Forecasting Electricity Demand in the Data-Poor Indian Context

by

Meia Alsup

B.S. Computer Science, Massachusetts Institute of Technology (2019)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Department of Electrical Engineering and Computer Science
August 14, 2020

Certified by .....
Robert Stoner
Deputy Director for Science and Technology, MIT Energy Initiative
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Forecasting Electricity Demand in the Data-Poor Indian Context

by

Meia Alsup

Submitted to the Department of Electrical Engineering and Computer Science
on August 14, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Electricity demand at the grid level is steadily growing in India. More areas are getting interconnected to the grid; and with rising incomes, electricity is highly affected by adoption of air conditioning systems and electric vehicles. Compared with the developed world context where electricity demand is approximately flat if not decreasing year to year, demand in India is growing. In this paper, we aim to examine forecasting methods and determine an optimal method for forecasts in India. Despite limited historical data, we improve forecasts of electricity demand in India out to the year 2050. The forecasts are in five year increments across three different GDP growth scenarios (not accounting for Covid-19). In addition, a layer of natural variation is added to the forecasts for the purpose of modeling the role of various energy technologies on the grid. The methodology to generate more realistic sample loads from predicted average scenarios is a key contribution.

Thesis Supervisor: Robert Stoner
Title: Deputy Director for Science and Technology, MIT Energy Initiative

Acknowledgments

I would like to thank Robert Stoner for serving as my M'eng Advisor, motivating this work originally, giving me the freedom to explore and learn several techniques, and providing support and guidance throughout the entire process. I would also like to thank Marc Barbar for his infinite help and willingness to serve as a sounding board constantly. This work also benefited from the technical advice of Pablo Duanes, Bentley Clinton, Stephen Lee, and Dharik Mallapragada. Finally, I would like to thank my family - Jim, Judy, and Jena - for their lifelong support and love, without whom all of this work would have been impossible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Future of Energy Storage Study Context

The MIT Energy Initiative (MITei)'s study, The Future of Energy Storage, will explore the increasingly important role of energy storage in a de-carbonized world. From electric vehicles to grid resiliency, energy storage technologies are critical to a low carbon society, and batteries will impact the ultimate success of numerous industries important in the transition. In particular, the study seeks to answer what role energy storage technologies will play in the near-term (2030) and beyond (to 2050) in efficient, low-carbon electricity systems. The study considers many storage technologies, cost curves, technological improvements, public policies affecting storage, and intersection with various industrial operations and transportation. Another key question the study tackles is the differing role of storage in OECD and developing countries, especially given the differences in existing infrastructure as well as demand growth. A key country considered in the developing world context is India.

GenX is a technology and optimization tool developed in the MIT Energy Initiative used to optimize electricity resource capacity expansion [15]. GenX is highly configurable electricity system modeling tool that solves a constrained optimization problem. Essentially, GenX gives as output, the mix of electricity generation, storage, and demand-side resource investments and operational decisions that are optimal, subject to a variety of power system operational and policy constraints, such as CO2

emissions limits [15]. The objective in the constrained optimization is lowest cost, and GenX is configured to determine this optimum for a future year, where an input to the model is the electricity demand curve for that future year.

Previous research conducted in the MIT Energy Initiative explores optimal technology suites for a decarbonized future in India using GenX. However, this research relied on a projected demand curve that was just a scaled up version of the current demand curve (ie a multiplicative factor) [21]. The future of demand in India, is not however, a simple scale up of what demand today looks like. While this still provides valuable insight, a much more nuanced understanding of the future grid and role of energy storage can be obtained by running the GenX optimization with a more accurate prediction of future electricity demand in India.

In particular, as explored in depth by the International Energy Agency (IEA), air conditioning is growing rapidly. Especially in hot countries such as India, air conditioning is a dominant component of electricity demand (20% of global electricity demand currently) [14]. As incomes in India rise and more of the population urbanizes and moves to cities, there will be a shift away from fans and towards air conditioning units in residential buildings. On top of that, the commercial sector in India is growing rapidly, and with the increase in office building space comes an increase in day-time air conditioning demand in urban areas. In addition, amongst countries in the world, India has some of the most limited access to cooling. Per capita energy consumption levels are 69kWh, while average world consumption is 272 kWh [18]. The combination of growing incomes, low current access, and hot climates means air conditioning demand and usage will increase dramatically in the next few decades [18].

Electrification of transportation infrastructure such as trains, as well as adoption of electric vehicles including both cars or motorcycles, are both projected to impact electricity demand greatly in the coming decades in India as well.

In addition to projected air conditioning and electrified transportation growth, India exhibits high GDP growth, population growth, increasing GDP per capita, and huge growth in the industrial and commercial sectors. All of these trends have been

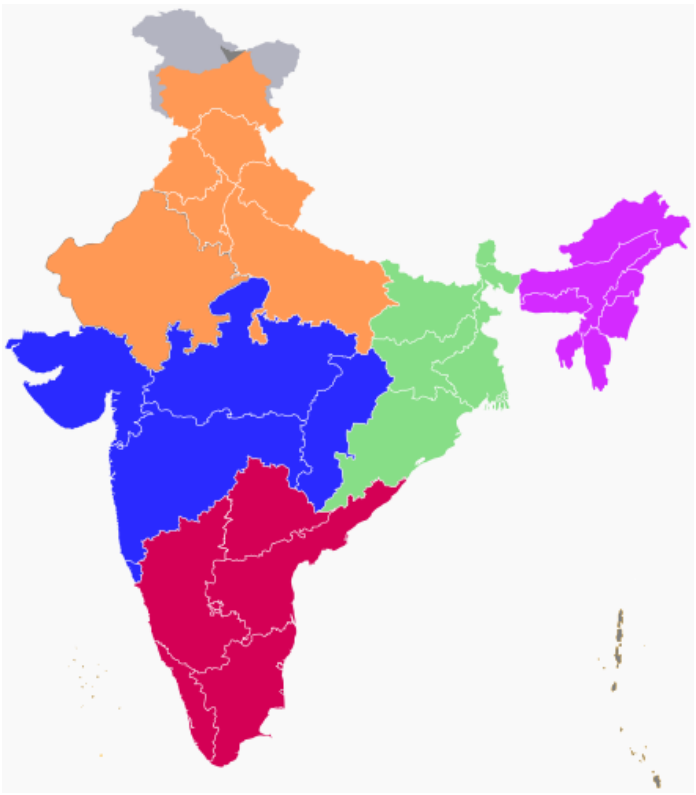

Figure 1-1: Regional Electric Grids in India

linked with increasing electricity demand. Given these trends, this research develops better demand projections in the Indian context that take into account shifting usage patterns as well as growth due to structural economic and social changes.

1.2 Demand Forecasting Goals

Electricity demand exhibits unique characteristics that are important to model including seasonal and daily patterns. Regardless of the time horizon of the forecasts, these characteristics are important to model. Forecasting can be classified into three categories of horizon: long term, medium term and short term. [20]. System planning, expansion, resource allocation, and policy decisions depend on long term forecasts while operational decisions including maintenance, energy management, and daily operations depend on short and medium term forecasts. The optimizations based on short term forecasts enable large amounts of economic savings, as well as secure

operation of the power systems [20]. Generally, forecasts that err too high lead to unnecessary capital expenditure and extra commitment of power generation units. If forecasts are too low, then in a deregulated market, power will be purchased at higher prices.

For the purpose of the Future of Energy Storage study, long horizon forecasts are needed for each of the five regional power grids. The grids are depicted in Fig 1-1. The long horizon predictions are made from 2020-2050 in five year increments. Predictions every five years are sufficient for the modeling purposes of the study. Three different GDP growth scenarios are considered in the projections. At the time of this writing, GDP predictions taking the effects of Covid-19 into account were unavailable, and are not included in the analysis. The GDP predictions are further discussed in 3.2.

1.2.1 Approach

Our approach to demand forecasting consists of four steps. The first step relies on projecting future evening peak and total consumption based on historical data and trends. This is the business as usual scenario that takes historical data and trends and projects into the future. The second step is to add natural variation to the projections from step 1. Since the end-use for the projections is for modeling the role of various energy technologies on the electric grid, it is more important that uncertainty, randomness, and outliers are captured in the projections than the absolute error from reality be minimized. Therefore, the second step adds in natural variation such that the projections into the future exhibit the same statistical characterizations and randomnes as the historical data. The third step, fits the projected peak and total consumption to a historical load curve featuring an evening peak. It is worth noting that load shapes have not shifted much historically, and thus trends that have existed historically are not expected to affect load shape in the future. The fourth step, more art than science, takes into account new trends where historical data lacks. This accounting results in an additive component to the electricity load. The sum of the output from the first three steps combined with the additions from the fourth step constitute our projections of electricity demand into the future.

In more depth, the first step considers econometric and time series approaches to forecasting peak and total consumption. The variables considered include temperature, humidity, rainfall, seasonality, wind speed, GDP growth, and other metrics as found to be important. The second step looks at the statistical difference between a projected year and the actual data. This is used to generate noise and natural variation which is added to the results.

The third and fourth steps considers growth of air conditioning, electric vehicles, energy efficiency trends, end use disagreggation patterns. These efforts were spearheaded by my colleague, Marc Barbar, an EECS PhD student in the MIT Energy Initiative. This thesis focuses on the first two steps of this process, and examines in detail the research process, methodology, and results.

