Towards Risk-Aware Energy Forecasting: A Case Study for Columbia University

Yi Xin (Allen) Hu | UNI : yh3468

December 20, 2023

1 Introduction

This project aims to utilize Bayesian Structural Time Series (BSTS) estimation to analyze electricity demand at Columbia University. The main contribution of this exercise seeks to utilize a probabilistic approach to model future electricity demand. Bayesian models are excellent at quantifying error in forecasts and allow for an increased understanding of the financial and operational risk of a concrete, steel-in-the-ground energy project. Additional gains can be seen with improved stakeholder communication through a statistical method (rather than deep learning) and financing, as the downsides are statistically quantified.

The paper first dives into the theory of structural time series and Bayesian extensions, as well as its applications so far. The paper then provides an in-depth discussion regarding the model formulation. Using a novel synthetic dataset combining NYC's and Columbia's data, we will approximate a BSTS using Maximum A Posteriori estimation. While we cannot achieve a 'full Bayesian Inference' model due to computational constraints, we can still nevertheless construct a probabilistic model with a bootstrapped result (as well as describe what would be needed in a fully Bayesian model). The report will then conclude with a discussion of the results as well as implications for Columbia University's BESS systems, as well as universities across the country.

2 Building Towards Bayesian Time Series Forecasting

2.1 Why Structural Time Series

Structural Time Series (STS) is a well-known and well-researched technique to explicitly model trends across time. What makes an STS 'structural' is that it requires the researcher to explicitly provide a *description* of the time series in terms of its components of interest. STS's key advantage over the empirically implied models such as ARIMA or Deep Learning techniques such as the RNN-based LSTM lies in the fact that the explicit model not only makes the underlying assumption clear but that, if properly formulated, has the expressiveness to capture great fluctuations in our data. (2) This means that structural models, by design, *forces* the researcher to clearly state the components of a series she suspects and the assumptions she holds. In doing so, structural models serve to bridge the gap between theory and empirical study, and allow us to discuss *why* theory does not explain data and *to how much* are we off. (3)

This explicit structural model, combined with a bayesian framework where we are also *forced* to demonstrate our priors and kernels (more on that later) and why we think they are correct is an excellent way to not only improve uncertainty quantification, but also communication and trust building between various stakeholders.

This project is heavily inspired by the work conducted by Mokilane, Paul, et al., where they utilized BSTS as a means to forecast electricity demand for South Africa.(1) This short project will be using similar techniques as well as adding a novel synthetic dataset to forecast demand at Columbia University.

2.2 Structural Time Series Foundations

2.2.1 Local Level Model

STS models are often understood by iteratively building upon a baseline *local level* model. Taking a simple local-level (random walk plus noise) model as an example, we have the following:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t & t &= 1, ...T & (1) \\ \mu_t &= \mu_{t-1} + \eta_t & t &= ..., -1, 0, 1, ... & (2) \end{aligned}$$

Think of equation 1 as 'observing' the data(y_t) from an underlying data-generating process (DGP) μ_t. Equation 2 then describes how the DGP shifts over time, and in this case, a random walk. This two-stage formulation is also why STS belong to the *state space models* family, where the data y_t comes from the unobserved variable μ_t, also known as the *state space* variable in our formulation. (2) This will be important because it will allow us to model the variability in electricity demand, as in the real world, it is not the case that everyone will turn on and off their appliances in a perfectly scheduled, ordered way. Since the parameters of this model are characterized by mean-zero errors ϵ_t and η_t, this will be useful when we begin to talk about the Bayesian approach to STS by viewing their covariance functions as kernels.

2.2.2 Local Linear Trend

Now that we have a local level model, we can add v_t, the local linear trend, to the DGP:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t & &\text{(unchanged)} & (3) \\ \mu_{t+1} &= \mu_t + v_t + \eta_t & & & (4) \\ v_t &= v_{t-1} + e_t & &\text{(linear trend)} & (5) \end{aligned}$$

This component gives our model to describe intertemporal dependencies. This addition simply means that the state space is now not only a random walk but can be used to describe trends in our data. For our particular project, this is extremely useful as people (generally) do not randomly switch on and off their appliances: rather, their usage generally is correlated by what they were using an hour (or some timestamp) before. Again, notice that we have one extra error term e_t that needs to be defined to characterize the equation. Once again, this will be useful when taking into account the Bayesian approach.

