Statistical Learning Models for Estimating Retail Tariffs of Electricity in the United States

by Alaeddine Mokri

B.S. in Mechanical Engineering, May 2009, University of Tlemcen
M.S. in Materials Science and Engineering, May 2011, Khalifa University

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Engineering

August 31, 2022

Praxis directed by

Joseph P. Blackford, DEng
Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Alaeddine Mokri has passed the Final Examination for the degree of Doctor of Engineering as of August 02, 2022. This is the final and approved form of the Praxis.

Statistical Learning Models for Estimating Retail Tariffs of Electricity in the United States

Alaeddine Mokri

Praxis Research Committee:

Joseph P. Blackford, Professorial Lecturer of Engineering Management and Systems Engineering, Praxis Director

Amir Etemadi, Associate Professor of Engineering and Applied Science, Committee Member

Timothy Blackburn, Professorial Lecturer of Engineering Management and Systems Engineering, Committee Member

© Copyright 2022 by Alaeddine Mokri
All rights reserved

Dedication

This work is dedicated in its entirety to my wife. Without Asma's unwavering support and dedication, this work would not have been possible.

My time at the university has been an extraordinary learning journey in the areas of applied mathematics and management. For that, I would like to express an immense amount of gratitude to my parents, Ghouti and Chafia, who gave me my earliest math and management lessons, and made every sacrifice to see me on this path. I am equally grateful to my siblings for their unconditional support.

To you all, you have given me what I will never be able to repay.

Acknowledgements

My heartfelt gratitude goes to my colleagues in this program and the entire GW community, especially my doctoral advisor. This praxis and the courses I took with J. P. Blackford have only expanded my understanding and appreciation for quantitative methods.

Prior to enrolling at GW, I have had countless conversations with Dr Cezar Ionescu who exposed me to the rigors of computation and the joy of computational research. Those teachings benefited me greatly during my time at SEAS.

I am deeply grateful for the support of my colleagues and managers at Shell: Ryan Hanley, Tamas Kerekjarto, and Jayaradha Natarajan.

Finally, I would like to thank the entire staff at GW and SEAS for creating this wonderful learning experience that has reshaped me in ways I never imagined.

Abstract of Praxis

Statistical Learning Models for Estimating Retail Tariffs of Electricity in the United States

Retail electricity tariffs represent how electric utilities charge their customers for the electricity they consume. They consist of several types of monetary charges: consumption-based, fixed, or based on minimum and maximum demand. This data is necessary for performing economic analysis on clean-energy projects (e.g., rooftop solar energy, home energy storage, etc.) In the United States, this data is scattered across thousands of electric utility websites. To mitigate this problem, in 2012, the US National Renewable Energy Laboratory launched a crowdsourcing website for collecting this data, organizing it, and making it available to the public for free. Despite this fruitful effort, the database includes outdated or no tariffs in thousands of jurisdictions. This makes it challenging to run economic analysis for clean-energy projects in those areas.

This praxis explores a variety of statistical learning methods to estimate missing tariffs based on available predictors, such as electricity prices in neighboring jurisdictions, the type and size of electric utility, the average price of electricity in the state, etc. To this end, several statistical inference approaches are considered: some that require ancillary data (deep neural networks, k-nearest neighbors, decision trees, linear regression, support vector machines, areal interpolation) and others that do not require any additional data beyond known tariffs in neighboring jurisdictions (inverse distance weighting, ordinary Kriging, average of the n nearest geographical neighbors, average within a radius). Eleven models are constructed to estimate the different type of charges: fixed charges and energy

charges for residential, commercial, and industrial customers while considering both time-dependent and time-independent rate structures.

Specifically, this praxis makes three key practical contributions. First, it studies the literature to determine which factors impact the different charges in electricity tariffs. Second, it determines how various inference approaches perform for estimating tariff charges anywhere in the US with a commentary on their advantages and limitations. Third, it proposes a probabilistic formulation to describe the inference error.

Table of Contents

Dedication ..... **iv**

Acknowledgements ..... **v**

Abstract of Praxis ..... **vi**

List of Figures ..... **xii**

List of Tables ..... **xiv**

List of Acronyms ..... **xvi**

Chapter 1 — Introduction 1

1.1 Background 1

1.2 Problem Statement 5

1.3 Thesis Statement 6

1.4 Research Questions and Hypotheses 6

1.5 Organization of Praxis 7

Chapter 2 — A Literature Review 9

2.1 Introduction 9

2.2 Inference Methods 9

2.2.1 K-Nearest Neighbors (KNN) 9

2.2.2 Decision Trees (DT) 10

2.2.3 Support Vector Machines (SVM) 10

2.2.4 Linear Regression (LR) 11

2.2.5 Artificial Neural Networks and other Machine Learning Methods 12

2.2.6 Spatial Interpolation Using Areal Features 14

2.2.7 Image Completion and Denoising Methods 14

2.2.8 Combined Methods 14
2.3 Predictors of Retail Electricity Prices 15
2.3.1 Market Deregulation 15
2.3.2 Renewable Energy Support Costs 16
2.3.3 Renewable Energy Generation 16
2.3.4 Fuel Prices 17
2.3.5 Utility Ownership 17
2.3.6 Meter Charges 18
2.4 Predictors of Meter Charges 18
2.4.1 Number of Customers 18
2.4.2 Local Power Generation 18
2.5 Summary and Conclusion 19
Chapter 3 — Methodology 20
3.1 Introduction 20
3.2 Data Collection and Cleaning 21
3.2.1 Retail Tariffs 21
3.2.2 Average Electricity Prices with Corresponding Zip Codes, Utility Unique Identifier, Ownership Type, State and Sector 23
3.2.3 State Electricity Data 25
3.2.4 Zip Codes, Geolocation and State 26
3.2.5 Extracting Charges from Tariffs 27
3.2.6 Identifying Neighboring Zip Codes 27
3.3 Outlier Analysis 28

3.3.1 Univariate Methods 28
3.3.2 Multivariate Methods 28
3.4 Model Inputs and Outputs 28
3.5 Model Performance and Validation 31
3.6 Inference Models 31
3.6.1 K-Nearest Neighbors (KNN) 32
3.6.2 Decision Trees (DT) 32
3.6.3 Linear Regression (LR) 32
3.6.4 Support Vector Machines (SVM) 33
3.6.5 Artificial Neural Networks (ANN) 33
3.6.6 Areal Interpolation 34
3.6.7 Point Interpolation 34
3.6.8 Inverse Distance Weighting 35
3.6.9 Ordinary Kriging 36
3.7 Finding the Closest Error Distributions 36
3.8 Summary 37
Chapter 4 — Results 38
4.1 Introduction 38
4.2 Outlier Detection 38
4.2.1 Univariate Outlier Detection 38
4.2.2 Multivariate Outlier Detection 38
4.2.3 Outlier Detection Analysis 40
4.3 Modeling 41

4.3.1 K-Nearest Neighbors (KNN) 41
4.3.2 Decision Trees 44
4.3.3 Support Vector Machines (SVM) 48
4.3.4 Linear Regression (LR) 49
4.3.5 Artificial Neural Networks (ANN) 52
4.3.6 Areal Interpolation 53
4.3.7 Point Interpolation 56
4.3.8 Ordinary Kriging 60
4.3.9 Analyzing Model Performance Results 62
4.4 Summary 68
Chapter 5 — Discussion and Conclusions 69
5.1 Discussion 69
5.2 Contributions to Body of Knowledge 71
5.3 Recommendations for Future Research 71
References 74
Appendix A 84

List of Figures

Figure 1-1. Zip codes with tariff data and their corresponding energy charges. 3

Figure 1-2. Zip codes with no tariff data. 3

Figure 1-3. The number of added and expired tariffs per year. 5

Figure 3-1. A scatter plot of energy and meter charges. 23

Figure 4-1. Histogram plots of energy and meter charges for all sectors. 39

Figure 4-2. Scatter plots of energy and meter charges for all sectors after outliers are removed. 40

Figure 4-3. Histogram plots of energy and meter charges without outliers. 41

Figure 4-4. MAE values for the K-Nearest Neighbors model with different values of hyper-parameter k. 42

Figure 4-5. Histograms of estimation errors for the K-Nearest Neighbors model. 43

Figure 4-6. Variable Importance Analysis results for the KNN model. 44

Figure 4-7. Histograms of estimation errors for the decision trees models. 46

Figure 4-8. The decision tree for residential energy charges. 46

Figure 4-9. The decision tree for commercial energy charges. 47

Figure 4-10. The decision tree for industrial energy charges. 47

Figure 4-11. The decision tree for residential meter charges. 47

Figure 4-12. The decision tree for commercial meter charges. 48

Figure 4-13. The decision tree for industrial meter charges. 48

Figure 4-14. MAE values for the SVM model with different kernels. 49

Figure 4-15. Histograms of estimation errors for the SVM model. 50

Figure 4-16. MAE values for the linear regression model with different regressors. 51
Figure 4-17. Results of Variable Importance Analysis for the Linear Regression model. 51
Figure 4-18. Histograms of estimation errors for the linear regression models. 53
Figure 4-19. Histograms of estimation errors for the artificial neural networks' models. 54
Figure 4-20. Histograms of estimation errors for the areal interpolation model. 55
Figure 4-21. MAE values for the point interpolation model. 56
Figure 4-22. Error distribution for the point interpolation model with the number of neighbors set to 5. 57
Figure 4-23. MAE values for the point interpolation model. 58
Figure 4-24. Error distribution for the point interpolation model with a radius of 15..... 59
Figure 4-25. MAE values for the IDW model with different parameter values. 60
Figure 4-26. Error distribution values for the IDW model with a parameter rho of 2.5... 61
Figure 4-27. Statistical distributions that are close to the distribution of the error in estimating energy charges with the 5 nearest geographic neighbors (NGN)..... 65
Figure 4-28. Statistical distributions that are close to the distribution of the error in estimating commercial meter charges with the 5 nearest geographic neighbors (NGN). 67

List of Tables

Table 2-1. Inference methods and their R-squared. 13

Table 3-1. Types of data issues and their counts. 22

Table 3-2. Ancillary data used as predictor variables with sample values. 26

Table 4-1. Results of the multivariate outlier detection method. 39

Table 4-2. Statistical summary of the errors for the K-Nearest Neighbors (k=2) model. 42

Table 4-3. Error values for the Decision Trees model with different split functions. 45

Table 4-4. Statistical summary of the error for DT model. 45

Table 4-5. Performance of the SVM model with different kernels. 48

Table 4-6. Statistical summary of the error for the SVM model. 49

Table 4-7. Statistical summary of the error for the linear regression model. 52

Table 4-8. Optimal ANN designs with the corresponding MAPE and MAE values. 52

Table 4-9. Estimation errors of the areal interpolation models. 54

Table 4-10. Statistical summary of the error for the areal interpolation model. 55

Table 4-11. Estimation errors of the point interpolation models with the number of neighbors set to 5. 56

Table 4-12. Estimation errors of the point interpolation models with the radius of 15. ... 58

Table 4-13. Estimation errors of the IDW model with a parameter rho of 2.5. 60

Table 4-14. Optimal variograms with the corresponding error values. 61

Table 4-15. Estimation errors for residential energy charges. 62

Table 4-16. Estimation errors for commercial energy charges. 63

Table 4-17. Estimation errors for industrial energy charges. 64

Table 4-18. Estimation errors for residential meter charges. 65

Table 4-19. Estimation errors for commercial meter charges. 66

Table 4-20.Estimation errors for industrial meter charges. 66

Table A-1. Mix of electricity generation in all US states in 2020. 84

Table A-2. The number of electricity customers in all US states in 2020. 86

Table A-3. Average price of electricity to customers for all US states in $/kwh. 88

Table A-4. Average electricity revenue per customer for all US states in US dollars..... 90

List of Acronyms

AI	Areal Interpolation
ANN	Artificial Neural Networks
DT	Decision Trees
EIA	Energy Information Administration
IDW	Inverse Distance Weighting
KNN	K-Nearest Neighbors
KWH	Kilowatt Hour
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
NGN	Nearest Geographical Neighbors
NNR	Nearest Neighbors within a Radius
NREL	National Renewable Energy Laboratory
OK	Ordinary Kriging
RMSE	Root Mean Squared Error
SVM	Support Vector Machines
URDB	Utility Rates Database
USURDB	United States Utility Rates Database

Chapter 1 — Introduction

1.1 Background

The American Society for Engineering Management (ASEM) defines Engineering Management, as "the art and science of planning, organizing, allocating resources, and directing and controlling activities that have a technological or systems component." (ASEM, 2022). Performing these tasks successfully often requires data which may not be always available. This is particularly true in the clean-energy industry in the United States where retail prices of electricity are not always accessible. This praxis is concerned primarily with addressing this problem of missing retail tariffs of electricity in the United States.