1.3 Contributions

This research makes four primary contributions.

- The primary contribution is on the forecasting front. Previous research with rigorous out of sample validation has predicted demand in India out to 2030. My work provides forecasts of daily peak and total consumption out to the year 2050 - in five year increments, for each of the five Indian electric grids.

- On top of the forecasts, an algorithm to add natural variation is developed and used. This is important for modeling efforts and investment decisions made on the predictions, which are not interested in the average scenario but instead in the peakiness and more indeterminate nature of electricity load.

- A framework for model evaluation in a data-poor context is also developed. This work could motivate future investigation from a learning theoretic approach, which is discussed more in the Future Work (Section 6.2).

- The Python code module used to acquire data from the NASA Merra-2 data-set has been open-sourced and made available to wider researcher community [1]. I

[1] github.com/meiaalsup/merra-2

found the Merra-2 interface difficult to work with, so this Python module can help prevent headache for many future researchers.

1.4 Paper Outline

In this paper, electricity demand of India is forecasted by regression methods which consider environmental factors and GDP. This paper is organized as follows: chapter 2 discusses the necessary background on related work, the data-sets, and provides a primer on model evaluation techniques, chapter 3 describes the data sets in depth, chapter 4 discusses the model results and demand projections, chapter 5 describes the methodology for adding natural variation and provides the results, and finally chapter 6 concludes with a discussion of challenges and future work.

Chapter 2

Background

2.1 Related Work

2.1.1 Demand Forecasting

Several models have been proposed and evaluated for forecasting electricity demand in various contexts. These models have utilized the Grey Model, time series forecasting techniques, regression, and machine learning.

Ivan Rudnick of the MIT Energy Initiative explored 24 scenarios of grid decarbonization in the Indian context using GenX out to 2037. The scenarios spanned varied clean energy and energy storage prices and characteristics, emissions limits, and gas prices. The input for electricity demand in 2037 was a scaled version of the 2015 electricity load profile [21]. A limitation acknowledged in this work is the deterministic nature of the load prediction.

A study put out by Brookings India, examines the growth of various end-use sectors out to 2030. Nine different cases are analyzed, representing low, medium and high GDP growth scenarios crossed with three scenarios for energy efficiency upgrades and conservation efforts [1]. A key driver of the projected commercial electricity demand is growth of air-conditioned square footage. The analysis further projects a decoupling of GDP growth from energy demand due to consumption efficiency and dominance of the services sector in electricity growth. The Brookings study uses a

bottom up approach of projecting demand in each different sector, ultimately adding them to get demand at the aggregate level.

Thomas Spencer of The Energy and Resource Institute (TERI) in India follows a similar approach to the Brookings Study in his projections of Indian Electricity Demand to 2030 [24]. Spencer considers three macroeconomic scenarios corresponding to GDP growth of 6.8%, 7.5%, and 8% GDP growth as well as a high and low energy efficiency scenario for a total of six scenarios. Base electricity projections are forecasted using econometrics based on historical scenarios on a sectorwise basis (agriculture, industrial, residential, services). A transportation network analysis as well as end use analysis of each sector is layered on top of that to fine-tune the projections. This is also a bottom up projection because the individual components of electricity demand are forecasted, and the results of each forecast are then summed together.

Since the 1960s, many models have been used to forecast and understand electricity demand. One popular field of models are time series models. In particular, Auto Regressive Integrated Moving Average (ARIMA) time series models which employ the Box-Jenkins method have been applied in several contexts including in Spain, India, Sri Lanka, South Africa, Colombia, and Turkey [2, 10, 23, 9, 16, 26, 27]. In these contexts, SARIMA does not take into account into exogenous variables, and used past seasonal and time based patterns to project the future. These were shown to have good short-term predictive potential. SARIMA is a class of models on top of ARIMA that adds in normalization for seasonal effects. In the South African context, Chikabvu et al. demonstrated that the SARIMA model produced more accurate short-term forecasts than many other models [9]. The results of Velasquez in Colombia also demonstrated the utility of the SARIMA model for short term forecasting [26]. A drawback of time series models is that they do not incorporate the effects of exogenous variables out of the box. The SARIMAX model was developed to include exogenous variables as well. SARIMAX has been used in a few demand forecasting contexts as well including in Japan and Italy [12, 25]. The use cases of these models were short term: including one to nine day ahead hourly forecasts.

In the Indian context, more time series models have been applied to forecast demand. Kumar et al. examined the Grey-Markov model, Grey-Model with rolling mechanism, and singular spectrum analysis (SSA) to forecast conventional energy consumption in India including crude-petroleum, coal, electricity, and natural gas. The models were hand selected based on the structure of each individual time series [16]. These models further demonstrated the potential utility of time series methods in demand forecasting. However, in the studies employing the Grey model so far, the performance was not evaluated out of sample [13, 17, 16]. Therefore, it is unclear how well the methods generalize to future predictions.

Another large category of models common in load forecasting are regression models. The largest advantage of regression methodologies are that they allow for modeling of specific relationships between the independent variables and the load. They are the simplest to implement and understand. However, they have limitations in that they cannot capture non-linear relationships, and do not capture seasonality well either. Linear regression models have been used for long-term forecasts in several contexts, including load demand in England, hydro electricity load in Quebec, demand in Italy, demand in Cypress, and electricity consumption in Hong Kong [27] [11] [7]. Bianco demonstrated that in the Italian context, predictions made with regression methodologies considering historical data from 1970-2007 matched experience and national government predictions well. This model took into account historical electricity consumption, gross domestic product (GDP), GDP per capita, as well as population. Egelioglu discovered that a model based on several econometric measures including tourism data could be used to predict demand in Cypress very accurately [11].

More recently, machine learning approaches have been applied to the problem of predicting energy consumption. Artificial Neural Networks in particular have received a lot of attention because of their ability to capture nonlinear relationships and capture multiplicative effects of multiple variables. Azadeh et al. propose an artificial neural network for forecasting annual demand in Iran with GDP and Population as independent variables [5]. Amjady uses fuzzy neural nets to improve day-ahead

forecasts in the Spanish market [3]. Saravanan et al. investigated artificial neural networks in the Indian context, but did not evaluate the model with out of sample validation [22]. Artificial neural networks have not demonstrated good predictive power for long term forecasts in the Indian context based on the literature review thus far. In Turkey, genetic algorithms, another are of machine learning research, have been used to predict two scenarios of low and growth for fifteen years out of electricity demand consumption [19]. Overall, neural network based approaches have included socio-economic indicators such as GDP, import and exports, and population, as well as household variables, pollution, air-conditioner counts, and temperature forecast electricity demand. Across the board, short-term forecasts were much more accurate than long-term forecasts. For each model, the obtained error for daily peak demand forecasts was higher than the average day ahead hourly forecasts.

In a 2017 review by Yildiz et al examining past experiments and methodologies, it was found that regression models performed fairly well compared with many more advanced machine learning methods. For long term forecasts, both of these models outperformed time series based approaches [27].

The models discussed so far can broadly be categorized in two buckets. The first bucket preserves the electricity demand shape, uncertainty, and statistical characteristics. These models scale either the entire load, or scale the load piece by piece. Either way, the randomness exhibited by the base year chosen is preserved in the projection. The second bucket involves models that project the "average" case by minimizing a loss function of choice. Effectively these models project a smoother, average curve from historical data. While these model minimizes error (for the error metric of choice), they do not capture the random fluctuations and uncertainty present in real electricity demand curves.

2.1.2 Model Performance Evaluation

The goal of performance estimation is to estimate how much error a model will and quantify the predictive loss into the future. Every forecasting effort includes performance evaluation in the pipeline to assess the how well the models generalize and

predict the future. In this work, I am interested in model performance evaluation specific to time series forecasting. The two generally most common methods of assessing model performance are cross validation and out of sample validation [8, 4].

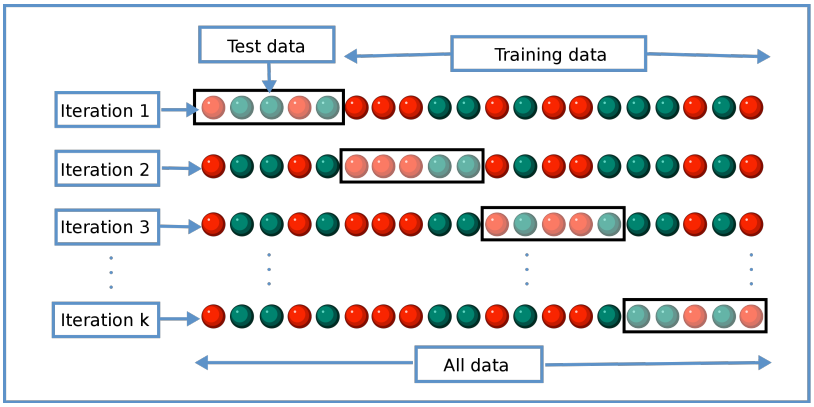

Figure 2-1: Cross Validation Data Splitting

Cross validation approaches enable the most efficient use of all the data. Cross validation works by minimizing loss, where loss is defined as the average loss from several separate training runs. For each training run, a subset of the data is held out at random. Figure 2-1 visualizes this methodology [1]. The assumption made here is that all observations are independent and identically distributed. While some strategies attempt to circumvent this, the time dependency of time series data leads to issues about the best way to not have information leak from the training set into the validation.