2.2.3 Seasonal Effects

Now we get to the most interesting component of our analysis. In particular, we will be looking at the seasonal (or cyclical) component of our STS. The seasonal component is given as:

$$\begin{aligned} y_t &= \mu_t + \tau_t + \epsilon_t & &\text{(observed)} \\ \mu_{t+1} &= \mu_t + v_t + \eta_t & &\text{(state space)} \\ v_t &= v_{t-1} + e_t & &\text{(linear trend)} \\ \tau_t &= -\sum_{s=1}^{S-1} \tau_{t-s} + \omega_t & &\text{(seasonal trend)} \end{aligned}$$

Where S is a dummy variable (1 for each season)

Unlike the linear trend, we did not add the seasonal trend in the state space model. This is because we want to explicitly state that the seasonal component is an "additional" factor affecting our observed y_t, which allows us to disentangle the linear effects from the seasonal effects. This idea will be helpful when describing our additive GP Model. The mean-zero parameter ω_t that can characterize this system will again be highly important in our Bayesian framework.[1]

[1] Note that there is also the exogenous component with form $y_t = \mu_t + \tau_t + \beta_t^T x_t + \epsilon_t$, but this will be beyond the scope of this short project

2.3 The Bayesian Approach to STS

Now that we have covered the standard approaches to STS, we will now dive into the Bayesian approach to STS modeling. In this paper, we will talk specifically about the STS model **without** a regression component to make the results a bit more explainable and a more seamless fit between the two concepts. Recalling that different components of our STS can be fully expressed as the zero-mean *error terms* from the state-space form, we can view the covariance function on these error terms as *kernel functions*.

2.3.1 Overview of Bayesian Methods

In any Bayesian framework, the crux of the problem lies in understanding how data changes our prior belief about the world, and with that gives us a *probability distribution* of the parameters. Then the maximum a posteriori is approximated. While we will not be approximating the entire posterior distribution through a full *Gaussian Process*, we will still take the Bayesian model approach to model our data and use the maximum a posteriori (MAP) to bootstrap sample our forecast. While this is not 100% Bayesian inference, we still yield a good forecast result.

We need to first define our *priors*, which then can be converted into *posteriors* over functions once we see *evidence*. This is described using the iconic expression of Bayes' Rule:

$$\underbrace{P(H|E)}_{\text{Posterior}} = \underbrace{P(H)}_{\text{Prior}} \cdot \frac{\overbrace{P(E|H)}^{\text{Likelihood}}}{\underbrace{P(E|H) + P(E|\neg H)}_{\text{Total Evidence (Marginal)}}}$$

More importantly, we do not need to define the *entire* function necessarily, rather just at finite, arbitrary values. (4) The main reason why this can give us the necessary results is that the GP assumes that the joint probability of observing all the evidence is jointly gaussian, meaning that we can characterize this function using some mean $\mu(X)$ and covariances $\Sigma_{ij}(X) = k(x_i, x_j)$ where k is positive definite. This function k is the *kernel function* that we have alluded to.

The kernel function simply is a description of how different values of x_i in the data relate to each other: if we define k so that x_i and x_j are similar enough, then their corresponding outputs $f(x_i)$ and $f(x_j)$ are also 'similar' as well. (9) Since a kernel is just the covariance function on the joint Gaussian distribution, we can *interpret* the error terms from our STS, $\epsilon, \eta, , \omega$, not as fixed values to estimate, rather *distributions* whose variance $\Sigma_{ij}(X)$ is described by the kernel functions $k(x_i, x_j)$ or simply $k(x, x')$. This is the key to bridging the gap between Bayesian methods and STS, creating the BSTS framework.

2.3.2 Bayesian Structural Time Series

For the BSTS, this project will extend upon an existing BSTS model formulation to forecast electricity prices in Australia(4). This original model was originally constructed to forecast upon hourly data. As hourly forecasts will be outside of the realm of this project (primarily due to the increase in timeline and thus increase in computing), we will adopt this model for daily energy consumption forecasting for Columbia University.

$$y_t = \overbrace{\mu_t}^{\text{Linear Trend}} + \underbrace{\tau_t}_{\text{Cyclical Trend}} + \overbrace{\gamma_t}^{\text{Medium Term Irregularity}} + \underbrace{\epsilon_t}_{\text{Residual Irregularity, Noise}}$$

This model explicitly the long-term, trend (for growing electricity usage throughout time), daily periodicity in usage (from day-night cycles), medium-term irregularities from usage spikes and/or supply imbalances (iconic for increased renewable generation), and residuals. As with state-space models, each one of these variables can be characterized by their error term, and consequently a covariance matrix. Lends itself nicely to a Bayesian approach by viewing these covariance matrices as kernel functions whose parameters need to be estimated.

