Data-Driven Generation of Synthetic Load Datasets Preserving Spatio-Temporal Features

Andrea Pinceti, Oliver Kosut, and Lalitha Sankar
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ, 85287

***Abstract*—A generative model for the creation of realistic historical bus-level load data for transmission grid models is presented. A data-driven approach based on principal component analysis is used to learn the spatio-temporal correlation between the loads in a system and build a generative model. Given a system topology and a set of base case loads, individual, realistic time-series data for each load can be generated. This technique is demonstrated by learning from a large proprietary dataset and generating historical data for the 2383-bus Polish test case.**

***Index Terms*—synthetic, historical, time-series data, generative models, spatio-temporal correlation, singular value decomposition, principal component analysis.**

I. Introduction

Synthetic grid models are an indispensable tool for power system researchers and engineers. Recently, projects like ARPA-E's GRID DATA [1] have tried to address the lack of publicly available large scale models by either publishing anonymized real data or creating synthetic but realistic systems. All public models must not contain critical electrical infrastructure information (CEII) while including basic data such as system topology, branch parameters, and generators and load characteristics. These details often only represent one operating case of the system and while they are important for many types of power system studies (powerflow, stability, fault analysis, etc), the usefulness of synthetic grids is greatly improved whenever historical load data is available. Knowing how each load varies over a day, week, or month not only benefits the aforementioned studies by providing different operating cases but it is also crucial for studies such as multi-temporal unit commitment and economic dispatch, transmission expansion planning, and long term reliability. In addition to these more traditional applications, historical data will also prove increasingly valuable in the emerging field of machine learning applied to power systems.

Currently, only a few of the grid models publicly available to researchers include historical load curves. For these models, the most common approach is to provide time-series data for the system demand at a net or zonal level and calculate the bus loads at each time step as a fixed fraction of the aggregate load. This method is simple to implement since many utilities and system operators publish historical net load information for their systems, often over multiple years [2] [3]. One of the most widely used system models which was developed following this approach is the IEEE RTS-96 case described in [4]. The drawback is that because a fixed load ratio is used for every bus, the variability in behavior that exists between different types of loads at different times of the day or week is not captured; having each load follow the same profile over time is not realistic. An alternative method consists in creating load data as a combination of prototypical load models at a bus level. This technique provides an overall more realistic dataset by creating individual profiles for each load, but it requires detailed geographical and/or demographical information to determine the load composition at each bus. For example, in [5], historical load data for the synthetic Texas model [6] is created starting from several typical load curves and combining them according to load types and actual population data of the state of Texas.

In this work, we develop an automatic, data-driven technique to generate bus-level historical load data for any given transmission-level grid model. The goal is to create a generative model which takes as only inputs the system topology and a set of base case loads and returns individual time-series load data for every bus in the system for an arbitrary period of time. To generate this synthetic data we need to first learn the spatio-temporal correlation which exists between the loads in a system. In our work, we introduce a technique based on principal component analysis (PCA) to model the temporal behavior of loads, and topology-based factors to model the spatial correlation between loads. The features learned from the real data are then used to generate realistic, individual temporal profiles for the loads of a new grid model. Examples of the application of PCA to the study of electrical loads can be found in load forecasting applications [7]. In [8]–[10], the authors describe the use of PCA for the processing of the data used in long and short-term forecasting models for a system's net load. Our approach differs in that we use PCA to extract temporal profiles from bus-level time-series data rather than identify the correlation between the variables governing the system net load.

We present a description of the real dataset on which our learning algorithm is tested in Section II. The details regarding the temporal and spatial characteristics of the data and the associated generative model are described in Sections III and IV, respectively. Finally, Section V presents a validation of our results to show that the generated data follows realistic spatio-temporal behaviors and that it represents feasible AC optimal power flow (ACOPF) test cases.

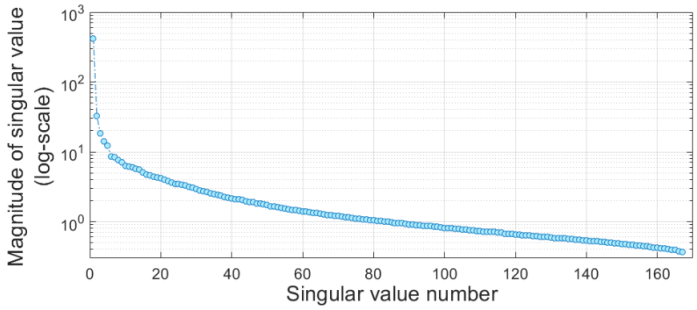

Fig. 1. Magnitude of the singular values.

