Seasonal decomposition-based stochastic generators for building-level electricity use: choosing seasonal components and remainder models

Gaëlle Faure, Ralph Evins

Energy in Cities group, Department of Civil Engineering,
University of Victoria, Victoria BC, Canada

Abstract

Stochastic generators are used to sample energy use time series, taking advantage of the availability of smart meter data and overcoming the challenge of their current small amount. The samples can be used to design building envelop or energy systems. This study focuses on the design of a seasonal decomposition-based stochastic generator for electricity use. This design implies choices such as the formulation of the model, the number and nature of seasonal components as well as the modelling of the stochastic part. 64 different variants or the model are proposed and compared using a new set of metrics including component strengths and white noise tests of the residual. The comparison show the avenues for the improvement of the generator, and demonstrate the usefulness of the new introduced metrics.

Introduction

Availability of smart meter data enables better knowledge of buildings electricity use. But the amount of available data is still too low to be directly used to design the building envelop or energy systems. Indeed, available data only cover a couple of years to date when building envelop and energy systems are optimized over several decades. To overcome this issue, the simplest method is to duplicate the data over 25 years, but this method fails to represent the inherent variability of energy use in buildings. This variability can be captured by stochastic models, which can then produce the amount of samples required by the simulation problem.

Several type of stochastic models have been used to generate energy use time series, ranging from Bayesian networks (Geraldi, Bavaresco, and Ghisi (Geraldi et al.)) to Generative Adversarial Networks (Sun et al. (2019)). Among them, *seasonal decomposition-based stochastic generator* are popular (Kegel et al. (2018),Moazeni et al. (2019), Patidar et al. (2019), Sun et al. (2019)). These models are based on the decomposition of the time series into a seasonal deterministic part and a stochastic part. They perform rather well, thanks to the strong seasonality of electricity use in buildings (Kegel et al. (2018), Sun et al. (2019)). However, the design of these generators imply a certain choices as the nature and number of seasonal components, or the modelling of the stochastic part. To the best of the authors knowledge, there is no study showing the significance of each choice on the final performance of the model. Moreover, metrics to assess and compare the performance of such models generally consist of looking at plots which makes it hard to automatize (Patidar et al. (2019), Sun et al. (2019)), or computing the average difference between raw and synthetic data (Geraldi, Bavaresco, and Ghisi (Geraldi et al.)), which does not make sense for a stochastic model.

In this paper, we propose to compare the performance of different variants of a seasonal decomposition-based stochastic generator for electricity use. The analyse will be performed over a set of buildings to obtain robust and generalized results. The different generators will be compared using a new set of metrics adapted to the assessment of a stochastic generator.

In the next section, seasonal decomposition-based stochastic generators are described, emphasizing the different design choices which have to be made. The methodology of comparison is then described, starting from the data set used for this study, followed by the list of the studied variants and the proposed set of metrics for the assessment. The results of this comparison are then presented. A section is dedicated to the discussion about this results before concluding about the study presented in this paper.

Seasonal decomposition-based stochastic generators

Seasonal-decomposition based stochastic generators assumes that energy use $Y(t)$ is composed of a deterministic $\bar{Y}(t)$ and a stochastic $\gamma(t)$ parts. The deterministic component $\bar{Y}(t)$ itself can be split into a trend $T(t)$ and n seasonal components $S_i(t)$, n generally ranging from 1 to 3.. There is two main ways of putting together the components, either by summing them (eq.1 and 2) or multiplying them (eq.3 and 4). The former is called *additive model* and the latter is a *multiplicative model.*

$$Y(t) = \bar{Y}^{+}(t) + \gamma^{+}(t) \tag{1}$$

$$\bar{Y}^{+}(t) = T^{+}(t) + \sum_{i=1}^{n} S_i^{+}(t) \tag{2}$$

$$Y(t) = \bar{Y}^{*}(t) \times \gamma^{*}x(t) \tag{3}$$

$$\bar{Y}^{*}(t) = T^{*}(t) \times \prod_{i=1}^{n} S_i^{*}(t) \tag{4}$$

Implementation of multiplicative models usually start by taking the logarithm of $Y(t)$. This logarithm is then treated as an additive model. The notations and equations for multiplicative models therefore become eq. 5 and 6. Note that this transformation is possible for energy use time series as energy use values are always strictly positive.

$$\log(Y(t)) = \bar{Y}^{*}(t) + \gamma^{*}(t) \tag{5}$$

$$\bar{Y}^{*}(t) = T^{*}(t) + \sum_{i=1}^{n} S_i^{*}(t) \tag{6}$$

Additive models are used when there is no correlation between the different components $\bar{Y}(t)$, $T(t)$ and $S_i(t)$, otherwise multiplicative models should be preferred (Hyndman and Athanasopoulos (2018)). Many authors use additive models for energy use without questioning this choice and in particular the independence of the different components (Sun et al. (2019), Moazeni et al. (2019)). Yet, this assumption is not straightforward as we can expect for example more variability between day and night consumption during the winter compared to spring or fall.

