Synthetic Time Series Data

MAT 280 Final Project
Noah Perry

March 24, 2023

1 Time Series Data

1.1 Overview

Time series data is where multiple data points are gathered about the same individual or object at different points in time. Since the data points are not independent of each other in time series data, the correlation between data points must be accounted for in any statistical analysis. This correlation introduces extra complexity but also facilitates prediction.

Time series data occur in a wide variety of contexts:

- Google and Apple collect smartphone location data repeatedly throughout the course of a day.
- Utilities and phone service providers collect customer-specific data on electricity or data usage over time for billing purposes.
- Cryptocurrency exchanges such as Binance and Kraken release ”tick” data where each row corresponds to a trade made on their exchange.
- An electrocardiogram (EKG) shows the electrical signals of a patient’s heart over time.
- A patient who is receiving treatment may make regular hospital visit and have their weight and blood pressure recorded at each appointment.

For some of these examples, the data are released as streams. From the time you read this sentence until you finish reading this report, Google or Apple may have collected your location data multiple times, your electrical usage will have increased, and the price of Bitcoin will likely have changed.

1.2 Statistical Properties

Some fundamental statistical properties of time-series data include:

1. **Stationarity:** Many statistical methods for time-series data are predicated on the assumption of stationarity. In many cases, non-stationary data can be transformed to become stationary so that the standard set of time-series tools may be applied. For example, stock prices are often non-stationary and the stationary log returns are analyzed instead:
$$\text{Log return: } log(\frac{y_t}{y_{t-1}}) = log(y_t) - log(y_{t-1})$$

2. **Autocorrelation:** Any patterns in the autocorrelation function and the rate of its decay are important for modeling and forecasting.

$$\text{Autocorrelation: } Corr(y_t, y_{t-k})$$

3. **Trend and Seasonality:** These components must be accounted for in modeling and may be eliminated in some cases.

2 Privacy and Data Release

2.1 Differential Privacy

The unique nature of time series data presents complications for privacy. The standard differential privacy (DP) definition involves randomized query responses on two datasets X and X' that only differ in one row. The basic idea is that for an individual included in the data, no additional information can be inferred about them based on the query response than if they were not included in the data (with high probability). For time series data, multiple rows correspond to the same individual, so the standard DP framework is incompatible and variations of DP have been developed.

It is important to note that while differentially private methods are theoretically attractive, a rigorous mathematical privacy guarantee does not necessarily imply "good" privacy results. For example, to maintain data utility, some high-profile users of DP methods (Google, U.S. Census Bureau) use very large ϵ values. By doing this, they are able to "check a privacy box," but the guaranteed level of privacy is essentially meaningless.

In addition, there are several limitations of a query response framework. First, a privacy budget is needed because privacy would be completely lost if an unlimited number of queries were allowed to be made. Second, the core of any statistical analysis is multivariate relationships. When variables in the data are correlated, the added noise must be adjusted accordingly as a query response about one variable may be disclosive about a correlated variable. This has a detrimental affect on the utility of the responses as noise levels would need to be increased.

Furthermore, there are also practical limitations. Using the DP Laplacian mechanism, the adequate amount of noise to be added to each response must be calibrated for each query. For simple queries like means or counts, calibrating noise and sending a response is not difficult. However, for more more complex statistical analyses, the implementation may not be trivial. For practitioners with time constraints, this may limit the range of possible queries that can be answer. For example, open source code implementing a DP response for regression coefficients may be available but quantile or ridge regression may not yet be available. Hence, the data analysis tools available to a query sender are limited by the development progress of the scientific community and the practical constraints of the organization holding the data.

2.2 Data Release

The issues mentioned above provide motivation for data release. Releasing an anonymized dataset allows data analysts to freely perform any statistical analysis desired without the limitations of a query-response framework. There are many established statistical disclosure control

(SDC) methods such as microaggregation and data swapping that aim to perturb the data in a systematic way so that the utility and statistical integrity of the data are approximately preserved while making it difficult to learn anything about a specific individual in the data. However, a major drawback is that the open-ended nature of these approaches make it difficult to provide rigorous mathematical guarantees for privacy. Some work has been made in this direction (e.g. probabilistic k-anonymity for rank swapping and microaggregation) but much is left to be desired.