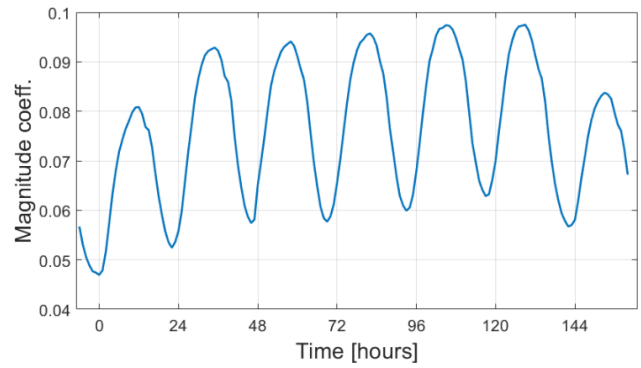

Fig. 2. Temporal profile corresponding to the largest singular value. Each hour multiple of 24 indicates midnight of the corresponding day.

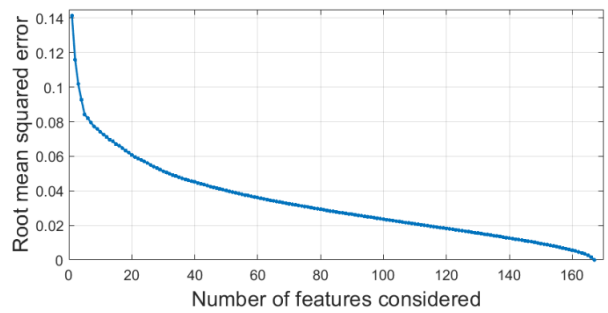

Fig. 3. Root mean squared error as a function of the number of features used to approximate the real loads.

II. Dataset description

The real data used in this work is proprietary and was provided by a large American independent system operator.

Our proposed technique requires two pieces of data: the bus-level historical load values and the topology of the system. The load data can be represented as a matrix $P \in \mathbb{R}^{n \times t}$, where n is the number of load buses in the system, and t is the number of time samples. The proprietary dataset contains more than 3500 loads, each sampled at hourly intervals for 167 consecutive hours (which is one hour short of a full week). The topology of the system can be represented as an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of all buses and $\mathcal{E}$ is the set of branches.

III. Temporal correlation

A. Principal component analysis

At a transmission level, the loads represent aggregates of residential, commercial, and industrial entities. Based on the assumption that the loads within each type behave similarly (especially residential and commercial), in a power system we can expect to observe common profiles among all loads. An effective way to identify and extract patterns from a dataset is by using PCA via singular value decomposition (SVD). The load matrix P can be factorized using SVD as $P = U\Sigma V^T$, where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{t \times t}$ are unitary matrices and $\Sigma \in \mathbb{R}^{n \times t}$ is an upper diagonal matrix. This factorization is able to extract and rank the common basis which, via linear combination, can reconstruct each load profile. In particular, the rows of V^T, which are vectors of size $1 \times t$, correspond to archetypal *temporal profiles* and they constitute the principal components. Each diagonal element of Σ, called a *singular value*, represents a scale factor which multiplies each corresponding principal component. Because the singular values are sorted from largest to smallest, they give an indication on the relative importance of each temporal profile contained in the V matrix. Fig. 1 shows the singular values obtained from the factorization of P, and it is clear that the first value is much larger than the following ones. Thus, the temporal profile corresponding to the first principal component, shown in Fig. 2, is the most dominant in determining the behavior of the loads. This profile shows the simplest and most common behavior: the load increases during the day, reaches a peak around noon and decreases in the evening. Moreover, considering that this data starts on a Sunday, we can see how the weekend peaks are lower than those of weekdays. The load profile at each individual bus is obtained as a linear combination of the principal components (columns of V) scaled by the singular values and multiplied by the corresponding coefficients in each row of U. These coefficients determine the composition of any given bus in terms of the archetypal profiles constituted by the principal components.

