sensors. This letter applies SRPCNN to recover the online incomplete input data of PV power generation. The SRP is an under-determined function that maps the flawed data with missing data x to the estimated complete data y, denoted as $f_\theta: x \to y$. The SRP mapping f_θ is implemented by a CNN which takes the flawed data (short vector) as the input features with a length of d_f and outputs the estimated complete data (long vector) with a length of d_c, $d_f < d_c$. Fig. 2 illustrates the architecture of SRPCNN, which consists of three parts: feature extraction, information supplement and reconstruction.

Given inputs $\mathbf{X} \in \mathbb{R}^{p \times d_f}$ with p instances and d_f features, the feature extraction part extracts features from $\mathbf{X}$ and the features of each instance are represented by m feature vectors, each of which has a length of d_f. These features $\mathbf{F}_i \in \mathbb{R}^{p \times m \times d_f}$ contain the abstract feature information of input $\mathbf{X}$. After that,

W. Liu and Y. Xu (corresponding author) are with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798. (email: eeyanxu@gmail.com)
C. Ren is with Interdisciplinary Graduate School, Nanyang Technological University, Singapore 639798.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2020.3029731, IEEE Transactions on Sustainable Energy

the information supplement part which contains one global residual connection and n local residual blocks supplements the missing information to the feature vectors. Inside each local residual block, the convolution layers perform non-linear mapping between the input and output feature spaces and the Rectified Linear Units (ReLU) are activation functions. The local residual connections and the global residual connection are applied to improve the performance by forcing the network to learn the residual functions.

Finally, the reconstruction part integrates the feature vectors of each instance into α sub-vectors, each of which has a length of d_f. These sub-vectors $\mathbf{F}_o \in \mathbb{R}^{p\times\alpha\times d_f}$ are then rearranged as the estimated full data $\mathbf{Y} \in \mathbb{R}^{p\times d_c}$. It is worth pointing out that the sub-vectors are generated in parallel using convolution operation, thus SRPCNN has high computational efficiency.

As described above, the output of SRPCNN is the estimated full data that is reconstructed from the original data. Therefore, the estimated full data can effectively represent the characteristics of the original input due to the strong feature learning ability of CNN.

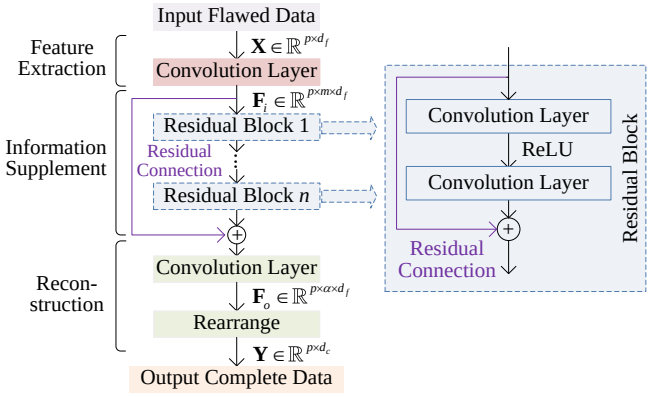

Fig. 2. The structure of SRPCNN

The mean squared error (MSE) is used to train the proposed network and the loss function is represented as

$$L\left(\mathbf{Y},\mathbf{Y}'\right) = \|\mathbf{Y}-\mathbf{Y}'\|^2 = \|\mathbf{Y}'-f_\theta(\mathbf{X})\|^2 \tag{1}$$

where $\mathbf{Y}'$ is the true complete data and θ represents the model parameters of SRPCNN.

The network is optimized by minimizing the above loss function using gradient descent-based method and the optimal model parameters θ can be computed as

$$\hat{\theta} = \min_{\theta}\|\mathbf{Y}'-f_\theta(\mathbf{X})\|^2 \tag{2}$$

C. Stochastic Configured Network for Regression

SCN is a randomized learning algorithm for single-layer feed-forward networks (SLFN) [5]. Unlike traditional SLFNs that suffer from computational cost caused by backpropagation, the input weights and biases at the hidden nodes of SCN are randomly assigned and the output weights are then analytically computed. As a novel randomized learning algorithm, SCN is different from existing algorithms (e.g. ELM) in the sense that the hidden nodes are incrementally added in the light of a supervisory mechanism.

Suppose an SCN with $L-1$ hidden nodes has been constructed, shown as follows:

$$f_{L-1}(x) = \sum_{j=1}^{L-1}\beta_j\cdot g_j = \sum_{j=1}^{L-1}\beta_j\cdot g(w_j^T x + b_j) \tag{3}$$

where w_j and β_j are the input and output weights of the j^{th} hidden node, respectively; $g(\cdot)$ denotes the activation function and b_j is the bias at the j^{th} hidden node.