An alternative to the traditional SDC methods is synthetic data. The main idea is the same: release a dataset that approximately captures the relationships and fundamental statistical properties of the original data without comprising the privacy of the individuals whose data is used. Releasing synthetic data may be preferable to releasing anonymized data because there is no direct correspondence between a row in the synthetic dataset and any particular individual in the original data.

Synthetic data is also useful for purposes beyond privacy preserving data release. It may also be generated for balancing datasets for improved model performance and fairness. A machine learning model applied to a classification task may not be as accurate when the dataset is highly imbalanced. For example, a bank that is developing a model to detect fraudulent transactions would typically have highly imbalanced training data because the majority of the transactions are not fraudulent. Therefore, creating synthetic fraudulent transactions to balance the training data may result in better model performance.

3 Generative Adversarial Networks

In a Generative Adversarial Network (GAN), two neural networks, a generator G and a discriminator D, compete against each other in a game to create and detect fake data. At each iteration, the generator provides the discriminator with a sample where some elements come from the true data and some do not. The discriminator responds by giving a probability that each element of the sample comes from the true data. During the training process, the generator's ability to create realistic-looking fake data and the discriminator's ability to distinguish between real and fake data are both honed at each iteration. The training process ends when the generator has become able to generate convincing enough fake data that the discriminator is essentially randomly guessing what samples are fake and real. Once this point has been reached, the GAN is prepared to generate synthetic data.

In summary, D attempts to maximize the following function V while G attempts to minimize it:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))]$$

4 Empirical Analysis

4.1 Data

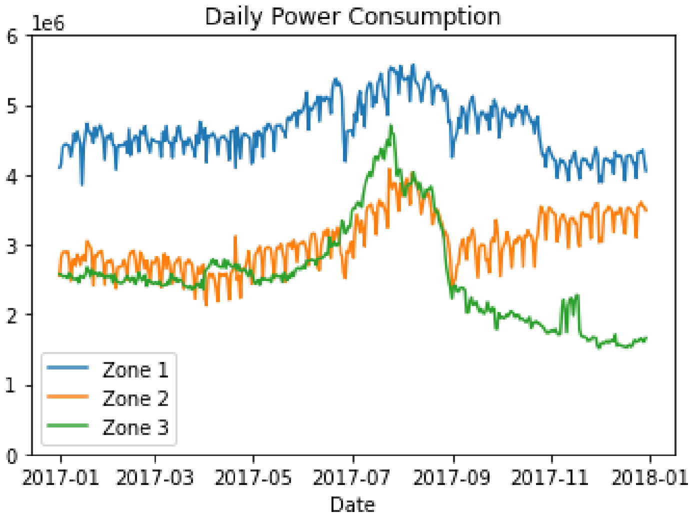
I used **ydata-synthetic**, a python module with implementations for a variety of GANs for synthetic data generation including one for time-series data: TimeGAN. TimeGAN differs from other GANs in its use of two additional networks and supervised loss in training. In addition to the generator and discrimator networks, TimeGAN also includes recovery and embedder networks. These four networks all interact through a latent space where the conditional distribution of the time series is learned with reference to the original data.

Using TimeGAN, I attempted to generate a synthetic version of a dataset containing weather conditions and power consumption in Tetouan, Morocco from January 1, 2017 to December 30, 2017. The original dataset from the UCI Machine Learning Repository gives data at 10-minute intervals (52,417 observations). From my initial experimentation with the TimeGAN, I found that the run time can be quite significant even for small datasets, so I aggregated the data to a daily level for my analysis (364 observations). Using the aggregated data, the joint networks training stage took around 10 hours to run using 500 steps.

Variable	**Aggregation Function**
Date	N/A
Temperature	Mean
Humidity	Mean
Wind speed	Mean
General diffuse flows	Sum
Diffuse flows	Sum
Zone 1 power consumption	Sum
Zone 2 power consumption	Sum
Zone 3 power consumption	Sum

Table 1: Variables in Tetouan Power Consumption Data

The following plot depicts the daily power consumption time series for each zone of the city in the original data. We see that the time series for zones 1 and 2 are quite different from zone 3. The power consumption in zones 1 and 2 varies noticeably by day of the week but experiences less extreme seasonal fluctuations. This may be because zones 1 and 2 are more industrial while zone 3 is more residential.