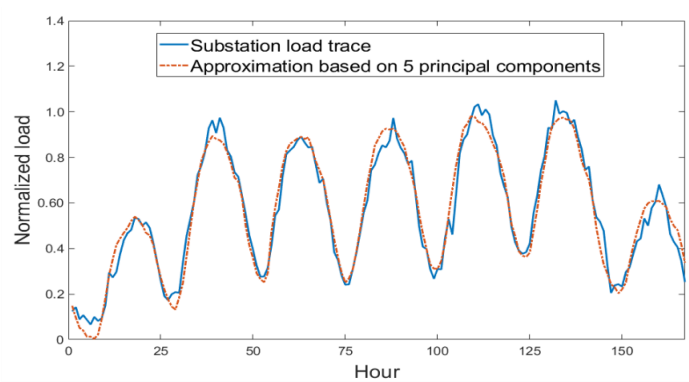

Fig. 4. Load trace across one week for an example substation. Also shown is the approximation based on the first 5 largest features.

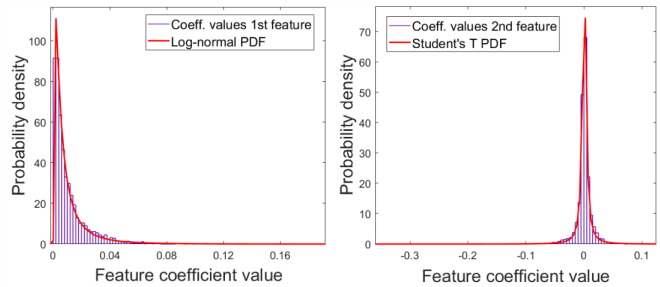

Fig. 5. Empirical and estimated probability density functions for the first feature (left) and the second feature (right) of the SVD load model.

B. Feature selection

As Fig. 1 shows, the first few rows of V^T have a significantly higher weight compared to the remaining ones, meaning that the original load matrix P can be approximated with good accuracy using only a subset of the basis vectors, or principal components. To demonstrate this fact, an approximation $\hat{P}$ of the original load matrix can be computed as $\hat{P} = U(:, 1:f)\Sigma(1:f:, 1:f)V^T(1:f, :)$, where $1 \leq f \leq t$ is the number of temporal profiles with the largest singular values.[1] The average root mean squared error (RMSE) as a function of the number of features used is shown in Fig. 3 and it allows for two observations. First, as one would expect, increasing the number of basis leads to a progressively better approximation which reaches an error of zero for $f = t$. Second, the error decreases sharply and almost linearly until $f = 5$, and then it slowly decays to zero. This means that the first five features can capture the main behavior of the load and they are sufficient to generate realistic synthetic load profiles. Fig. 4 is an example showing a real load and its approximation using a limited number of temporal profiles (notice that the load values have been normalized for anonymity reasons). It can be seen that the behavior of the load is captured very accurately while the small magnitude variability is smoothened out; this drawback is addressed by adding random noise, as explained in the next section.

C. Temporal generative model

Having identified some typical patterns, a new load profile can be created by generating a vector of coefficients and multiplying it by the set of base profiles contained in V. To compute these new coefficients we need to learn the distribution of the coefficients in the original data (e.g. the columns of U). The probability distribution functions (PDFs) are estimated using the Matlab Distribution Fitter App. Each column of U is analyzed independently and the best PDF for each is determined. Fig. 5 shows a histogram representation of the empirical coefficients for the first two features and the respective fitted PDFs as an example.

Since it was determined that the first five features will be used to generate the new data, this fitting procedure is performed for the first five columns of U. Table I shows the selected PDFs and their defining parameters. It is interesting to notice that the first coefficient is best approximated by a log-normal function, while all the successive ones follow a Student's t distribution.

[1]Notation: $U(:, 1:f)$, $\Sigma(1:f:, 1:f)$, and $V^T(1:f, :)$ indicate the first f columns of U, the first f columns and rows of Σ, and first f rows of V^T respectively.