The trend component $T(t)$ represents the long-term tendencies of the time series. It is extracted using linear regression or a low-pass filter, depending on the amount of information one wants to keep in this component. Low-pass filters lead to trend components which still depict a lot of variability. It is therefore not straightforward to use them in a generator. Therefore, the trend is modelled by a straight line and obtained via linear regression in this paper.

Seasonal components $S_i(t)$ are the cyclic and repeated tendencies of the time series. Each seasonal component is characterized by its length and its granularity. The length correspond to the total duration of one period of the cycle represented by the component. Some authors aggregates the initial time series into a longer time step before fitting each seasonal component. For example, to represent the "year" season, Kegel et al. (2018) propose to use monthly aggregated data, resulting in one value for the whole month. The time step used for a season is called granularity. Seasonal components are obtained by (1) slicing the data into season's length samples, (2) aggregating them to the proper granularity if required, (3) taking the average of the samples. To perform the sampling, each seasonal component is repeated until achieving the total required length of the sample.

Constructing a new model starts by extracting the deterministic component $\bar{Y}(t)$ from the data, i.e. the trend $T(t)$ and the different seasons $S_i(t)$ by increasing order of length.
The stochastic component $\gamma(t)$ is then fitted on the remainder $R(t)$, obtained using eq. 7 for an additive model, and eq. 8 in case of a multiplicative model.

$$R^{+}(t) = Y(t) - \bar{Y}^{+}(t) \tag{7}$$

$$R^{*}(t) = log(Y(t)) - \bar{Y}^{*}(t) \tag{8}$$

In summary, elements to choose in order to design a seasonal decomposition-based stochastic generator are:

- Additive or multiplicative model;
- Number, length and granularity of the different seasonal components;
- Model of the stochastic part $\gamma(t)$.

Methodology of comparison

In the previous section, we described the model studied in this paper: a seasonal decomposition-based stochastic generator. We also listed the different parameters which can be varied in order to optimize the model for a specific data set. We will now analyse the impact of these parameters in the case of the modelling of individual building energy use. Figure 1 presents the general methodology developed to perform this study. This work flow was applied to each building and for each model variant:

1. The generator is built, first by extracting the deterministic component, then by fitting a stochastic model on the remainder.
2. The stochastic model is used to generate 10 samples with the same length of the initial time series. The 10 corresponding residuals $e(t)$ are then obtained using equation 9 for an additive model and 10 for a multiplicative model.
3. Assessment is performed on the different parts of the model.

$$e(t) = R^{+}(t) - \gamma^{+}(t) \tag{9}$$

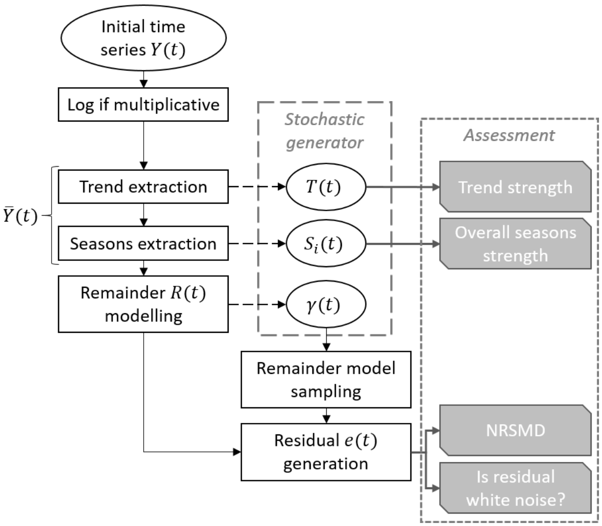

Figure 1: Main steps of the methodology proposed in this paper. This process is applied for each couple of season decomposition and remainder model, for each building of the data set.

$$e(t) = \exp\left(R^*(t) - \gamma^*(t)\right) \tag{10}$$

In the following, the data set employed in this paper is presented. The different variants of the model are then described and listed, as well as the metrics employed for the assessment of each variant.

Data set

We use in this study data from the Pecan Street Dataport project Pecan Street (Pecan Street). It consists of hourly total electricity use of residential houses in United States. This electricity use encompasses specific electricity use (lighting, appliances...) as well as air conditioning and heating system powered by electricity.