4.2 Privacy Scenarios

Two hypothetical scenarios to motivate the use of an aggregated power consumption data for a privacy analysis:

1. If two competing companies have large factories in different zones of Tetouan, they may attempt to use the power consumption data to learn something about the production activities of their competitor. If an important piece of equipment breaks down in one factory, this may be reflected in the power consumption data as a short period where power consumption in that zone is uncharacteristically low for that time of year.

2. Companies who are unsatisfied with their profit margins sometimes decide to collude with their competitors to restrict supply of a product to drive up its price (a violation of competition law in many countries). As there is often some economic incentive to cheat on the agreement, participants in these conspiracies often attempt to detect cheating by other members of the cartel. Since frequent communications may leave an undesirable paper trail, the conspirators may resort to more indirect means. One possible method could be monitoring power consumption data for increases in their competitor's zone of the city.

4.3 Analyses

We perform several basic statistical analyses to compare the real and synthetic data and check how well TimeGAN has captured the fundamental properties of the time series. All the referenced plots are at the end of the report.

4.3.1 Time Series Plots

We plot each variable's time series in both the real (blue) and synthetic (orange) data. We immediately notice that the orange lines are consistently much smoother than the blue lines. In a predictive modeling context, this would likely be regarded as a positive attribute - the model appears to have avoided overfitting to the training data. This may also be advantageous for privacy in some ways. Outlying values are challenging to provide privacy to as they tend to be among the most easily identifiable data points. The smoothness of synthetic data may help obfuscate rare events like the one mentioned in the first privacy scenario above.

While the synthetic data tends to follow the general trends in real data, the dates where sharp changes or local max/min values occur in the synthetic data is not exactly contemporaneous with those events in the real data. For example, in the zone 2 power consumption chart, we see that a sharp downward increase occurs around late August, whereas in the synthetic data a similarly sharp downward change does occur but not until 1-2 months later. In the second privacy scenario, this would help conceal the timing of the activities of each factory from their competitors.

We noted earlier that for zones 1 and 2 of the city, the amount of power consumed varies by day of the week. This is an important property of these time series, but the synthetic data appears to not have picked up on it at all.

4.3.2 Autocorrelation Plots

We plot the sample autocorrelation function for power consumption of each zone. In all three plots, the synthetic data captures the general shape of the autocorrelation functions and decays to 0 at a very similar rate. The day of the week effect in zones 1 and 2 is quite visible in these plots as the sample autocorrelation jumps up every 7th lag in the real data but this pattern is completely missing in the synthetic data.

4.3.3 Pairwise Correlations

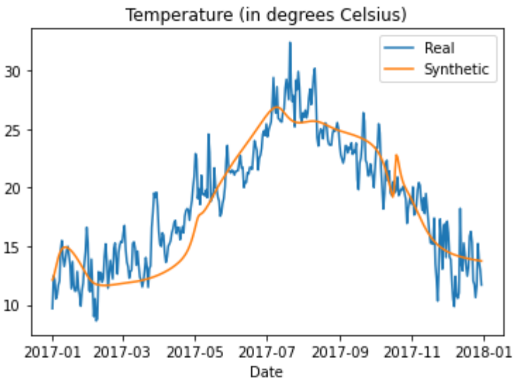
We also examine the correlations between the different variables in the data. As seen in the color-coded correlation matrices below, the general correlation structure of the data is preserved in the synthetic data. When applying many DP or SDC methods, correlation structure typically degrades and correlations fade to 0 as the level of noise or perturbation increases. For our synthetic data, however, we see that some correlations are even stronger than in the original data. This is likely due to the smoothness of the synthetic time series that we observed above.

