SolarGAN: Multivariate Solar Data Imputation Using Generative Adversarial Network

Wenjie Zhang, Yonghong Luo, Ying Zhang, and Dipti Srinivasan, *Fellow, IEEE*

***Abstract*—Photovoltaic (PV) is receiving increasing attention due to its sustainability and low carbon footprint. However, the penetration level of PV is still relatively low because of its intermittency. This uncertainty can be handled by accurate PV forecasting, which requires high-quality solar data. Nevertheless, up to 40% of solar data can be found missing, which significantly worsens the quality of solar data. This letter proposes a novel solarGAN method for multivariate solar data imputation, in which necessary modifications are made on the input of generative adversarial network (GAN) to effectively tackle the relatively independent solar time series data. Case studies on a public dataset show that the proposed solarGAN outperforms several commonly-used machine learning and GAN based data imputation methods with at least 23.9% reduction of mean squared error.**

***Index Terms*—Solar data imputation, generative adversarial network (GAN), PV forecasting, machine learning, smart grid.**

I. Introduction

PHOTOVOLTAIC (PV) is a promising alternative to fuel fossil-based generations due to its low carbon footprint. Consequently, many governments have promoted solar panel installations [1]. However, because of the inherent intermittency of PV power generation, power system utilities are often reluctant to incorporate more PV installations. Therefore, to properly integrate PV generation in the context of smart grids and to further improve the reliability of smart grids, accurate forecasts and adequate quantification of PV generation uncertainties are imperative [2], [3]. PV forecasting requires high-quality solar data, such as solar irradiance, temperature, wind speed and direction, humidity, and past PV generation records [4]. Nevertheless, it is not uncommon that up to 40% of solar data are missing or incorrect because of data transmission losses and noise interference. For example, 66% of samples are missing in the record of daily observations made at the National Solar Observatory at Sacramento Peak [5]. The low-quality data can evidently reduce the accuracy of PV forecasting, which further hinders the integration of more PVs. Therefore, imputing the missing solar data is necessary.

Solar data imputation is handled in an unsupervised manner as there are no ground truths. There have been various unsupervised data imputation methods in the literature, which are elaborated in Section III. Recently, an unsupervised learning framework, called generative adversarial network (GAN) [6], has become very popular. GAN is a neural network (NN) based structure, which has shown convincingly superior performance in synthesizing images, changing image styles, and repairing images. Repairing images is equivalent to matrix completion since images are sets of matrices. As solar data are stored as numerous time series data, which can be treated as matrices, GAN should have the potential to impute missing solar data. In [7], a GAN based imputation method that uses a hint vector to impute the missing values is proposed. In [8], one variation of GAN – Wasserstein GAN (WGAN) [9] has shown outstanding performance in medical data imputing. In this letter, we migrate the method in [8] to solar data imputation. Compared to the medical data, solar data are collections of more independent time series data. Therefore, necessary modifications on the GAN proposed in [8] are made to adequately handle the more independent time series data.

The contributions of this letter are summarized as follows:

1) GAN is introduced into solar data imputation, which provides a solid unsupervised framework for the solar data imputation.
2) A modified WGAN for solar data imputation, namely solarGAN, is proposed to handle the more independent solar time series data. The proposed solarGAN is derived from the GAN proposed in [8], but uses the summation of random noise and real samples as inputs to the generator in GAN, instead of pure random noise in [8]. The modified inputs can tackle the relatively independent solar time series data more efficiently.
3) Case studies show that the proposed solarGAN is more accurate than commonly-used machine learning based and two GAN based data imputation methods.

The rest of the letter is organized as follows. The methodology of the proposed method is described in Section II. Case studies, results, and discussions are presented in Section III. Lastly, conclusions are drawn in Section IV.

Manuscript received June 17, 2019; revised December 8, 2019 and April 30, 2020; accepted June 9, 2020. Date of publication June 24, 2020; date of current version December 16, 2020. This work was supported by the Energy Market Authority of Singapore under Grant R-263-000-C97-279. *(W. Zhang and Y. Luo contributed equally to this work.) (Corresponding authors: Wenjie Zhang; Ying Zhang.)*

Wenjie Zhang and Dipti Srinivasan are with the Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore 117576, Singapore (e-mail: wenjie_zhang@u.nus.edu; dipti@nus.edu.sg).