In cross validation, the method for choosing the particular sequence of data splits is an important question. Each split should be random and independent, as well as take into account some of the underlying data structure. Specifically, if the data is stratified across certain dimensions, the proportions of the different strata should be about equal in each training and test partition [4]. Still, there is no settled methodology to estimate performance with cross validation in the time series setting given the time dependence of the train and test set . Despite these limitations, in data-poor settings, cross validation is popular because it enables all the data to be used [8].

In contrast, out of sample methods preserve the temporal order of observations

[1] //en.wikipedia.org/wiki/Cross-validation_(statistics)

and instead hold out a final time period on which to evaluate the predictions which were trained on only the earlier time period. The disadvantage is that the newest data cannot be used in the model training. However, this method accounts for potential temporal correlations between consecutive time series values and doesn't have the issues that cross-validation techniques have.

Given the plethora of forecasts that utilize both of these methodologies, it is useful to understand their merits for time series forecasting. In a 2019 paper, Cerquiera compared cross validation and out of sample approaches on real world and synthetic time series data. They found that the cross validation approaches perform well on stationary synthetic time series. However, in the 63 real-world scenarios examined, the out-of-sample methods produced more accurate estimates [8].

2.2 Data

Data was collected from a variety of sources including POSOCO (the major Indian utility)[2], the Energy and Resources Institute (TERI)[3], The Ministry of Statistics and Programme Implementation by the Government of India (MOSPI)[4], NASA (specifically their Merra 2 dataset)[5], the World Bank[6] and the Reserve Bank of India's Handbook of Statistics on the Indian Economy[7]. There were many challenges associated with obtaining data that would be useful for the model. The process of data acquisition and searching for data is explained in depth in the Discussion in section 6.1.1.

For the full four step process, more data was considered including a 2015 full hourly load profile for India broken down by region, daily total electricity consumption and daily peak from 2014 to 2019, sector-wise consumption (See Figure 2-2), and end-use load dis-aggregation for specific regions and sectors. Some of the economic metrics

[2] //posoco.in/
[3] //www.teriin.org/
[4] //mospi.nic.in/data
[5] //gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
[6] //data.worldbank.org/country/india
[7] //www.rbi.org.in/Scripts/AnnualPublications.aspx?head=Handbook+of+Statistics+on+Indian+Economy

available from the Reserve Bank of India (RBI), including quarterly GDP growth, state-wide consumer price index, per capita state net domestic product, amongst others were considered as well.

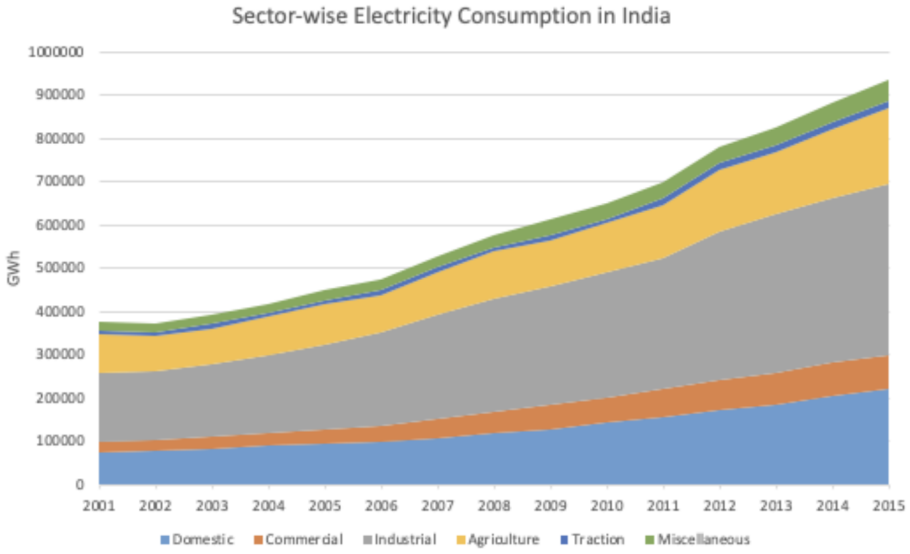

Figure 2-2: Sector-wise Consumption in India from 2001 to 2015

For the first two steps that are discussed in this thesis, the data used in the model was all obtained from POSOCO reports, NASA's Merra-2 dataset and from the Reserve Bank of India for statewise GDP. The augmentations and forecasts on top of this historical data is discussed in Section 3.

All of the data used in this analysis, along with our code and our results can be found on the project github[8].

2.3 Model Evaluation

2.3.1 Evaluation Techniques

In this work, given that there are no accepted cross validation methodologies for time series data, the evaluation metric is based on the error on an out of sample test set. The peak and consumption daily data used in the model are from 2014-2019. As such,

[8] //github.com/barbarmarc/india-load

models were trained on 2014-2018 and then evaluated for error on predictions of 2019 versus actual 2019 values. Cross validation evaluation techniques were considered, but not used in practice since the literature does not support its utility in time series scenarios.

2.3.2 Types of Error

In model training, it is useful to distinguish how a selected model or hypothesis leads to errors on the test data. There are two main ways a hypothesis leads to errors: structural error and estimation error[9]. Structural error occurs when there is no hypothesis from the family of hypotheses under consideration that will perform well. An example of this is data generated and pulled from a log distribution is attempted to be fit with a line. Estimation error occurs when there is insufficient data for the model to learn the representation and make good predictions.

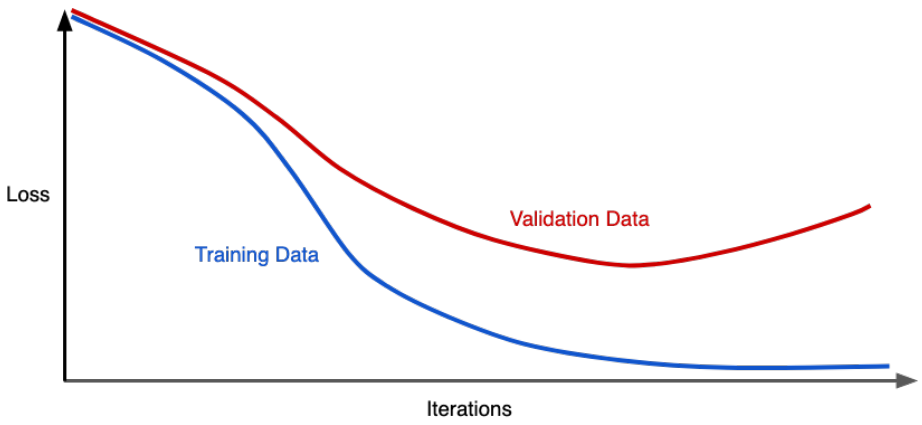

Figure 2-3: Generalization Curve

One method of minimizing structural risk is to penalize the model. Essentially, this prevents the model from overfitting to the data in the training set. Overfitting means the model learns relationships that are specific to the data used in training and do not generalize well. Figure 2-3 shows a model in which training loss initially decreases gradually along with decreasing loss in the validation set, but eventually loss

[9] //developers.google.com/machine-learning/crash-course/regularization-for-simplicity/l2-regularization

in the validation set starts to increase. This increase is because the model has learned relationships that exist only in the training set and are just noise. The standard method for preventing overfitting is regularization, which penalizes complexity in models to keep them simple and general.

2.3.3 Regularization

Regularization adds a penalty term to objective that is trying to be minimized. Basically, rather than solely minimizing loss the minimization is for the loss of a model (error) added to the model complexity (multiplied by a tunable parameter, usually λ). The model complexity term is the regularization term. There are two common regularization techniques in model fitting, the L1 and L2 regularization. L1 regularization penalizes the number of features with non-zero weights included while L2 regularization penalizes the magnitude of the weights.

L1 regularization is the number of non-zero coefficients. L2 regularization is the sum of squared weights. In L1 regularization, outlier weights don't have a huge impact, but many non-zero weights or very small weights have high impact. In L2 regularization, a weight with high absolute value affects model complexity a lot more but weights close to zero have little impact on complexity.

Elastic Net is a regularization technique that combines L1 and L2 regularization linearly via a tunable parameter for the relative ratios of both. In this research, elastic net is used to help prevent overfitting.

Chapter 3

Data Acquisition

The GDP data used was put together by Marc Barbar, and details can be found in our working paper [6]. Environmental variables were secured from the NASA Merra-2 data set. Specifically, the variables obtained were specific humidity, temperature, eastward wind and northward wind all 2m above the surface and 10m above the surface. Precipitable ice water, precipitable liquid water, and precipitable water vapor were also included. These values were obtained from the instantaneous two dimensional collection "inst1_2d_asm_Nx (M2I1NXASM)". Detailed descriptions of these variables are available in the Merra-2 file specification[1] provided by NASA. The environmental variables available from the NASA MERRA-2 dataset were given on an hourly basis. The energy data used as baseline truth for peak and total consumption was obtained from the daily reports Posoco made available with statewise data.

