INL/CON-19-52623-Revision-0

Correlated Synthetic Time Series Generation using Fourier and ARMA

Paul W Talbot, Cristian Rabiti, Andrea Alfonsi, Cameron J Krome, Ross Kunz, Aaron S Epiney, Congjian Wang, Diego Mandelli

June 2019

The INL is a U.S. Department of Energy National Laboratory
operated by Battelle Energy Alliance

INL/CON-19-52623-Revision-0

Correlated Synthetic Time Series Generation using Fourier and ARMA

Paul W Talbot, Cristian Rabiti, Andrea Alfonsi, Cameron J Krome, Ross Kunz, Aaron S Epiney, Congjian Wang, Diego Mandelli

June 2019

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy

Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Correlated Synthetic Time Series Generation using Fourier and ARMA

Paul Talbot, Cristian Rabiti, Andrea Alfonsi, Cameron Krome, M. Ross Kunz, Aaron Epiney, Congjian Wang, Diego Mandelli

Nuclear Engineering Methods Development, Idaho National Laboratory, Idaho Falls, Idaho, paul.talbot@inl.gov

INTRODUCTION

As the contribution of renewable energy grows in electricity markets, the complexity of nuclear's place within the energy mix increases, and likewise the need for robust simulation techniques. While decades of wind, solar, and demand profiles can sometimes be obtained, this is too few samples to provide a statistically meaningful analysis of a system with baseload, peaker, and renewable generation. Synthetic time series generation presents itself as a suitable methodology to meet this need. One approach for synthetic series generation is training a model using Fourier series decomposition for seasonal patterns and Auto-Regressive Moving Average models (ARMA) to describe time-correlated statistical noise about the seasonal patterns. When combined, the Fourier plus ARMA (FARMA) model has been shown to provide an infinite set of independent, identically-distributed sample time series with the same statistical properties as the original data [1]. See also references in [1] for examples of how this work has been applied in other fields.

When considering an energy mix with renewable electricity production, several time series of energy, grid, and weather measurements are needed for each synthetic year modeled to statistically comprehend the efficiency of any given energy mix. These cannot be considered independent series in a given synthetic year. To capture and reproduce the correlations that might exist in the measured histories, the ARMA can further be extended as a Vector ARMA (VARMA). In the VARMA algorithm, covariance in statistical noise is captured both within a history as part of the autoregressive moving average, and with respect to the other variables in the time series.

In this work the implementation of the Fourier VARMA in the RAVEN uncertainty quantification and risk analysis software framework [2] is presented, along with examples of correlated synthetic history generation.

Hybrid energy systems (HES) composed of multiple energy generators have been proposed as an effective strategy to mitigate some challenges on the grid that an increase renewable energy penetration bears [3] and provide scenarios to increase the viability of nuclear energy as part of an energy mix [4]. As the penetration of renewable energy in markets grows, the complexity of the mix needed to satisfy consumer demand also grows. This results in an involved, tightly-coupled system with various energy-producing elements including baseload, peaker, and renewable generators.

Prior work [5] has been focused on dynamic modeling, simulation, control, and optimization for HES. However, such prior work models renewable generation using historical weather conditions, for which there are a limited number of measurements. To expand the sample size, some effort has been made to generate synthetic weather scenarios, with artificial measurements that have the same fundamental statistics as the historical measurements but are independent and identically-distributed. For instance, combining Fourier series expansion with auto-regressive and moving-average (ARMA) models can capture both seasonal trends as well as sub-hourly statistical randomness [3]. The combined Fourier and ARMA (FARMA) methodology has previously been implemented as part of the Risk Analysis Virtual Environment (RAVEN) developed at Idaho National Laboratory [2]. One limitation of the existing implementation is the independence of various time series.