These prices are what customers pay for electricity and what appears on their electricity bill. Before one can assess the effectiveness of clean-energy measures (e.g., installing a rooftop solar system) in reducing the bill amount, electricity rates are necessary. Equations 1 and 2 show how the different tariff charges are considered to calculate the electricity bill where Equation 1 uses time-dependent energy charges and Equation 2 uses time-independent energy charges.

$$Bill(start, end) = Fixed\ Charges + Demand\ Charges + \sum_{t=start}^{t=end} Consumption(t)\ .\ Energy\ Charges(t) \quad (1)$$

$$Bill(start, end) = Fixed\ Charges + Demand\ Charges + Energy\ Charges\ .\ \sum_{t=start}^{t=end} Consumption(t) \quad (2)$$

Studies have shown that these rates influence the overall economics of clean-energy projects. An economic analysis on similar medium-sized office buildings with roof-top solar in 25 American cities reported bill savings between 7% and 25% under different

conditions of climate and electricity rates (Ong et al., 2010). In a 2009 study, the break-even cost, the point where the cost of power from solar equals the cost of power from the electric utility, for residential solar in the largest 1,000 utilities in the United States differed by a factor of 10 despite a much smaller variation in solar irradiance (Denholm et al., 2009).

This problem is particularly important because the United States is the second largest emitter of CO2 in the world. Effective decarbonization of the US energy market will have a major impact on the reduction of greenhouse emissions globally. Furthermore, 65% of greenhouse gas emissions in the US are the result of burning fossil fuels for transportation, electricity, and heating and cooling (US Environmental Protection Agency, 2022). To ensure the decarbonization of these sectors is effective, the economics of prospective clean-energy initiatives ought to be accurate. To perform any economic assessment of this kind, one must know the applicable retail electricity prices at the location where the project is considered. Without this data, the project developer will not be able to estimate the return on investment for these interventions; and consequently, miss valuable investment and decarbonization opportunities.

Retail electricity tariffs are on the websites of load serving entities (LSEs), commonly called electric utilities. Based on the tariffs data published by the US National Renewable Energy Laboratory (NREL), there are 2,841 electric utilities in the United States as of 2019 (US Department of Energy, 2019). These tariffs also have different structures and applicability rules depending on geography, type of customer (industrial, residential, or commercial), peak demand, season, time of day, day of week, season, etc. These tariffs also change as electric utilities respond to changes in fuel costs, regulations, etc.

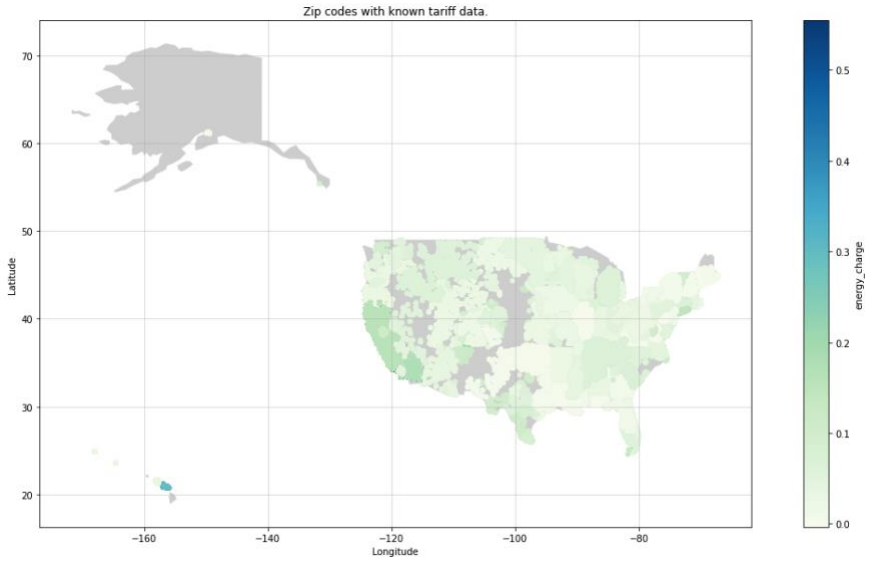

Figure 1-1. Zip codes with tariff data and their corresponding energy charges.

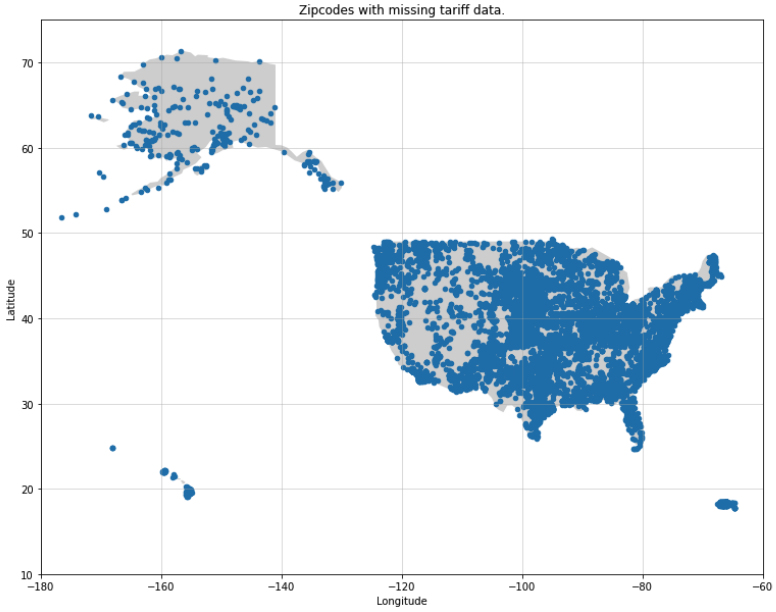

Figure 1-2. Zip codes with no tariff data.

Additionally, these prices change as electric utilities respond to a variety of forces (e.g., changing fuel prices, inflation, new regulations, etc.) A survey of 31 utilities in North America has shown that utilities respond to these forces differently based on several factors (e.g., ownership, size, etc.) (Langlois-Bertrand et al., 2018). This means that the tariffs in the database are not always up to date. In fact, as it will be discussed in the next chapter, 39% of tariffs in the database expired in 2020 or earlier. All these factors make this data challenging to work with by clean-energy developers.

To address this problem, in 2012, NREL launched the US Utility Rate Database (https://apps.openei.org/USURDB/), a crowdsourcing website where internet users can submit electricity rates. A team at NREL reviews these rates and makes them publicly available in both a human-readable and computer-readable format, for free. This team also collects data when needed. At the time of publication of the dataset, tariffs were missing for about 15% of customers (Ong et al., 2012). Upon close examination of the data, no valid tariff data was found for several zip codes (e.g., 99359, 99701, 57025, etc.) The maps in Figure 1-1 and 1-2 show the areas where tariffs are available and unavailable, respectively.

To ensure this dataset remains up-to-date, records continue to be updated, but manually. This is both labor intensive and not scalable to cover every active tariff in the United States. The engineers leading this effort were interviewed as part of this praxis and they shared that updating a single tariff can take anywhere from 1 hour to several days and updating all tariffs from a single utility takes about eight hours. Because this process is time consuming, the team focuses on the largest 150 utilities and update their tariffs regularly. The unreliability of this method is reflected in the data update patterns. Upon the

examination of how frequently tariffs are added and expired (see Figure 1-3), the net number of tariffs added every year is much smaller in recent years than when the database was first released. The team at NREL mentioned that attempts to automate the process were unsuccessful.

From the perspective of the end user, alternative options may include using a commercial vendor which is both costly and unable to guarantee full data coverage.

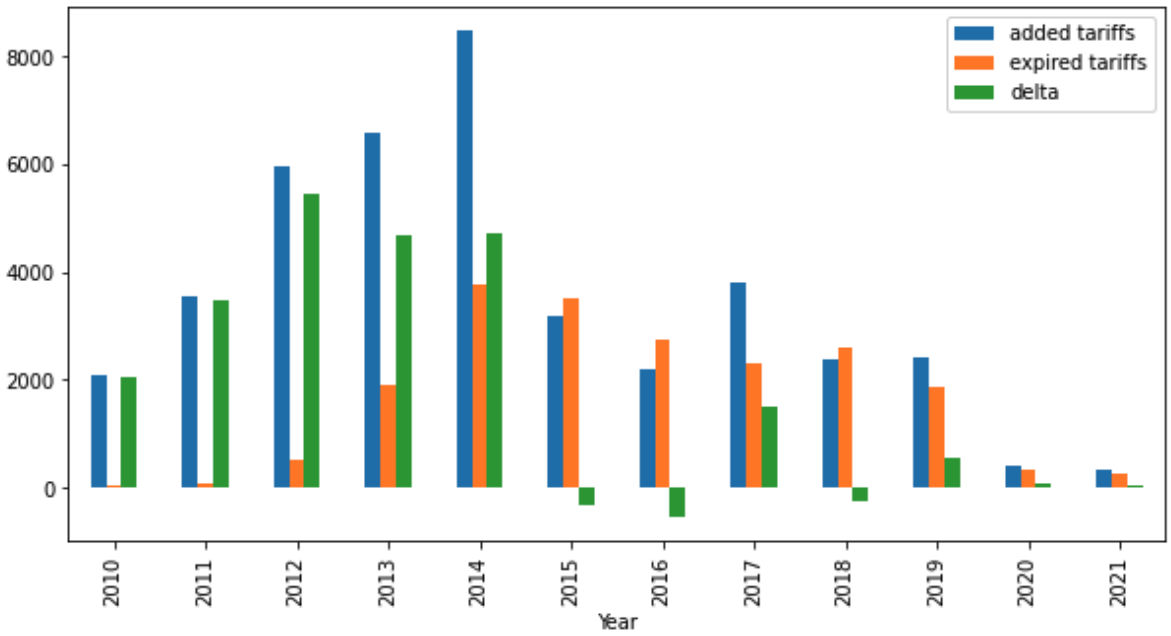

Figure 1-3. The number of added and expired tariffs per year.

Considering the above discussion on the inefficacy of existing solutions, and by drawing inspiration from solutions to similar problems (e.g., estimating the potential sale value of real-estate properties, estimating the composition of soil in inaccessible areas, etc.), the aim of this research is to develop a data inference model to address the gaps in tariff data.

1.2 Problem Statement

NREL's Utility Rate Database (URDB), an important source of retail tariffs for developing clean-energy projects, doesn't have full geographical coverage in the United States. Thousands of jurisdictions have no up-to-date tariffs and there are no effective alternatives to address this data gap.

1.3 Thesis Statement

Statistical inference models have proven very successful in real-estate where property values can be estimated based on neighboring property values (Comber et al, 2019; Tchuente and Nyawa, 2021). Such statistical models proved successful in environmental studies as well, where properties of certain geographies (e.g., soil composition, solar irradiance, rain fall patterns, etc.) could be inferred based on those same properties in neighboring regions (Rehman and Ghori, 2000; Cheng et al., 2008, Li et al., 2011; Kuzyakova, 2001). Another class of analogous problems can also be found in image reconstruction where missing parts of an image can be inferred based on what is already represented (Besag, 1986) or image de-noising where noise pixels are detected based on the color of neighboring pixels (Lu and Jiang, 2001). The problem at hand is analogous to this class of problems.

From these analogies, statistical inference models or image reconstruction models can estimate unknown retail electricity prices based on known predictors (e.g., electricity tariffs in neighboring jurisdictions, characteristics of the local electricity provider, etc.)

1.4 Research Questions and Hypotheses

The research questions addressed in this praxis are the following:

- **RQ1:** What predictors determine local electricity prices?

- **RQ2:** What statistical methods perform best for inferring retail electricity prices in the United States?
- **RQ3:** Is there a mathematical formulation to model the inference error?

These questions are approached with the following hypothesis:

- **RH1:** Electricity charges are correlated with 1) rates available in neighboring jurisdictions, 2) prices provided by electric utilities with similar profiles (investor owned or cooperative or public, number of customers, energy sources, etc.,) 3) the state's energy profile (e.g., energy mix, average price of electricity, average electricity revenue per customer, etc.)
- **RH2:** All missing tariffs in the National Database of Utility Rates can be inferred statistically by using open data within reasonable error ranges that depend on the type of the charge.
- **RH3:** The inference error can be modeled mathematically.

1.5 Organization of Praxis

In the present chapter, the scope of the praxis was provided. Chapter 2 offers a review of the literature on how various data inference methods have been used to tackle similar problems in various areas: real-estate, environmental studies, image reconstruction and denoising, etc. It also reviews the literature on which factors influence electricity charges. The methods and predictors explored in the literature review are leveraged in Chapter 3, the methodology chapter, where the datasets are described, the data cleaning steps are outlined, the outliers detected and removed, the model validation procedure described, and the different models constructed and run to answer the research questions listed above. Chapter 4 looks at the performance of the models individually and compares

their error figures. A probabilistic description of the estimation error is also proposed. Chapter 5 summarizes the factors that influence tariff charges, comments on which inference methods perform best on which type of charges. Recommendations for future work are also made.