By looking at the data, we observed that there are no labeled missing values but a lot of zeros. We treated the zero values as missing data. For the purpose of our study, we selected only buildings located in Austin, Texas, with more than two years of available data. At the end of the cleaning and selection process, the data set is composed of 348 buildings whose characteristics are depicted in Figure 2.

Studied variants of the generator

As we saw before, three parameters can be varied inside a seasonal decomposition-based stochastic generator: additive or multiplicative formulation of the overall model, the choice of the seasons and their granularity for the seasonal decomposition, the choice of model for the stochastic part. In the following paragraphs, the options chosen in this study for the two latter elements are detailed.

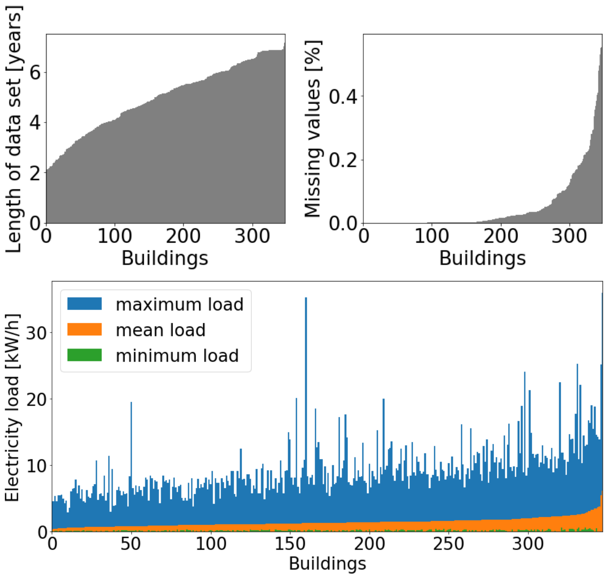

Figure 2: Characteristics of the data set employed in this study: length of each time series (top left), fraction of missing values (top right) and energy use statistics, ordered by increasing mean (bottom).

Seasonal decomposition: Three different seasons are regularly used for energy use data: the year, which represents variation due to air conditioning and heating uses ; the day, which encompasses night-day cycles ; the week due to behavior changes between weekdays and week-end. The associated granularities are: hourly (H), daily (D) and monthly (MS) aggregations. Table 1 lists the seasonal decomposition variants studied in this paper. Each variant name is composed of the list of the different seasons, from shortest to longest. Each season is depicted as a number and one or two characters describing its length and granularity. For example, variant "24H7D" is a variant with one season of 24 hours (day season with hourly time step) and one season of 7 days (week season with daily time step).

Table 1: List of seasonal decomposition variants used in this study. Each variant name is a list of season lengths and granularities.

Nb seasons	**1**	**2**	**3**
List of variants	24H 168H 7D 8760H 365D 12MS	24H168H 24H7D 24H8760H 24H365D 24H12MS	24H168H8760H 24H168H365D 24H168H12MS 24H7D365D 24H7D12MS

Only variants with decreasing granularity when season length is increasing are retained, as the contrary would not make sense (for instance a week season with daily time step and a year season with hourly time step). Previous analysis of energy use time series showed that the predominant seasons are day and year. Therefore, the authors chose to ignore variants containing only a week and a year season, as adding the week seasonality to a year season does not add a lot of information.

Modelling of the stochastic part: Two simple models are chosen to model the stochastic part, based on the strong assumption that almost all the information contained in each time series can be captured by the deterministic component:

- *a single distribution (SD)*: it is assumed that the remainder is close to white noise and therefore can be modelled by a single distribution. This distribution is fitted using kernel density estimation.
- *a Markov chain (MC)*: here we assume that some correlation between two consecutive values is still present in the remainder. Markov chain models are designed to represent this kind of stochastic process. In our implementation, the previous state is discretized in 30 bins and a distribution is fitted for each bin using kernel density estimation.

Each seasonal decomposition variant is tested with both models for the stochastic part. In addition, each of these tuple (seasonal decomposition, stochastic model) is tested using an additive and a multiplicative formulations of the model. In total, 64 different models are compared in this study.

Metrics for assessment

A good model can fully replicate the information contained in the time series used to fit it. The proposed set of metrics follow this idea. Each variant of the generator is fitted for each of the 348 buildings of the data set. Therefore, quantitative numerical metrics i.e. metrics which give one number per building must be employed. In this way, the performance of each variant can be analyzed through the distribution of each metric over the whole data set. Three subsets of metrics are proposed in order to assess each part of the generator:

- the strengths of the deterministic components,
- the normalized root-mean-square deviation (NRSMD) of the residual,
- the probability that the residual is white noise.

