Stochastic Synthetic Data Generation for Electric Net Load and Its Application

M. Vivienne Liu
Cornell University
ml2589@cornell.edu

Patrick M. Reed
Cornell University
patrick.reed@cornell.edu

C. Lindsay Anderson
Cornell University
cla28@cornell.edu

Abstract

The increasing integration of renewable energy in electric power systems focuses attention on realistic representation of "net load" because it aggregates the information from both demand and the renewable supply side; net load is the remaining demand that must be met by non-renewable resources. However, the net load data is not readily accessible because of cost, privacy and security concerns. Furthermore, even if historical data is available, multiple stochastic scenarios are often needed for a wide range of power system applications. To address these issues, this paper proposes a stochastic synthetic net load profile generation approach. A seasonal detrending technique is combined with the modified Fractional Gaussian Noise method to deal with the complex multi-periodic seasonal trends in the net load profile. A thorough statistical validation and temporal correlation check are performed to show the quality of the synthetic data. The benefits of the synthetic data are demonstrated by a microgrid energy management problem.

1. Introduction

The system load profile is essential to a wide range of power system management applications including energy resource planning, electricity market clearing, risk assessment, reliability analysis, and policy design [1, 2]. However, as the effects of climate change intensify, integration of intermittent renewable energy resources is accelerating[3, 4]. Under these conditions, the load profile is insufficient to readily inform most power system operation and planning applications. As a result of the increasingly distributed nature of renewable energy resources, they are often consumed or stored locally (behind the meter) and invisible to the system operator [5]. Thus, instead of using load profile, net load profile, which is defined as the difference of load and any behind-the-meter energy, is increasing in importance[6].

Net load profile is not easily accessible by researchers and the reason is two-fold; first, it is expensive to install and maintain ubiquitous equipment necessary to collect electricity demand and renewable generation data with high spatial and temporal resolution; second, renewable generation is relatively nascent at large scale, and as result, historical data is scarce (for example, only 3 years' data is available from CAISO) [7]. Thus, a novel modeling strategy to capture the stochastic time-varying behavior of both the electricity demands and the renewable supply would fill an important need in power systems planning and operations research.

In addition to lack of availability of historical net load data, the growing interest in stochastic optimization and sequential time-series solutions [8, 9], and approximate dynamic programming, policy-based decision making approaches, [10, 11, 12, 13] requires a large set of plausible scenarios of net load input data to inform and enhance the decision making process. The set of stochastic net load profiles must preserve the statistical properties and temporal correlation of historical records [14].

Significant attention has been paid to time-varying synthetic renewable energy profile generation [15, 16, 17, 18, 19, 20, 21, 22, 23]. The Markovian state transition property has been an underlying assumption of the renewable energy behavior in much of this literature; first and second order Markov Chain-based methods have been proposed by [16, 17, 19, 20, 21, 23]. Most of these methods were able to preserve the probability distribution of the historical records, but temporal correlation has only been considered in [16, 20, 20] though these efforts leave a significant gap between the synthetic and historical data. Authors in [22] used a Fourier series and auto-regressive moving average model to capture the characteristics of historical data. The synthetic profiles showed promising performance in quantile and cumulative density function validations, though the temporal correlation was not explored as a key feature for time-series data generation.

URI: https://hdl.handle.net/10125/70998
978-0-9981331-4-0

Relative to interest in synthetic renewable energy profiles, load/net load profile generation has not received sufficient attention. The methods used to generate renewable energy profiles are not readily applicable to load/net load profiles directly because of the complex multi-periodical seasonal trends in electricity demands. In [24] Liu and Maldonado, performed a probabilistic analysis of masked loads with PV penetration behind-the-meter using gaussian processes to model the spatial correlation and a stochastic differential equation to capture temporal correlation. The theoretical foundation is promising, but no statistical characteristics or temporal correlation analysis were shown to validate the effectiveness of the model. Pillai et al. focused on using weather and load data as inputs for an artificial neural network (ANN) to generate synthetic data [25]. ANN was effective in reducing the root mean square error (RMSE) between the real and synthetic data but failed to maintain the temporal characteristic given the underlying assumption of independent and identically distributed (iid) inputs to the model. In [26], Pinceti et al. concentrated on generating transmission grid level synthetic load dataset that maintain spatio-temporal features. While the spatial correlation was well preserved, the dataset was limited to one consecutive week and therefore was unable to model the monthly and seasonal changes over a longer time period. Overall, less attention has been paid to synthetic load/net load profile generation. For existing models, a thorough validation process is needed to assess the effectiveness of each approach and to better understand the advantages of using the synthetic data to better capture the internal variability and extremes when net load is treated as a stochastic process.

Given the dearth of studies of temporal correlation and thorough statistical validation [27] of long-term synthetic load/net load profile generation, this paper seeks to introduce a long-term stochastic synthetic net load profile generation approach by embedding a seasonal detrending technique into the modified Fractional Gaussian Noise (mFGN) method [28]. "Comparing to other synthetic data generation methods such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), the mFGN method is easy and fast to avoiding the conceptual and computational complexity of identifying and training network architecture as well as feature selection, which are potentially both problematic with highly temporally correlated data." The mFGN method has shown very promising results for producing long-term weekly-streamflow data and replicates the temporal correlation. However, the hourly net load has a higher temporal resolution and a much more complex multi-periodical seasonality than weekly-streamflow data. Thus a seasonal detrending process should be embedded into the mFGN method to accommodate these trends in net load data. The derived synthetic samples should be statistically compared to the historical record and applied to a real-world case study to test its effectiveness. The key contributions of this paper are:

- development of a generalizable approach to generate long-term synthetic time series data with complex, multi-periodical seasonal trends, with the ability to replicate characteristics of historical data
- thorough statistical validation and temporal correlation evaluation on the synthetic net load profiles, and
- demonstration of the superior performance of using stochastic synthetic net load profiles on a multiobjective microgrid energy management problem under uncertainty

The paper is organized as follows: Section 2 provides an overview of the CASIO data set and some special characteristics of the net load data. Section 3 introduces the mFGN method and the seasonal detrending technique to generate the synthetic records. The statistical validation is performed in Section 4 to prove the fidelity of the synthetic data. Section 5 demonstrates an application of the synthetic profiles and Section 6 concludes the paper.