For example, consider a HES consisting of several components, including solar and wind farms. For simulating many weather scenarios, it is desirable for each sampled synthetic year to be independent of other synthetic years. However, in a particular synthetic year, it is not generally statistically accurate to treat solar availability, wind speeds, and consumer demand as three independent time series. For instance, during summer months in the northern hemisphere when the wind speed is low and solar availability is high, demand will also tend to be higher than when either wind speed is high or the solar availability is low. To address this limitation, the work described in this paper demonstrates development of a mechanism for training FVARMA models with correlated input series and demonstrate generation of synthetic scenarios with correlation between series preserved in all independent samples. Because the Fourier trend in multiple series are deterministic, the correlations between weather measurements will be mathematically preserved in the seasonal trends. To quantify the correlation between sub-seasonal "noisy" data, a Vector ARMA (VARMA) is trained from the correlated actual measurement data and preserved when generating synthetic histories.

In order to validate the Fourier plus VARMA (FVARMA) model, both consumer demand and air temperature are taken as typical energy system histories to consider. Air temperatures are key along with Global Horizontal Irradiance (GHI) for solar generators, while the consumer demand drives the energy generation within an electrical grid system. A single synthetic scenario includes a correlated sampling of both temperature and demand histories. Key statistics (mean and standard deviation) as well as qualitative correlation observations are used to validate the synthetic scenario generation. The contribution of this paper is summarized as follows: development of a computation model, combining Fourier series with correlated ARMA, proposed to synthesize energy and weather measurements consistent with recorded observations.

METHODOLOGY

The Fourier plus VARMA treatment for signal processing and training consists of two major steps: detrending the signal through Fourier analysis and characterizing the remaining "noisy" signal using Auto-Regressive Moving-Average (ARMA) algorithms. The choice of which Fourier frequencies

to remove as trends from the signal before training the ARMA is strongly impactful. Choosing too many Fourier frequencies can overfit the signal, wiping out moment-to-moment randomness and failing to produce independent synthetic scenarios; i.e. all sampled histories will be identical. Selecting too few Fourier frequencies, however, can lead to too much variance in the ARMA training. This variance translates to unpredictable and unrealistic scenario generation during synthetic history production; in other words, the sampled histories will be too noisy compared to the trained data.

Furthermore, the choice of autoregressive terms to include is also important to the performance of the ARMA. If too few autoregressive or moving average terms are included, then the resulting signal can oscillate much more than the original training data and produce data that does not closely resemble the training data. However, the computational cost of training the ARMA increases quickly with an increased number of autoregressive and moving average terms, exponentially so in the case of the correlated VARMA.

In the following sections, the algorithms for applying the Fourier detrending and ARMA characterization are detailed. Furthermore, the extension from independent to correlated variables using the VARMA is described. Examples of the results of these methods is considered in the sections that follow.

Fourier for Seasonal Trends

The time history of energy or weather measurements can be assumed as composed of two parts: a superposition of seasonal trends (or "signal"), and some statistical deviation (or "noise") from the trends. The ARMA model expects data with noise that is distributed according to a standard normal distribution, which is suitable to address the statistical noise. However, the measurements often have seasonal effects that should not be considered part of the statistical noise. For example, higher solar irradiation measurements in summer months should not be considered statistical outliers, but a periodic seasonal effect. To capture such effects, the signal is modified by removing some Fourier modes as defined in Eq. 1.

$$F(t) = \sum_{c} \sum_{i=0}^{k} a_i \sin\left(\frac{2\pi f_i}{c} t\right) + b_i \cos\left(\frac{2\pi f_i}{c} t\right), \tag{1}$$

where c are a set of base characteristic time periods over which a time series has cyclical pattern; f_i are integer frequencies that subdivide the characteristic time lengths; and a and b are calculated to fit $F(t)$ to the original signal. For example, if a characteristic time length is 30 days and k is 3, then seasonal effects are captured for 10-day, 15-day, and 30-day periods. The Fourier signal $F(t)$ is the superposition of all requested bases and orders, and is removed from the original signal to produce a noisy residual.