Strength of the deterministic components. Variance is one of the common measures of quantity of information: the bigger the variance over the values of a time series, the more information is contained in this time series. Based on this idea, Hyndman and Athanasopoulos (2018) proposed two metrics, eq. 11 and 12, to compute the strength of trend and season components of an additive seasonal decomposition-based model.

$$F_T = max\left(0, 1 - \frac{Var(R^+(t))}{Var(T^+(t) + R^+(t))}\right) \tag{11}$$

$$F_{Si} = max\left(0, 1 - \frac{Var(R^+(t))}{Var(S_i^+(t) + R^+(t))}\right) \tag{12}$$

Each strength is measured as 1 minus the ratio between information contained in the remainder alone versus the sum of the remainder and the studied component. The more information is captured by the component, the smaller variance of the remainder will be and therefore, the bigger the strength. Strength values range between 0 and 1.

For the purpose of this study, we propose a variant to compute the strengths of all the season components at once (eq. 13), knowing that the sum of all the season components and the remainder is equivalent to the difference between the initial time series and the trend component. Variants for a multiplicative model are also defined (eq. 14 and 15).

$$F_S = max\left(0, 1 - \frac{Var(R^+(t))}{Var(Y(t) - T^+(t))}\right) \tag{13}$$

$$F_T = max\left(0, 1 - \frac{Var(\exp{(R^*(t))})}{Var(\exp{(T^*(t) + R^*(t))})}\right) \tag{14}$$

$$F_S = max\left(0, 1 - \frac{\exp{(R^*(t))}}{Var(\exp{(\log{(Y(t))} - T^*(t))})}\right) \tag{15}$$

In this paper, trend strength and overall season strength for both additive and multiplicative models are used.

NRMSD of the residual. NRMSD is not strictly speaking a measure of the remaining information but it is widely used. It represent the deviation between the initial data set and the generator. To be meaningful in the case of a stochastic model, its distribution needs to be analysed over several samples. For each sample j, NRMSD is obtained by the usual formula:

$$NRMSD_j = \frac{1}{\mu_Y}\frac{1}{N}\sqrt{\sum_{i=1}^{N} e_j(t_i)^2} \tag{16}$$

where μ_Y is the mean of the initial time series, t_i is the time at time step i and N is the number of time steps in $Y(t)$.

In this paper, we will work with the average value of NRMSD over the J samples:

$$NR\bar{M}SD = \frac{1}{J}\sum_{j=1}^{J} NRMSD_j \tag{17}$$

Probability that the residual is white noise. White noise is a signal purely random, i.e. containing no information. Assessing if the residual can be assimilated to white noise is a common way to perform model diagnostic, in particular for ARIMA models (Mahan et al. (2015)). Ideally the deviation between a model and the initial time series must be white noise, indicating that this remaining deviation is only due to random unpredictable fluctuations. It is however currently not widely applied to analyse the performance of other stochastic models.

By definition, a time series can be assimilated to *white noise* if it is a sequence of serially uncorrelated random variables with zero mean and finite variance[1]. Inspired by the work of Mahan et al. (2015), we propose a series of three tests to check each element of this definition:

1. Test for non correlation using Ljung-Box test.
2. Test for constant variance using Levene's test.
3. Test for zero mean using one-sample t-test.

Table 2 summarizes the used statistical tests as well as their null and alternative hypothesis. Each statistical tests return a p-value (a mean p-value in the case of the one-sample t-test). This p-value can then be compared to a chosen significance level α (typically 5% or 1%). If the p-value is less than the chosen significance level (α), that suggests that the null hypothesis may be rejected.

Results

This section presents the results obtained by applying the methodology described in the previous section. Each figure contained in this section shows the distribution of the values obtained for one metric for each variant of the models over the whole data set. This distribution is depicted as a box plot. Acronym used for each variant is constructed as follow:

(seasonal decomposition) _ (stochastic model) _ (formulation)

Seasonal decomposition acronyms come from Table 1, stochastic model can be either "SD" (single distribution) or "SMC" (Markov chain), model formulation is "add" for an additive model or "mul" for multiplicative one.

Strengths of the different components

Figure 3 and 4 respectively presents the distribution of trend and overall season strengths over the data set for the different variants of the seasonal decomposition.

Trend strength is on average quite low (very close to 0) but multiplicative models show more variance over the data set: 25% of the values generally lays between 0 and almost 40%. Overall season strength values are higher and present more diversity. The highest values are obtained for variants with a high granularity (hourly time step). Generally the more seasonal components, the better. The variants "7D_add" and "7D_mul" seem to really under perform the other models, which is expected as week season is generally not very strong. Finally, multiplicative models give slightly better results than additive models, for a same choice of seasons.