Chapter 2 — A Literature Review

2.1 Introduction

Inference methods have been applied successfully to address gaps in geospatial datasets in a variety of disciplines: real-estate (Ho et al., 2021), environmental studies (Li et al., 2011), social studies (Dorji et al., 2019), image reconstruction (Besag, 1986), etc. Although this literature review did not find prior work on addressing the gaps in electricity prices, the problems on which these methods have been applied are analogous in nature: given known information about a certain geographic location, how can we infer missing information in a different geographic location. Therefore, the scope of this chapter is the different data inference methods that can be applied to address gaps in geospatial datasets.

Furthermore, to answer the first research question on which factors influence retail electricity prices and determine the predictors to construct the model, this chapter covers research on what factors were shown to influence retail electricity prices globally.

2.2 Inference Methods

2.2.1 K-Nearest Neighbors (KNN)

Suchenwirth et al. (2014) used KNN to estimate the concentration of organic carbon from soil and vegetation in the Danube Floodplane by using satellite data only, and then adding ancillary data such as elevation, ground water level and historic measurements of carbon stocks. The estimates were validated with ground measurements and compared to estimates from a Self-Organizing Maps (SOM) model and those in the literature. The RMSE values for both models were close and comparable to those reported in several studies. For the KNN model, the use of auxiliary data led to a reduction in RMSE. The

number of neighbors from one to thirty were considered and five was used because RMSE did not improve noticeably when higher values were used.

In a separate study, Jalali et al. (2013) performed a benchmarking analysis of KNN against several methods (co-Kriging, simple Kriging, universal Kriging, ordinary Kriging, inverse distance weighting, radial basis function, local polynomial, and global polynomial) to estimate soil saturated hydraulic conductivity in Bojnurd in Iran. The optimal number of neighbors for the KNN was 10 and it led to the highest Pearson's correlation coefficient and the RMSE values were close to those obtained from other methods.

2.2.2 Decision Trees (DT)

Balk and Elder (2000) used decision trees to estimate the distribution of snow in the Rocky Mountains National Park by using the following independent variables: solar radiation, elevation, slope, and vegetation cover type. The results showed that the R^2 values improved from 54-65% to 60-85% when the estimates from the DT model were combined with the output of a Kriging interpolation model. The combined model was setup for the DT to estimate large-scale variations in snow depth and the Kriging model to estimate small-scale variations

In a separate paper, Alaboz et al. (2021) used DT for estimating soil erosion parameters from soil properties in the Isparta Province in Turkey. Three algorithms were considered: Classification and Regression Trees (CART), Chi-square Automatic Interaction Detection (CHAID), Exhaustive CHAID. The CART algorithm was shown to perform best based on prediction-percentage accuracy values with an R^2 value of 71%.

2.2.3 Support Vector Machines (SVM)

Sitharam et al. (2008) used SVM to estimate rock depth in Bangalore in India. Three kernel functions were considered: radial basis function (RBF), polynomial and spline. The spline kernel led to the highest R value of 0.835. The SVM results were compared with those obtained from an Ordinary Kriging (OK) model and an Artificial Neural Network (ANN) model. The authors did not report the error values but provided five groundtruth values with their estimates. Based on those five points, the MAPEs for the ANN, SVM and OK were 0.08%, 0.23% and 0.8% respectively.

In a separate research, Wohlberg at al. used SVM to study geologic properties of the subsurface environment where several kernels were considered: polynomial, exponential radial basis, Gaussian radial basis, and sigmoid (Wohlberg at al., 2005). The Gaussian radial basis kernel was found to perform best. A Kriging model was constructed, and its results were compared with the SVM model at different data sampling densities. The SVM model consistently led to lower errors than the Kriging model particularly at low data sampling densities (0.25%) where the difference in error rates was the largest.

2.2.4 Linear Regression (LR)

Linear regression has been particularly popular in the estimation of the value of real-estate. Chen was able to derive a linear function to estimate property value in the city of Zhaoqing in China (Chen, 2022). The function takes as arguments: income of urban residents, rate of employment among residents, land price, the investment in real estate development, population, population density, and the proportion of urban population. R^2 values as high as 0.99 were achieved.

Liu also used linear regression to estimate the value of real-estate properties in China but by using different predictors (monthly income, disposable income per-capita,

housing expenditure, and surface area) (Liu, 2022). The author reported a maximum prediction error of 8%.

2.2.5 Artificial Neural Networks and other Machine Learning Methods

Given the variety of machine learning methods and the different ways that they can be implemented (e.g., hyper-parameters, architecture of neural networks, dataset size, etc.) different studies focused on comparing these methods to determine which one performed bast and they all reported different levels of success. One key study on the effectiveness of machine learning methods for geo-spatial inference is the one conducted by Li et al. (2011), where 23 different algorithms were considered to study mud content along Australia's southern coast.

Machine learning methods were considered by Gholami et al. to simulate water quality in Iran (self-organizing maps, artificial neural networks, and co-active neuro-fuzzy inference systems) and they reported R^2 values ranging between 0.73 and 0.89 (Gholami, 2022). In a separate study that explored groundwater drawdown in the same area reported an R^2 of 0.81 by using artificial neural networks (Gholami and Sahour, 2022).

In addition to the generic ML models adapted to geo-statistics problems, one method that has been used heavily on geo-spatial data is geographically weighted regression. Shahneh et al. considered several variations of this method on different types of datasets and reported R^2 values as high as 0.94 (Shahneh et al., 2021). The results of these ML methods are summarized in Table 2-1. Li et al. results are not included in the table because no R^2 results were reported.

Table 2-1. Inference methods and their R-squared.

Reference	Independent Variables	Dependent Variable	Method	R^2
Gholami et al., 2022	distance from industries, groundwater depth, and transmissivity of aquifer formations	groundwater quality index	Self-Organizing Maps	0.8
Gholami et al., 2022	distance from industries, groundwater depth, and transmissivity of aquifer formations	groundwater quality index	Artificial Neural Networks	0.73
Gholami et al., 2022	distance from industries, groundwater depth, and transmissivity of aquifer formations	groundwater quality index	Co-active Neuro-fuzzy Inference Systems	0.89
Gholami and Sahour, 2022	groundwater depth, annual precipitation, annual evaporation, the transmissivity of the aquifer formation, elevation, distance from the sea, distance from water sources (recharge), population density, and groundwater extraction in the influence radius of each well (1000 m)	groundwater drawdown	Artificial Neural Network	0.81
Shahneh et al., 2021	Room type, minimum nights that a guest can stay, number of reviews, reviews per month, amount of listing per host, and the availability of the unit along with the location	AirBnB listing price	Geographically Weighted Regression	0.76

Shahneh et al., 2021	Number of bedrooms, number of bathrooms, square footage of the apartment's interior living space, number of floors, house condition, and grade, the year built, and the location	House price	Geographically Weighted Regression	0.94

2.2.6 Spatial Interpolation Using Areal Features

These methods consist of taking a property of a large area (e.g., average price of electricity in a state) and estimating the same property within that area (e.g., average price of electricity in a jurisdiction in the state). Comber et al. (2019) reviewed these methods and grouped them into those that use ancillary data to guide the interpolation and those that don't.

2.2.7 Image Completion and Denoising Methods

These methods look at pixels of images with missing parts and attempt to fill the colors for the missing pixels based on the colors of pixels in their proximity in addition to information about the pixels (e.g., color distribution, color ordering, etc.) (Besag, 1986).

Lu and Jiang (2001) demonstrated how this technique can be extended to restoring corrupted images and found that it performed better than other denoising techniques at shape transitions.

2.2.8 Combined Methods

Li et al. (2011) studied different combinations of methods and found that combining the machine leaning method random forest with inverse distance squared and Ordinary Kriging performed better than any of the 23 methods evaluated.

Sekulic et al. (2020) proposed a method that combines random forests with traditional geo-statistic methods and found that it performed better than any of the geo-statistics methods for estimating precipitation and temperature values in Spain and Croatia, respectively.

2.3 Predictors of Retail Electricity Prices

2.3.1 Market Deregulation

Although deregulation of the power markets was intended to increase competition and reduce prices, several studies reported no evidence on the effect of deregulation on retail power markets, or even an increase in prices (Moreno et al., 2012; Taber et al., 2005; Razeghi et al., 2017; Zarnikau et al., 2006; MacKay and Mercadal, 2022). A literature review on the effect of deregulation on prices globally showed inconclusive results between the markets studied, with several studies suggesting that the effect on prices changes over time: decreasing in the short term and increasing in the long term (Necoechea-Porras et al., 2021). In the US market, MacKay and Mercadel reported an increase in prices because of an increase in markups (MacKay and Mercadal, 2022).

Taber et al. examined this question by studying the evolution of retail prices in regulated and deregulated markets in the United States and found no evidence to support the expectation that deregulation would lower prices (Taber et al., 2005). The study found that even if most customers in deregulated markets saw declines in prices, similar trends

were seen in regulated states. Furthermore, these price differences remain even after considering the effects of climate, generation mix and fuel prices (Taber et al., 2005).

More geography-focused studies conducted in the state of California (Razeghi et al., 2017), the state of Texas (Zarnikau et al., 2006), and 27 countries in Europe (Moreno et al., 2012) reported similar findings, where prices have declined historically, but there was no evidence that those cost declines were driven by deregulation policies and market restructuring.

2.3.2 Renewable Energy Support Costs

A thorough review of the factors that influence retail electricity prices in 27 European countries showed that support prices for promoting renewable power generation were the most significant contributing factor to increasing prices (Moreno et al., 2012). A similar study, which focused on six European countries, reported that taxes and levies associated with the promotion of renewable power generation led to increased retail prices in all six countries (Ktena et al., 2019). In a similar study focused on the effect of support costs in Europe, the authors reported different impacts on different customer types where industrial customers saw higher price increases than residential customers (Trujillo-Baute et al., 2018).

2.3.3 Renewable Energy Generation

The effect of renewable electricity share on retail electricity prices was studied in 34 OECD countries and the results showed that the influence is positive and statistically significant. Countries with higher renewable energy mix were associated with higher retail electricity prices (Oosthuizen et al., 2022).

In a separate study which looked specifically on the impact of renewable energy generation on price volatility, it was reported that Germany and Denmark specifically responded differently partly because the shares of various energy resources were different (Rintamäki et al., 2017).

2.3.4 Fuel Prices

The dynamic between electricity prices and three fossil fuel prices – coal, natural gas and crude oil – was studied by using annual data in the US between 1960–2007. The study showed a stable long-run relation between retail prices and coal. However, there were insignificant long-run relations between electricity and crude oil and natural gas prices (Mohammadi, 2009).

A similar study conducted in Mexico reported a different conclusion by looking at data between the years 2006 and 2016. The study found a positive correlation between retail prices and the prices of all three fuels (Bernal et al., 2019). This discrepancy can be tracked down to the differences in energy mix and the timelines considered (International Renewable Energy Agency, 2021). The contribution of coal to the US energy mix has declined consistently in the favor of gas since the 1960s. This explains why the impact of gas prices is more pronounced when recent timeframes are considered.

2.3.5 Utility Ownership

A doctoral thesis examined the effect of ownership on rates and reliability of the power grid in the state of Florida where 19 electric utilities operated: seven investor-owned and twelve public (Pazzalia, 2022). The study bucketed the rates in tiers based on monthly energy consumption and found that publicly-owned utilities had statistically significant lower residential rates while investor-owned utilities had statistically significant lower

industrial rates. For the commercial rates, public utilities had lower rates for the lowest tiers (i.e., lowest monthly consumption limit) and investor-owned utilities offered more competitive rates for the highest tiers (i.e., lowest monthly consumption limit.) The study found no statistically significant difference in the measures of reliability between the two types of ownership.

2.3.6 Meter Charges

Deason and Schwartz (2016) studied rate structures in the US and reported that when utilities increase their fixed charges, they also reduce their energy charges to control the overall increase in the electricity bill.

2.4 Predictors of Meter Charges

2.4.1 Number of Customers

Faruqui and Leyshon surveyed 37 utilities in the US to understand how they set meter charges. It was revealed that they run cost-of-service studies that consider multiple costs. The costs fall broadly in two categories: customer-related (e.g., billing, metering, customer support) and distribution-related (e.g., transformers, poles, etc.) The authors identified a trend of increasing meter charges, except that some states like California may have regulations on a maximum fixed charge. The survey showed that the number of customers in the system is a key indicator in setting meter charges.

2.4.2 Local Power Generation

Deason and Schwartz (2016) studied rate structures in the US and reported that several electric utilities raised meter charges as a response to the increase in distributed renewable generation in their network. This is because, when customers generate their electricity locally, the utility's revenue drops. One way to recover the revenue lost is to

increase the fixed charges. Sometimes, these increases were applied to all customers, and in other times, they were applied selectively to customers with local power generation.

2.5 Summary and Conclusion

In this chapter, inference methods used to solve geo-statistical problems like the one at hand were reviewed. The methods ranged between traditional machine learning methods which are adapted to the problems in geo-statistics, pure geo-statistics ones and their combination. The insights learned from those model implementations will be used in Chapter 3.