It is probable that the residual signal has an arbitrary distribution, while the ARMA algorithm expects standard normally-distributed ("white") noise. To standardize the noise, a transformation of variable through a standard normal distribution is performed:

$$y(t) = \Phi^{-1}\left[f(x - F(t))\right], \tag{2}$$

where Φ is the standard normal distribution's cumulative distribution function, f is the empirical cumulative distribution of the detrended residual signal $x - F(t)$, and y is the standardized residual signal.

ARMA for Statistical Noise

After removing the Fourier trends and converting the data to a standard normal distribution, the residual signal can be statistically captured by the ARMA model. The Auto-Regressive (AR) model is given as

$$y_t^{AR} = \sum_{i}^{P} \phi_i y_{t-i} + \epsilon_t, \tag{3}$$

where y_t is the whitened, Fourier-reduced signal at time step t; P is the maximum number of AR lag terms; ϕ_i is the linear extrapolation term; and ϵ is random Gaussian noise. Furthermore, a Moving Average (MA) is added to the auto-regressive model to yield the ARMA,

$$y_t = \sum_{i}^{P} \phi_i y_{t-i} + \epsilon_t + \sum_{j}^{Q} \theta_j \epsilon_{t-j}, \tag{4}$$

where Q is the maximum number of lag terms in the moving average and *theta* is a moving weight associated with the moving average lag term. The terms ϕ and θ are fitted to maximize a likelihood function, as described in [1] and [6]. It is noted that some tuning of Fourier periods is often required to result in a signal that is stationary; the ARMA model is only valid for stationary data, or data whose joint probability distribution does not change when shifted in time. A stationarity check is useful in determining if sufficient signal has been captured through Fourier detrending.

VARMA for Correlated Series

The VARMA is an extension of the ARMA to consider not just regression (or "lag") terms within a series, but other correlated series as well. That is, one value for a variable at a particular time may be influenced not only by values of the variable in previous times, but also by values of other variables at current and previous times. In this instance, the random Gaussian noise (ϵ in Eq. (4)) is taken from a multivariate normal distribution defined by the time-evolving correlation between time series. The vector auto-regression is given by

$$\boldsymbol{Y}_t = \sum_{i}^{P} \boldsymbol{\Phi} \boldsymbol{Y}_{t-i} + \boldsymbol{\mu}_t, \tag{5}$$

where $Y \in \mathbb{R}^{n \times k}$ with n as the number of time steps in the histories and k the number of correlated histories; and each element in $\boldsymbol{\mu} \in \mathbb{R}^{n \times k}$ is determined by the correlation between time series in the moving average,

$$\boldsymbol{\mu}_t = \epsilon_t + \sum_{j}^{Q} \Theta_j \epsilon_{t-j}. \tag{6}$$

with random noise term $\epsilon \in \mathbb{R}^{n\times k}$. The dependency of the moving average term on the vector auto-regression determines the vector auto-regression equation. Letting L be the lag operator, so that $Ly_t = y_{t-1}$, then the VARMA representation is

$$\boldsymbol{Y}_t - \sum_{t}^{\infty} \Pi_i \boldsymbol{Y}_{t-i} = \Theta(L)^{-1}\boldsymbol{\Phi}(L)\boldsymbol{Y}_t, \tag{7}$$

and the Π matrices are given by

$$\boldsymbol{I}_k - \sum_{k}^{\infty} \Pi_i L^i = \Theta(L)^{-1}\boldsymbol{\Phi}(L). \tag{8}$$

SUMMARY RESULTS

By way of demonstration, energy and weather measurement data was collected for the first full week in June for the Electric Reliability Council of Texas (ERCOT) North-Central region, around Forth Worth, Texas [7]. Measurements taken include global horizontal irradiance (GHI), wind speed, air temperature, and consumer demand. The air temperature and demand are used to train a correlated FVARMA. The implemented algorithms are applied using the RAVEN code [2].