2. Dataset Description

The dataset used in this paper is drawn from the California Independent System Operator (CAISO) [7]. Interested reader is referred to find the data under "System Demand" category. The historically recorded hourly wind generation, solar generation, and load profile for year 2017-2019 is used and a small number of missing points are filled with the mean of adjacent points. Figure 1 shows the wind generation, solar generation, and load profile for January 1st, 2019. Net load is calculated, and indicated by the red line. The shape of the aggregated net load is primarily driven by load, with additional uncertainty introduced by renewable energy. California data is selected for use in this paper for two reasons; first, as previously mentioned, real renewable energy data is very limited and CAISO has three full years of data; second, the California data is strongly influenced by solar energy, which creates a very deep valley in the middle of the day. The resulting twin-peak of net load profile

adds an additional approximate 12-hour period to the net load profile. This additional period makes the dataset even more challenging to model and provides a perfect opportunity to test the robustness of the proposed method.

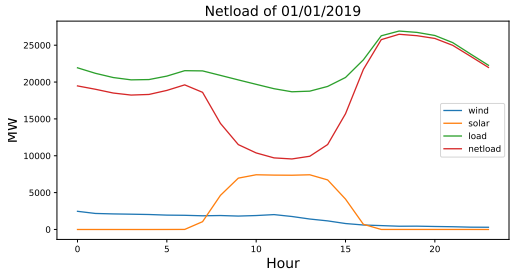


Figure 1. Wind, solar generation and load, net load profile on Jan 1, 2019.

3. Methodology

In this section, the modified Fractional Gaussian Noise (mFGN) method [28], which has shown promising performance to generate stochastic synthetic streamflow records, is introduced. However, since the streamflow data is recorded weekly, the seasonal complexity is not as high as that of the net load data. Thus, a seasonal detrending technique is implemented to further stabilize the net load historical records for the mFGN method.

3.1. Modified Fractional Gaussian Noise

Fractional Gaussian Noise [29] method was first proposed in the late 1960s, for use in hydrological data applications, due to its ability to preserve correlations of normally distributed datasets without seasonal trends. Kirsch et el. improved it to be able to deal with a relatively low level of seasonal patterns in the correlation structure [28]. This characteristic makes it effective for climatological systems, most notably streamflow data. Following the mFGN method in [28], the historical time series net load data that usually stored in a vector y is reformulated into a matrix $\mathbf{Y}$. Given N days' record, matrix $\mathbf{Y}$ will be $N \times 24$ with each row representing hourly data for a single day. Each column of $\mathbf{Y}$ is normalized according to:

$$Y_{i,j} = (Y_{i,j} - \overline{Y_j})/\sigma_j \tag{1}$$

where $\overline{Y_j}$ is the mean of column j and σ_j is the standard deviation of column j. If there is no significant multiple seasonal trends in the data, after such normalization, the data will be an approximate $\mathcal{N}(0,1)$ distribution. In the case of the net load data in this paper, the existence of multiple seasonal trends requires further detrending, which will be introduced in the following sub-section. The rest of this sub-section focuses on describing the mFGN method in the net load dataset context.

To generate synthetic data, a matrix $\mathbf{X}$, which has the same dimension as $\mathbf{Y}$, is bootstrapped through an intermediate matrix $\mathbf{M}$. Each entry of $\mathbf{M}$ is sampled with replacement from the set $[1, 2, ..., N]$. Then, $\mathbf{X}$ is formed such that:

$$X_{i,j} = Y_{(M_{i,j}),j} \tag{2}$$

Since $\mathbf{X}$ is independently bootstrapped through $\mathbf{M}$, the temporal correlation is lost. To reintroduce the correlation to $\mathbf{X}$, the sample covariance matrix of $\mathbf{Y}$ is calculated and then decomposed with Cholesky decomposition [30]:

$$\mathbf{Corr(Y)} = \mathbf{QQ^T} \tag{3}$$

where $\mathbf{Q}$ is the upper triangular matrix that contains the correlation information for each row of $\mathbf{X}$. By applying the following equation:

$$\mathbf{Z} = \mathbf{XQ} \tag{4}$$

the upper triangular property of $\mathbf{Q}$ ensures that for $X_{i,j}$, namely the net load data of day i hour j would take into consideration the information carried by the net load data of hour 1 to $j-1$. As a result, matrix $\mathbf{Z}$ preserves the autocorrelation within each row of data but not between each row. It could be problematic when one wants to generate time series data longer than one day because only intra-daily correlation is maintained (not inter-daily correlation). Such a problem can be solved by constructing a matrix with a longer time horizon in each row. However, the dimension of the sample correlation matrix $\mathbf{corr(Y)}$ increases as the number of columns of $\mathbf{Y}$ increases. Furthermore, the computational complexity of Cholesky decomposition is $\mathbf{O}(n^3)$ where n is the number of columns of the correlation matrix. It is therefore intractable to generate year-long hourly data.