Yonghong Luo and Ying Zhang are with the College of Computer Science, Nankai University, Tianjin 300071, China (e-mail: luoyonghong@dbis.nankai.edu.cn; yingzhang@nankai.edu.cn).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSTE.2020.3004751

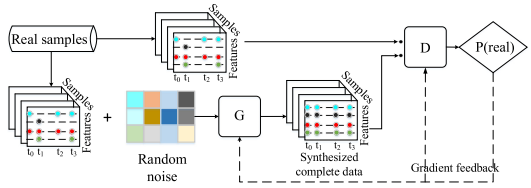

Fig. 1. The structure of solarGAN

II. Methodology

A. General GAN Architecture

GAN has revolutionized many unsupervised learning areas since Ian Goodfellow put forward the framework [6], on which most variations of GAN rely. GAN consists of an NN-based generator (G) synthesizing data and an NN-based discriminator (D) classifying whether the fed data are real or synthesized. Denoting the mapping of G and D as $G(\bullet)$ parameterized by θ_G and $D(\bullet)$ parameterized by θ_D, respectively, we can formulate the GAN training as a min-max game problem, which is shown in (1). In (1), P_d represents the real distribution, and P_g represents the Gaussian distribution. The maximization over θ_D provides a measurement of the difference between the synthesized probabilistic distribution and the real one, while the minimization over θ_G pushes the synthesized probabilistic distribution to be as close as possible to the real one. In other words, the objective of (1) is training the generator to synthesize fake data that can fool the discriminator.

$$\min_{\theta_G}\max_{\theta_D} E_{x\sim P_d}\left(log\left(D(x)\right)\right) + E_{z\sim P_g}\left(log\left(1-D\left(G(z)\right)\right)\right) \tag{1}$$

B. WGAN

The original GAN suffers from some problems, such as non-convergence, mode collapse, and diminished gradients. These are due to the fact that the maximization game over θ_D in the original GAN estimates the Jensen–Shannon (JS) divergence. In the case that the synthesized probability distribution has no overlap with the real one, JS divergence will saturate. In other words, the gradients vanish. WGAN is proposed to alleviate the gradient vanishing problem. WGAN utilizes the non-saturating earth mover distance to quantify the difference between two probability distributions [9], [10]. The formulation of WGAN is shown in (2).

$$\min_{\theta_G}\max_{\theta_D} E_{x\sim P_d}\left(D(x)\right) - E_{z\sim P_g}\left(D\left(G(z)\right)\right) \tag{2}$$

The min-max game of WGAN should follow the condition that D belongs to the set of 1-Lipschitz functions, , which is the inequality (3).

$$\left\|D(x)\right\|_L \leq 1, \text{ for all } x. \tag{3}$$

For the sake of easier training, we adopt WGAN in the proposed solarGAN.

C. Proposed Solargan

1) SolarGAN Structure and Modifications: The proposed solarGAN for solar data imputation is shown in Fig. 1. In Fig. 1, samples from a real database are extracted to train D, which tells us how likely the fed data are real samples. For the other branch, the summations of incomplete samples and random noises are fed to G to generate synthesized complete data, which will also be discriminated by D. Generally, the data fed to G should be purely random noise. However, in our experiments, it turns out that pure random noise cannot preserve the inter-column information efficiently. Therefore, we further modify the GAN by adding real samples and random noise together to result in improved information flow throughout G. In other words, $z + x$ is used to replace the pure z input to G.

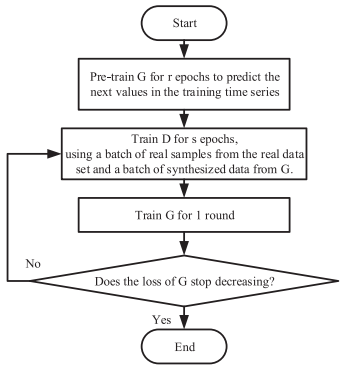

Fig. 2. The training scheme of solarGAN

More specifically, G contains a gated recurrent unit for data imputation (GRUI) layer (details of GRUI can be found in Ref. [8]) and a fully-connected layer with dropout. The time series data (incomplete samples plus random noise) are fed to the GRUI layer, which produces hidden states for the corresponding time indexes. The current hidden state of the GRUI layer is further fed to a fully-connected layer to generate complete solar data for the current time index. It is noted that before being fed to D, all the outputs from the fully-connected layer are concatenated and batch normalized. Similar to G, D also consists of a GRUI layer and a fully-connected layer (only one neuron in this layer with a sigmoid activation). The last hidden state of the GRUI layer is fed to the fully-connected layer to generate the probability of real data $P(real)$.