First, we consider the air temperature and demand for the region for the first full week of June in Figure 1. Hour 0 corresponds to midnight on Monday morning. A light-dark background pattern is provided in the figure to help identify day-night cycles. As expected, in general the temperature and demand are highest during the day and drop to their lowest at night. Interestingly, the air temperature shows that Friday for this week was much cooler than other days, and the temperature did not drop as far on Thursday night. The demand instead rose from Monday through Thursday, then dropped on Friday before returning on the weekend.

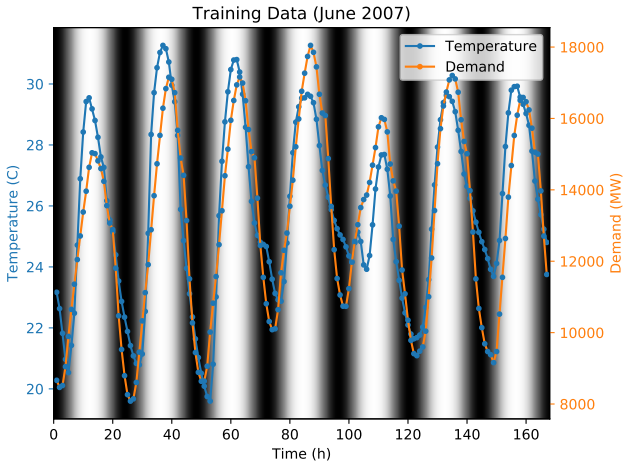

Fig. 1. Measurement Data for Demand, Temperature

Figure 2 shows the original signal as a solid black curve along with ten synthetically-generated histories represented by colored dots with dashed-line connections. In general, the trends of the synthetic samples follow the same daily trends as the original signal, with peaks near noon of each day and valleys during the night. There are several interesting behaviors to note. While the morning and afternoon demand ramps show relatively little variance, the peaks of the demand synthetic histories have significant spread. This is not surprising given the behavior of the peaks in the original data; there is difference of approximately 3000 MW between the training data's Monday and Thursday demands, which is reflected as variance in the synthetic samples. Note that some of this variance might actually be a trend over multiple days or a week that could be captured with Fourier detrending instead of by the VARMA. In this work the signals not captured by Fourier series with a period of one year, three months, one week, two days, or one day are considered to be statistical variance. Note that despite analyzing a signal only one week in length, contributions from much larger-period Fourier trends can be captured. Depending on the period, these may appear as constant or linear signals with small curvature.

The temperature synthetic histories show a wider variance from the training data than in the demand synthetic histories. This is understandable when the training data is considered. In particular, the transition between afternoon and evening on Wednesday and Thursday have a shape quite unlike the other days, which is read by the VARMA as a large variance in the signal. The inconsistency of peak height and width for the temperature synthetic histories similarly yield a wide range of possible realizations for the synthetic samples.

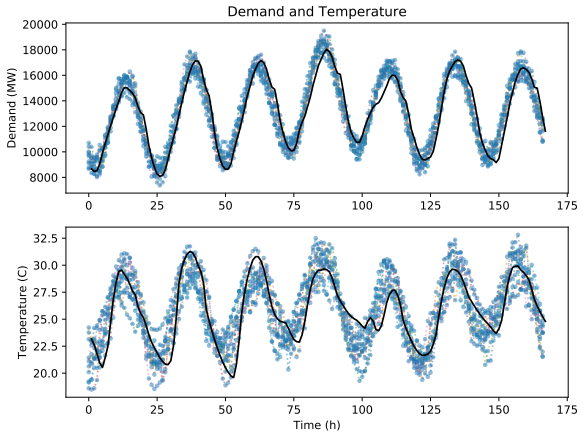

Fig. 2. Synthetic Samples compared to Training Data

Figure 3 shows one of the synthetic signals contrasted with the original training data. While the diurnal trends are obvious, there are key differences in the two signals. For example, the air temperature and demand are lower on Wednesday and higher on Thursday than the training data. As seen by the cloud of samples in Figure 2, a wide variety of synthetic days are produced by sequentially evaluating the trained model, each with unique characteristics that provide statistical challenges while maintaining the fundamental components of the original training signal. It is also noted that the synthetic signal shows more frequent inflection than the training signal, leading to a more spiked, noisy appearance. This indicates that longer lag terms P and Q might be able to be employed to obtain a smoother signal.