Predictors of energy and meter charges were also reviewed. It was shown that, contrary to common wisdom, in the power sector, there is no evidence that rates are lower in states with deregulated markets. Other factors were shown to influence energy rates (energy mix, fuel prices, ownership, meter charges, and clean-energy support costs) and meter charges (local power generation, number of cusomers). When considering predictors for building estimation models in the next chapter, these findings will be further explored.

Chapter 3 — Methodology

3.1 Introduction

This chapter starts with describing the datasets collected, the cleaning procedures to remove incomplete rows, the outlier detection methods used, the problem formulation, and the description of the models used. It is worth noting that the modeling work was divided to estimate six dependent variables:

- meter charges in the residential sector (USD/month),
- meter charges in the commercial sector (USD/month),
- meter charges in the industrial sector (USD/month),
- energy charges in the residential sector (USD/kwh),
- energy charges in the commercial sector (USD/kwh),
- energy charges in the industrial sector (USD/kwh).

In this praxis, demand charges are not considered because of several reasons. First, they are not always needed because they are only applicable when the project involves changing the maximum monthly power demand. Second, demand charges vary by the amount of power used and different utilities set those limits independently. That said, those charges needed to be standardized to be able to compare them across all tariffs in the database. This is a heavy enough exercise that could derail the scope of this research. Third, the predictive factors that we discussed in the literature review influence energy charges, but demand charges have their own dynamics and own predictors. Studying those predictive factors could further derail this research. For the combination of these reasons, only meter and energy charges are studies in this praxis.

There are two reasons why dependent variables are considered separately for each sector. First, the value ranges are different by orders of magnitude. The means of meter charges are 15.37 $/month, 316.89 $/month and 2,339.49 $/month for the residential, commercial, and industrial sectors respectively. Second, while energy charges are typically influenced by fuel costs, meter charges are determined by different factors. Meter charges are driven the utility's way to recover costs for their operations and are driven by the size of the utility and its type of ownership among other factors (Willis and Philipson, 2018).

3.2 Data Collection and Cleaning

3.2.1 Retail Tariffs

The raw data has been downloaded in a CSV format from US Utility Rates Database on Feb 17, 2022. The file includes 50,343 tariffs and 669 attributes. These tariffs are provided by 2,841 electric utility and are grouped in the following four sectors: industrial (18%), lighting (13%), commercial (46%), residential (23%).

Upon reviewing the data, many tariffs had issues: expiry date set in the past, a creation date set to a decade ago which means that the tariff may no longer be in use, missing fields, inconsistent units of measurement, etc. Table 3-1 shows the issues observed and the corresponding count of tariffs removed or updated.

After cleaning the data, 5,107 complete tariffs remained. These represent 10.14% of the published tariffs. These tariffs cover all three sectors: commercial (41%), industrial (16%) and residential (29%).

A scatter plot of meter and energy charges is shown in Figure 2-1 and a few outliers are noticed. Before the data can be used, the outliers will be removed.

Table 3-1. Types of data issues and their counts.

Issue	Count	Percent	Comment
Expiry date set to before 01/01/2021	19,738	39.21%	Tariffs expiring in 2021 and 2022 are left in the data because they reflect current rates.
Added before 01/01/2017	37,429	74.35%	Tariffs older than five years are removed because they no longer reflect current prices.
The field “sector” is missing.	7	0.01%	The tariff sector will be used as an independent variable.
The field “fixedchargefirstmeter” is missing.	4,797	9.51%	The meter charges are part of the tariff (i.e. dependent variable).
The field “fixedchargefirstmeter” is set to 0.	48	0.10%	The meter charges should not be 0. There is no reason to believe why this could be the case.
The field “fixedchargeunits” is set to “$/day” instead of “$/month”.	229	0.45%	For consistency, these values are converted to “$/month” by multiplying them by 30.
The fields “energyweekdayschedule” or “energyweekendschedule” are missing.	5,909	11.74%	These fields specify the time periods during which energy charges apply. Without this data, energy charges cannot be linked to the times during which they are applicable.
Field “energyratestructure/period0/tier0rate” is missing or set to 0.	8,050	15.99%	This field represents energy charges which is a dependent variable in the model.
Field “sector” is set to “Lighting”	6,341	12.60%	This research is only concerned with commercial, residential, and industrial rates because they pertain most to the development of clean energy projects.

3.2.2 Average Electricity Prices with Corresponding Zip Codes, Utility Unique Identifier, Ownership Type, State and Sector

This data set consists of 18 CSV files downloaded from the US Energy Information Administration. It includes the zip codes, the corresponding states, the utilities operating in those zip codes, their ownership type as well as average rates for residential, industrial, and commercial customers in those zip codes. This dataset is needed for two reasons. First, it maps electric utilities to zip codes. This data is necessary to associate tariffs with zip codes in the NREL dataset. Second, it provides the average price of electricity for each sector in all zip codes. This data can be used as a predictor because it approaches the energy charges applied in any given zip code. Additionally, the data includes the utilities that operate in any given zip code with their type of ownership

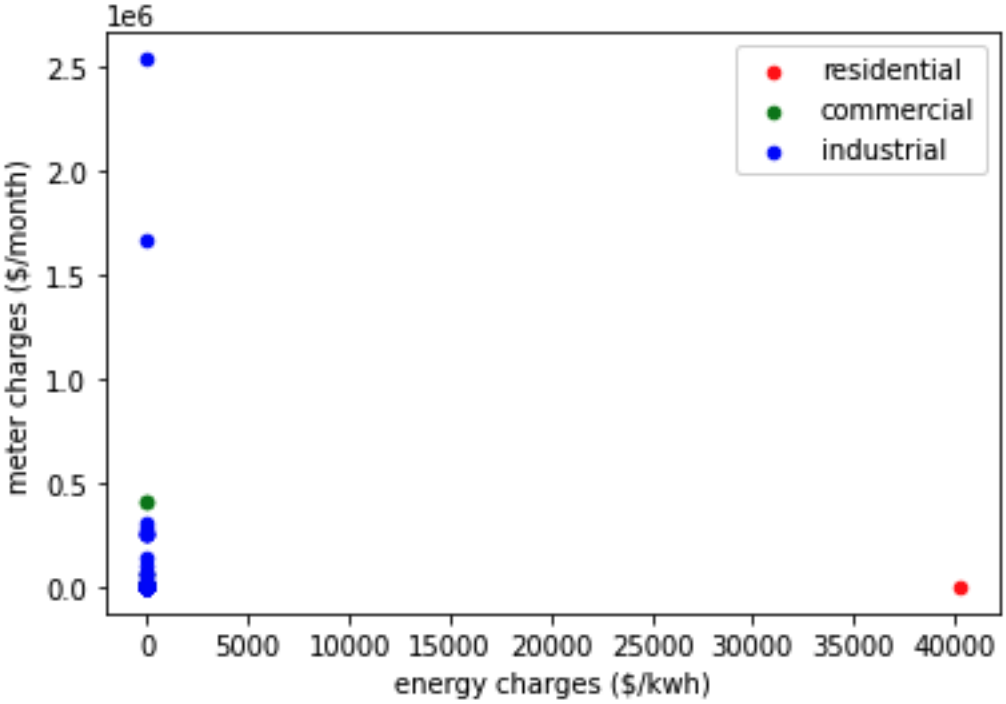

Figure 3-1. A scatter plot of energy and meter charges.

The number of zip codes covered by investor-owned utilities are 52,246 and those covered by non-investor owner (municipal, public, etc.) amount to 36,494. The dataset includes the following columns: zip, utility id, utility name, state, service type, ownership, and the average electricity rates for commercial, industrial, and residential sectors.

This data is merged with the tariff data on the unique identifier of the electric utility. As a result, tariffs could be associated with the zip codes where they are applicable, the average price of electricity, the ownership type of the utility providing the tariff. Table 3-2 shows sample values of these variables in their respective columns: zip, eiaid (unique identifier of the electric utility), comm_rate (average energy charge in the commercial sector), ind_rate (average energy charge in the industrial sector), res_rate (average energy charge in the industrial sector), public (whether the utility is public), investor_owned (whether the utility is investor owned), cooperative (whether the utility is a cooperative).

Utility types were modified from the original data in two steps. In a first step, the number of utility types was reduced from seven (federal, municipal, state, political subdivision, cooperative, investor owned, retail power marketer) to three top-level sectors: public, investor-owned and cooperative. This step was performed because, for the purposes of this exercise, the types federal, municipal, state, and political subdivision all represent a large category that represents public ownership. Similarly, the power marketer category is like ownership by investors, the difference being that power marketers buy the power and do not generate it themselves. For the purposes of this research, those two categories are the same and named conveniently: investor-owned. In a second step, the variables were converted into binary variables because not all models support categorical variables. For

consistency, all categorical variables that do not convey ordering are converted into binary variables.

3.2.3 State Electricity Data

The EIA provides data on average electricity rates in all US states including:

- The energy mix: this is represented in percent values in the columns nuclear, coal, natural gas, petroleum, hydro, geothermal, solar, wind, and other as shown in Table 3-1. Table A-1 in Appendix A contains all the values.
- The number of electricity customers in the state: this is represented in the columns num_cus_res, num_cus_ind and num_cus_com for the number of electricity customers in the residential, industrial, and commercial sectors respectively (see Table 3-1). The unit is customer. Table A-2 in Appendix A contains all the values.
- The average electricity rate for each sector: this is represented in the columns num_cus_res, num_cus_ind and num_cus_com for the average rate of electricity for customers in the residential, industrial, and commercial sectors respectively (see Table 3-1). The unit was originally provided in cents/kwh and it was converted to USD/kwh for consistency. Table A-3 in Appendix A contains all the values.
- The utilities' revenue per customer: this is represented in the columns rev_res, rev_ind and rev_com for average revenues in the residential, industrial and commercial sectors respectively. The unit is USD per customer. Table A-4 in Appendix A contains all the values.

3.2.4 Zip Codes, Geolocation and State

For the inference methods that require the calculation of proximity between zip codes (e.g., inverse distance weighting, Ordinary Kriging, average of n nearest neighbors, etc.), mapping from zip code to their geolocation coordinates is needed. This data was obtained from Stanford Center for Population Health Sciences.

Additionally, because different states have different energy policies and the rates of electricity differ amongst them, this variable is used as predictor as well. To ensure the data can be used in models that do not support categorical variables, the value of the state was converted to binary variables as well.

Table 3-2. Ancillary data used as predictor variables with sample values.

zip	29804	29003	29512	29445	29056
eiaid	162	1170	1566	1613	1613
state	SC	SC	SC	SC	SC
comm_rate	0.117	0.125	0.110	0.121	0.121
ind_rate	0.071	0.113	0	0.054	0.054
res_rate	0.140	0.124	0.130	0.133	0.133
public	0	1	1	0	0
investor_owned	0	0	0	0	0
cooperative	1	0	0	1	1
avg_price_res	0.1278	0.1278	0.1278	0.1278	0.1278
avg_price_com	0.1035	0.1035	0.1035	0.1035	0.1035
avg_price_ind	0.0598	0.0598	0.0598	0.0598	0.0598
num_cus_res	2,377,020	2,377,020	2,377,020	2,377,020	2,377,020
num_cus_com	395,288	395,288	395,288	395,288	395,288
num_cus_ind	3,714	3,714	3,714	3,714	3,714
nuclear	0.558	0.558	0.558	0.558	0.558
coal	0.127	0.127	0.127	0.127	0.127
natural gas	0.248	0.248	0.248	0.248	0.248
petroleum	0.001	0.001	0.001	0.001	0.001
hydro	0.025	0.025	0.025	0.025	0.025
geothermal	0	0	0	0	0

solar	0.018	0.018	0.018	0.018	0.018
wind	0	0	0	0	0
other	0.024	0.024	0.024	0.024	0.024
rev_res	2	2	2	2	2
rev_com	5	5	5	5	5
rev_ind	404	404	404	404	404

3.2.5 Extracting Charges from Tariffs

Tariffs consist of three components: energy charges, demand charges and meter charges. Demand charges are out of the scope for this praxis and are therefore ignored. Meter charges are a single value associated with a tariff. Demand charges might be a single value or multiple values that are applicable either during or outside the weekend period, at specific hours during the 24-hour day, and specific seasons represented by a start and end months.

For meter charges, it is straightforward to map them to the zip codes where they are applicable. For energy charges, 2 x 24 x 12 new columns were created to hold binary values that represent when a certain charge is applicable. For instance, for a given charge, if hour_1 is 1, weekday is 1 and month_1 is 1, that means the charge is applied during the weekday as opposed to the weekend, in the month of January and during the first hour of the day (from midnight to at 1 am local time).

3.2.6 Identifying Neighboring Zip Codes

Some of the inference methods discussed require data about charges available in neighboring zip codes. To identify the nearest neighbors to each zip code, the geocoordinate data is used to calculate the distance in miles between every zip code in the data and every other zip code in the data. Only those that fall within 20 miles are retained. To calculate the distance, Earth is assumed to have a perfect spherical shape.