TABLE I
PROBABILITY DISTRIBTUION FUNCTIONS

Feature number	PDF	μ	σ	ν
1	log-normal	-5.13	1.15	-
2	Student's t	7.55×10^{-4}	3.7×10^{-3}	1.16
3	Student's t	1.29×10^{-4}	4.7×10^{-3}	1.26
4	Student's t	1.51×10^{-3}	3.5×10^{-3}	1.08
5	Student's t	1.01×10^{-3}	4.3×10^{-3}	1.18

At this point, to generate new profiles it is sufficient to create a new coefficient matrix U_{new}, where each entry is sampled from the appropriate distribution, and multiply it by $\Sigma \in \mathbb{R}^{f \times f}$ and $V^T \in \mathbb{R}^{f \times t}$, where f is the chosen number of basis to be used ($f = 5$ in our case). The remaining uncertainty which is not captured by using a limited number of features is approximated by adding random noise to the profiles resulting from the above generative process. The noise has been empirically estimated to be normally distributed, with zero mean and $\sigma = 0.02$. The resulting generative model for m new loads can be written as

$$P_{\text{new}} = U_{\text{new}} \Sigma V^T + W \tag{1}$$

where $P_{\text{new}} \in \mathbb{R}^{m \times t}$, $U_{\text{new}} \in \mathbb{R}^{m \times f}$, $\Sigma \in \mathbb{R}^{f \times f}$, $V^T \in \mathbb{R}^{f \times t}$, and $W \in \mathbb{R}^{m \times t}$ is the matrix whose entries are sampled from $\mathcal{N}(0, \sigma^2)$.

The model is tested by generating 3000 synthetic load profiles using the first five basis of the real data, both with and without noise. Each resulting load matrix is then decomposed and approximated via SVD using an increasing number of features in the same way as done on the original data in Section III-B. The average root mean squared error of the synthetic data with and without noise is shown in Fig. 6. We can see that in the absence of noise, as one would expect, the approximation error reaches zero when five features are used to reconstruct the data. When noise is included, the error follows a similar curve to that of the original data, shown in Fig. 3. Thus, we can confirm that the generative model is able to capture both the predominant, long-term load behaviors as well as the short-term, high variability and randomness of the data. An example of load profile generated using the model in (1) is shown in Fig. 7.

IV. Spatial correlation

The profiles resulting from (1) are generated independently of each other, which, in general, is not a valid assumption about the loads in a power system. Realistically, loads that are geographically close should show some degree of correlation in their temporal behaviors. This is true because of two factors:

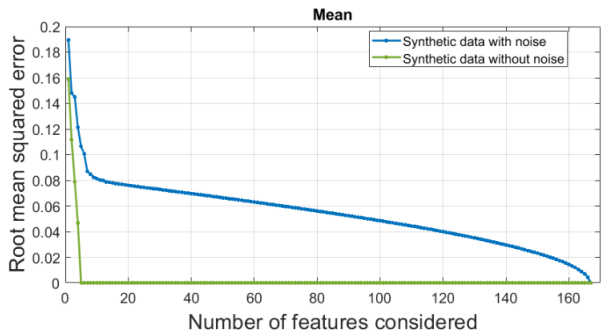

Fig. 6. Root mean squared error as a function of the number of features used to approximate synthetic data generated with and without the addition of noise.

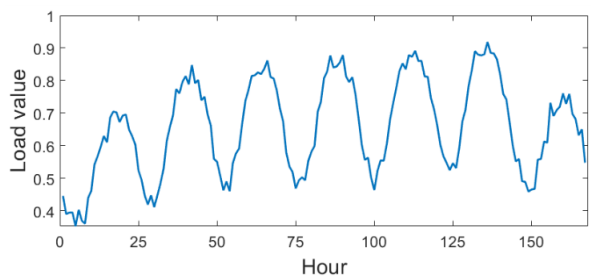

Fig. 7. Example load trace across one week generated using the model described by (1).

(i) nearby loads are likely to be of the same type (residential, commercial, industrial, etc..), and (ii) geography-dependent factors (such as weather conditions) will affect neighboring loads in similar ways. For this reason, it is important for any generative model to take into account the spatio-temporal correlation between loads that can be learned from a real dataset.

Let us define $b_i = \{b_{i,1}, b_{i,2}, ..., b_{i,t}\}$ as the load vector for bus i from time 1 to t; the correlation coefficient $r_{i,j}$ between the load vectors of buses i and j is:

$$r_{i,j} = \frac{\sum_{k=1}^{t}(b_{i,k} - \bar{b}_i)(b_{j,k} - \bar{b}_j)}{\sum_{k=1}^{t}(b_{i,k} - \bar{b}_i)^2 \sum_{k=1}^{t}(b_{j,k} - \bar{b}_j)^2} \tag{2}$$