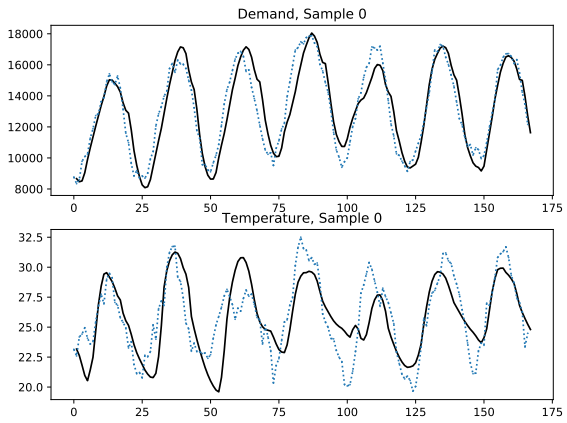

Fig. 3. One Synthetic Sample compared to Training Data

SUMMARY CONCLUSIONS

The ability to synthesize any number of independent but meaningful time series allows model testing to be approached statistically in a manner not possible using historical data. Whether the time series represent power consumption or demand, fuel or electricity prices, or operational power histories, the ability to test many possible scenarios statistically enables a more thorough analysis of simulation models. We have demonstrated a synthetic time series generation method that can preserve correlation between multiple data sets, in order to provide samples that are physically as well as statistically meaningful.

Since the predicating methodologies of this work have been employed in industry-laboratory research collaborations [3][8], it is expected that each enhancement to the capability to produce synthetic histories will enable increasingly complex and accurate analyses. Providing a method for correlating multiple signals provides a means for analyses with multiple variable energy sources along with demand in a more realistic manner.

REFERENCES

1. J. CHEN and C. RABITI, "Synthetic wind speed scenarios generation for probabilistic analysis of hybrid energy systems," *Energy*, **2016**, 1–11 (2016).
2. C. RABITI, A. ALFONSI, J. COGLIATI, D. MANDELLI, and R. KINOSHITA, "RAVEN, a nNew Software for Dynamic Risk Analysis," *PSAM*, , *12* (2014).
3. C. RABITI and ET AL, "Status report on modeling and simulation capabilities for nuclear-renewable hybrid energy systems," Tech. Rep. NL/EXT-17-42441, Idaho National Laboratory (2017).
4. J. PARSONS, J. BUONGIORNO, M. CORRADINI, and D. PETTI, "A fresh look at nuclear energy," *Science*, **363**, *6423*, 105–105 (2019).
5. H. GARCIA, J. CHEN, J. KIM, M. MCKELLAR, W. DEASON, and R. VILIM, "Nuclear hybrid energy systems - region studies: west Texas and northeastern Arizona," Tech. Rep. NL/EXT-15-34503, Idaho National Laboratory (2015).
6. G. BOX, G. JENKINS, and G. REINSEL, *Time Series Analysis: Forecasting and Control*, Prentice-Hall (1994).
7. E. R. C. OF TEXAS, "ERCOT Planning Guide section 3: Regional Planning," (2018).
8. A. S. EPINEY, C. RABITI, P. W. TALBOT, J. S. KIM, S. M. BRAGG-SITTON, and J. RICHARDS, "Case Study: Nuclear-Renewable-Water Integration in Arizona," Tech. Rep. NL/EXT-18-51359, Idaho National Laboratory (2018).