3.1 NASA Merra-2

For each of the five electric grids, the largest cities in each regions were identified using population data made available by the United Nations[2]. Then, the latitude and longitude of the city was extracted from Google Maps, and used to pull down

[1] //gmao.gsfc.nasa.gov/pubs/docs/Bosilovich785.pdf
[2] //population.un.org/wpp/

the corresponding environmental data from the Nasa Merra-2 data set.

The cities used for each of the five regions are listed here:

- Northern: Delhi, Jaipur, Lucknow, Kanpur, Ghaziabad, Ludhiana, Agra
- Western: Bombay, Ahmedabad, Surat, Pune, Nagpur, Thane, Bhopal, Indore, Pimpri-Chinchwad
- Eastern: Kolkata, Patna, Ranchi[3]
- Southern: Hyderabad, Bangalore, Chennai, Visakhapatnam, Coimbatore, Vijayawada, Madurai
- Northeast: Guwahati, Agartala, Imphal

For each of the environmental variables considered, there were 24 data points for each of the hours of the day. Since the variable being predicted, peak and consumption, was at the daily scale, the hourly data was encoded in the model in three variables: daily minimum, daily maximum, and daily average.

For each region, there are three multiplied by the number of cities multiplied by the number of environmental variables considered (11 total) environmental variables originating from the NASA data set in the model.

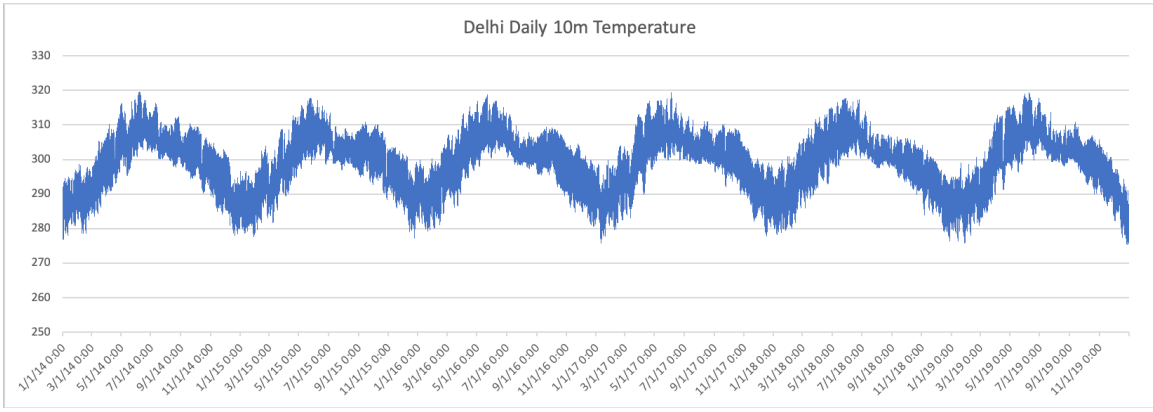

Figure 3-1: Delhi Hourly Temperature 2014-2019

[3]Howrah was ignored because the environmental factors are the same as Kolkata

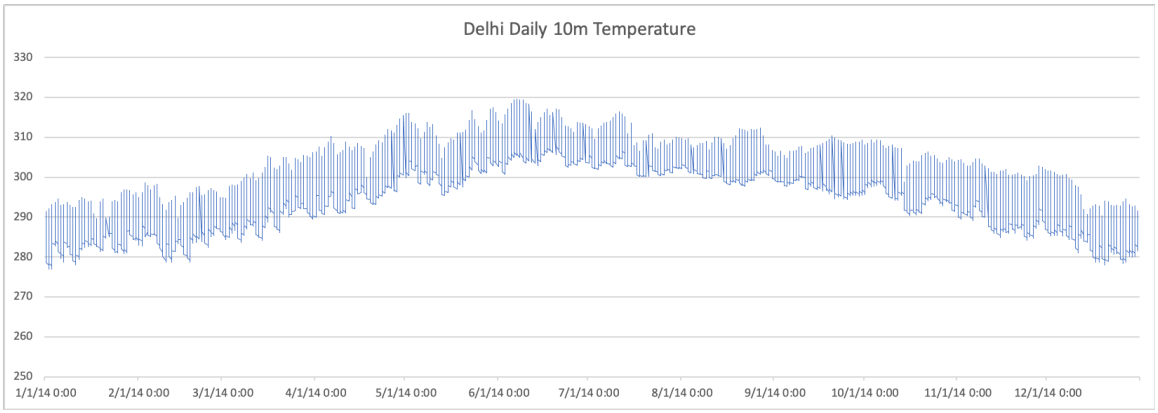

Figure 3-2: Delhi Hourly Temperature 2014

An example of the daily temperature 10m above the surface in Delhi is shown in Figure 3-1 for 2014-2019 and for just 2014 in 3-2.

For each of the environmental variables, given that there are no forecasts for temperature, humidity, precipitation, etc at a citywide basis out to 2050, forecasting of these variables was simply the average of each variable from 2014-2019. An improvement to this that includes taking climate change into account is discussed in Section 6.2 on Future Work.

3.2 GDP

Statewise historical GDP was obtained from the Reserve Bank of India. Three different growth scenarios were built. Through linear regression we project GDP per capita based on the available historical data. The GDP per capita from historical data was adjusted via weighted sum of purchasing power parity per capita and kilowatt-hour consumption per capita. For the medium growth stable scenario, the weight are nominal. For the rapid and slow cases, the weights are adjusted to follow an S-curve above and below the stable case, respectively. The values are chosen from a normal distribution with a mean of the highest historical growth rate for the rapid case, and lowest for the slow case. The approach to GDP forecasting is detailed in the working paper [6].

3.3 Daily Peak and Consumption

The daily peak and consumption, the variables being predicted out to 2050, were available historically from POSOCO in daily reports. The daily report pdfs were scraped for these two variables of interest.

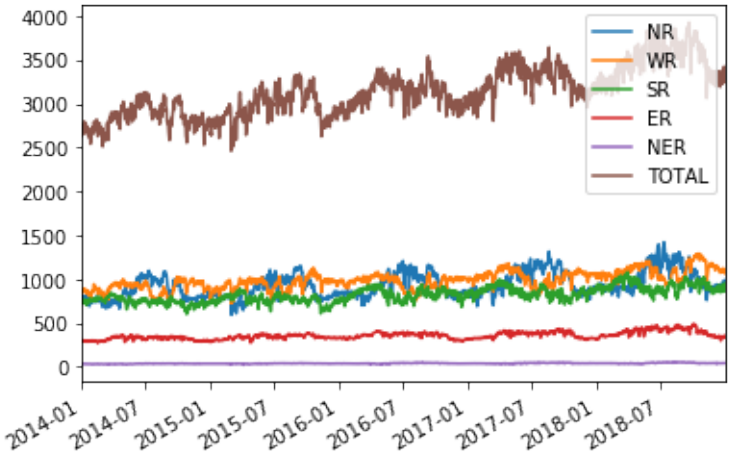

Figure 3-3: Daily Consumption for All Regions 2014-2018

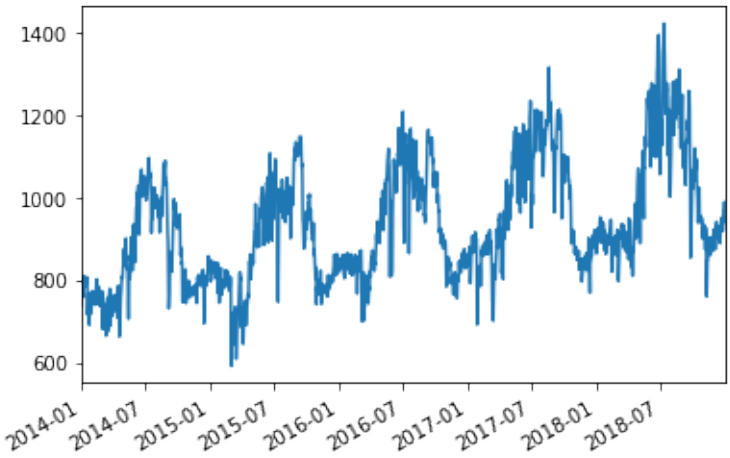

Figure 3-4: Daily Consumption for the Northern Region 2014-2018

Consumption data for the five regions from 2014-2018 can be visualised in Figure 3-3. Figure 3-4 presents the Northern Region daily consumption as a sample close up look.

Chapter 4

Demand Forecasting

4.1 Experimentation and Model Selection

The top priorities and trade-offs considered in model selection were explain-ability, interpret-ability, and accuracy. Three model families were considered: deep learning neural net based models, time series models, and regression models.

4.1.1 Machine Learning Models

Initially, machine learning models were explored by the forecasting team member, Marc Barbar, in the Texas power grid context. The feasibility of machine learning methods was investigated. Model performance appeared decent, though we caveat with the fact that the machine learning model was not actually compared with any other models in the Texas context. However, when the work was presented with the broader team, other team members asked about the importance of different variables including GDP in the predictions. Since Machine Learning models are black box models that cannot be interpreted in this way, this was not a question we could answer.

For the purposes of the Future of Storage study, it was important to be able to quantitatively understand the importance of different variables including GDP and temperature in the results. Thus our research shifted direction to prioritize time series

and regression based methods, and de-prioritized machine learning models.