3.3 Outlier Analysis

3.3.1 Univariate Methods

The meter and energy charges in each sector were assumed to have a known distribution. The mean parameter was moved from the minimum value to the maximum value, and at each iteration, the extreme values from either end of the charges were removed one by one until the remaining data returned a p value higher than 0.001 for the goodness of fit test. This approach detects outliers that make the distribution deviates from what it is supposed to be. It requires knowing the underlying distribution (Aggarwal, 2017).

Pazzalia (2022) used a different approach by assuming the data was normal, and when the test failed, the data was transformed to pass the normality test. Values at either end of the curve were considered outliers if they fell beyond a certain number of standard deviations.

3.3.2 Multivariate Methods

Jasinka and Preweda (2021) looked at detecting outliers in real estate data in Poland where multiple multi-variate methods were considered. The study demonstrated the effectiveness of the Mahalanobis Distance method which is based on calculating the distance between a point and a distribution. For each charge, a Mahalanobis distance is calculated. The distances are supposed to follow a chi distribution. A goodness of fit is then performed for each distance and the charges that fall outside of the distribution with a value lower than 0.001 are eliminated.

3.4 Model Inputs and Outputs

After the data has all the necessary columns, it is split into three datasets: commercial charges, industrial charges, and residential charges. Each of the datasets only

includes the variables that are associated with that sector. For instance, the commercial dataset includes the following columns:

- 54 columns that hold binary values for all 50 states, DC and large territories. The two-letter designation of each is used as the column name (e.g., “AK”, “AL”, etc.) A value of 1 means the zip code is in that state.
- 24 columns that hold binary values for each hour in the day (e.g., “hour_1”, “hour_2”, etc.)
- 12 columns that hold binary values for each month in the year (e.g., “month_1”, “month_2”, etc.)
- Two columns that hold binary values to represent whether the tariff is applicable to weekdays or weekends. These are named “weekday” and “weekend”, respectively.
- A column for latitude and another one for longitude. They are named “Latitude” and “Longitude”, respectively.
- A column named “zip” for storing the zip code where the charge applies.
- A column named “neighbors” for storing the list of zip codes located within a 20 mile radius.
- Three columns for the type of ownership: “public”, “cooperative”, “investor_owned”.
- A column for the state-average electricity rate named “avg_price_com”.
- A column for the zip-average electricity rate in the commercial sector named “comm_rate”.

- A column for the number of commercial electricity customers in the state named “num_cus_com”.
- A column for the state-average revenue by commercial electricity customer “rev_com”.
- Nine columns to represent the sources of electricity with their share in the state’s energy mix. They are named: “coal”, “geothermal”, “hydro”, “natural gas”, “nuclear”, “petroleum”, “solar”, “wind”, “other”.
- A column to hold the unique identifier of the tariff. This is needed to link the charges that belong to the same tariff.
- Two columns that hold the dependent variables: “energy_charge” and “meter_charge”.

After performing the operations above, and null values removed (e.g., charges with unknown utility id, etc.), the three datasets associated with each sector had the fooling shapes:

- Residential charges: 1,960,821 rows and 115 attributes.
- Commercial charges: 3,609,968 rows and 115 attributes.
- Industrial charges: 970,786 rows and 115 attributes.

The data in its current shape makes a few assumptions:

- Federal and state holidays are not considered. Although electric utilities may have special rates for holidays, these are not considered in the modeling (e.g., fourth of July, Thanksgiving, etc.)

- Special-application tariffs are not considered. These are tariffs that are only applicable to customers using the electricity for pre-determined applications such as street lighting, electric-vehicle charging, agriculture, etc. In NREL's tariff data, there are no labels for distinguishing these tariffs and they could not be identified as a result.
- State-average data has a uniform distribution over the entire state (e.g., energy mix is the same everywhere in the state.)

This data will be used to determine the following variables:

- Meter charges in the residential sector.
- Meter charges in the commercial sector.
- Meter charges in the industrial sector.
- Energy charges in the residential sector.
- Energy charges in the commercial sector.
- Energy charges in the industrial sector.

3.5 Model Performance and Validation

For machine learning models, the model validation is performed by dividing the dataset for each sector into two parts: 80% for training and 20% for testing. Rows from the original dataset were added randomly to either dataset.

For all models, the root mean squared error (RMSE), and mean absolute error (MAE) are used to evaluate their performance. These are selected because they give an estimate of how far the estimated value from the actual value is.

3.6 Inference Models

Several of the methods used in this praxis have been documented in detail in several textbooks (Hastie at al., 2009; Elpaydin, 2020; Webster and Oliver, 2007). Therefore, in this chapter, the methods covered in those texts will not be described again.

The models are implemented using the open-source Python packages Scikit-Learn (Pederegosa et al., 2011; Geron, 2019) for machine leaning methods and PyKrige for the Kriging methods. Custom Python code was written to import the necessary modules from the afore mentioned packages, tune hyper-parameters, calculate the errors, etc.

3.6.1 K-Nearest Neighbors (KNN)

The key parameter in constructing a KNN model is selecting the optimal number of neighbors k. To determine the optimal value of k, for all six cases, k is incremented in steps of 2, between 2 and 18. The value of k is chosen based on the values of the MAE. This approach is consistent with what is reported in other studies (Jalali et al., 2013; Suchenwirth et al., 2014).

3.6.2 Decision Trees (DT)

The performance of decision trees depends on the function used for measuring the quality of the split at each node in the tree (Hastie at al., 2009; Elpaydin, 2020; Geron, 2019). Three splitting functions are considered: the mean absolute error (MAE), the mean squared error (MSE) and Friedman's squared error (Friedman MSE). The split function that leads to the lowest error figures is selected.

3.6.3 Linear Regression (LR)

The performance of linear regression models depends on the regressors used (Hastie at al., 2009; Elpaydin, 2020; Geron, 2019). In this praxis, three were considered: linear, ridge and Bayesian ridge. These methods have demonstrated good performance in

the literature. Shi et al. (2016) used Bayesian ridge to study the relationship between traffic congestion and the frequency of car accidents in Florida and reported R^2 values of 0.77. Anderson (1981) used ridge regression for estimating real-estate value and showed that it performed better than least squares and that it was less sensitive to multicollinearity.

3.6.4 Support Vector Machines (SVM)

With support vector machines, the performance depends on the kernel which can be linear, polynomial, radial, sigmoid, among others (Hastie at al., 2009; Elpaydin, 2020; Geron, 2019). In the problem at hand, two kernels are considered: polynomial and radial because they were shown to lead to superior performance in the literature (Wohlberg et al., 2005; Sitharam et al., 2008; Kanevski et al, 1999).

3.6.5 Artificial Neural Networks (ANN)

The performance of deep neural networks depends on the architecture of the network and the hyper-parameters (Tuba et al., 2021). Due to the high number of possible architectures and hyper-parameters, optimization techniques are often used to find the optimal values (Yu and Zhu, 2020). These techniques include genetic algorithms (Aszemi and Dominic, 2019), bayesian optimization (Ranjit at al., 2019), particle swarm optimization (Foysal et al., 2021), among several others (Yu and Zhu, 2020).

In this praxis, the conventional neural network (CNN) architecture is used with one, two and three hidden layers. A brute-force approach is used where the number of nodes in each layer is incremented from 10 to 170 to determine the optimal network design with no more than 600 nodes. After this limit, the model takes several days to run on a standard laptop machine.

3.6.6 Areal Interpolation

Areal interpolation consists in finding areal features over a large area and estimating the same property for a smaller area within the boundaries for the large area. Several variations of this method have been reviewed by Comber and Zeng (2019). Jing et al. (2020) used this technique on an analogous problem to the one at hand where urban population was mapped in residential neighborhoods in China by using GIS and crowdsourced data. The study showed that MAPE values decreased as resolution increased from a subdistrict level to a neighborhood level.

In this praxis, the zip-average and state-average are both considered to be the known areal features and are assumed to be distributed uniformly across the entire area. The energy charge estimates are set as the corresponding zip-average or state-average electricity rates. One major limitation of this method is that it cannot estimate meter charges because no averages are published.

3.6.7 Point Interpolation

For each zip code, all the neighboring zip codes which are within 20 miles are identified, their respective distances calculated, and their average energy and meter charges calculated by averaging all the applicable charges in the zip code. Time of use periods are considered when averaging the charges. With this data, unweighted and weighted averages can be calculated based on the interpolation method used.

These methods are inspired from the interviews conducted with the maintainers of the URDB dataset, who mentioned that these are some of the heuristics used when no tariffs are available.

3.6.7.1 The Average of the n Nearest Neighbors

Ping et al. (2004) used this method with the number of neighbors increasing from 4 to 10 to estimate cotton yield. The model with four neighbors led to the highest figure of spatial correlation, the measure of performance in this study.

In this praxis, estimates are generated by considering the average charges at the nearest one, two, three, four and five zip codes.

3.6.7.2 The Average of the Nearest Neighbors within a Radius

Estimates are generated by averaging all applicable charges within 10, 15 and 20 miles.

3.6.8 Inverse Distance Weighting

This approach uses the same data that includes the nearest geographical neighbors, except that weighting is applied so the closer the neighbor the more impact it has on the estimate. Chen and Liu (2012) used it to estimate spatial distribution of rainfall in Taiwan by using data from 46 rainfall stations. The authors used the parameter values between 0 and 5 with 0.1 increments and the optimal value was reported to be 2.0. Bekele at al. (2003) used the same method with the parameter values 1, 2 and 3 to map soil potassium levels. The parameter value of 2 lead to the lowest MSE value. Ping et al. (2004) used this method with the parameter values 1 through 5 to study cotton yield, and found that the parameter value 1 lead to best performance, which was described in terms of spatial autocorrelation.

Lu and Wong (2008) proposed an adaptive approach for setting the parameters, and it assigned lower values to points in clustered areas and higher values in disperse areas. This technique was applied in two problems: estimating precipitation in Taiwan and surface elevation in the state of Virginia. In both cases, parameter values of 1, 2 and 3 were

considered. Additionally, a separate model was set up to dynamically pick the parameter between 1 and 3 based on the local level of clustering. This model with dynamic selection of the parameter led to lower percent error values than when the parameter values were constant.

Considering the findings of these studies, the parameter rho is varied between 1 and 2.5 with 0.5 increments where 1 is a special case where no weighting is applied, and a higher rho gives the closest neighbor more impact on the estimate. The increment was set to 0.5 because small increments were shown to lead to similar error figures. The dynamic assignment of parameters was not considered because it makes the model slower to run. Also, while this method was shown to perform better than assigning constant parameter values, the difference in percent error was within 1%.

3.6.9 Ordinary Kriging

This approach requires a variogram, which is a function that characterizes the relationship between the distance between any arbitrary pair of points and the difference in charges between them. This variogram is obtained by fitting the data to a function, which can be challenging as discussed by Wackernagel (2003). Various methods to find the variogram were discussed by Goulard and Voltz (1993) and Wackernagel (2003). One way to circumvent this is to make assumptions about the fit. Five types of variograms are considered: one linear and four non-linear (power, gaussian, exponential, and spherical). The ones with the lowest error figures are selected.

3.7 Finding the Closest Error Distributions

To determine the closes distribution of the error values, the Python package Fitter is used. Fitter compares a distribution to 80 standard distributions. The data is fitted against

each distribution and a sum of squares error is calculated. The distribution with the lowest error value is returned. The package has the necessary functions to obtain the parameters of the distribution with the lowest sum of squares error.

3.8 Summary

In this chapter, the different methods for detecting and removing outliers from the data are described. Additionally, the data inference models are described with a commentary on which parameters affect their performance and which parameter values were used in the literature. Based on that review, the parameters used in this study are selected. In Chapter 4, the results are presented and discussed.

Chapter 4 — Results

4.1 Introduction

In this chapter, the outlier detection results are reported, followed with the estimation errors for each inference method, and finally a formulation of the estimation error.

4.2 Outlier Detection

Histogram plots in Figure 4-1 show that energy charges are all between 0 and 0.8 $/kwh with a left-skewed distribution (a single energy charge of 40,000 $/kwh was removed because it is nowhere near a typical value of an energy charge). For meter charges however, the ranges are much larger which could be an indicative of a presence of outliers.

4.2.1 Univariate Outlier Detection

The model that removes outliers from the univariate variable of meter charges was set up by assuming a normal distribution and shifting the mean between the minimum and maximum values in the range until the remaining points pass the normality test. The normal distribution was assumed because there is no reason to believe why meter charges would be skewed in either direction. The ones with the least number of outliers were retained. The model returned a high number of outliers. No univariate outlier detection was performed on energy charges because the values were well within the expected range.

4.2.2 Multivariate Outlier Detection

The Mahalanobis Distance was calculated for energy and meter charges in all sectors and those that did not fit in the chi distribution were removed. The results are shown in Table 4-1, Figure 4-2, and Figure 4-3.