To deal with the challenge of creating longer data sets, a matrix manipulation technique is used to reconstruct the inter-daily correlation. Figure 2 is an adjusted visualization taken from [28] to show the matrix manipulation process as it is applied for net load data. As shown in Figure 2, matrix $\mathbf{Y'}$ takes the second half of the first day, i.e. hour $13-24$ of day one, and the first half of the second day, i.e hour $1-12$ of day two as the first synthetic day stored in row 1 of $\mathbf{Y'}$. The remaining rows of $\mathbf{Y'}$ are formed following the same rule. Thus $\mathbf{Y'}$ is a $(N-1) \times 24$ matrix because the first

half of day 1 and the second half of day N are lost. The loss of one day's data can be considered negligible for a year-long time series. $\mathbf{X}'$ is created in the same way as $\mathbf{X}$. $\mathbf{Q}'$ is decomposed by the sample correlation matrix of $\mathbf{Y}'$, which thus preserves the correlations between any two days (rows). Then, Equation (4) is applied to these reorganized matrices to derive $\mathbf{Z}'$:

$$\mathbf{Z}' = \mathbf{X}'\mathbf{Q}' \tag{5}$$

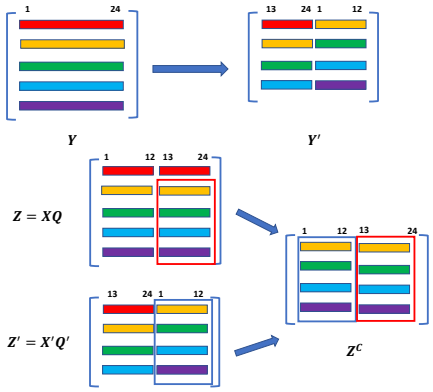


Figure 2. Visualization of the matrix manipulation to form $\mathbf{Y}'$ and $\mathbf{Z}^{\mathbf{c}}$

To assemble the final synthetic data, the second-half columns of $\mathbf{Z}$, which capture the intra-daily correlation, and the second half-columns of $\mathbf{Z}'$, which captures the inter-daily correlation, are combined into a new matrix $\mathbf{Z}^{\mathbf{c}}$ as highlighted by the blue and red boxes in Figure 2. By de-normalizing $\mathbf{Z}^{\mathbf{c}}$ and reformulating it into a vector $\mathbf{Z}^{\mathbf{c}}$, a stochastic synthetic trajectory is created. A more detailed description of the mFGN method and the comparison of mFGN with Auto-regression methods can be found in [28].

3.2. Seasonal Detrending

Figure 3 shows the autocorrelation of the net load from January 1st to Dec 31st, 2017. Peaks in autocorrelation occur every 24 hours, with higher peaks after one day and one week (highlighted by the red box), indicating a $T = 24$ and $T = 168$ periodical change. It is also worth noting that a small peak exists at around lag $12 \times x$, which represents the twin-peak of the net load data, previously mentioned in Section 2. Furthermore, the net load for 2017 shown in Figure 4 exhibits a distinct difference between summer (roughly Jun-Oct) and non-summer (roughly Nov-May) seasons of the year.

The idea of seasonal detrending or "whitening" the normalized net load data is to remove the impact caused by the summer and non-summer season and the weekly

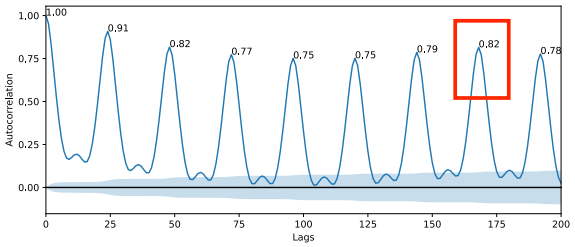


Figure 3. Autocorrelation of net load for 2017

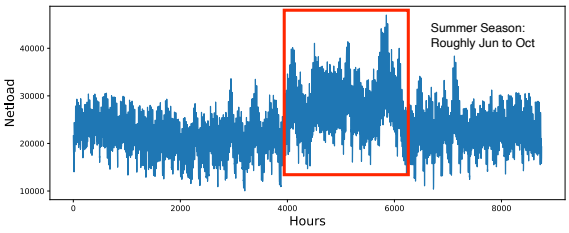


Figure 4. Net load for 2017: the red box highlighted the obvious difference for summer period from June to October

periodic impact. Thus, the "whitening" process has two steps. First, a year-round seasonal curve is fitted by:

$$y = a\sin(\frac{2\pi t}{c} - b) + d\cos(\frac{2\pi t}{f} - e) \tag{6}$$

where $a - f$ are the parameters to be estimated from the historical data. The combination of a sine and cosine functions is chosen to keep the formulation as simple as possible to avoid over-fitting, and to be easily generalizable for most of other seasonal time series data. Study cases of different horizons will be tested and presented in Section 4 to ensure that seasonal periodicity is captured reasonably well. Figure 5 shows an example of this curve fitting, with the red markers representing the normalized data and the blue markers illustrating the fitted seasonal curve. Figure 6 shows the data after the first seasonal detrending step. An approximately white noise signal is shown with several spikes. The spikes represent extreme events where peak demand happens. As there is no magnitude difference within the dataset, such spikes are kept to ensure the model's capability to replicate extreme events in the synthetic records.