In summary, according to this set of metrics, best performing variants are those with hourly time step, a multiplicative formulation, and several seasonal components.

Analysis of residuals

NRMSD distributions obtained for the different variants are depicted in Figure 5. As a remainder, in this case, the best variants are those which have the lowest values. Overall, main differences are observed between multiplicative and additive formulation, and between the models to represent the stochastic part (single distribution or Markov chain): best variants are the tuples (multiplicative formulation, Markov chain), then additive variants come with both stochastic models. Finally multiplicative formulations with single distribution are far worse than the other options. Among the different choice of seasons, variants including a year season with hourly time step ("8760H") give lower NRSMD.

Figure 6 and 7 present the distribution of the p-values obtained for each test which are part of white noise testing. The significance value α of 5% is also showed as a red line. P-values above this line indicates that the null hypothesis can be retained with a confidence of 95%. Results for Levene's test are not showed as they are less interesting: all the variants have the same behavior and pass the test.

All the variants fail LJung-Box test, which means that there is still correlation in the remainder. The use of more complex models for the stochastic part seems to be the first major step to improve our model. It has to be noted that multiplicative formulations with a stochastic part modelled using a single distribution passed the test for some buildings.

All the variants pass one-sample t-test except multiplicative models with a stochastic part modelled using a single distribution. It is not really surprising as multiplicative models works with the logarithm of the initial data, therefore the trend does not remove the constant of the initial data set but of the modified one. Which is more surprising is that using a Markov chain model seems to remedy the problem... The highest p-values are obtained for additive models with single distribution. These variants also show more consistency in the results.

The choice of seasonal components does not significantly affect the result of any test associated to white noise. The formulation of the model as well as the model used for the stochastic part have a bigger impact on the results. No variant successfully pass all the test for white noise, indicating that efforts still need to be done to improve the modelling of energy use time series. The most promising avenue seems to be the development of a more complex model for the stochastic part.

It may be interesting to look more closely at variants including multiplicative formulation and single distribution as model for the stochastic part, as they

[1] This definition corresponds to the notion of *weak white noise*. *Strong white noise* requires that the random variables are i.i.d.

Table 2: List and description of the different statistical tests used for white noise assessment.

Tested property	**Non-correlation**	**Constant variance**	**Zero mean**
Name	Ljung-Box test	Levene's test	One-sample t-test (two-sided)
Null hypothesis	H_0 : the data are not correlated.	$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$	$H_0 : \mu = 0$
Alternative hypothesis	H_a : the data are correlated.	$H_a : \sigma_i^2 \neq \sigma_k^2$ for at least one pair (i, j)	$H_a : \mu \neq 0$