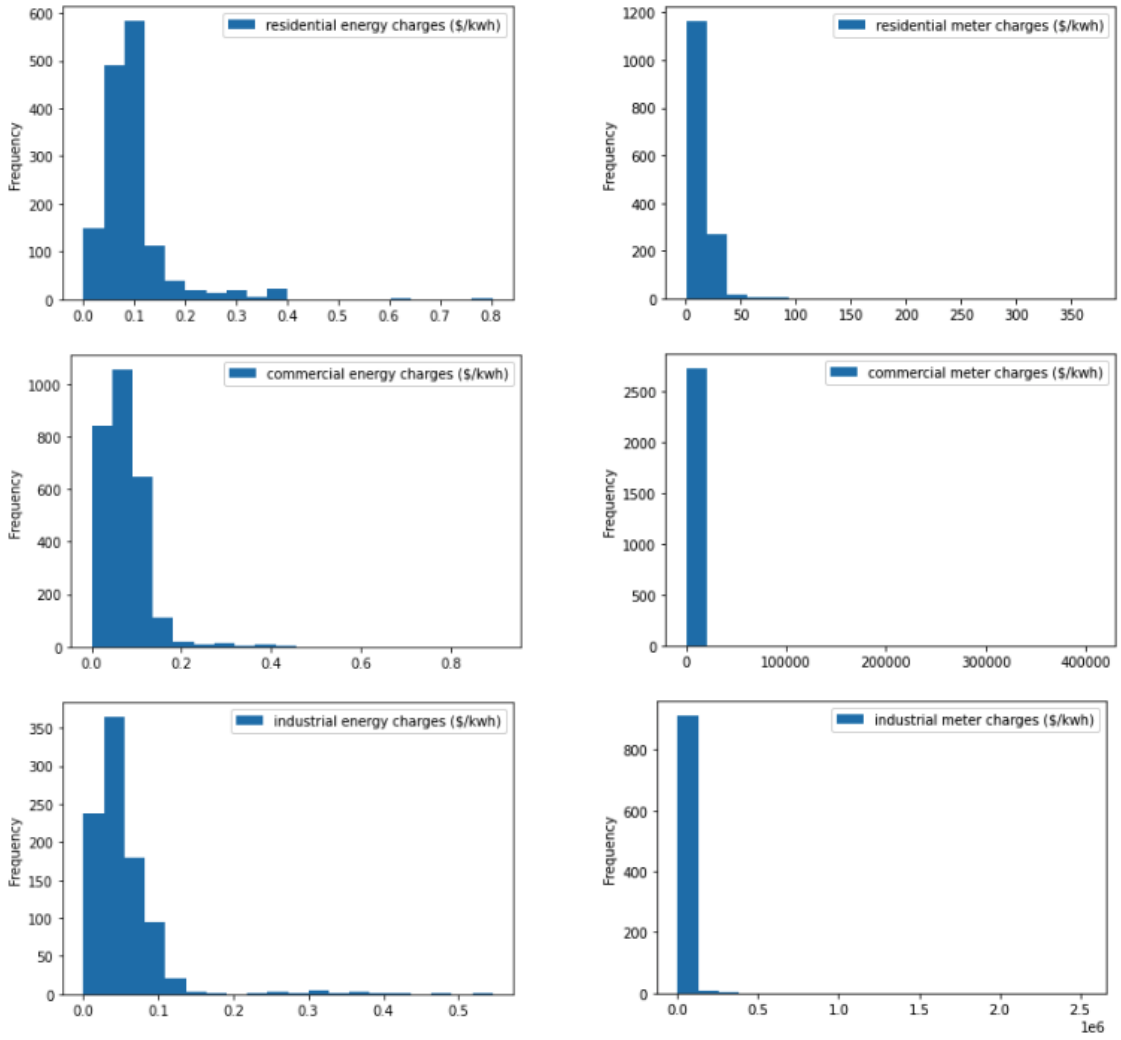

Figure 4-1. Histogram plots of energy and meter charges for all sectors.

Table 4-1. Results of the multivariate outlier detection method.

Variable	Number of charges including outliers	Number of outliers (percent)	Minimum	Median	Maximum
residential meter charges	1,456	32 (2.2%)	0.56	11.89	49.5
commercial meter charges	2,727	259 (9.5%)	0.34	30.12	17,264
industrial meter charges	924	30 (3.2%)	1.00	255.70	251,527

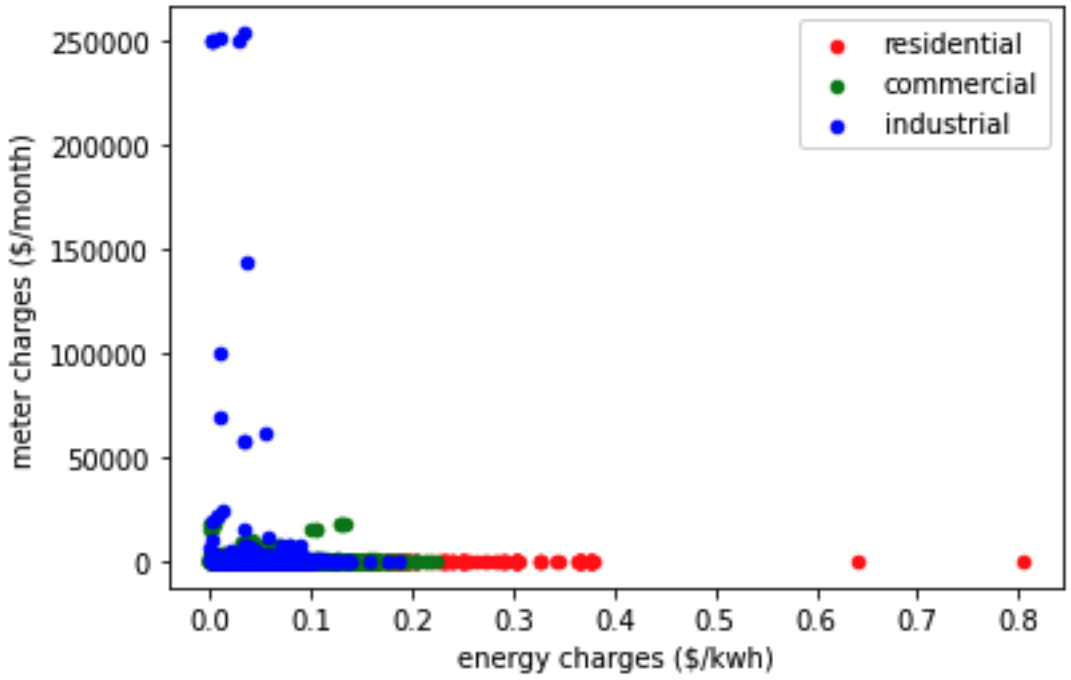

Figure 4-2. Scatter plots of energy and meter charges for all sectors after outliers are removed.

4.2.3 Outlier Detection Analysis

The univariate method eliminated a significant number of meter charges in the dataset as outliers. There are also charges that fall below the minimum values in Table 4-1 but were verified to be correct in the electric utility's website. Although this method has been applied successfully elsewhere (Jasinska and Preweda, 2021), the results of the multivariate method leave out fewer outliers and the outcome is consistent with what was verified manually by checking the tariffs in the websites of the electric utilities. That said, for the remaining of this praxis, the dataset used is the output of the multivariate outlier detection analysis only. The result is that all charges have a right-skewed distribution (Figure 4-3).

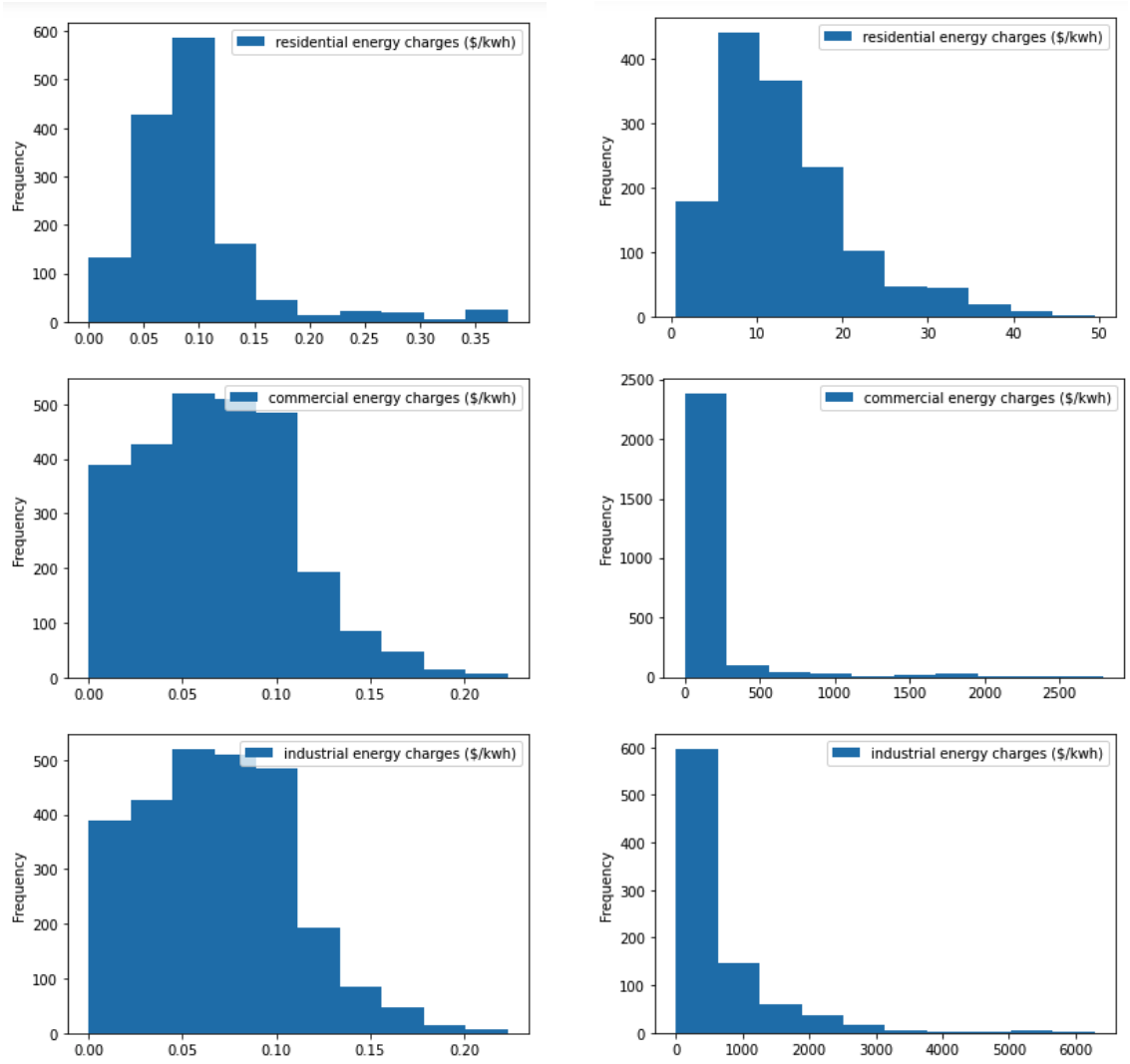

Figure 4-3. Histogram plots of energy and meter charges without outliers.

4.3 Modeling

4.3.1 K-Nearest Neighbors (KNN)

The KNN model requires the hyper parameter k, the number of neighbors. To determine the optimal value of k, the model was run 9 times by changing the value of k between 2 and 18 and incrementing it by 2 in each run. The results are shown in Figure 4-4. The lowest MAE values are associated with a k value of 2. Energy charges are estimated with MAE values of 0.01, 0.02 and 0.03 USD/kwh for the industrial, residential, and

commercial sectors respectively. Meter charges are estimated with MAE values of 2.10, 361.48 and 1,156.58 for the residential, commercial, and industrial sectors respectively.

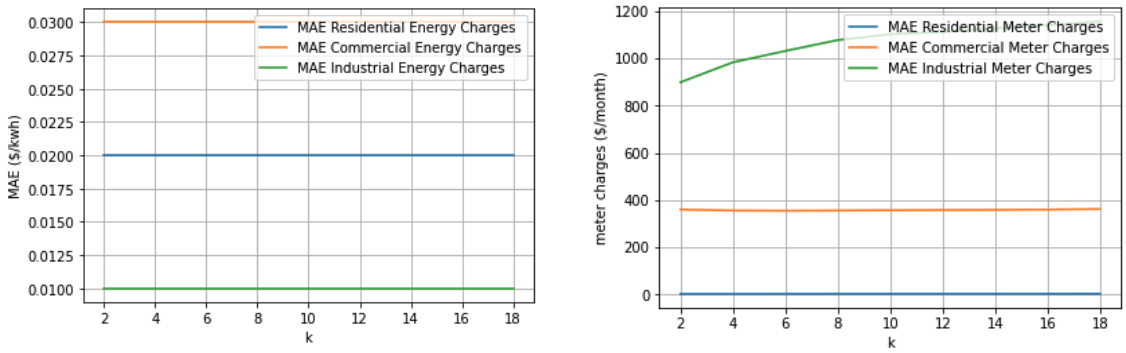

Figure 4-4. MAE values for the K-Nearest Neighbors model with different values of hyper-parameter k.