The second step of the "whitening" process is to model the weekly seasonality by further detrending the data shown in Figure 6 with the following function:

$$y = ax + b + c\sin(\frac{2\pi t}{T} - d) + d\cos(\frac{2\pi t}{T} - f) \tag{7}$$

where t is the hour of the day, $T = 168$, and $a -$

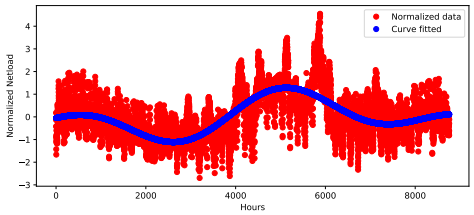


Figure 5. Curve fitted for yearly seasonality

f are parameters to be estimated. The reason only weekly periodic is included is that the mFGN method already takes care of the inter-daily correlation, so there's no need to include another sine or cosine function with daily periodic and causing potential over-fitting of the data. The fitted curve will be reapplied back to the "whitened" data before normalizing back. It will be shown in Section 4.3 that the seasonal detrending technique can stabilize the data and significantly improve the autocorrelation preservation.

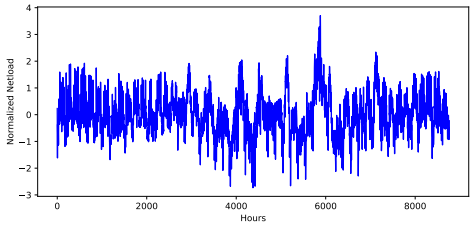


Figure 6. "Whitened" data after the first seasonal detrending step

4. Statistical Validation

In this section, a series of statistical validation tests are performed to confirm the efficacy of synthetic records in approximating the properties of the historical data. To have a better understanding of the impact of the time horizon on results, a short-term case, a mid-term case, and a long-term case (4, 8 and 12 months respectively) will be compared in terms of mean, standard deviation, cumulative distribution, autocorrelation, and percentile performance. Finally, a concatenation approach will be performed to further improve long-term data generation.

4.1. Mean and Standard Deviation

To compare the performance of the original mFGN method and with the additional seasonal detrending technique, 1000 samples are generated for short-term (4 months), mid-term (8 months) and a long-term (12 months), respectively. The means and standard deviations of the 1000 synthetic samples should be statistically similar to the historical mean and standard deviation. Welch's t-test [31] and Levene's test [32] are performed for mean and standard deviation respectively. The results are shown in Tables 1 and 2.

Table 1. Percentage of synthetic samples with statistically similar mean to historical data using t-test. Significance set at $p < 0.05$

Term	mFGN	mFGN+deseasonalized
Short	71.7%	93.5%
Mid	65.5%	85.2%
Long	65.4%	83.1%

Table 2. Percentage of synthetic samples with statistically similar standard deviation to historical data using Levene test. Significance set at $p < 0.05$

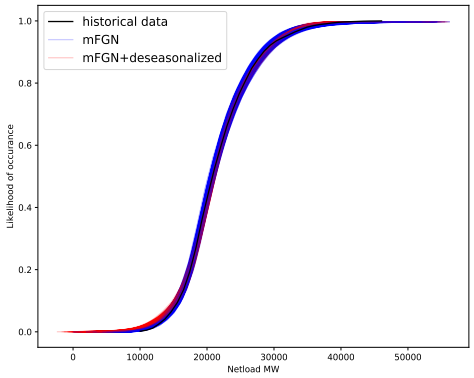
Term	mFGN	mFGN+deseasonalized
Short	80.5%	80.9%
Mid	77.3%	90.5%
Long	75.7%	86.1%

A much higher percentage of samples have staticically similar mean and standard deviation when the extra "whitening" process is involved, which means the synthetic data generated is more similar to the historical data in general.

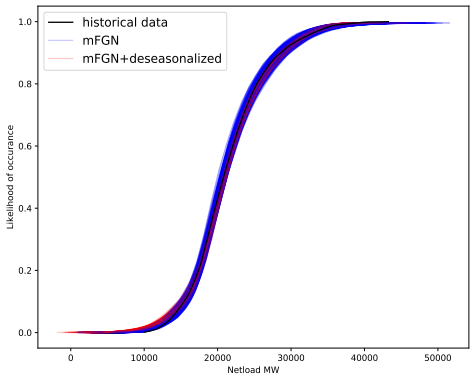
4.2. Cumulative distribution

In addition to the ability to preserve the mean and standard deviation of the historical data, it is also necessary to explore whether the synthetic data is fully representing the statistical distribution of the historical data. The two-sample Kolmogorov-Smirnov (K-S) [33] test is one of the goodness to fit test for comparing the empirical cumulative distribution functions of the historical data and synthetic samples. The K-S test indicates that all 1000 samples derived from both approaches are drawn from the same underlying distribution as the historical data, which means that the synthetic records generated by both approaches replicate the historical cumulative distribution very well. Figure 7 shows the empirical cumulative distribution of 1000 short, mid and long term synthetic samples with mFGN in blue, seasonal detrending technique embedded mFGN in red, and historical data in black. The tails of both ends are fully covered by the synthetic data distribution, which means the synthetic data is able to reproduce rare events. Upon closer inspection, the extended tails of the synthetic distribution show a wider range of values for net load profile than historical data

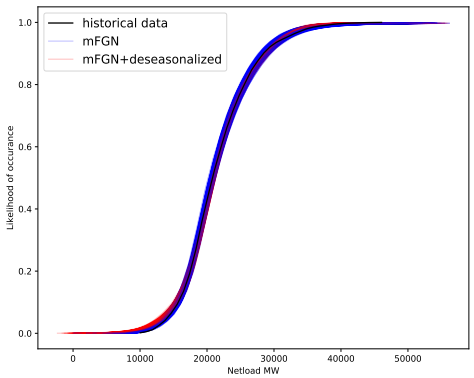
with a few samples showing negative values for net load. While this may seem a lack of fit at first glance, with 1000 "parallel universes", it is reasonable to see a few "universes" showing more extreme values at the tails. One could easily drop the samples that are not reasonable for specific study cases.