2) SolarGAN Training: The training of solarGAN shown in Fig. 2 follows the standard GAN training, except for pre-training G for r epochs to get a relatively reasonable initialization of G. r and s are both hyperparameters to be tuned. The stopping criterion of the training is that the parameter set of G, θ_g, converges. Specifically, the early stopping technique is utilized to determine when the training of the proposed solarGAN should be stopped. During training, a training set and a validation set are used, the epoch to stop the training (the stopping epoch) is chosen to be the epoch where the G achieves the lowest metric error on the validation dataset. The G attained at the stopping epoch is the generator used to generate imputations. It is noted that, for most missing rates shown in the case study done in the letter, it takes about 20 epochs to converge. Notably, there is another ideal case indicating that the GAN is well trained, which is the $P(real)$ converges to 50%, implying that the D makes random guesses as it cannot discriminate whether the fed data are real or not.

D. z Determination and x Imputation

Once the training of solarGAN is done, which means that θ_G and θ_D are available, we need to feed $z + x$ to the trained G

to generate the corresponding complete $\widehat{x}$. However, as the z is random, there is no conditioning constraint on the $G(z+x)$. Our objective is to generate $\mathrm{G}(z+x)$ such that $\mathrm{G}(z+x)$ is most similar to x. Therefore, we need to condition the objective. In the literature, the conditioning can be done by conditional GAN, whose training is more complicated and unstable than GAN. Aiming at determining the z while simplifying the training, we formulate the z determination problem as shown in (4),

$$\min_{z} \|x \circ M - G(z+x) \circ M\|_2 - \lambda D(G((z+x))) \quad (4)$$

where M denotes the indexes of the non-missing values. For example, if $X = [0.5\ None]^T$, then $M = [1\ 0]^T$. $\circ$ denotes the entry-wise products. The first term in (4) is to minimize the difference between $G(z+X)$ and x for the non-missing part, and the second term is to maximize the probability that D classifies $G(z+x)$ as real samples. λ is a hyperparameter providing a balance between the first and second terms.

Let $\tilde{z}$ be the optimum of (4), then the complete sample $\widehat{x}$ can be imputed as shown in Eq. (5).

$$\widehat{x} = x \circ (1-M) + G(\tilde{z}+x) \circ M \quad (5)$$

III. Case Study

In this section, a case study is presented based on a publicly available solar dataset from the GEFCom 2014 solar track [11]. The dataset is a prevalent and reliable dataset for solar forecasting, which contains 13 different time series, including solar irradiance, humidity, temperature, cloud index, wind speed, wind direction, precipitation, PV power, etc.. Therefore, it is a data set rich of diverse variates.

A part of data from GEFCom 2014 is randomly deleted on purpose according to different missing rates to generate train and test datasets. Therefore, the data in the data set are missing at random. In practice, data imputation should be unsupervised, which means that ground truths are never available. Therefore, direct comparisons with the ground truths are impossible. For the ease of understanding, we use a complete and reliable dataset as the basis to generate an incomplete dataset and then directly compare imputed data with the corresponding ground truths. The usage of the complete dataset helps to present readers our proposed method and its superiority in a more straightforward and self-evident manner.

The code is available on [12] to facilitate the reproducibility of the case study done in the letter.

A. Evaluation Metrics and Benchmark Methods

A widely accepted evaluation metric for data imputation is the mean squared error (MSE) between missing data and the corresponding imputations. MSE directly measures the difference between imputed data and the ground truths for the missing parts. It is noted that using MSE or similar root MSE for evaluating the accuracy of imputation is a common practice in the literature [7], [8], [13], [14].

For validating the performance of the proposed solarGAN, numerous popular data imputation methods used in academia and industry are used as benchmark methods, which are listed as follows.

TABLE I
MSE OF DIFFERENT IMPUTATION METHODS

Missing rate	Mean	Last	MF	KNN	MICE	GAIN	GAN-Z	solarGAN
10%	1.027	0.173	0.207	0.222	0.161	0.222	0.255	**0.160**
20%	1.003	0.194	0.211	0.233	0.179	0.349	0.332	**0.168**
30%	1.002	0.238	0.216	0.245	0.199	0.363	0.357	**0.181**
40%	0.998	0.259	0.219	0.253	0.217	0.378	0.369	**0.196**
50%	0.999	0.310	0.240	0.285	0.236	0.380	0.372	**0.216**
60%	1.002	0.398	0.254	0.302	0.266	0.385	0.398	**0.245**
70%	0.996	0.472	0.295	0.343	0.314	0.394	0.420	**0.291**
80%	1.005	0.621	0.433	0.431	0.387	0.421	0.490	**0.373**
90%	1.001	0.813	0.485	**0.312**	0.536	0.605	0.647	0.541