If the current residual error does not reach a predefined tolerance level, a new hidden node with w_L and β_L is added to the current model and the output weights β_L are analytically evaluated. As a result, the new model with L hidden nodes is constructed with decreased residual error, shown as

$$f_L = f_{L-1} + \beta_L\cdot g_L = f_{L-1} + \beta_L\cdot g(w_L x + b_L) \tag{4}$$

With the increasing number of hidden nodes, the residual error will converge to zero if w_L and β_L are generated to satisfy the following supervisory mechanism (inequality constraint):

$$\langle e_{L-1}^*, g_L\rangle^2 \geq b_g^2(1-r-\mu_L)\langle e_{L-1}^*, e_{L-1}^*\rangle^2 \tag{5}$$

where e_{L-1}^* denotes the residual error with $L-1$ nodes, r is a given constant and $0<r<1$, $\{\mu_L\}$ is a sequence satisfying $0<\mu_L\leq 1-r$ and $\lim_{L\to+\infty}\mu_L = 0$, b_g is one upper bound of the norm of the activation function $g(\cdot)$.

The supervisory mechanism described above ensures the universal approximation property of SCN, which is proved by theoretical results and verified with simulation results in [5]. In general, the two main parts of SCN can be summarized as:

1) *Configuration of Hidden Parameters*: Randomly assigning the input weights and biases to meet the constraint (5), then generating a new hidden node and adding it to the current learner model.
2) *Determination of Output Weights*: Analytically determining the output weights of the current learner model through solving a least-squares problem with *Moore–Penrose* generalized inverse operation.

III. Case Study

A. Dataset Description

The proposed method is tested with an open dataset collected at St. Lucia Campus, University of Queensland, Australia [7]. This dataset covers the period from December 2015 and November 2017 with 1-minute time resolution, including four seasons. For each season, the data from the former and latter years are used for training and testing, respectively. Note that only PV generation between 7.00 am and 5.59 pm each day is used for the case study. In this letter, the 10-minute ahead PV generation is predicted using the PV generation of the last 60 minutes.

B. Testing Results with The Proposed Method

The normalized root of mean squared error (nRMSE) is used to measure the forecasting accuracy and is defined as

$$nRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2}\,/P_n \tag{6}$$

where y' and y are the estimated and true PV generation respectively, and P_n is the nominal capacity of the PV site.

A series of ten missing-data rates ranging from 0% to 95% are implemented to test the effectiveness of the proposed method at various missing-data rates. 0% missing-data rate indicates that there is no missing data in the testing dataset. Without loss of generality, for any missing-data rate, the corresponding amount of data in the testing dataset are randomly removed. Therefore, there are no presumptions on the pattern of missing

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSTE.2020.3029731, IEEE Transactions on Sustainable Energy

data in this test. But other missing data patterns can be considered if necessary.

The overall forecasting results under different missing-data rates are summarized in Table I. To show the effectiveness of the developed SCN model, backpropagation neural networks (BPNN) [3] and long short-term memory networks (LSTM) [8] are also implemented using complete data. One can see that with complete data, the average errors of three approaches are all over 10%, indicating strong volatility of the dataset. Comparatively, the proposed method outperforms the other two methods with an average error of 10.21%. It is also shown that among four seasons, the prediction error in winter is the lowest, implying that PV generation in winter is less stochastic.

When 95% of the input testing data is missing, the average error only increases by less than 0.75% (from 10.21% to 10.96%) compared to the error obtained with the complete dataset. Interestingly, it is noted that the forecasting accuracies under low missing-data rates are even slightly higher than those obtained with the full data. This accuracy improvement is due to the strong feature learning capability of the SRPCNN, so the reconstructed data may better represent the characteristic of the original data.

TABLE I.
NORMALIZED RMSE (%) OF THE PROPOSED METHOD AND BENCHMARKS

Test cases		Seasons				Average
		Summer	Fall	Winter	Spring	
Benchmarks	BPNN [3]	13.25	12.25	6.40	12.18	11.02
	LSTM [8]	12.33	11.17	6.14	12.55	10.55
Proposed method under missing-data rates	0%	12.12	11.24	6.06	11.42	10.21
	50%	11.99	10.84	5.88	11.1	9.95
	67.7%	11.84	10.89	5.92	10.99	9.91
	75%	11.86	10.84	6.04	11.03	9.94
	80%	11.92	10.97	6.22	11.04	10.04
	83.3%	11.78	11.15	5.89	11.07	9.97
	90%	12.02	10.97	6.19	11.05	10.06
	91.7%	12.14	10.92	6.09	11.10	10.06
	93.3%	12.44	11.31	6.37	11.74	10.46
	95%	13.51	11.55	6.88	11.88	10.96

The proposed method is implemented using a laptop with 8.0 GB RAM and Intel(R) Core(TM) i3 CPU @ 2.00 GHz processor. Taking 50% missing-data rate and winter as an example, the offline and online stages take 551.2 and 1.1 seconds, respectively. Therefore, the proposed method can be applied online as 10-minute ahead forecasting is carried out.