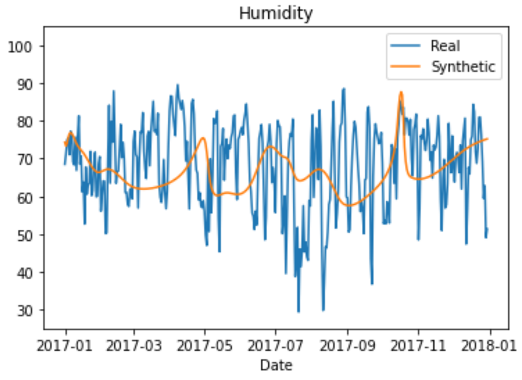
5 Summary of Findings

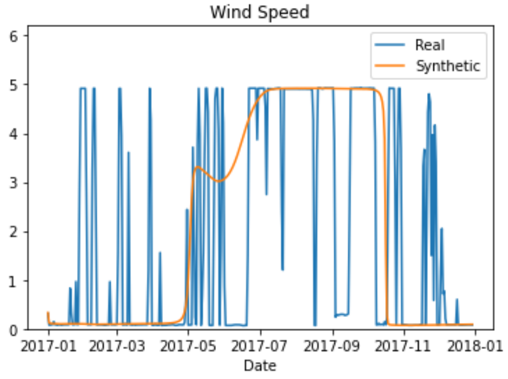
We performed a variety of basic statistical analyses and visualizations using a synthetic version of the Tetouan power consumption data generated by a time series-oriented GAN. Any privacy preserving data release method must carefully balance the competing goals of utility and privacy. Overall, our findings are positive as the synthetic data captured many of fundamental properties of the time series (good for utility) without exactly mimicking the original data (good for privacy).

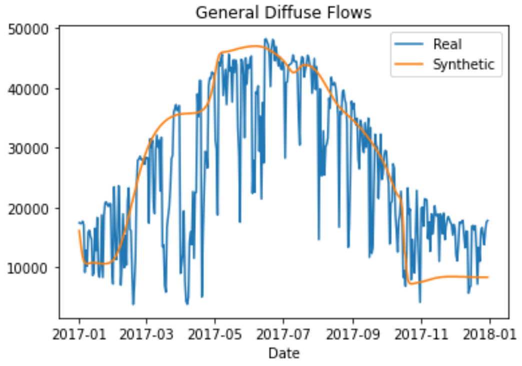
6 Ideas for Future Work

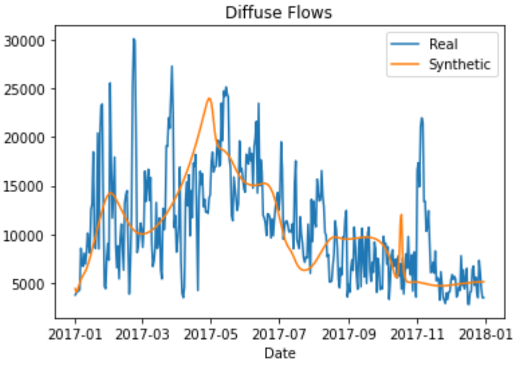
- Run TimeGAN using more training steps to see if how sensitive the results are to this aspect of the algorithm.
- Compare the performance of time series models such as Vector Autoregression on the real and synthetic data.
- Simulate a random walk and then create a synthetic version using TimeGAN. Apply a Augmented Dicky-Fuller test to whether the synthetic version is still a random walk.
- Compare the performance of TimeGAN with the much simpler Fourier Perturbation Algorithm, which is differentially private.
- Financial market data has some unique and throroughly studied features such as volatility clustering. Apply TimeGAN to financial market data and examine how well these features are captured.

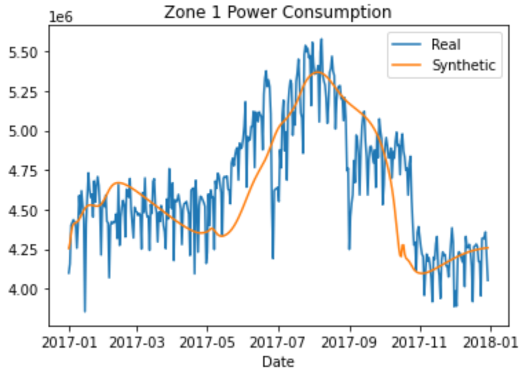

Temperature (in degrees Celsius)
Real
Synthetic
30
25
20
15
10
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date