1) Mean Value Filling: It is a naive data imputation method that replaces the missing values with the mean of the nearest k values.
2) Last Value Filling: It is another naive method which uses the last available value as substitutions.
3) Matrix Factorization (MF) [15], [16]: MF has been successfully applied in imputing missing traffic speed values [16]. In short, MF factorizes incomplete data matrix into the product between low-rank matrices U and V, which is done by gradient descent. During the gradient descent, the L1 penalty on U and the L2 penalty on V are also included in the corresponding optimization objective. After the gradient descent, the missing value will be fulfilled by the parameters that minimize the optimization objective (error).
4) K- Nearest Neighbor (KNN) [17]: KNN can match a data point with its closest k neighbors in a high-dimensional space. Applying KNN in data imputation is based on the assumption that the missing value can be approximated by its nearest neighbor or the mean of its k nearest neighbors.
5) Multivariate Imputation by Chained Equation (MICE) [18]: MICE is a principled and advanced method for data imputation, which has been used for missing data imputation in several areas, such as solar radiation [19], traffic sensing and monitoring [20], and galactic cosmic rays [21]. It is also known as sequential regression imputation. In MICE, a chain of regression equations is created to impute the data one by one. More details about MICE can be found in Ref. [18].
6) Generative Adversarial Imputation Nets (GAIN) [7]: GAIN is a GAN based imputation method that uses a hint vector to impute the missing values. The proposed solarGAN differs from GAIN in terms of the used NN types in G and D (the proposed solarGAN uses recurrent neural network while GAIN uses fully-connected NNs), GAN structure, and the input to the G in GAN.
7) GAN-Z [8]: GAN-Z is the GAN for data imputation shown in [8]. It is the GAN that the proposed solarGAN bases on. The GAN in [8] is named as GAN-Z as it only takes pure noise z as the input to G. Unlike GAN-Z, the proposed solarGAN takes the summation of noise and real samples as the input to G, which helps to adequately handle relatively more independent solar time series data.

B. Results and Discussions

The MSE comparisons between the proposed solarGAN and benchmark models are shown in Table I, with the best performance in the corresponding missing rates in bold. It is

evident that for most missing rates, the proposed solarGAN has the smallest MSE, which implies that the proposed solarGAN is more accurate than the benchmark methods in solar data imputation. As compared to traditional GAN-based methods, the proposed solarGAN is superior, implying that changing the input from pure noise to the summation of real samples and noise does contribute to improving the accuracy of imputations. The improvement of the proposed solarGAN can also indicate that the new input to the generator can tackle the relatively independent solar time series data more efficiently. As compared to the commonly-used machine learning based methods and GAN based methods, the error reduction is at least 23.9%. It can be inferred from the diversity of time series data in the used solar data set that the proposed method can be extended to many other time series data, such as wind speed and temperature data.

For the missing rate of 90%, the solarGAN doesn't perform well due to the lack of data. As the 90% missing rate is very high, the real data left for training GAN are extremely limited. GAN is a data-driven method and its performance highly relies on the volume of data. However, the performance at the 90% missing rate should not weaken the credibility of solarGAN as the 90% missing rate is very rare in practice.

IV. Conclusion

Incorporating more PV generations in smart grid relies heavily on accurate PV generation forecasting, which requires solar data of high quality. However, solar data are mostly incomplete, whose missing rate can be up to 40%. This letter has proposed a novel GAN-based method – solarGAN for multivariate solar data imputation. Case studies have shown that the proposed solarGAN outperforms other existing methods (including two naive methods, three machine learning based methods, and two GAN based methods) for a wide range of missing rates. Considering the shown superiority, the solarGAN should have great potential in facilitating more accurate PV generation forecasting. It is noted that the proposed solarGAN can be extended to other time series data imputation, such as wind speed and temperature data. The proposed solarGAN is publicly available at [12]. Further research will focus on exploiting the proposed methods with more evaluating metrics and integrating PV forecasting in the solarGAN framework.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which improve the quality of the letter.

References

[1] C. Feng, M. Cui, B.-M. Hodge, S. Lu, H. Hamann, and J. Zhang, "Unsupervised clustering-based short-term solar forecasting," *IEEE Trans. Sustain. Energy*, vol. 10, no. 4, pp. 2174–2185, Oct. 2019.