4.1.2 Time Series Models

Time Series methodologies were explored starting by with examining auto-correlations in the data and the ARIMA model.

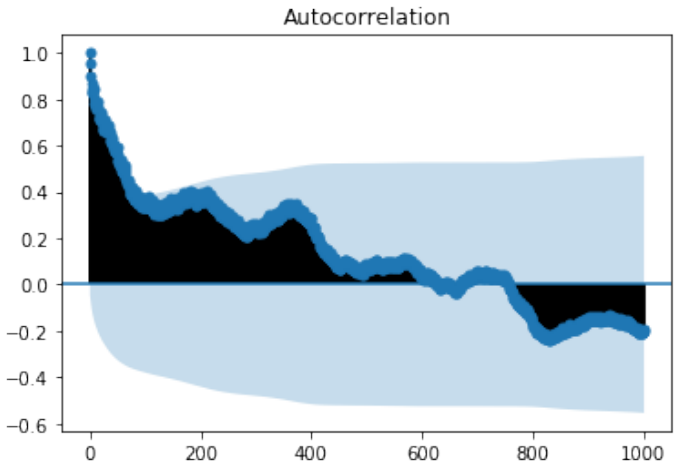

Figure 4-1: Autocorrelation of Southern Region Consumption Data

The auto-correlation plot in Figure 4-1 shows that there is large seasonality effects in the data. This pattern was seen across all ten scenarios (the cross product of five regions with peak and consumption). The plot for the Southern region shown functions as an example and is representative of the rest of the data.

Given the seasonality of the data, SARIMA was determined to be a better model since seasonal effects could be controlled for. Initially the model was given a yearly and weekly period to account for both trends in the data. In order to include the effects of exogenous variables including GDP, SARIMA-X was used. However, there was a limitation that the existing Python libraries for SARIMA-X could only support a single periodicity. Since the data has both yearly and weekly patterns, this would not work. Instead, a method of taking advantage of fourier transforms was explored.

For this new approach, instead of using a python time series library for sarima, a regression model was used instead. Sin and cosine waves were added as independent variables, from which the model could construct the periodic waves. Environmental

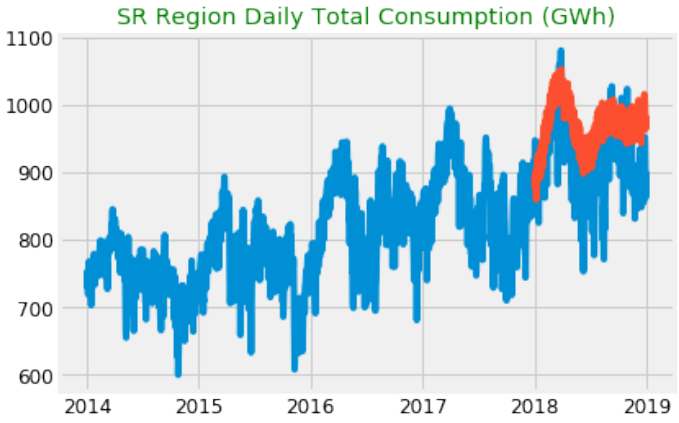

Figure 4-2: Linear Regression Model on fourier components for Southern Region Consumption Data

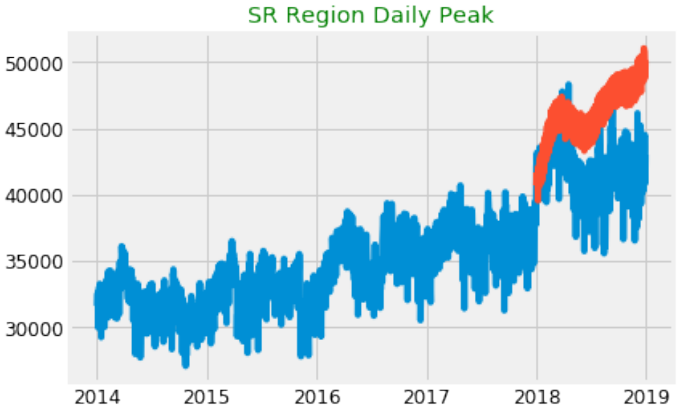

Figure 4-3: Linear Regression Model on fourier components for Southern Region Peak Data

variables and GDP were also included as independent variables. A few sample results from this model for the Western and Southern regions are shown in Figures 4-2, 4-3, 4-4, 4-5. These predictions were all made out of sample.

The predictions out to 2050 that the model made exhibited some strange behaviors. A few samples are shown in Figures 4-6, 4-7, and 4-8. The model learned a negative relationship with some independent variables in a few cases, which led to results that did not make sense.

A further investigation into the literature suggested that seasonality based models generally perform well in the short term, and have not seen success for long horizon forecasts in the energy context. Further, the approach via fourier coefficients was not

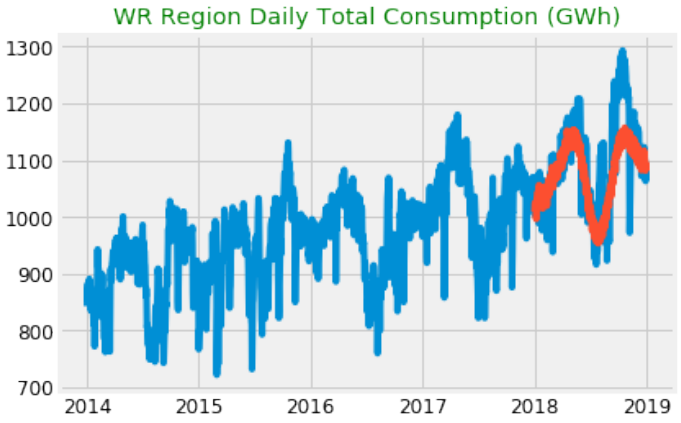

Figure 4-4: Linear Regression Model on fourier components for Western Region Consumption Data

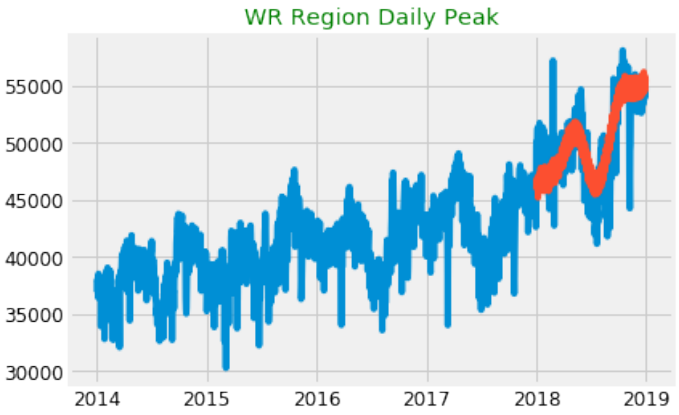

Figure 4-5: Linear Regression Model on fourier components for Western Region Peak Data

well supported by other experiments in the literature.

4.1.3 Regression

Plots of temperature against electricity demand showed that most periodic seasonality was actually well captured in the environmental data. Therefore, there was no additional need for the fourier approach to add periodicity in the regression approach. A sampling of results from the regression are shown in the next two sections for the Peak and Consumption data.

The elastic net sklearn python module[1] was used to train the linear regression

[1] //scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html

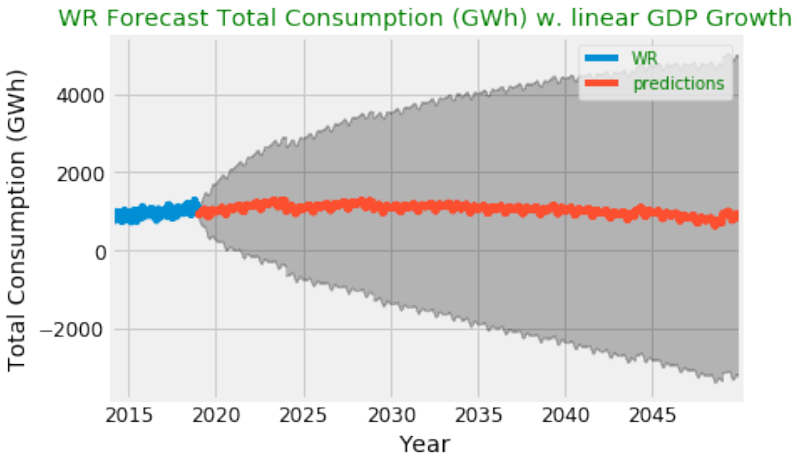

Figure 4-6: Linear Regression Model on fourier components for Western Region Peak Data Forecasts

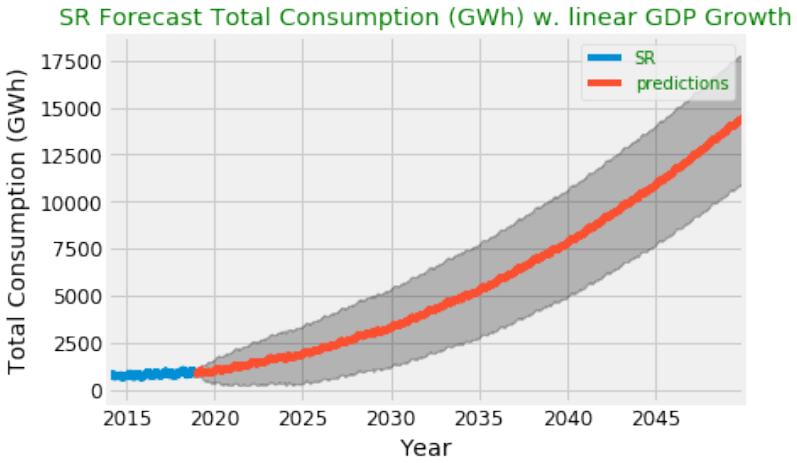

Figure 4-7: Linear Regression Model on fourier components for Southern Region Peak Data Forecasts

model and tune the ratio of l1 to l2 regularization. 2019 data was held out for out of sample validation. A table showing the average negative log likelihood of the results for the ten scenarios can be found in the Appendix in Table A.1. This relationship is also shown in Fig 4-9. Based on these results, a L1 ratio of .9 was chosen in model training. This indicates that the model performed better when more weights were sent to 0 and weight magnitudes were not heavily penalized.