$$\begin{aligned} y_t \sim\ &\mathbf{GP}_{Linear}(0, \eta_1^2 k_1(t,t')) + \mathbf{GP}_{Cyclical}(0, \eta_2^2 k_2(t,t')) + \\ &\mathbf{GP}_{MediumIrr.}(0, \eta_3^2 k_3(t,t')) + \mathbf{GP}_{Noise}(0, \eta_4^2 k_4(t,t')) \end{aligned}$$

Where t is a one-day time-stamp. What these kernels really are saying is how one time stamp is similar to other time stamps. This will be the main function that we will be using to estimate future demand for Columbia.[2]

3 Data and Modeling Approach

Two years' worth of Columbia University's electricity demand (15-minute, kWh, between 2018 and 2020) will be used to conduct our exercise. We will hold out the last month of electricity demand as our forecasting period. Additionally, we will '*augment*' Columbia's data with New York Independent System Operator's (NYISO) 10-minute, MWh electricity demand data across the NYC region between 2017 and 2018. Then using the augmented data, we will train a BSTS model and create a 1-month ahead forecast with upper and lower confidence bounds. The main goal is to model and forecast day-level data, so an aggregation on daily mean was conducted to synchronize the levels of the two series.

3.1 Model Formulation

3.2 Kernel Selection

A key component of a our model is selecting the kernel that describes the data. Recalling that one of the key advantages of a Bayesian view of this question is being able to explicitly state assumptions about the data through kernels, the following section hopes to not only provide the kernels being used in the model but also the rationale behind choosing them for the specific task.

1. Linear Trend Kernel

$$\eta_1^2 k_1(x,x') := \eta_1^2 \mathbf{RBF}(\mathbf{x},\mathbf{x'}) = \eta_1^2 \exp\left(-\frac{\|x-x'\|^2}{2\ell^2}\right)$$

The reason for using an RBF kernel for our linear trend (instead of, say, a linear kernel) is that preserves the flexibility to generalize and has a stronger tolerance to perturbations. While electricity demand trends are generally stationary, they still exhibit strong perturbations that a stricter kernel might not be able to capture. Furthermore, while RBFs are not suitable to capture discontinuities in our data (perhaps Matérn would be better), we see that in the *long term*, electricity consumption is smooth and does not have sharp breaks in the trend that warrants a more complicated Matérn kernel. Note that ℓ denotes our "length-scale", which can be seen as the length of perturbations in our function; and η^2 is our output variance that governs the average distance our data deviates from the mean (which is essentially a scalar) (5).

2. Cyclical Kernel

$$\begin{aligned} \eta_2^2 k_2(x,x') :&= \eta_2^2 \mathbf{LocalPer}^*(x,x') \\ &= \eta_2^2 \left(\mathbf{Periodic}(\mathbf{x},\mathbf{x'}) \cdot \mathbf{Matern_{5/2}}(\mathbf{x},\mathbf{x'})\right) \end{aligned}$$
$$= \eta_2^2 \exp\left(-\frac{\sin^2(\pi\|x-x'\|\frac{1}{T}}{2\ell^2}\right) \cdot \left(1 + \frac{5\sqrt{5(x-x')^2}}{\ell} + \frac{5(x-x')^2}{3\ell^2}\right) \exp\left[-\frac{\sqrt{5(x-x')^2}}{\ell}\right]$$

This is perhaps the most complicated kernel in our model, where we have a circular kernel multiplied by a Matérn kernel. While initially experimenting with a single Circular kernel,(W_π $(\mathrm{dist_{geo}}(\mathrm{x},\mathrm{x'}))$ (where W_π is the Weinland function W_π $=$

[2] Recall that we denote $\mathbf{GP}$, we unfortunately will not be running a full Gaussian Process with sampling, rather using MAP to estimate parameters to bootstrap

$(1+\tau\frac{t}{c})(1-\frac{t}{c})_+^\tau$ to preserve positive definite for covariance matrices), I found that the function was not nearly as expressive as I had hoped as it failed to follow the small-scale perturbations in electricity demand. Instead, a custom locally periodic kernel $\big(\mathbf{Periodic}(\mathbf{x},\mathbf{x}')\cdot\mathbf{Matrn_{5/2}}(\mathbf{x},\mathbf{x}')\big)$ is constructed using a periodic kernel and a Matérn kernel to better capture movements in our data. (3). As most electricity usage is not purely periodic and don't repeat itself exactly, multiplying kernels together (seen as an "AND" function) gives our model the flexibility to not only exhibit periodic behavior but also allow the repeating part to change over time. (3)