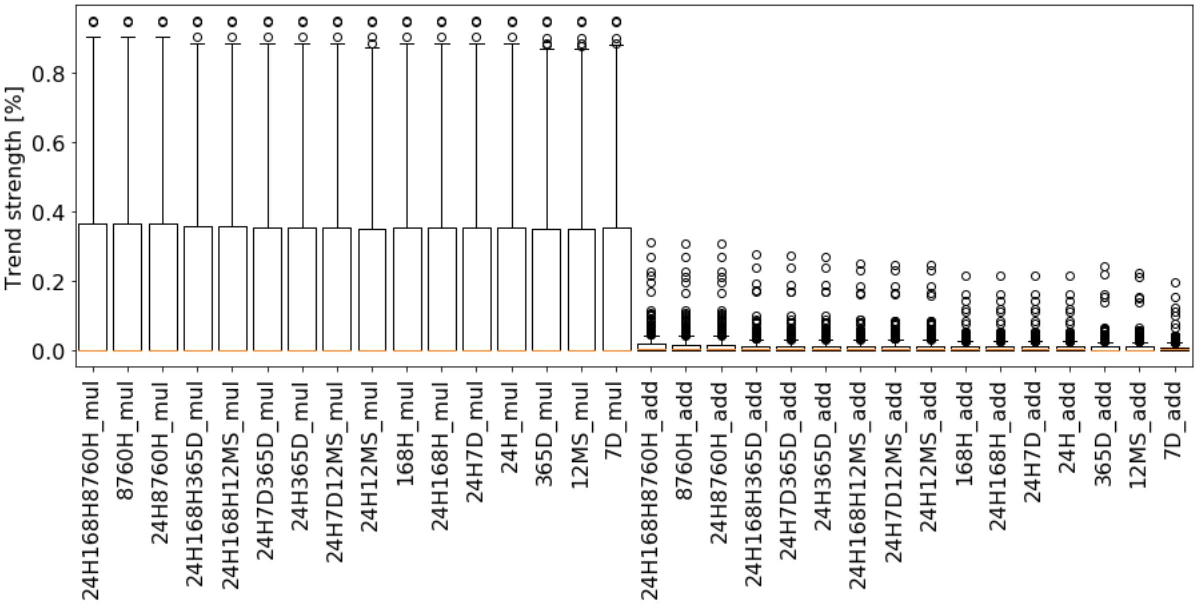

Figure 3: Trend strengths of the different seasonal decompositions, sorted by decreasing mean.

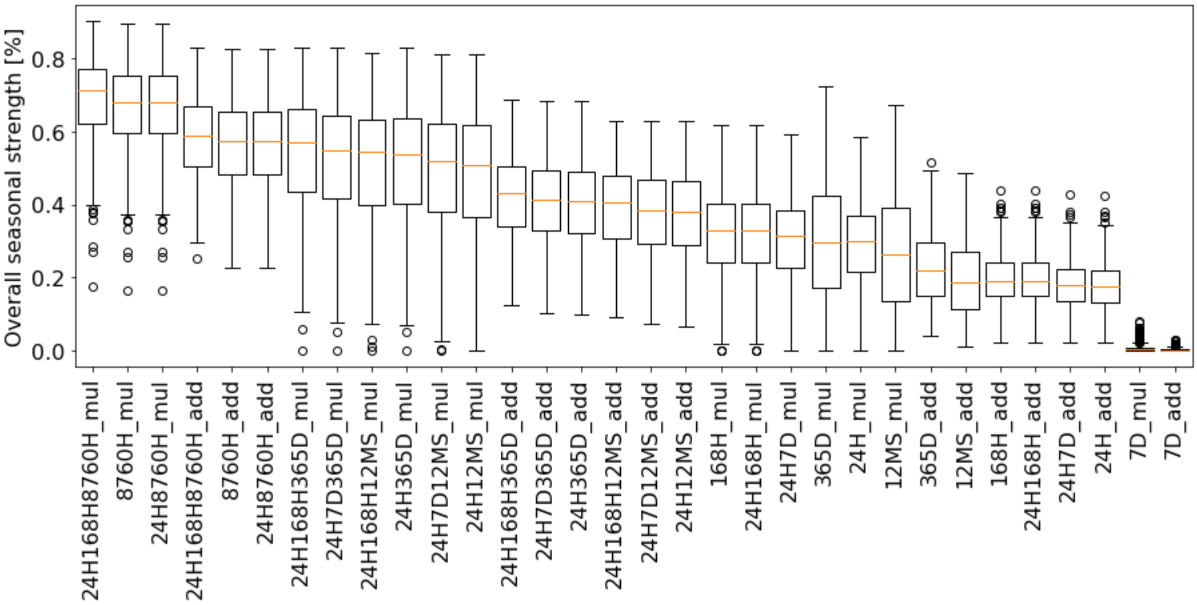

Figure 4: Overall seasons strengths of the different seasonal decompositions, sorted by decreasing mean.