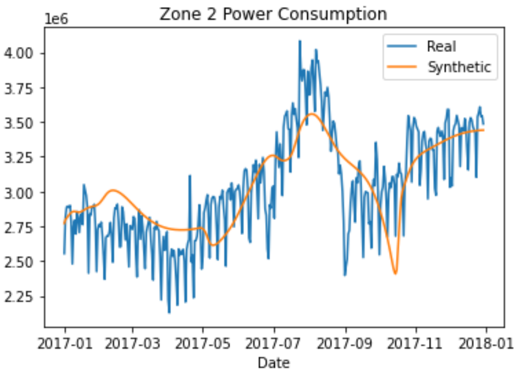

Humidity
Real
Synthetic
100
90
80
70
60
50
40
30
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date

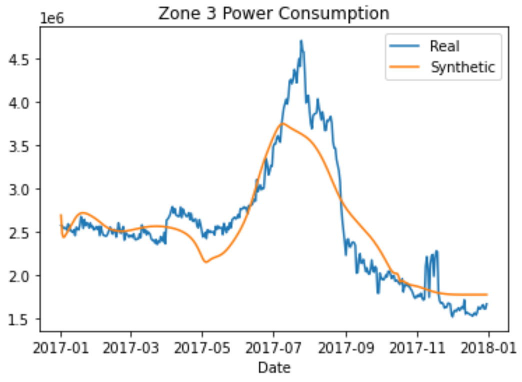

Wind Speed
Real
Synthetic
6
5
4
3
2
1
0
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date

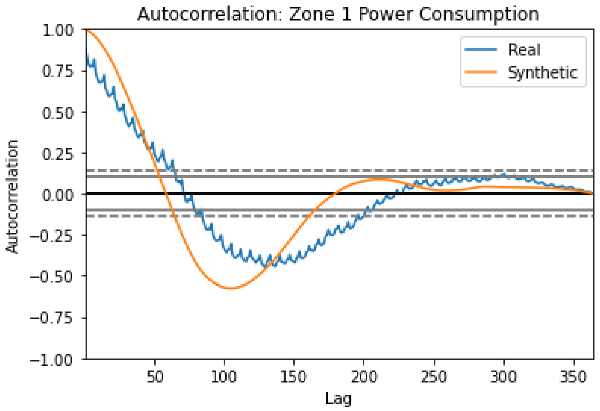

General Diffuse Flows
Real
Synthetic
50000
40000
30000
20000
10000
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date


Diffuse Flows
Real
Synthetic
30000
25000
20000
15000
10000
5000
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date

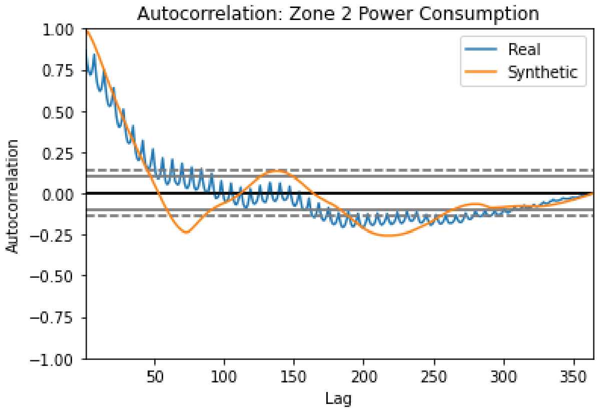

1e6
Zone 1 Power Consumption
Real
Synthetic
5.50
5.25
5.00
4.75
4.50
4.25
4.00
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date


1e6
Zone 2 Power Consumption
Real
Synthetic
4.00
3.75
3.50
3.25
3.00
2.75
2.50
2.25
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date

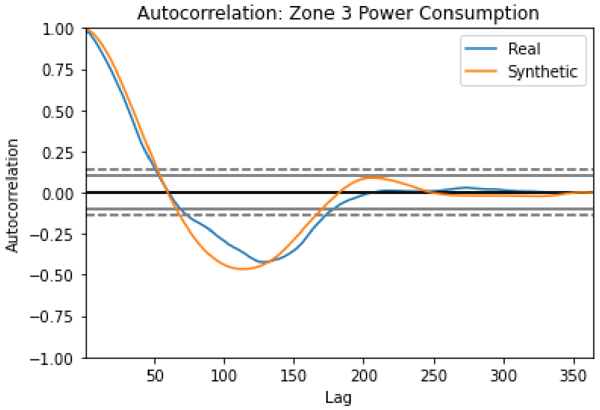

1e6
Zone 3 Power Consumption
Real
Synthetic
4.5
4.0
3.5
3.0
2.5
2.0
1.5
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
Date