3. Medium Irr. Kernel

$$\eta_3^2 k_3(x,x') := \eta_3^2\mathbf{RQ}(\mathbf{x},\mathbf{x}') = \eta_3^2\left(1+\frac{(x-x')^2}{2\alpha\ell^2}\right)^{-\alpha}$$

   The rationale behind an $\mathbf{RQ}$ kernel is because of its theoretical ability to vary smoothly across many length scales. Since the $\mathbf{RQ}$ kernel is essentially adding together many $\mathbf{RBF}$ kernels together, we can theoretically provide our function the flexibility to capture more uncertain medium-term fluctuations in the electricity demand. In addition to the ℓ parameter and our η_3 parameter, we also have a new parameter α that determines the relative weighting the kernel puts on long-scale vs. short-scale variations (3). This means that we have more expressiveness and allows our function to investigate perturbations in demand at various length scales.

4. Noise Kernel

$$\begin{aligned}\eta_4^2 k_4(x,x') &: = \eta_4^2\mathbf{WN}(\mathbf{x},\mathbf{x}')\cdot\mathbf{Matern_{3/2}}(\mathbf{x},\mathbf{x}')\\ &= \eta_4^2\mathbf{I}(x,x')\cdot\left(1+\frac{\sqrt{3(x-x')^2}}{\ell}\exp\left[-\frac{\sqrt{3(x-x')^2}}{\ell}\right]\right)\end{aligned}$$

   Lastly, the noise model is a white noise model multiplied by the Matérn (3/2) kernel. While I experimented with a Matern (1/2) kernel, which I initially thought would be more descriptive of the true movement of the error, the combined kernel proved to give a better prediction result. One conclusion that can be drawn by the experiments in the cyclical kernel and the noise kernel is that compound kernels are able to be far more expressive than trying to leverage a single kernel to capture all the movements in the data.

Finally, combine these kernels additively to create our model

$$\begin{aligned}y_t \sim \mathbf{GP}_{Linear}(0,\eta_1^2 k_1(t,t')) + \mathbf{GP}_{Seasonal}(0,\eta_2^2 k_2(t,t')) +\\ \mathbf{GP}_{MediumIrr.}(0,\eta_3^2 k_3(t,t')) + \mathbf{GP}_{Noise}(0,\eta_4^2 k_4(t,t'))\end{aligned}$$

3.3 Prior Selection

3.3.1 Kernel Priors

For the kernel prior selection, we will rely heavily on the gamma distribution with varying alphas and betas. Since we will remain agnostic against how the distributions should look like, gamma priors are used for our model. The main reasoning behind gamma priors is that they are very flexible and can be transformed into other well-known families of models by modifying the α and β terms. (7) This flexibility in prior allows our model to use the provided data and take on a posterior shape without interference or prior input.

3.3.2 Error Scaler η, σ Priors

For the error scaler priors, I have elected to use half Cauchy and half normal priors. While the default is usually normal distributions, high variability in the electricity demand should be accounted for using heavier tails.(8) Specifically, the η values on the non-noise components are all Cauchy and the ones on the noise component are normal. This design choice stems from the fact that, relative to a normal distribution, Cauchy distributions have a heavier tail. Since our electricity demand data moves with substantial intra-year variability, the model should be able to capture the extreme values should they occur.

Table 1: Prior Distributions

Parameter	Distribution	Prior Params
ℓ_{pdecay}	Gamma	$\alpha = 4$, $\beta = 0.5$
ℓ_{psmooth}	Gamma	$\alpha = 4$, $\beta = 3$
period(T)	Normal	$\mu = 1$, $\sigma = 0.5$
ℓ_{med}	Gamma	$\alpha = 1$, $\beta = 0.75$
α	Gamma	$\alpha = 5$, $\beta = 2$
ℓ_{trend}	Gamma	$\alpha = 4$, $\beta = 0.1$
ℓ_{noise}	Gamma	$\alpha = 4$, $\beta = 4$
η_{per}	Half-Cauchy	$\beta = 2$
η_{med}	Half-Cauchy	$\beta = 1.0$
η_{trend}	Half-Cauchy	$\beta = 3$
σ	Half-Normal	$\sigma = 0.25$
η_{noise}	Half-Normal	$\sigma = 0.5$