C. Comparison with Existing Methods

Two existing methods for missing-data problems are also tested with the same dataset as benchmarks: 1) linear interpolation (LI) and 2) random forest with surrogate splitting (RFSS) [9]. To normalize the effects of missing-data rates on forecasting accuracy at different seasons, error increasing rate (EIR) is used in this letter, defined as

$$EIR(\%) = (nRMSE_i - nRMSE_0)/nRMSE_0 \times 100 \quad (7)$$

where $nRMSE_i$ is the error when the missing-data rate is i and $nRMSE_0$ denotes the error without missing data.

A comparison of the proposed method with these two benchmarks is presented in Fig. 3. As is shown, all three methods can tackle low missing-data rates. The forecasting error only increases slightly and the EIRs of all the three methods at low missing-data rates are approximately the same. However, when the missing-data rate grows higher, the EIRs of LI and RFSS will get significantly larger while the proposed SRPCNN can still maintain a relatively low EIR. Taking spring as an example, the EIRs of LI and RFSS at 95% missing-data rate are 37.2% and 50.7%, respectively. As a comparison, the error of SRPCNN with 95% missing data only grows 4.1% compared to the error with complete data availability. The small EIR at such a high missing-data rate demonstrates the excellent missing-data tolerance of the proposed method.

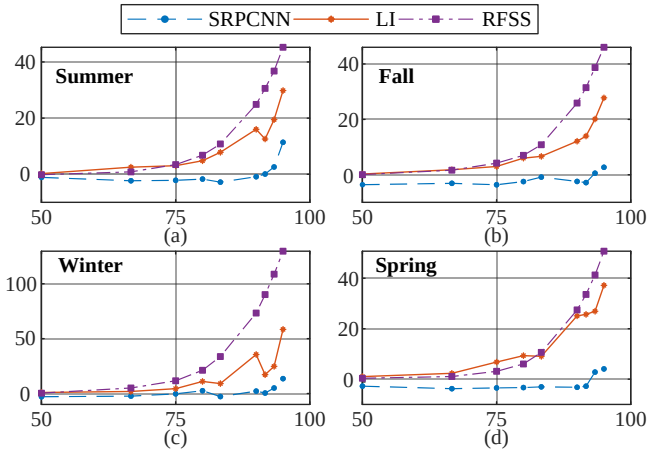

Fig. 3. Comparison of different methods for different seasons: (a) Summer; (b) Fall; (c) Winter and (d) Spring. X-axis: Missing-data rates (%); Y-axis: EIR (%).

IV. CONCLUSION

A novel framework is designed in this letter for PV generation forecasting with missing input data, where SRPCNN is used for missing data imputation and SCN is utilized as the regression model. The testing results on an open dataset show that at low missing-data rates, the proposed method can even slightly improve the forecasting accuracy compared to the results with complete data due to the strong feature learning capability of SRPCNN. Under extremely high missing-data rate, e.g. 95%, the errors of the proposed method only increase marginally while the errors of conventional methods LI and RFSS grow significantly.

REFERENCES

[1] M. J. Sanjari and H. B. Gooi, "Probabilistic Forecast of PV Power Generation Based on Higher Order Markov Chain," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 2942–2952, Jul. 2017.

[2] H. Yang, Q. Yu et al., "Optimal Wind-Solar Capacity Allocation With Coordination of Dynamic Regulation of Hydropower and Energy Intensive Controllable Load," *IEEE Access*, vol. 8, pp. 110129–110139, 2020.

[3] J. Liu, W. Fang, X. Zhang, and C. Yang, "An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data," *IEEE Trans. Sustain. Energy*, vol. 6, no. 2, pp. 434–442, Apr. 2015.

[4] T. Kim, W. Ko, and J. Kim, "Analysis and Impact Evaluation of Missing Data Imputation in Day-ahead PV Generation Forecasting," *Appl. Sci.*, vol. 9, no. 1, p. 204, Jan. 2019.

[5] D. Wang and M. Li, "Stochastic Configuration Networks: Fundamentals and Algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3466–3479, Oct. 2017.

[6] J. Gu, G. Liu, G. Liang, and J. Zhao, "Super-Resolution Perception for Industrial Sensor Data," arXiv preprint arXiv: 1809.06687.

[7] "PV Power Generation, Ambient Temperature and Solar Irradiance Data." [Online]. Available: http://solar.uq.edu.au/user/ reportPower.php.

[8] X. Qing and Y. Niu, "Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM," Energy, vol. 148, pp. 461–468, Apr. 2018.

[9] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.