where $\bar{b}$ indicates the sample mean. To understand the spatial characteristics of the original dataset, the correlation coefficient between buses is computed for every combination (i, j) with $1 \leq i \leq m$ and $1 \leq j \leq m$. Furthermore, each value $r_{i,j}$ is paired with the distance between the two corresponding buses, indicated as $\text{dist}_{i,j}$ and defined as the number of branches along the shortest path connecting buses i and j. The correlation coefficients are then collected as a function of their associated distance and the following metrics are computed: mean, standard deviation, and 25th, 50th, and 75th percentiles. This process allows us to understand at a general level how the similarity between buses varies as a function of their relative distance. Fig. 8(a) shows the statistics computed for the real loads dataset. We can see that, on average, as the distance between two buses increases the correlation slowly decreases, confirming the previous hypothesis that some spatial correlation exists.

To capture this behavior, the individual coefficients of matrix U_{new} from the model in (1) must be modified to take into account the values of the neighboring buses. To do so, the coefficients are first generated by randomly drawing from the estimated distributions as described in Section III-C. Then, each row of U_{new} is modified by adding to the coefficients vector of each bus a linear combination of the randomly generated vectors of the neighboring buses. Simulations have shown that the best results are obtained when each bus is modified by taking into account its neighbors within a maximum distance of 3. Moreover, the scaling factor that multiplies the coefficients of the neighbors is defined as a function of the distance, such that the greater the distance between two buses, the smaller the scaling factor. Formally, the model is rewritten as

$$P_{\text{new}} = (DU_{\text{new}})\Sigma V^T + W \tag{3}$$

where $D \in \mathbb{R}^{m \times m}$, and each entry is computed as

$$d_{i,j} = \begin{cases} 1, & \text{if } i = j \\ e^{-2\text{dist}_{i,j}}, & \text{if } \text{dist}_{i,j} \leq 3 \text{ and } i \neq j \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

V. Testing of the Generative Model on the Polish Test Case

We test our proposed generative model by creating load profiles for the publicly available grid model for the country of Poland [11]. This test case, commonly referred to as the 2383 Polish case, contains 2383 buses and 1822 loads. For each load, random coefficients are sampled from the five distributions described in Table I and the matrix D is computed based on the topology of the Polish system. Equation (3) is used on this data along with the singular values and principal components from Sections III-A and III-B respectively.

A. Spatial correlation

The spatial characteristics of the generated data are shown in Fig. 8. Plot (b) shows the statistics of the correlation coefficients as a function of the distance between buses for the synthetic Polish load data generated using (3). Plot (c) shows the same metrics for synthetic data resulting from (1), which does not take into consideration the spatial correlation between loads. It is clear how this latter graph greatly differs from that of the real data, while the data generated considering the spatial characteristics of the loads closely matches the behaviors of real loads. These results prove that the weight matrix D with exponential coefficients in (4) closely approximates the correlation between load profiles observed in real world data.

B. Loads scaling

The last step in creating realistic historical load data consists in scaling the profiles obtained from (3) to the values of the base case loads. This can be done by scalar multiplication of P_{new} by the column vector $S \in \mathbb{R}^m$. The entries of this

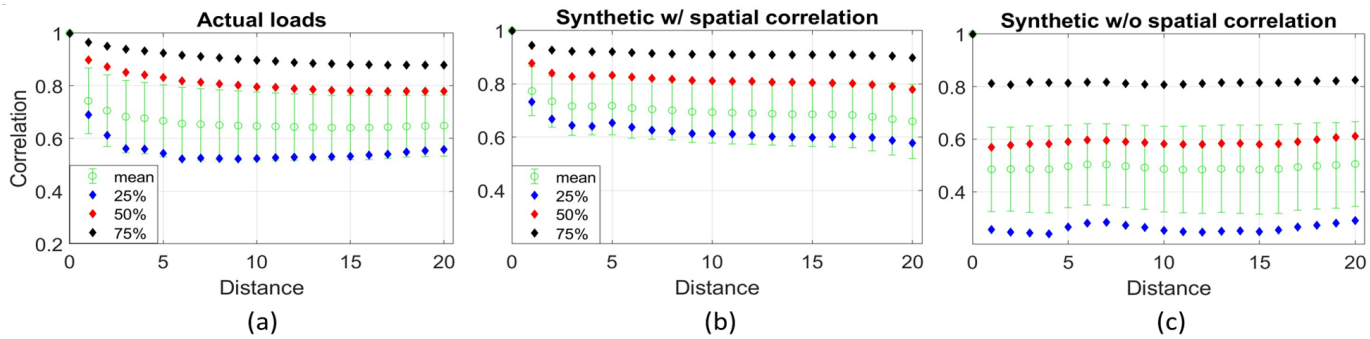

Fig. 8. Statistics of the correlation coefficients between load profiles as a function of the distance between buses, for the original data (left) and synthetic data generated considering the spatial correlation (center) and without considering it (right).