3.4 Optimization Rule

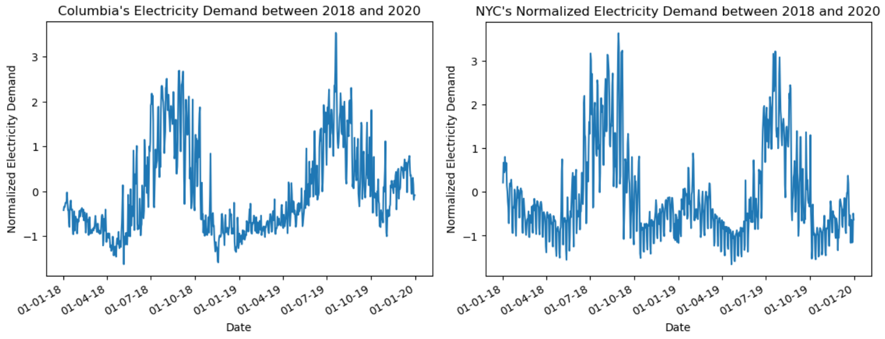
While we would ideally prefer to sample entire distributions through MCMC-based techniques of our posteriors, the computational resources that this project was able to muster were not able to effectively do so. Therefore, our Bayesian implementation is ’partial’ in the sense that we will be directly extracting the maximum a posteriori (MAP) from a gradient-based algorithm instead of using sampling to infer the distribution of our parameters. While we fall shy of the ’full’ BSTS operation by predicting based on a sample from a trace, we will still be able to predict confidently with quantified errors. The MAP is given by:

$$\hat{\theta}_{MAP}(X) = \arg\max f(X|\theta)g(\theta)$$

Where $g(\theta)$ is our prior distribution.

3.5 Data Augmentation

To supplement the two years worth of Columbia’s data, NYC’s normalized electricity data will be used, ’augmenting’ the Columbia data by giving it one more year of data to work with.



(a) Normalized Columbia Electricity Demand **(b)** Normalized NYC Electricity Demand

Figure 1: Normalized Electricity Demand of Columbia and NYC

Notice that the general features between the two figures are quite similar, specifically the peaks and troughs. Since Columbia is a subset of NYC and can be seen as a typical establishment (whereby the primary energy usage is similar to the overall profile of the city). The hypothesis is that adding on the overall NYC data to our training will allow us to further reduce the forecast error. For a better time series forecast, we will be attaching one year of normalized NYC data (between January 2017 and December 2017) to the normalized Columbia demand data. The joint synthetic data is given below.

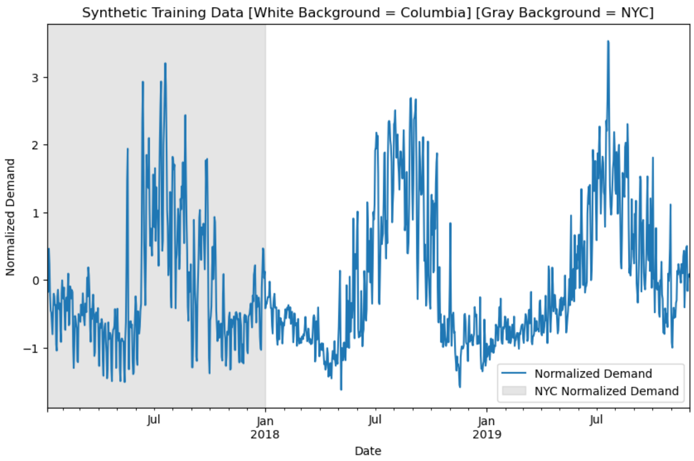
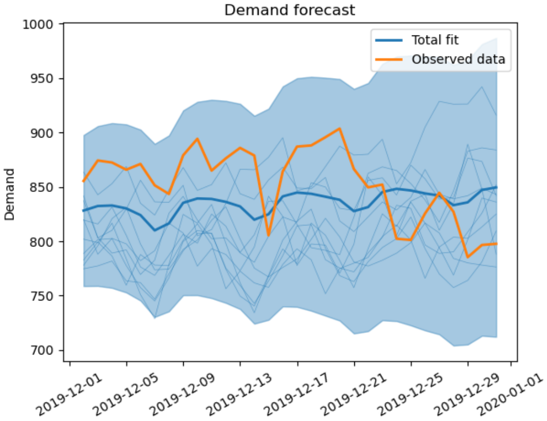


Figure 2: Synthetic training data combining normalized NYC data and Columbia's data