Table 4-2. Statistical summary of the errors for the K-Nearest Neighbors (k=2) model.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	0.00	0.00	0.00	-0.09	-3.99	1547.15
Std. dev.	0.04	0.04	0.02	3.25	1,349.22	12,507.39
Min	-0.72	-0.39	-0.48	-32.00	-17203.00	1.00
25%	-0.01	-0.02	0.00	0.00	-37.00	156.00
50%	0.00	0.00	0.00	0.00	0.00	354.00
75%	0.01	0.02	0.00	0.00	72.00	930.00
Max	0.65	0.40	0.48	31.00	17,203.00	251,528.00
RMSE	0.04	0.04	0.02	3.27	1,338.86	8,821.67
MAE	0.02	0.03	0.01	2.10	361.48	1,156.58

Figure 4-5 and Table 4-2 show the error distribution which is centered around a mean and median of 0 for all energy charges with standard deviation values between 0.02 and 0.04. The average absolute error for energy charges is between 0.01 and 0.03 USD/kwh. For meter charges however, both mean and median divert from the 0 value

which is symptomatic of a bias. The error values are also much higher than those reported for energy charges. For the case of commercial and industrial charges, the RMSE is a multiple of the MAE. This is indicative of the presence of large errors.

The results of the variable importance analysis (Figure 4-6) show that the number of customers, revenue per customer, latitude and longitude are important predictors for all six charges.

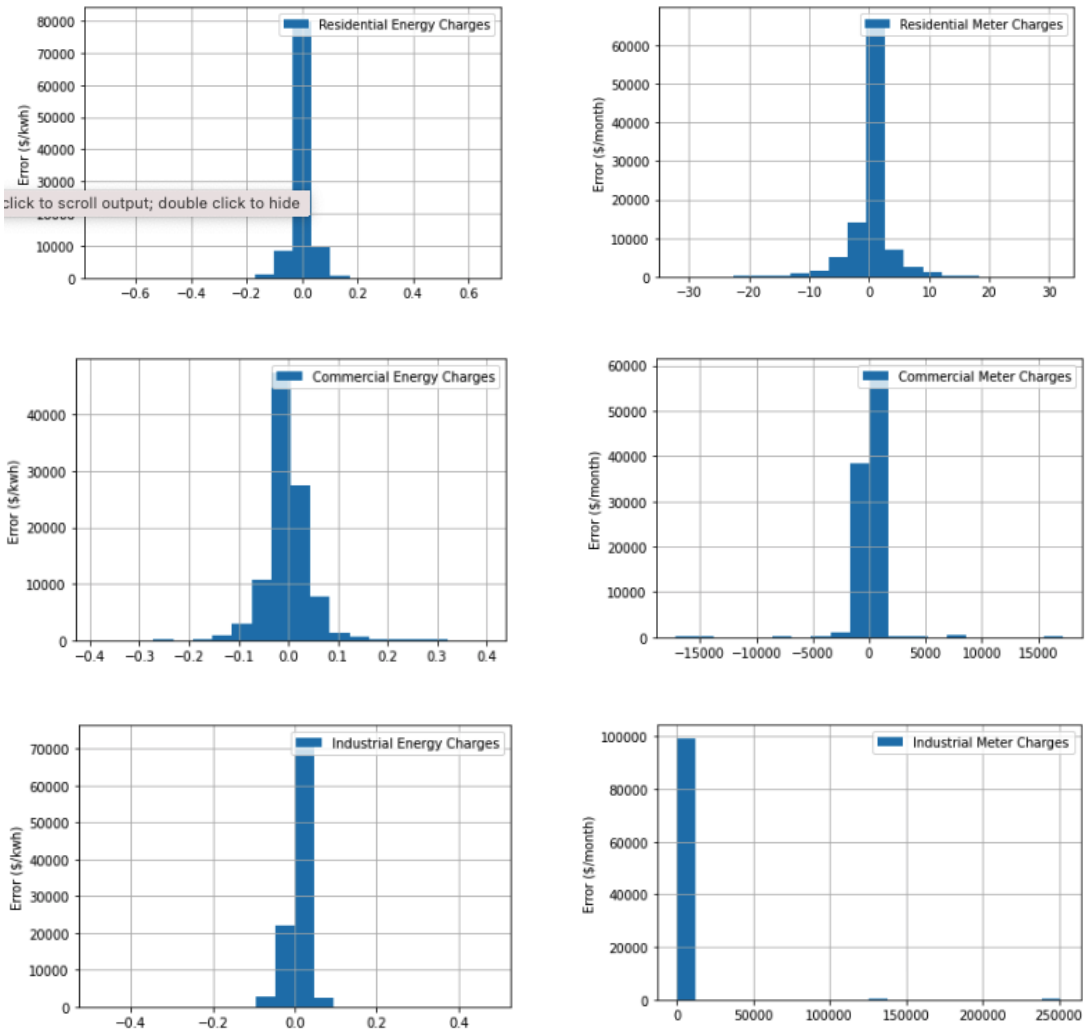

Figure 4-5. Histograms of estimation errors for the K-Nearest Neighbors model.

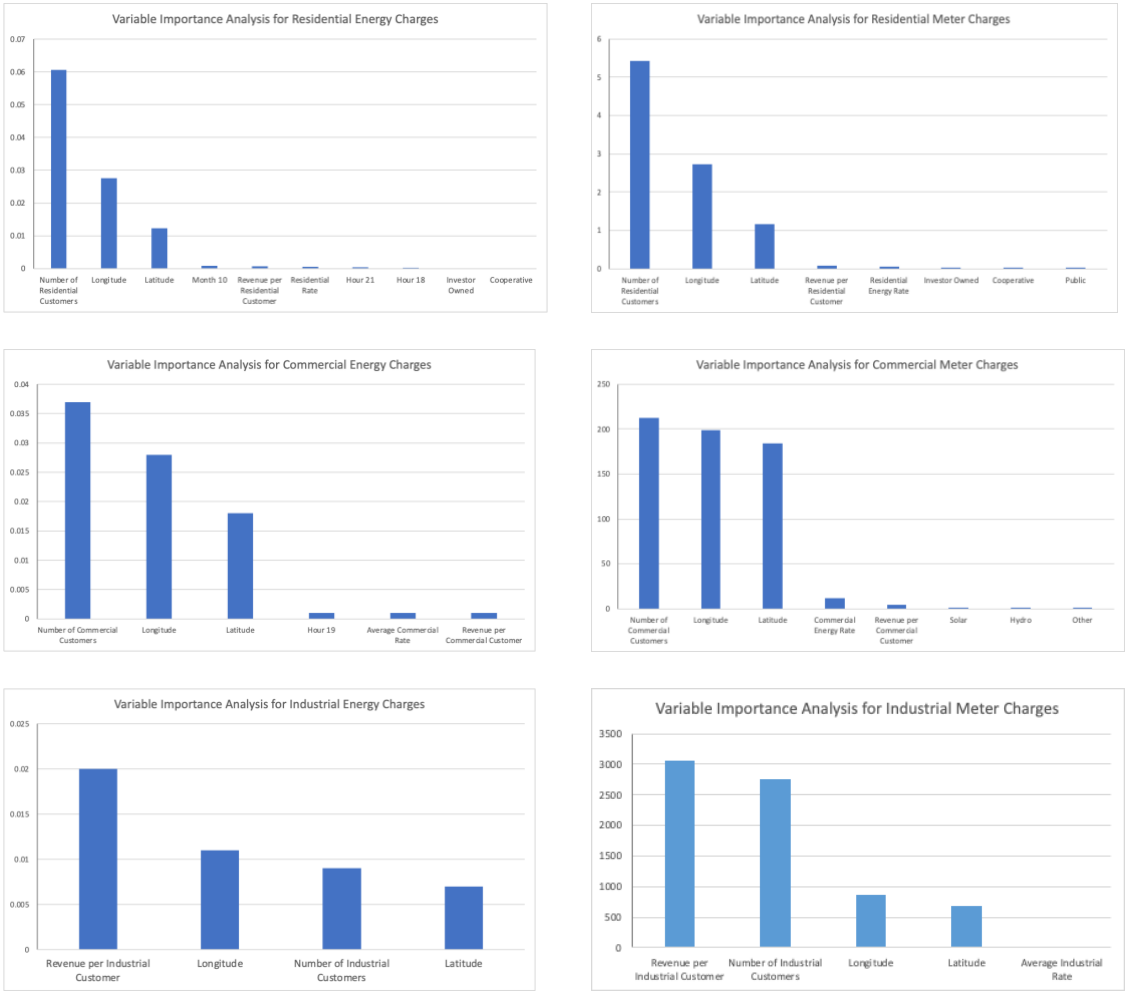

Figure 4-6. Variable Importance Analysis results for the KNN model.

4.3.2 Decision Trees

Three different functions were used to measure the quality of the split: mean absolute error (MAE), mean squared error (MSE) and Friedman mean squared error (Friedman MSE). This function is also referred to as the criterion. The error values are shown in Table 4-3 and 4-4. For energy charges, the results show similar performance for all split functions. For meter charges however, the MSE split function leads to lower error values. Therefore, the MSE split function is used for the remainder of the study.

Figure 4-7 and Table 4-4 show a similar performance to the KNN model. The error distribution is centered around a mean and median of 0 for all energy charges with standard deviation values between 0.02 and 0.04. The MAE for energy charges is between 0.01 and 0.05 USD/kwh. For meter charges, the mean and median divert from the 0 value which is symptomatic of a skew (i.e., bias). The error values are also much higher than those reported for energy charges.

Table 4-3. Error values for the Decision Trees model with different split functions.

Error	Split function	Energy Charges			Meter Charges		
		Residential	Commercial	Industrial	Residential	Commercial	Industrial
RMSE	Friedman MSE	0.04	0.05	0.03	3.81	1,547.97	10,013.41
	MAE	0.04	0.05	0.03	3.87	1,560.20	11,283.42
	MSE	0.04	0.05	0.03	3.80	1,546.61	9,995.00
MAE	Friedman MSE	0.02	0.03	0.01	1.78	380.97	875.02
	MAE	0.02	0.03	0.01	1.81	382.14	997.42
	MSE	0.02	0.03	0.01	1.78	379.76	877.32

Table 4-4. Statistical summary of the error for DT model.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	0.00	0.00	0.00	-0.11	12.65	1,544.91
Std. dev.	0.04	0.04	0.02	3.13	1204.76	11,006.72
Min	-0.72	-0.36	-0.47	-38.00	-17126.00	1.00
25%	-0.01	-0.02	-0.01	-1.00	-31.00	175.00
50%	0.00	0.00	0.00	0.00	9.00	383.00
75%	0.01	0.02	0.01	1.00	116.00	988.00
Max	0.42	0.29	0.37	1	16,561	249,722
RMSE	0.04	0.05	0.03	3.80	1,546.61	9,995.00
MAE	0.02	0.03	0.01	1.78	379.76	877.32

The top 3 levels of the trees are in Figure 4-8 through Figure 4.13. The trees are all different from each other in terms of variables used at each split. This is different from what was observed in the KNN model where the most important predictors where the same for estimating all six charges.