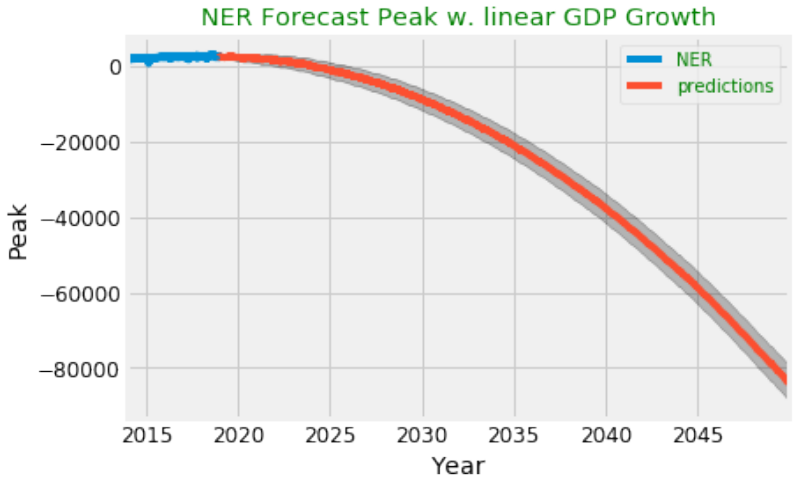

Figure 4-8: Linear Regression Model on fourier components for Northeastern Region Peak Data Forecasts

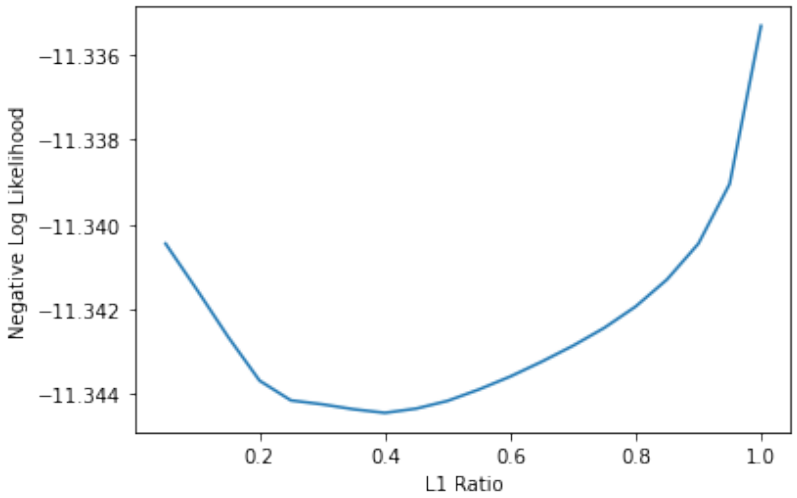

Figure 4-9: Relationship between ratio of L1 to L2 regularization and model performance

4.2 Results

The full results can be found on github[2]. A sampling of Peak and Consumption results are depicted in the remainder of this section.

[2] //github.com/barbarmarc/india-load/tree/master/step1/linear_regression

4.2.1 Peak

Forecasts for the Northern and Western Regions with linear GDP growth in 2035 are shown.

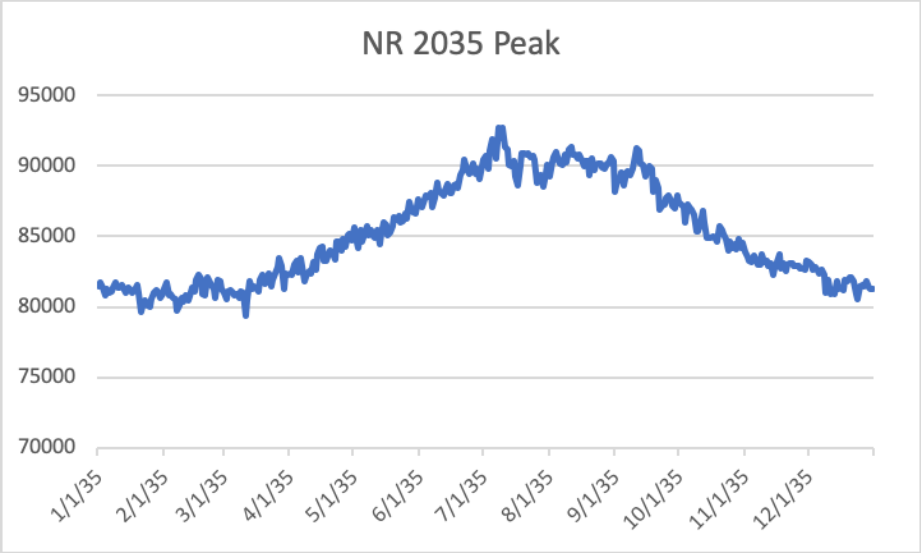

Figure 4-10: Northern Region Peak Forecasts in 2035

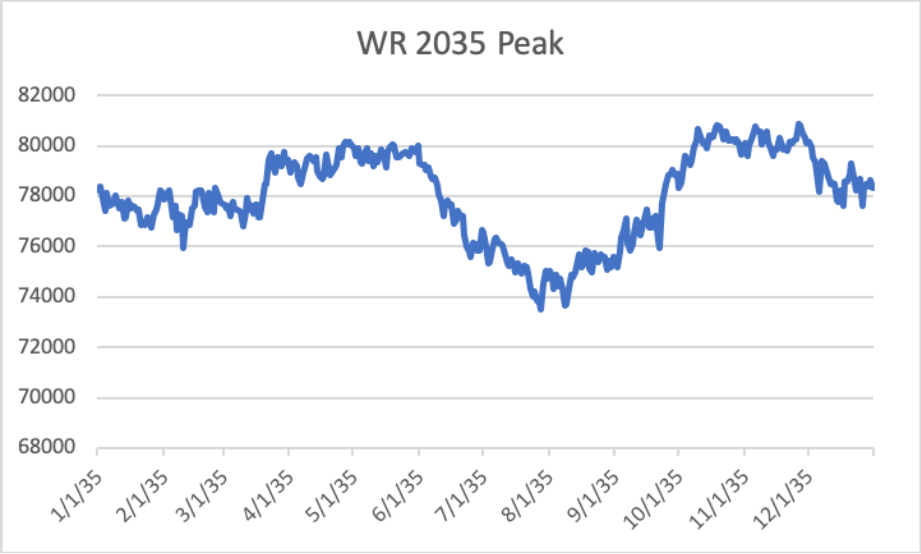

Figure 4-11: Western Region Peak Forecasts in 2035

4.2.2 Consumption

Forecasts for the Northern and Western Regions with linear GDP growth in 2035 are shown.