4 Forecasting Results

The primary bootstrapped forecasting results and parameters of our model are given below:



(a) Forecast from model trained on synthetic dataset

Parameter	MAP Value
period	0.875
α	2.190
η_1	0.763
η_2	0.264
η_3	0.076
η_4	0.504
σ	0.093
ℓ_{med}	1.463
ℓ_{noise}	1.517
ℓ_{pdecay}	42.061
ℓ_{psmooth}	0.532
ℓ_{trend}	43.507

(b) Maximum A Posteriori Values

Figure 3: Normalized Electricity Demand of Columbia and NYC

Notice that our model can capture weekly dynamics within our data. The first day of our prediction period (2019-12-01) was a Sunday, and we can see that as the weak goes on, our model predicts a subsequent rise during the workweek, then falling as we approach the weekend. This is a very intuitive result from our model, and suggests that our model can capture the weekly dynamics correctly. However, it is also worth noting that since this is December, we would intuitively expect that the overall trend of electricity usage to be decreasing (as the holidays approach), yet our model predicts a very slight upward trend. This is mainly due to the fat that our model has not fully captured the year-level seasonality that should also be taken into account. Regarding the error bars, we see an intuitive 'fanning-out' of our errors as the period increases, characteristic of time-series forecasting. Coming to the (critical) error estimation of our model, we see that our model can simultaneously be within a 7% error (initially) at the 95% confidence level as well as capture 100% of the true data within this month. Operationally speaking, if Columbia were to install a BESS system and had to forecast out the charge-discharge states a month ahead (which is already plenty), operators would be able to account for the uncertainties in the electricity demand from our model, providing greater operational stability and financial confidence should we arbitrage the electricity demand using this model.

5 Conclusion

In this investigation, we were able to walk through the importance of BSTS in electricity demand forecasting, introduce the foundations of STS and Bayesian methods, construct our BSTS and provide our reasoning for our model design, and generate a forecast that can be used to understand Columbia's future electricity usage. Beyond building upon existing literature, we introduced a novel data augmentation approach to create synthetic data from the publicly available superset of electricity demand and use it to augment our target demand. One key extension of this short project would be to provide the entire distribution of parameters rather than just optimizing our parameters with MAP, leaving us with a 'partial' Bayesian method rather than the full-blown Bayesian approach. This also points towards a key limitation of these models, which is the expensive computation required to train even a small model. Nevertheless, the model results are a first step in creating an electricity demand forecasting method that can be used by institutions to better understand their electricity usage, forecast their usage with operational confidence, and potentially pave the way for certainty in financing, thus paving the way for a more energy-conscious and sustainable energy usage portfolio.

References

[1] Mokilane, Paul, et al. "Bayesian Structural Time-Series Approach to a Long-Term Electricity Demand Forecasting." Applied Mathematics Information Sciences, vol. 13, no. 2, 1 Mar. 2019, pp. 189–199, https://doi.org/10.18576/amis/130206.

[2] A. Harvey, Forecasting, structural time series models and the Kalman filter. 1990. doi: 10.1017/cbo9781107049994.

[3] P. Reiss and F. A. Wolak, "Chapter 64 Structural Econometric Modeling: Rationales and Examples from Industrial Organization," in Handbook of Econometrics, 2007, pp. 4277–4415. doi: 10.1016/s1573-4412(07)06064-3.

[4] I. Katsov, "BSTS Forecasting with PyMC," GitHub, May 27, 2022. https://github.com/ikatsov/tensor-house/blob/master/_basic-components/time-series/bsts-part-4-forecasting-pymc3.ipynb

[5] D. Duvenaud, "Kernel Cookbook," www.cs.toronto.edu. https://www.cs.toronto.edu/ duvenaud/cookbook/

[6] A. Jones, "The Matérn Class of Covariance Functions," Andy Jones, Jul. 31, 2021. https://andrewcharlesjones.github.io/journal/matern-kernels.html

[7] L. Leemis, "Univariate Distribution Relationship Chart," Wm.edu, 2023. https://www.math.wm.edu/ leemis/chart/UDR/UDR.html

[8] "Variability in Electricity Demand Highlights Potential Roles for Electricity Storage." Www.eia.gov, www.eia.gov/todayinenergy/detail.php?id=13131.

[9] K. Bailey, "Gaussian Processes for Dummies ·," katbailey.github.io. https://katbailey.github.io/post/gaussian-processes-for-dummies/ (accessed Dec. 20, 2023).