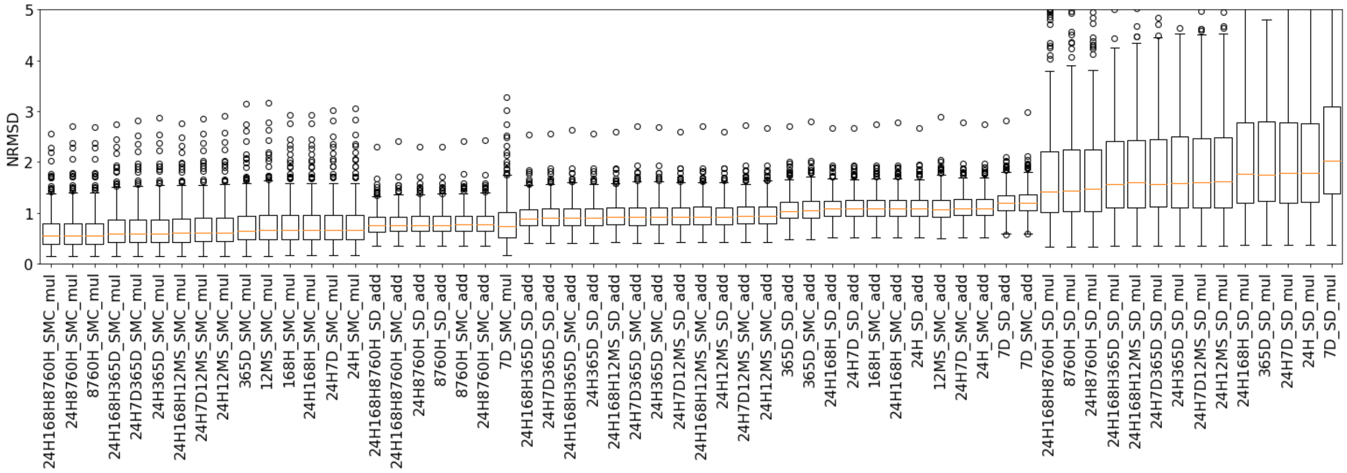

Figure 5: Normalized root-mean-square deviation, sorted by decreasing mean.

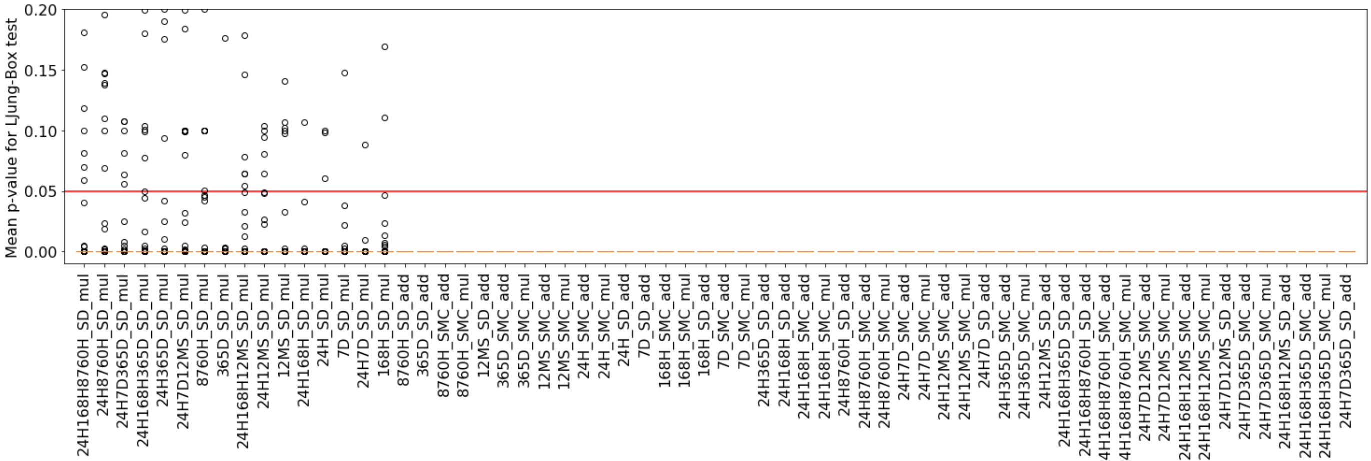

Figure 6: P-values obtained for Ljung-Box test, sorted by decreasing mean. The red line corresponds to a significance value α of 5%.

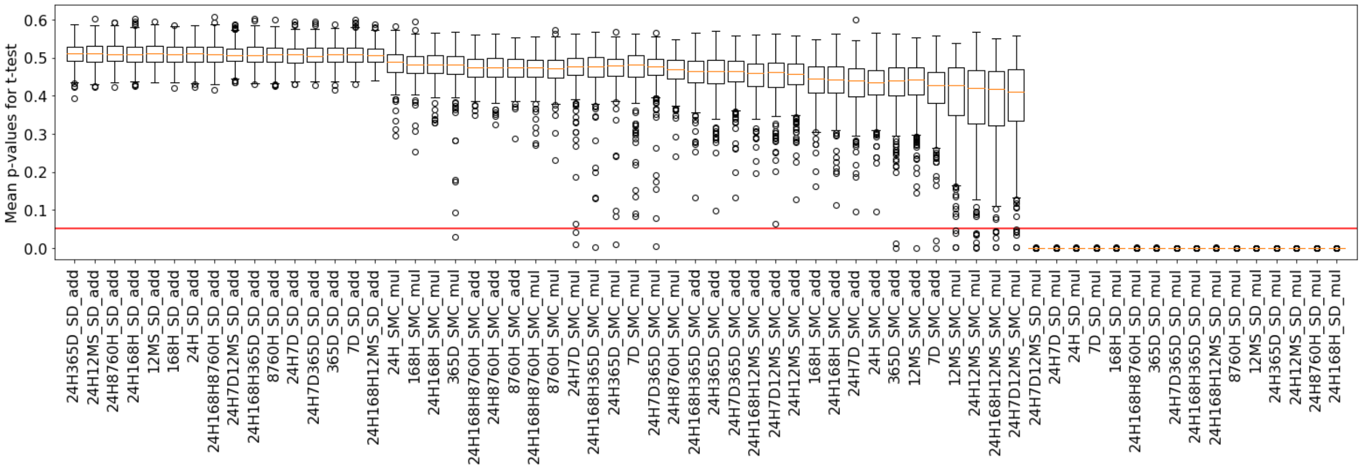

Figure 7: P-values obtained for one-sampled t-test, sorted by decreasing mean. The red line corresponds to a significance value α of 5%.

express a different behavior as the other variants.