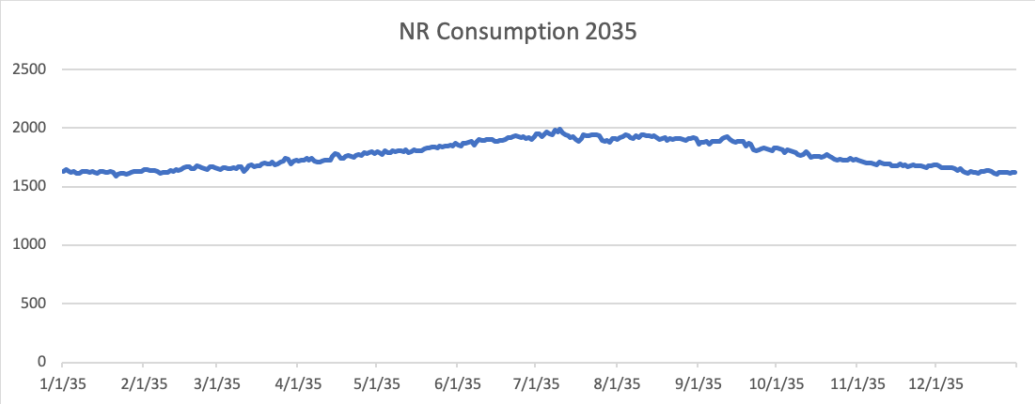

Figure 4-12: Northern Region Consumption Forecasts in 2035

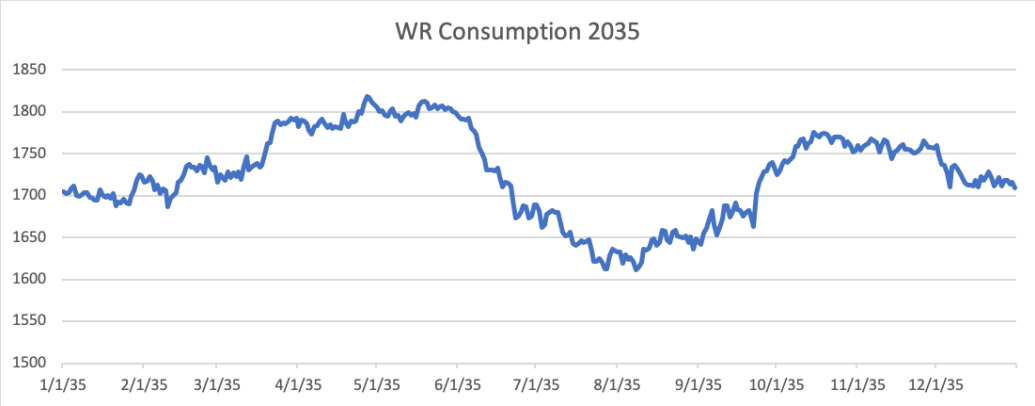

Figure 4-13: Western Region Consumption Forecasts in 2035

Chapter 5

Adding Natural Variation - Noise

5.1 Motivation

The primary use case for the daily peak and daily total consumption is to generate forecasts as inputs into MITei's GenX system for modeling. Because of this, the natural variation in electricity demand is desired.

5.2 Methodology

To add natural variation to the predictions, the amount of natural variation was first estimated. Then the natural variation distribution was modeled and noise values were drawn from that distribution to add to the model. This estimation was done by subtracting the predictions made for 2019 out of sample from the ground truth 2019 values and tracking the absolute value of the differences. The mean and standard deviation of the differences was calculated. For each future year, values were drawn from a normal distribution with the mean and standard deviation calculated, and randomly multiplied by a true false bit for positive or negative noise.

5.3 Results

All results can be found on github in the corresponding "with noise"[1]. A few sample plots of the predictions before and after noise are shown in Fig 5-1 and 5-2. These plots show the peak and consumption forecasts for the Northern region with average GDP growth.

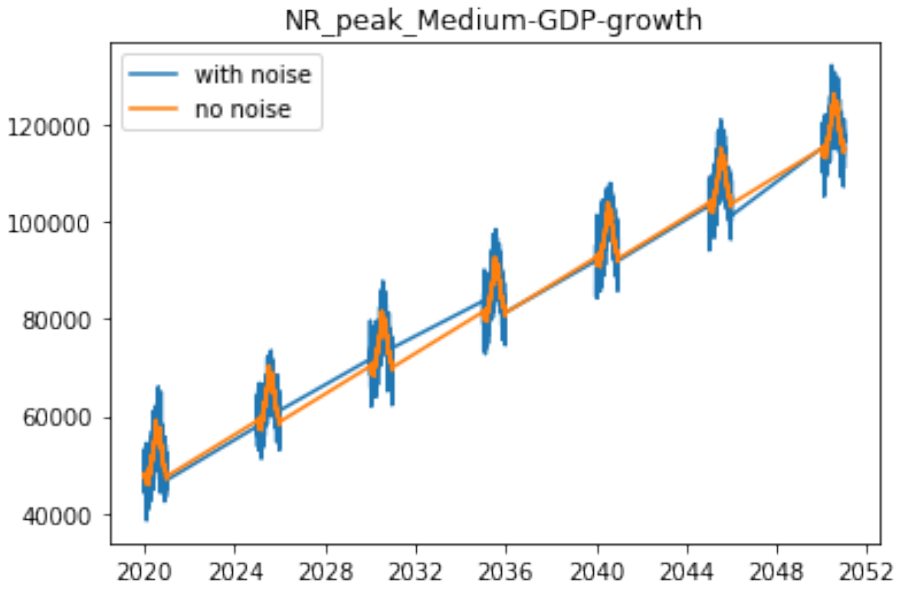

Figure 5-1: Northern Region Peak Forecasts with natural variation

A closer up view at just the year 2035 is depicted for the same scenarios in Fig 5-3 and 5-4.

The same natural variation (noise) was added to the three growth scenarios. While the natural variation was generated randomly, the three scenarios when modeled downstream will be more comparable with the same variation. Thus noise was generated once and shared amongst the three growth scenarios.

[1] //github.com/barbarmarc/india-load/tree/master/step1/linear_regression

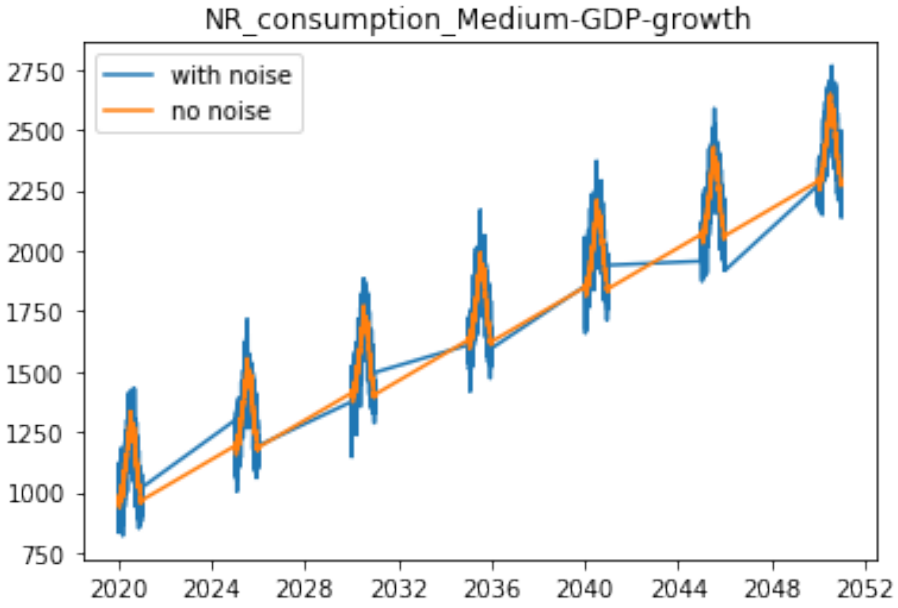

Figure 5-2: Northern Region Consumption Forecasts with natural variation

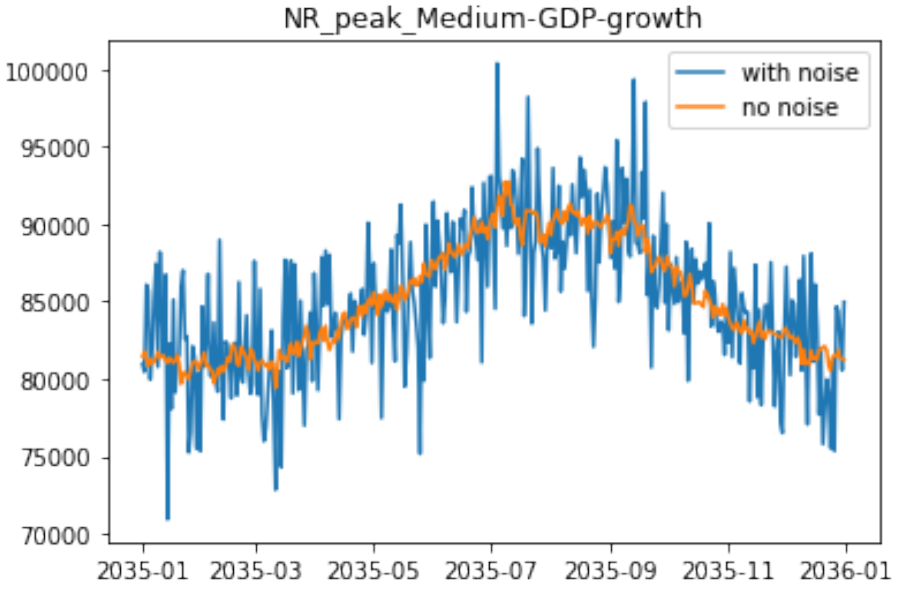

Figure 5-3: Northern Region Peak Forecasts with natural variation for 2035

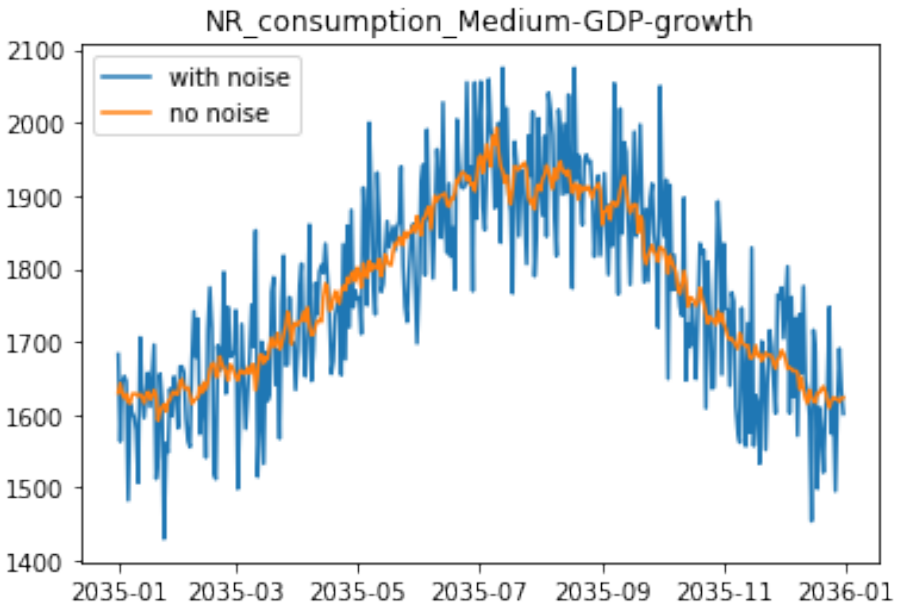

Figure 5-4: Northern Region Consumption Forecasts with natural variation for 2035