(a).Short-term



(b).Mid-term



(c).Long-term

Figure 7. Cumulative distribution of 1000 synthetic samples from mFGN and mFGN+seasonal detrending. Short-term (a), Mid-term (b), Long-term (c)

4.3. Temporal Correlation

Temporal correlation plays an important role in maintaining the structure of time series data, and has not always received sufficient consideration in the past. Figures 8, 9, and 10 compare the temporal correlation preservation for a range of different time horizons. In each figure, the temporal correlation of 1000 samples generated from mFGN, and the seasonal detrending technique embedded in mFGN, are shown in blue and red, respectively. The temporal correlation drawn from the historical data is shown as the thick black line and the light blue shaded area is the 95% confidence interval. The x-axis shows the lags from current time t and y-axis shows the autocorrelation between the current time and different lagged times. For example, assuming the current hour is t, the value at lag 1 represents the autocorrelation between time t and $t-1$. If a value at lag a is with in the 95% confidence interval, then at the 95% confidence level, the sample autocorrelation is considered to be 0, which means the autocorrelation between time t and $t-a$ is 0. As shown in the figures, data generated with the seasonal detrending process outperforms the data generated with mFGN only in all three cases. The time horizion of the autocorrealtion checking is chosen to be 400 to cover the weekly periods twice. Although as the horizon increases, the red curves depart from the historical record occationally, the detrending technique reconstructs almost identical temporal correlation for the short-term (4 months) scenario. Another observation is that the 12-hour period caused by the twin-peaks is well captured by the synthetic data. As the time series gets longer, the temporal structure starts to deteriorate a bit more. To fill the gap a bit more, a concatenation approach will be demonstrated in Section 4.5 taking advantage of the superior performance of the short-term case.

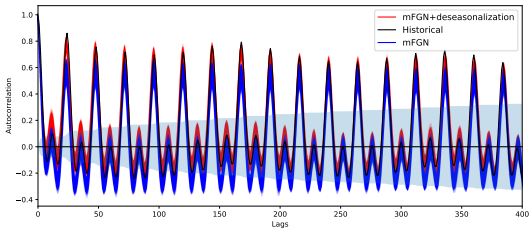


Figure 8. Comparing temporal correlation of synthetic data and historical data for the short-term (4 months) case.

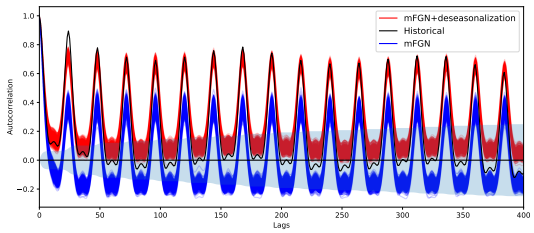


Figure 9. Comparing temporal correlation of synthetic data and historical data for the mid-term (8 months) case.

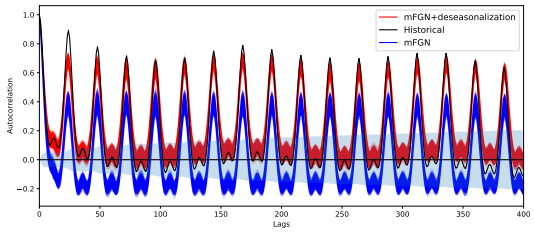


Figure 10. Comparing temporal correlation of synthetic data and historical data for the long-term (12 months) case.

4.4. Percentile

In previous sections, the aggregated performance of the stochastically generated synthetic data was assessed based on statistical comparison with the historical data set. In this section, we consider the detailed performance over three consecutive days. Figure 4 shows that the net load profile has very different behavior for the summer and non-summer periods. Thus, a non-summer three-day case is randomly chosen to compare with the summer three-day case where the peak net load of the year is included in the middle day. As the mFGN augmented with a detrending technique outperformed the original mFGN method, the synthetic samples of both summer and non-summer cases are generated with the seasonal detrending embedded mFGN.

Figure 11 shows the non-summer case and Figure 12 shows the summer case with the heavy black line representing the historical data and the percentile of the synthetic data in the color scale. During the non-summer season, the historical data lies in the $30-50$ percentile range of the synthetic data. However, percentile ranges are much broader for the middle of the day when net load is driven by solar generation resulting in more uncertainty. The performance is very different for the peak net load summer days, where the net load is dominated by the electric consumption associated with cooling demand. The model shows a reasonably robust performance to re-generate the extreme events. As shown in Figure 12 , the historical line settles in the $60-80\%$ range, which means out of the 1000 samples, 200-400 of the synthetic data is able to recreate the most rare events in historical data.

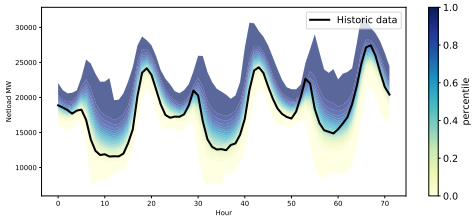


Figure 11. Percentile plot for a consecutive 3 non-summer season days

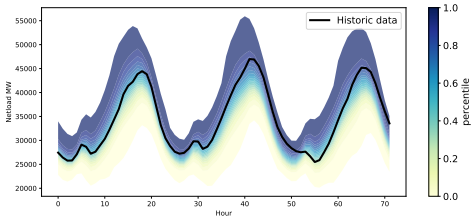


Figure 12. Percentile plot for a consecutive 3 summer season days

4.5. Concatenation approach

To further improve the temporal correlation reconstruction for the long-term scenario, the data for year 2017 is decomposed into summer and non-summer seasons. The rationale behind this approach is that, the summer season has a very unique moving trend within itself as shown in Figure 5 and such a trend cannot be captured by a simple periodic fixed sine and cosine function combination. Thus, a separate seasonal detrending and mFGN process are performed on summer season (roughly Jun-Oct) and the summer season synthetic data is put back to the whole year sequence after the separate process. Such a concatenation approach is able to be performed because when bootstrapping the intermediate matrix $\mathbf{M}$, the random seed could be fixed for the summer and non-summer season. In other words, $\mathbf{M}$ is sampled once for the full year, but the detreading curves are fitted separately for the summer and non-summer seasons. The random consistency in $\mathbf{M}$ ensures the consistency of the concatenated re-sampled trajectories. The result of the concatenation approach is shown in Figure 13, where the gap between the synthetic data and the historical records is significantly reduced.