scaling vector can be determined in various ways, depending on the nature of the available base case loads. For example, s_i can be defined as $s_i = P_{\text{base},i}/f(P_{\text{new},i})$, where $f(P_{\text{new},i})$ is any function of the profile i, such as average, minimum etc. In the case of the Polish system, the loads provided in the model represent peak hour values so we have selected the scaling function to be $f(P_{\text{new},i}) = \max(P_{\text{new},i})$. In this way, the maximum load values of the synthetic historical data will coincide with the original base case loads.

C. Power flow validation

Having demonstrated the realistic nature of the generated loads in terms of spatio-temporal characteristics, it is important to verify that this data actually represents feasible cases from a power system perspective. To this end, ACOPF is run on every hour of the synthetic historical data to check for convergence. The results of this study show that all the 167 synthetic cases lead to converging solutions with bus voltages within the limits and moderate levels of congestion. This confirms that the generated loads can be supplied without exceeding transmission capacity, while leading to valid operating cases.

VI. Conclusion

We have presented a method to generate realistic historical load data for any grid model starting from a real dataset. The learning algorithm based on principal component analysis has been demonstrated to extract typical temporal patterns which can be linearly combined to generate new time-series data. The final generative model includes weighting factors which guarantee that the spatial characteristics that were observed in the real data are maintained in the synthetic data. Finally, we have verified that all the resulting cases represent valid power system operating conditions.

One of the drawbacks of this generative model is the fact that the length of the synthetic data in terms of time is limited to the length of the real data that is available. In our case, the data is generated at hourly interval for one consecutive week. In our future work we intend to devise a strategy to be able to extend the time-length of the data to multiple weeks while capturing the monthly and seasonal changes throughout the year. A fairly straightforward approach would be to scale the generated weekly loads according to the changes of the net or zonal loads, for which data is readily available as discussed in the introduction.

VII. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1449080.

References

[1] ARPA-E, "GRID DATA." [Online]. Available: https://arpa-e.energy.gov/?q=arpa-e-programs/grid-data

[2] CAISO, "CAISO historical EMS hourly load." [Online]. Available: http://www.caiso.com/planning/Pages/ReliabilityRequirements/Default.aspx#Historical.

[3] PJM, "PJM metered load data." [Online]. Available: https://dataminer2.pjm.com/feed/hrl_load_metered/definition.

[4] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, "The IEEE reliability test system-1996. A report prepared by the reliability test system task force of the application of probability methods subcommittee," *IEEE Transactions on Power Systems*, vol. 14, no. 3, pp. 1010–1020, Aug 1999.

[5] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," in *2018 IEEE Texas Power and Energy Conference, TPEC 2018*, 2018.

[6] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 3258–3265, July 2017.

[7] A. Khotanzad, R. Afkhami-Rohani, T.-L. Lu, A. Abaye, M. Davis, and D. J. Maratukulam, "ANNSTLF - A neural-network-based electric load forecasting system," *IEEE Transactions on Neural Networks*, vol. 8, no. 4, pp. 835–846, July 1997.

[8] I. Nadtoka, S. Vyalkova, and F. Makhmaddzonov, "Maximal electrical load modeling and forecasting for the Tajikistan power system based on principal component analysis," in *2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, May 2017, pp. 1–4.

[9] L. Xiao-fei and S. Li-qun, "Power system load forecasting by improved principal component analysis and neural network," in *2016 IEEE International Conference on High Voltage Engineering and Application (ICHVE)*, Sept 2016, pp. 1–4.

[10] F. M. Bianchi, E. D. Santis, A. Rizzi, and A. Sadeghian, "Short-term electric load forecasting using echo state networks and PCA decomposition," *IEEE Access*, vol. 3, pp. 1931–1943, 2015.

[11] R. D. Zimmerman, C. E. Murillo-Sanchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, 2011.