[2] M. Cui, J. Zhang, B.-M. Hodge, S. Lu, and H. F. Hamann, "A methodology for quantifying reliability benefits from improved solar power forecasting in multi-timescale power system operations," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6897–6908, Nov. 2018.

[3] A. Bracale, G. Carpinelli, and P. De Falco, "A probabilistic competitive ensemble method for short-term photovoltaic power forecasting," *IEEE Trans. Sustain. Energy*, vol. 8, no. 2, pp. 551–560, Apr. 2016.

[4] J. Wang, H. Zhong, X. Lai, Q. Xia, Y. Wang, and C. Kang, "Exploring key weather factors from analytical modeling toward improved solar power forecasting," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1417–1427, Mar. 2019.

[5] T. D. de Wit, "A method for filling gaps in solar irradiance and solar proxy data," *Astron. Astrophys.*, vol. 533, 2011, Art. no. A29.

[6] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[7] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5689–5698.

[8] Y. Luo *et al.*, "Multivariate time series imputation with generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1596–1607.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[10] X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang, "Improving the improved training of Wasserstein GANS: A consistency term and its dual effect," in *Proc. Int. Conf. Learn. Representations*, 2018.

[11] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecasting*, vol. 32, no. 3, pp. 896–913, 2016.

[12] W. Zhang and Y. Luo, "Solargan: Multivariate solar data imputation using generative adversarial network," 2019. [Online]. Available: https://github.com/stephenzwj/SolarGAN, Accessed: Dec. 8, 2019.

[13] M. Amiri and R. Jensen, "Missing data imputation using fuzzy-rough methods," *Neurocomputing*, vol. 205, pp. 152–164, 2016.

[14] B. K. Beaulieu-Jones and J. H. Moore, "Missing data imputation in the electronic health record using deeply learned autoencoders," in *Proc. Pacific Symp. Biocomputing*, 2017, pp. 207–218.

[15] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[16] X.-Y. Huang, W. Li, K. Chen, X.-H. Xiang, R. Pan, L. Li, and W.-X. Cai, "Multi-matrices factorization with application to missing sensor data imputation," *Sensors*, vol. 13, no. 11, pp. 15 172–15 186, 2013.

[17] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," Dept. Statist. Stanford Univ., Stanford, CA, USA, Tech. Rep., 1999.

[18] S. V. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Statist. Softw.*, vol. 45, no. i03, 2011.

[19] C. Turrado, M. López, F. Lasheras, B. Gómez, J. Rollé, and F. Juez, "Missing data imputation of solar radiation data under different atmospheric conditions," *Sensors*, vol. 14, no. 11, pp. 20 382–20 399, 2014.

[20] K. Henrickson, Y. Zou, and Y. Wang, "Flexible and robust method for missing loop detector data imputation," *Transp. Res. Record*, vol. 2527, no. 1, pp. 29–36, 2015.

[21] R. Fernandes, P. Lucio, and J. Fernandez, "Data imputation analysis for cosmic rays time series," *Adv. Space Res.*, vol. 59, no. 9, pp. 2442–2457, 2017.

Wenjie Zhang received the B.Eng. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China, in 2015, and the Ph.D. degree in electrical and computer engineering from the National University of Singapore (NUS), Singapore, in 2020. He served as a Visiting Scholar in Stanford University in 2019. Now he is also actively involved in multiple AI-oriented projects in NUS. His current research interests include uncertainty quantification, machine learning, and deep learning in smart cities and smart grids.

Yonghong Luo received his bachelor and master degree in computer science from Nankai University, China. He is currently working in Pony.ai. His research interests include the imputation of time series data, clustering of streaming data and designing of big data mining algorithms.

Ying Zhang received the Ph.D. degree from Nankai University in 2013. From 2011 to 2013, she was a Visiting Scholar with the Department of Computer Science, Purdue University. She is currently an Associate Professor with the College of Computer Science, Nankai University. Her research interests include natural language processing, multimodal data analysis, and machine learning.

Dipti Srinivasan (Fellow, IEEE) received the M.Eng. and Ph.D. degrees in electrical engineering from the National University of Singapore in 1991 and 1994, respectively. She is currently a Full Professor in the Department of Electrical and Computer Engineering, National University of Singapore. Her research interest is in the application of soft computing techniques to engineering optimization and control problems. Dr. Srinivasan is currently serving as an Editor of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Evolutionary Computation, IEEE Transactions on Sustainable Energy, and IEEE Computational Intelligence Magazine.