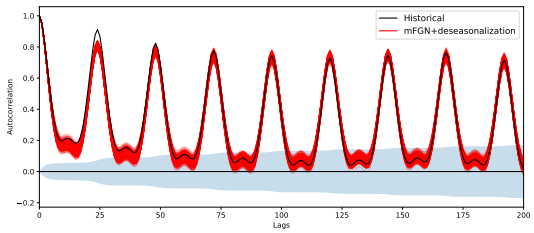


Figure 13. Improved temporal correlation with the concatenate approach.

5. Application

In this section, a microgrid energy management problem is used to show an example application of the synthetic data. In this section, we compare the performance of the strategies (also called policies) developed by the model with historical data, to those developed with the synthetic data. After a description of the microgrid system, we introduce the Evolutionary Multi-Objective Direct Policy Search (EMODPS) framework, which is used to solve the microgrid energy management problem. The section will conclude with a comparison of policies arising from the historical and synthetic datasets.

5.1. Microgrid Energy Management

The context of this analysis is a microgrid that consists of a local diesel generator, a wind farm, a solar farm, and a battery storage unit. The microgrid is connected to the main utility grid and can buy or sell energy with the utility grid. A demand response program is implemented to provide limited flexibility in shifting peak load across the day. Distinct from the traditional economic dispatch setting, where only cost (or revenue) is considered as the optimization objective, in this case, four objectives are modeled explicitly including: expected revenue, expected CO_2 emissions, reliability, and expected daily average ramping. Such explicit multi-objective modeling enables the exploration of the complex trade-offs between different stakeholder perspectives and could provide deep insights on engaging consumer participation and system operations under uncertainty [12]. The goal of EMODPS is to find the best policies to control the daily local generation, battery charging/discharging, and energy exchange with the utility grid. In this case, CAISO data is downscaled to be appropriate for the microgrid context.

5.2. Evolutionary Multi-Objective Direct Policy Search

EMODPS [34] is a simulation-based optimization framework that contains two major components: a simulation model that mimics the microgrid operation dynamics and a multi-objective evolutionary algorithm (MOEA) that searches for the best control parameters based on the simulation results. Borg [35] is the MOEA implemented in this instance given its robust performance over a diverse set of multi-objective problems, where its performance is identical or superior to other state-of-the-art MOEAs [36, 37, 38]. In the simulation model, instead of estimating individual sequential hourly decisions for the control variables, an approximate mapping is used to characterize the relationship between the current system states and the control actions shown as equation 8:

$$X = \mathbf{F}(S_t) \tag{8}$$

where S_t is the system state and X is the control action. $\mathbf{F}$ could be any universal mapping functions that can capture non-linear, non-convex complex system behavior, such as radial basis functions or artificial neural networks. Radial basis functions are used here given its ability to model the multi-mode property. The approximate mapping will be referred to as the *policy* for the rest of the paper. Given the space limitation of this paper, the detailed algorithm description is omitted here, though the interested reader is referred to [12]. The parameterized policies could be generalized on unforeseen system conditions, which provides a suitable test-bed to compare the generalizability of the synthetic net load profiles and the set of historical profiles.

5.3. Result Comparison

Historical data from Jan 2017, and 1000 synthetic samples for the same month are each used to train the operation policies, separately. At this point, it is worth noting that the multiobjective optimization does not result in a single point solution, but instead a *non-dominated* Pareto frontier. These solutions are non-dominated in that there is no other solution that improved one objective without some loss in one or more of the others. Using Borg, 140 policies are identified with the historical data, henceforth referred to as the historical policy, and 228 policies result from the synthetic data, which will be referred to as synthetic policy. All the policies are then re-evaluated on the Feb 2017 data, which was not used in training. In Figure 14, the combined Pareto frontier of the historical policies (in purple) and synthetic policies (in yellow) are shown.

Each marker represents the expected performance of one policy on Feb 2017, with 114 non-dominated policies resulting. Of these policies on the combined Pareto frontier, 31 are from historical policies and 83 are from the synthetic policies. This data indicates that the synthetic may outperform the historical policies, as 72.8% of the Pareto frontier arises from synthetic policies. It is likely that one-month of historical data is not enough to fully represent the underlying uncertainty of renewable energy and load profiles. Conversely, well-characterized synthetic data could provide a better sampling of the state space of the stochastic process, and generalize to unforeseen system states.

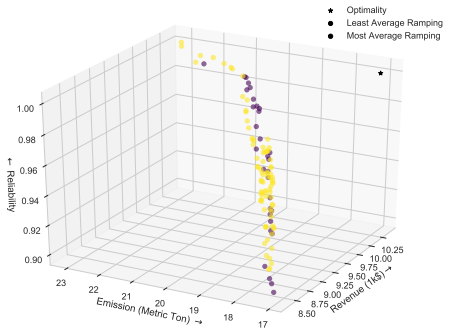


Figure 14. Combined pareto frontier of synthetic and historical policy .