Chapter 6

Discussion

6.1 Challenges

6.1.1 Data Acquisition

The beginning of my research into electricity forecasting in India consisted of an exhaustive literature review as well as many calls and video meetings with key stakeholders and potential partners, many of whom were in India. The goal of all of these meetings was to acquire more data that could be useful in the predictions. This proved excruciatingly difficult since most parties were unwilling to share data without an NDA, and MIT research policies strongly frown upon such agreements.

6.1.2 Data Quantity

Further, the data collected was often for a single year, or for a certain subset of time. The overlap of data-sets on the time axis was poor, meaning it was very difficult to line up and utilize multiple data-sets, even if they were available.

In particular, census data was available in 10 year increments. State-wise population was available up until 2012. A full annual demand curve in 15 minute increments was available for the year 2015 for each of the five regions. Daily total consumption and daily peak were available for 2014-2019. Demographic information related to household size, urbanization percentages, rural versus urban populations and more

were available for 2012 and 2015, and with a prediction for 2030. End use profiles were available broken down by relative hourly usage in some cities. Sector growth for service, industry, and agriculture were available for 2001-2015, with no additional more recent data nor future predictions.

Other metrics related to the Indian economy were available from the Indian government on an annual or quarterly basis. However, these were only available at the aggregate India level, and not broken down by regional grid. As a result, these were also not very informative for modeling different growth patterns in each of the five regions.

On the surface, what appeared like a lot of data to pull from, was actually not very useful using traditional modeling methods. Figuring out a way to effectively to combine data from non-overlapping time-frames is outside the scope of this research. Since no demand data was available prior to 2014, any data from before 2014 was useless for modeling. Further, data that was only available for a small subset of 2014-2019 were also insufficient to serve as explanatory variables because a model cannot learn anything useful from only a couple data points.

Learning a trend from 6 years of data is already a challenge. Learning a relationship with data for only a subset of those 6 years is nearly impossible.

6.2 Future Work

6.2.1 Additional Data

The primary lever that would improve the quality of predictions would be increased access to data. Potential avenues to consider would be agreeing to signed NDAs with the Indian utilities and government, or more persuasive conversations with these key stakeholders to publicize and share their data. More work and creative asks in this area could greatly impact the accuracy and downstream impact of these predictions.

6.2.2 Learning Theory

The field of computational theory studies performance bounds, time complexity, and the feasibility of learning. In particular, a "learn-able" model indicates that a model class can be learned in polynomial time. Leslie Valiant proposed Probably Approximately Correct (PAC) algorithms which provide δ and ϵ bounds on the deviation of the observed samples from underlying distribution and the probability of learning the model.

The theoretical understanding from the field of learning theory could be applied to data poor modeling in general, and specifically to the energy context. This could help quantify the expected probability at arriving at a model that describes the underlying distribution, based on the number of samples in previous data. Namely, less samples means there is a higher probability that the observed samples don't describe the true underlying distribution well and a model cannot be learned.

6.2.3 Revised GDP Forecasts

The global health pandemic caused by Covid-19 has impacted economies and growth worldwide. As a result, GDP predictions out to 2050 should be revised down, especially in the near-term. The model should be rerun with revised GDP predictions to be as accurate as possible.

6.2.4 NASA Merra-2 Dataset forecasts

The forecasts for the Merra-2 datasets take the average of the previous 6 years from 2014-2019. These forecasts are therefore an average, and do not exhibit the natural variation and fluctuations present in real data. This is part of the reason that natural variation adding was necessary in this research. It would be useful to explore modeling different future scenarios that take into account changing climate patterns. In particular, it would be especially valuable to consider global warming in the temperature and humidity forecasts.

Appendix A

Tables

Table A.1: Average Negative Log Likelihood for different regularization splits

0.05	-11.34045
0.15	-11.34265
0.25	-11.34414
0.35	-11.34435
0.45	-11.34434
0.55	-11.34389
0.65	-11.34323
0.75	-11.34243
0.85	-11.34129
0.95	-11.33904

Bibliography

[1] Sahil Ali. The future of indian electricity demand, Oct 2018.

[2] Himanshu A Amarawickrama and Lester C Hunt. Electricity demand for sri lanka: a time series analysis. *Energy*, 33(5):724–739, 2008.

[3] Nima Amjady. Day-ahead price forecasting of electricity markets by a new fuzzy neural network. *IEEE Transactions on power systems*, 21(2):887–896, 2006.

[4] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[5] A Azadeh, SF Ghaderi, and S Sohrabkhani. A simulated-based neural network algorithm for forecasting electrical energy consumption in iran. *Energy Policy*, 36(7):2637–2644, 2008.

[6] M Barbar and M Alsup. Electricity demand forecasting of india. *working paper*.

[7] Vincenzo Bianco, Oronzio Manca, and Sergio Nardini. Electricity consumption forecasting in italy using linear regression models. *Energy*, 34(9):1413–1421, 2009.

[8] Vitor Cerqueira, Luis Torgo, and Igor Mozetic. Evaluating time series forecasting models: An empirical study on performance estimation methods. *arXiv preprint arXiv:1905.11744*, 2019.

[9] Delson Chikobvu and Caston Sigauke. Regression-sarima modelling of daily peak electricity demand in south africa. *Journal of Energy in Southern Africa*, 23(3):23–30, 2012.

[10] Zafer Dilaver and Lester C Hunt. Industrial electricity demand for turkey: a structural time series analysis. *Energy Economics*, 33(3):426–436, 2011.

[11] Fuat Egelioglu, AA Mohamad, and H Guven. Economic variables and electricity consumption in northern cyprus. *Energy*, 26(4):355–362, 2001.

[12] Niematallah Elamin and Mototsugu Fukushige. Modeling and forecasting hourly electricity demand by sarimax with interactions. *Energy*, 165:257–268, 2018.

[13] Che-Chiang Hsu and Chia-Yon Chen. Applications of improved grey prediction model for power demand forecasting. *Energy Conversion and management*, 44(14):2241–2249, 2003.

[14] IEA. The future of cooling – opportunities for energy efficient air conditioning, 2018.

[15] Jesse D Jenkins and Nestor A Sepulveda. Enhanced decision support for a changing electricity landscape: The genx configurable electricity resource capacity expansion model, Nov 2017.

[16] Ujjwal Kumar and VK Jain. Time series models (grey-markov, grey model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in india. *Energy*, 35(4):1709–1716, 2010.

[17] Yi-Shian Lee and Lee-Ing Tong. Forecasting energy consumption using a grey model improved by incorporating genetic programming. *Energy conversion and Management*, 52(1):147–152, 2011.

[18] Forest Climate Change Government of India Ministry of Environment. India cooling action plan, 2019.

[19] Harun Kemal Ozturk and Halim Ceylan. Forecasting total and industrial sector electricity demand based on genetic algorithm approach: Turkey case study. *International journal of energy research*, 29(9):829–840, 2005.

[20] Srinivasa Rao Rallapalli and Sajal Ghosh. Forecasting monthly peak demand of electricity in india—a critique. *Energy policy*, 45:516–520, 2012.

[21] Ivan Rudnick. Decarbonizing the indian power sector by 2037: Evaluating different pathways that meet long-term emissions targets. Sept 2019.

[22] SKCTS Saravanan, S Kannan, and C Thangaraj. India's electricity demand forecast using regression analysis and artificial neural networks based on principal components. *ICTACT Journal on soft computing*, 2(4):365–70, 2012.

[23] Arunesh Kumar Singh, S Khatoon Ibraheem, Md Muazzam, and DK Chaturvedi. An overview of electricity demand forecasting techniques. *Network and Complex Systems*, 3(3):38–48, 2013.

[24] Thomas Spencer and Aayushi Awasthy. Analysing and projecting indian electricity demand to 2030.

[25] Agostino Tarsitano and Ilaria L Amerise. Short-term load forecasting using a two-stage sarimax model. *Energy*, 133:108–114, 2017.

[26] Juan David Velásquez Henao, VIVIANA RUEDA MEJIA, and Carlos Jaime Franco Cardona. Electricity demand forecasting using a sarima-multiplicative single neuron hybrid model. *Dyna*, 80(180):4–8, 2013.

[27] Baran Yildiz, Jose I Bilbao, and Alistair B Sproul. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, 73:1104–1122, 2017.