Discussion

Analysis of strength of model components shows discrepancy mostly based on the choice of the seasonal components, whereas analysis on the residual emphasizes the impact of the choice of the formulation and the modelling of the stochastic part. As the overall performance of the model is pictured by the behavior of the residual, we can conclude that the formulation and the modelling of the stochastic part are the most significant choice to make when designing a seasonal decomposition-based stochastic generator for energy use. NRMSD and white noise tests does not give the same results, but NRMSD is conducted only over 10 samples, which is probably not enough to be statistically significant. Analysis of the residuals using white noise tests enable to extract useful information for the improvement of the model.

The best model is the one which captures the main characteristics but which also stays general enough as we know that there are not enough data to be fully representative. This is the common trade-off between under- and over-fitting encountered in machine learning techniques. The set of metrics proposed in this paper enables to check that a model does not underfit, but the problem of over-fitting is not addressed. It could be interesting to apply the techniques used to assess neural networks over-fitting to this generator.

Another interesting question is to understand if it is more interesting to capture the most of the information in the deterministic part of the model, or if it is better to keep more flexibility. Applying white noise tests to the remainder could help answering this question, as we could compare the results obtained for the remainder and the residual for each variant.

Conclusion

In this paper, we analysed the impact of the design choices of a seasonal decomposition-based stochastic generator for electricity use. We proposed a methodology to compare different variants of the model. In particular, new metrics to assess the amount of information captured by the model as component strengths and white noise test of the residual are proposed. We applied this method to a data set of several hundreds of buildings and showed that none of the proposed variant models correctly the electricity use. In particular, the variants fail to encompass the auto-correlation contained in the time series.

This study enabled to highlight the most promising avenue to improve our model, i.e. developing a more advanced model for the stochastic part. Moreover, additive VS multiplicative formulation are probably to restrictive and there are a lot of intermediate options which could be studied. The methodology should also be applied to other data sets (in particular for different locations) to check the consistency of the results. The developed methodology can also be applied to other data sets with strong seasonalities as for instance weather data.

This study is a first step towards the development of performing generators which will enable to overcome the problem of scarcity of the data in the field of building science.

Acknowledgment

We are grateful to Pecan Street Project for providing energy use via their Pecan Street Dataport platfrom (https://dataport.pecanstreet.org). This work used the BESOS platform (https:// besos.uvic.ca) supported by CANARIE Grant RS-327.

References

Geraldi, M. S., M. V. Bavaresco, and E. Ghisi. Bayesian Network for Predicting Energy Consumption in Schools in Florianópolis – Brazil. pp. 8.

Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice* (OTexts: Melbourne ed.). Australia.

Kegel, L., M. Hahmann, and W. Lehner (2018). Feature-based comparison and generation of time series. In *Proceedings of the 30th International Conference on Scientific and Statistical Database Management - SSDBM '18*, Bozen-Bolzano, Italy, pp. 1–12. ACM Press.

Mahan, M., C. Chorn, and A. Georgopoulos (2015). White Noise Test: detecting autocorrelation and nonstationarities in long time series after ARIMA modeling. Austin, Texas, pp. 97–104.

Moazeni, S., A. H. Miragha, and B. Defourny (2019, May). A Risk-Averse Stochastic Dynamic Programming Approach to Energy Hub Optimal Dispatch. *IEEE Transactions on Power Systems 34*(3), 2169–2178.

Patidar, S., D. P. Jenkins, A. Peacock, and P. McCallum (2019, September). Time Series decomposition for simulating electricity demand profile. In *Proceedings of Building Simulation 2019: 16th Conference of IBPSA*, Rome (Italy).

Pecan Street. Dataport: https://www.pecanstreet.org/dataport/.

Sun, S., F. Kazhamiaka, S. Keshav, and C. Rosenberg (2019). Using Synthetic Traces for Robust Energy System Sizing. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems - e-Energy '19*, Phoenix, AZ, USA, pp. 251–262. ACM Press.