To have a better understanding of the daily-level performance of the policies, two policies that have similar revenue and average ramping results are taken from the historical and synthetic policy sets, respectively. The objective values are shown in Table 3 and the daily behavior of one day (Feb 1st) is shown in Figure 15. The upper panel of Figure 15 shows the diurnal profile of the emission rate and the local marginal price of the utility grid. The middle panel shows the synthetic policy behavior and the lower panel shows the historical policy behavior. These results show that the synthetic policy achieves lower expected emissions, while maintaining similar revenue outcomes relative to the historical policy. Examining the red curves, it can be seen that the synthetic policy charges before the price peaks and discharges during the peak to reduce the electricity purchases from the utility grid. Furthermore, the synthetic policy also takes advantage of the demand response for load shifting, so that the generation (shown in brown) more closely follows the net load with demand response to avoid continuous generation at a higher level. Conversely, the historical policy fails to use the battery efficiently and does not coordinate the generation and energy exchange policy effectively, resulting in an inferior performance to the synthetic policy.

Table 3. Compare objectives of the synthetic and historical policy

Type of policy	Revenue (k$)	Emission (M-Ton)	Reliability (%)	Average Ramping (MW)
Historical	10.03	21.90	99.10	2.65
Synthetic	10.05	21.21	99.85	2.65

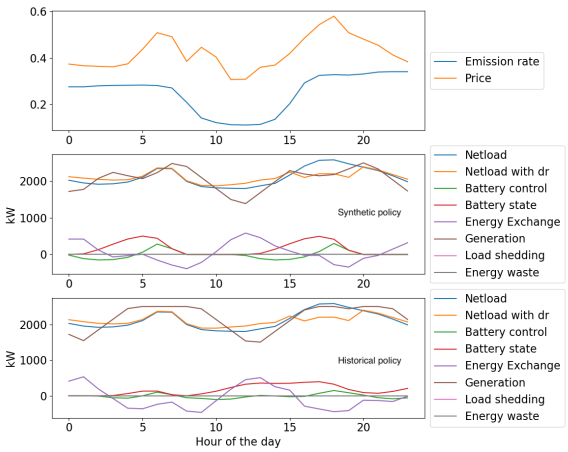


Figure 15. Daily behavior comparison of synthetic and historical policy.

6. Conclusion

This paper proposes a seasonal detrending technique to augment the mFGN method to improve suitability for time series data with complex multi-periodic trends. The method is tested on CAISO net load data which has a 12-hour period, a daily period, a weekly period, and a seasonal year long period. Statistical validation shows that the seasonal detrending process enhances the mFGN to generate synthetic data that is more statistically similar to the historical data. In terms of temporal correlation, the complex twin-peak of the CAISO data has been replicated effectively and the "whitening" process improved the temporal correlation to be almost identical to the historical data for a year-long horizon.

Other than the promising results related to statistical validation, there are several potential advantages. First, the proposed method is easily generalizable to other time series data with appropriate adjustment to the period of the sine and cosine functions, estimated through the autocorrelation analysis. Second, no additional data is required to replicate the process; only the time series itself is needed. In addition, the computational complexity grows linearly to the length of the time series input, instead of cubicly as a result

of the sample correlation matrix, which has a fixed dimension instead of growing with the horizon of the data.

Analysis of a microgrid energy management application shows that the synthetic data set can characterize the underlying uncertainty well, resulting in policies that are more generalizable to unforeseen system states than the historical data. Such a result indicates that this approach could be beneficial to many power system applications that make decisions under significant uncertainties. Future work could focus on considering the spatial correlation to generate wind, solar data for a regional space with multiple wind or solar farms. This work also provides a foundation for future work on network level net load profiles with spatial correlation, while extra attention might be needed to deal with the less smooth data at nodal level.

References

[1] H. Li, A. L. Bornsheuer, T. Xu, A. B. Birchfield, and T. J. Overbye, "Load modeling in synthetic electric grids," *2018 IEEE Texas Power and Energy Conference, TPEC 2018*, vol. 2018-February, pp. 1–6, 2018.

[2] T. Xu, A. B. Birchfield, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Application of Large-Scale Synthetic Power System Models for Energy Economic Studies," *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, pp. 3123–3129, 2017.

[3] A. Hirsch, Y. Parag, and J. Guerrero, "Microgrids: A review of technologies, key drivers, and outstanding issues," *Renewable and Sustainable Energy Reviews*, vol. 90, no. September 2017, pp. 402–411, 2018.

[4] W. Su, J. Wang, and J. Roh, "Stochastic energy scheduling in microgrids with intermittent renewable energy resources," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1876–1883, 2014.

[5] Y. Wang, N. Zhang, Q. Chen, D. S. Kirschen, P. Li, and Q. Xia, "Data-driven probabilistic net load forecasting with high penetration of behind-the-meter pv," *IEEE Transactions on Power Systems*, vol. 33, no. 3, pp. 3255–3264, 2018.

[6] S. Borlase, *Smart grids: Advanced technologies and solutions*. CRC Press, 2017.

[7] "Caiso," *California ISO*, p. http://oasis.caiso.com/mrioasis/logon.do, 2020.

[8] M. Manbachi and M. Ordonez, "AMI-based Energy Management for Islanded AC/DC Microgrids Utilizing Energy Conservation and Optimization," *IEEE Transactions on Smart Grid*, vol. 3053, no. c, 2017.

[9] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, "Dynamic Energy Management System for a Smart Microgrid," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 8, pp. 1643–1656, 2016.

[10] D. F. Salas and W. B. Powell, "Benchmarking a scalable approximate dynamic programming algorithm for stochastic control of multidimensional energy storage problems," *Dept Oper Res Financial Eng*, 2013.