Autocorrelation: Zone 1 Power Consumption

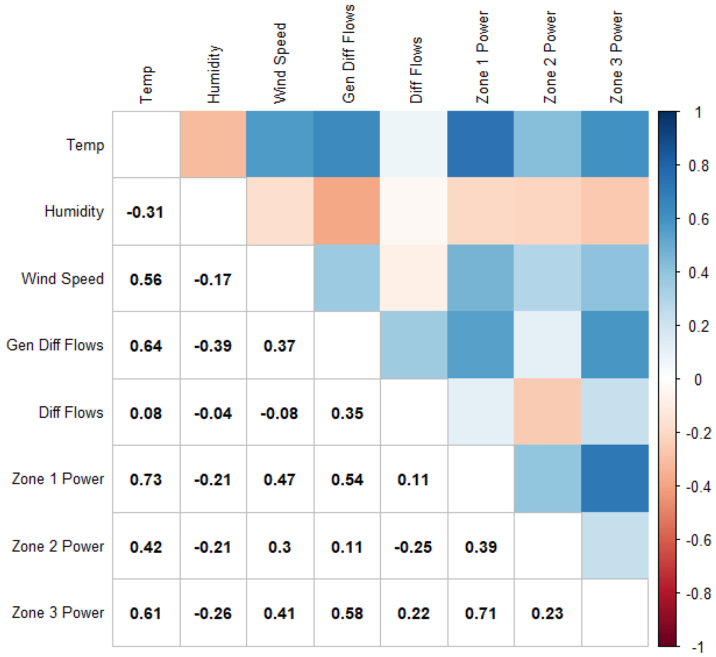


Autocorrelation: Zone 2 Power Consumption

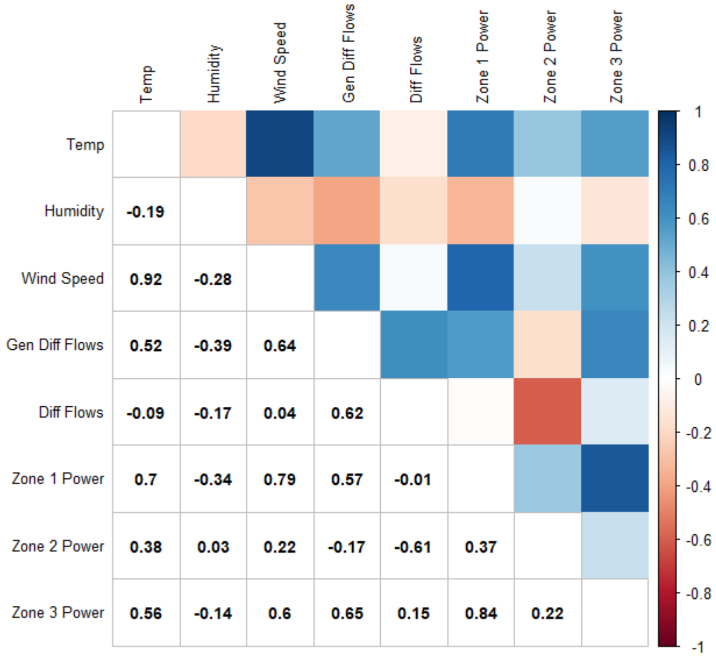


Autocorrelation: Zone 3 Power Consumption

Real Data



Synthetic Data


Temp
Humidity
Wind Speed
Gen Diff Flows
Diff Flows
Zone 1 Power
Zone 2 Power
Zone 3 Power
Temp
Humidity -0.19
Wind Speed 0.92 -0.28
Gen Diff Flows 0.52 -0.39 0.64
Diff Flows -0.09 -0.17 0.04 0.62
Zone 1 Power 0.7 -0.34 0.79 0.57 -0.01
Zone 2 Power 0.38 0.03 0.22 -0.17 -0.61 0.37
Zone 3 Power 0.56 -0.14 0.6 0.65 0.15 0.84 0.22