[11] D. R. Jiang, T. V. Pham, W. B. Powell, D. F. Salas, and W. R. Scott, "A comparison of approximate dynamic programming techniques on benchmark energy storage problems: Does anything work?," in *2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pp. 1–8, IEEE, 2014.

[12] A. Gupta, M. Liu, D. Gold, P. Reed, and C. L. Anderson, "Exploring a direct policy search framework for multiobjective optimization of a microgrid energy management system," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[13] L. Zéphyr and C. L. Anderson, "Stochastic dynamic programming approach to managing power system uncertainty with distributed storage," *Computational Management Science*, vol. 15, no. 1, pp. 87–110, 2018.

[14] J. D. Herman, H. B. Zeff, J. R. Lamontagne, P. M. Reed, and G. W. Characklis, "Synthetic drought scenario generation to support bottom-up water supply vulnerability assessments," *Journal of Water Resources Planning and Management*, vol. 142, no. 11, pp. 1–13, 2016.

[15] R. Turner, X. Zheng, N. Gordon, M. Uddstrom, G. Pearson, R. de Vos, and S. Moore, "Creating synthetic wind speed time series for 15 new zealand wind farms," *Journal of applied meteorology and climatology*, vol. 50, no. 12, pp. 2394–2409, 2011.

[16] D. A. Halamay and T. K. Brekken, "A methodology for quantifying variability of renewable energy sources by reserve requirement calculation," in *2010 IEEE Energy Conversion Congress and Exposition*, pp. 666–673, IEEE, 2010.

[17] C. O. Inácio and C. L. T. Borges, "Stochastic model for generation of high-resolution irradiance data and estimation of power output of photovoltaic plants," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 952–960, 2017.

[18] C. M. Fernández-Peruchena and M. Gastón, "A simple and efficient procedure for increasing the temporal resolution of global horizontal solar irradiance series," *Renewable energy*, vol. 86, pp. 375–383, 2016.

[19] A. Shamshad, M. Bawadi, W. W. Hussin, T. Majid, and S. Sanusi, "First and second order markov chain models for synthetic generation of wind speed time series," *Energy*, vol. 30, no. 5, pp. 693–708, 2005.

[20] H. Aksoy, Z. F. Toprak, A. Aytek, and N. E. Ünal, "Stochastic generation of hourly mean wind speed data," *Renewable energy*, vol. 29, no. 14, pp. 2111–2131, 2004.

[21] B. O. Ngoko, H. Sugihara, and T. Funaki, "Synthetic generation of high temporal resolution solar radiation data using Markov models," *Solar Energy*, vol. 103, pp. 160–170, 2014.

[22] J. Chen and C. Rabiti, "Synthetic wind speed scenarios generation for probabilistic analysis of hybrid energy systems," *Energy*, vol. 120, pp. 507–517, 2017.

[23] F. O. Hocaoglu, O. N. Gerek, and M. Kurban, "The effect of markov chain state size for synthetic wind speed generation," *Proceedings of the 10th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2008*, pp. 113–116, 2008.

[24] S. Liu, D. A. Maldonado, and E. M. Constantinescu, "Probabilistic analysis of masked loads with aggregated photovoltaic production," 2020.

[25] G. G. Pillai, G. A. Putrus, and N. M. Pearsall, "Generation of synthetic benchmark electrical load profiles using publicly available load and weather data," *International Journal of Electrical Power and Energy Systems*, vol. 61, pp. 1–10, 2014.

[26] A. Pinceti, O. Kosut, and L. Sankar, "Data-Driven Generation of Synthetic Load Datasets Preserving Spatio-Temporal Features," *IEEE Power and Energy Society General Meeting*, vol. 2019-August, 2019.

[27] R. A. Davis, S. H. Holan, R. Lund, and N. Ravishanker, *Handbook of discrete-valued time series*. CRC Press, 2016.

[28] B. R. Kirsch, G. W. Characklis, and H. B. Zeff, "Evaluating the impact of alternative hydro-climate scenarios on transfer agreements: Practical improvement for generating synthetic streamflows," *Journal of Water Resources Planning and Management*, vol. 139, no. 4, pp. 396–406, 2013.

[29] B. B. Mandelbrot and J. W. V. Ness, "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Review*, vol. 10, no. 4, pp. 422–437, 1968.

[30] N. Higham, "Cholesky factorization," Sept. 2009.

[31] B. L. WELCH, "THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED," *Biometrika*, vol. 34, pp. 28–35, 01 1947.

[32] H. Levene, "Contributions to probability and statistics," *Essays in honor of Harold Hotelling*, pp. 278–292, 1960.

[33] M. A. Stephens, "Edf statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association*, vol. 69, no. 347, pp. 730–737, 1974.

[34] D. P. Bertsekas, *Reinforcement learning and optimal control*. Athena Scientific Belmont, MA, 2019.

[35] D. Hadka and P. Reed, "Borg: An auto-adaptive many-objective evolutionary computing framework," *Evolutionary Computation*, vol. 21, no. 2, pp. 231–259, 2013.

[36] D. Hadka, P. M. Reed, and T. W. Simpson, "Diagnostic assessment of the borg moea for many-objective product family design problems," in *2012 IEEE Congress on Evolutionary Computation*, pp. 1–10, 2012.

[37] D. Hadka and P. Reed, "Diagnostic assessment of search controls and failure modes in many-objective evolutionary optimization," *Evolutionary Computation*, vol. 20, no. 3, pp. 423–452, 2012.

[38] J. Zatarain Salazar, P. M. Reed, J. D. Herman, M. Giuliani, and A. Castelletti, "A diagnostic assessment of evolutionary algorithms for multi-objective surface water reservoir control," *Advances in Water Resources*, vol. 92, pp. 172 – 185, 2016.