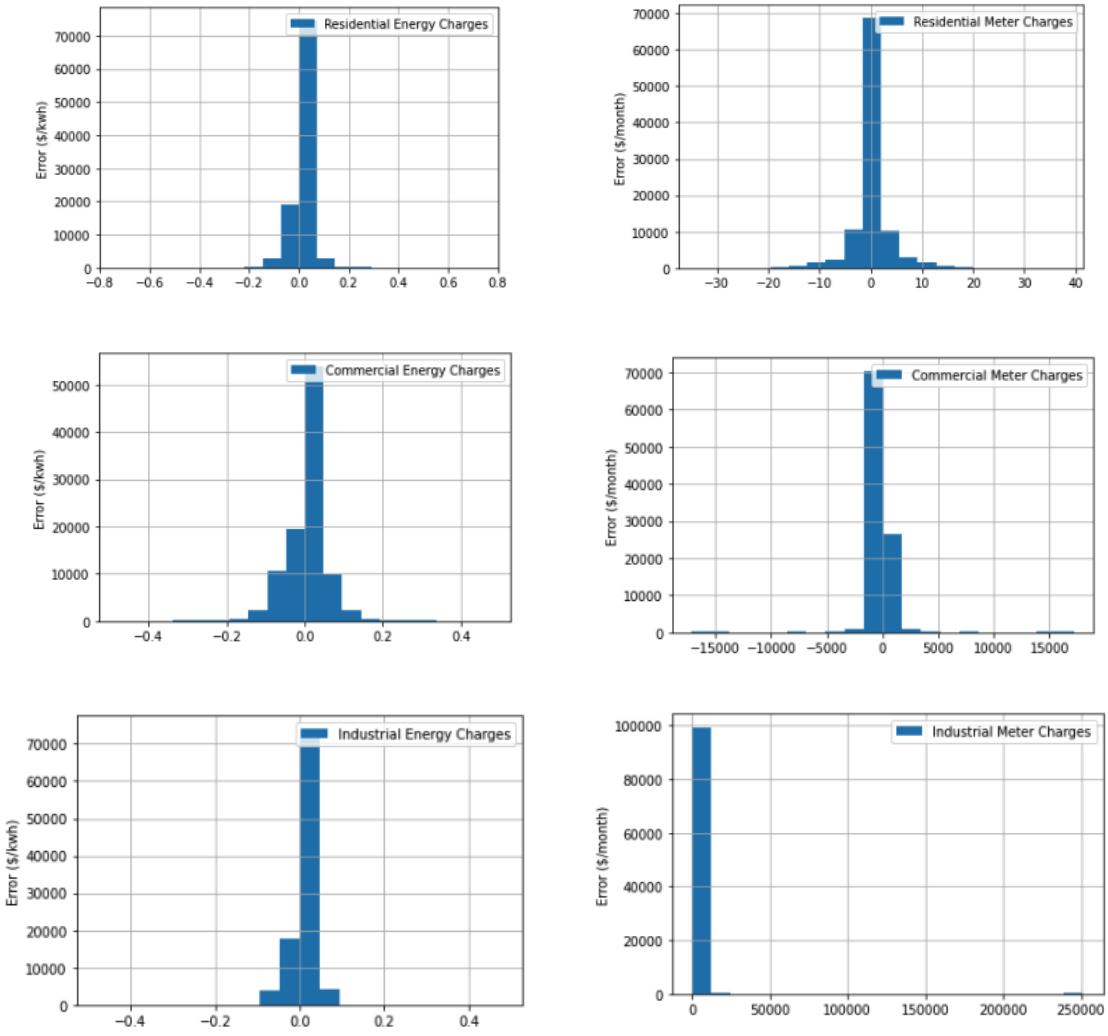

Figure 4-7. Histograms of estimation errors for the decision trees models.

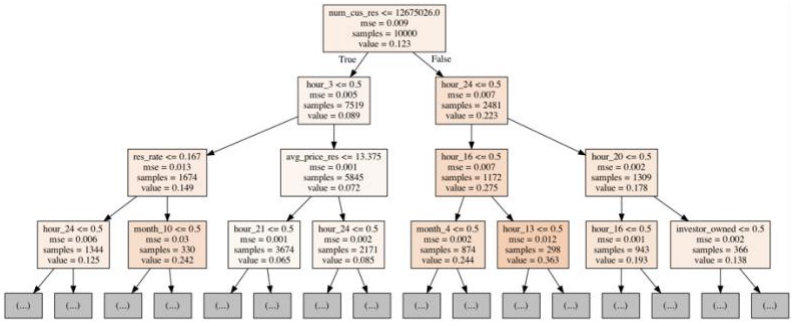

Figure 4-8. The decision tree for residential energy charges.

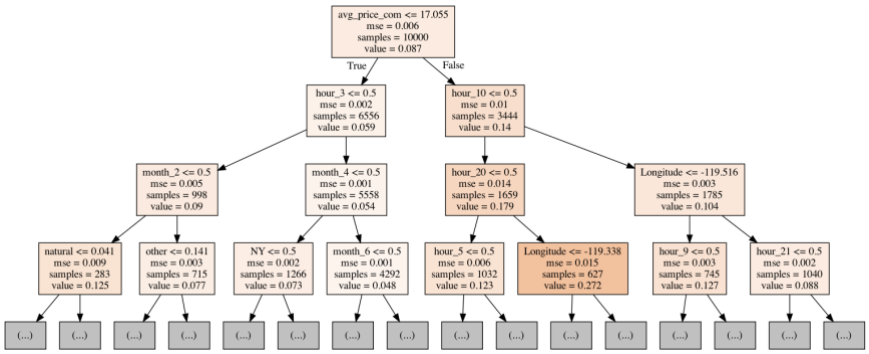

Figure 4-9. The decision tree for commercial energy charges.

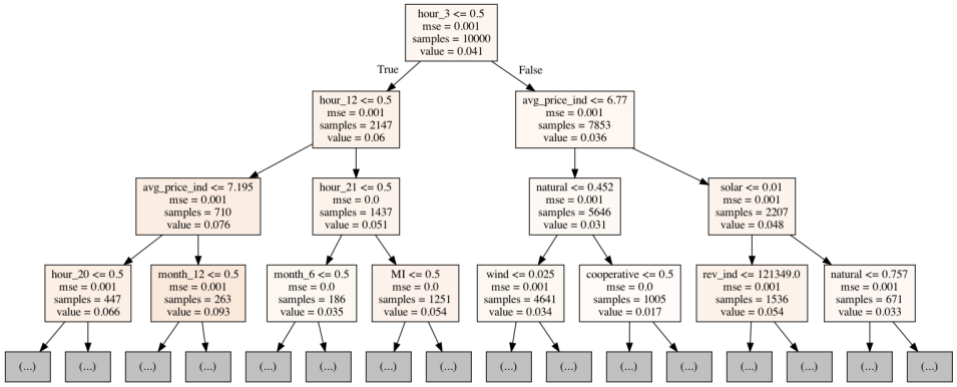

Figure 4-10. The decision tree for industrial energy charges.

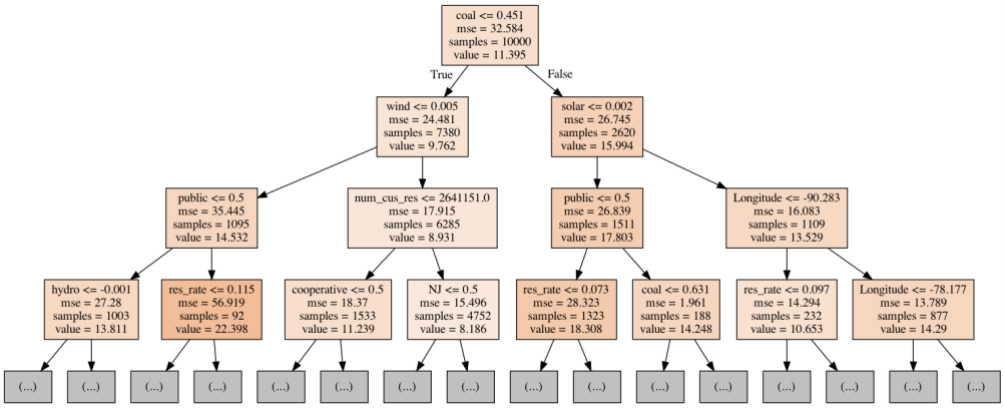

Figure 4-11. The decision tree for residential meter charges.

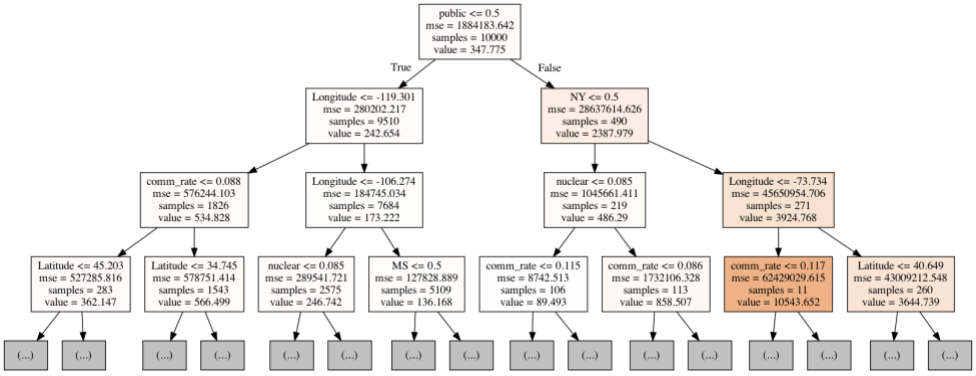

Figure 4-12. The decision tree for commercial meter charges.

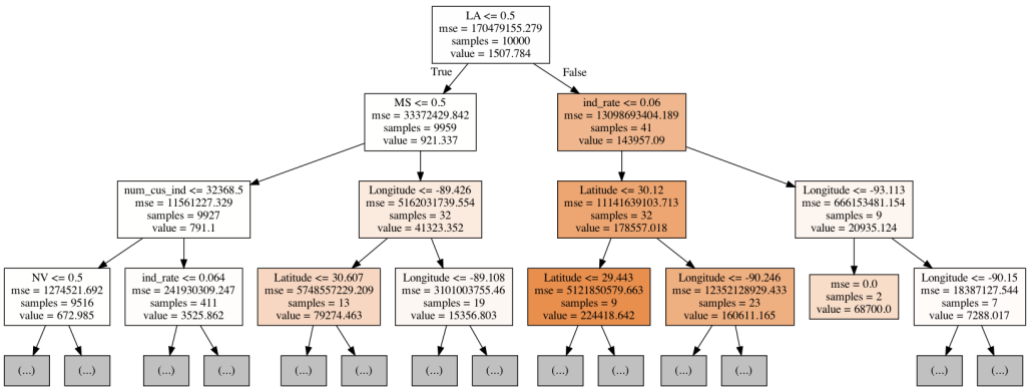

Figure 4-13. The decision tree for industrial meter charges.

4.3.3 Support Vector Machines (SVM)

The model was constructed with two kernels: polynomial and radial basis function (RBF). For energy charges, all kernels lead to the same MAE values (Table 4-5). However, for the meter charges, the RBF kernel leads to lower RMSE and MAE values.

Table 4-5. Performance of the SVM model with different kernels.

Error	Kernel	Energy Charges			Meter Charges		
		Residential	Commercial	Industrial	Residential	Commercial	Industrial
RMSE	Polynomial	0.05	0.05	0.03	6.59	1,137.24	1,555.19
	RBF	0.06	0.06	0.03	5.86	1,122.13	1,481.79
MAE	Polynomial	0.05	0.05	0.03	4.04	336.66	597.66
	RBF	0.05	0.05	0.03	3.43	331.42	512.34

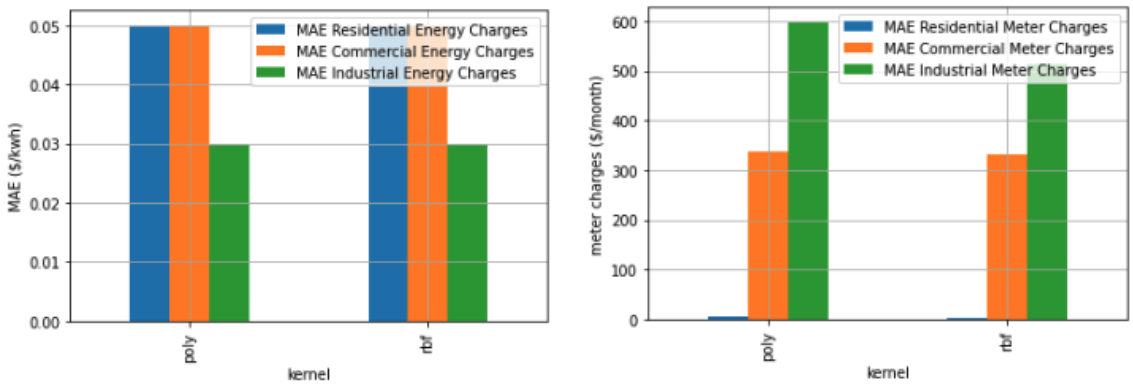

Figure 4-14. MAE values for the SVM model with different kernels.

Table 4-6. Statistical summary of the error for the SVM model.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Kernel	Polynomial	Polynomial	Polynomial	RBF	RBF	RBF
Mean	0.03	0.03	0.02	-0.97	-316.27	240.29
Standard deviation	0.04	0.04	0.02	5.81	1092.4	31.88
Min	-0.22	-0.1	-0.09	-35	-8497	54
25%	0.01	0	0	-1	-212	239
50%	0.04	0.03	0.02	0	0	250
75%	0.06	0.05	0.04	2	19	250
Max	0.1	0.1	0.09	30	47	252
RMSE	0.05	0.05	0.03	5.86	1,137.24	1,555.19
MAE	0.05	0.05	0.03	3.43	336.66	597.66

Figure 4-15 and Table 4-6 show the error distribution is biased in either direction for all charges. For energy charges, the RMSE is equal to the MAE which is indicative of small errors in the model. For meter charges, the MRSE values are much higher than the MAE which is symptomatic of large estimation errors.

4.3.4 Linear Regression (LR)

The model was constructed with three different regressors: linear, ridge and Bayes ridge. For energy charges, all regressors led to the same MAE values (Figure 4-16). For meter charges, all regressors showed close MAE and RMSE values for the commercial

and residential sectors but for the industrial sector the ridge regressor performed better than the linear regressor which in turn performed better than the Bayesian regressor.

The coefficients of the linear model show that hour of day, month and state are the most important predictors for energy charges (Figure 4-17). For meter charges, state, energy mix and utility ownership have the highest coefficients (Figure 4-17).

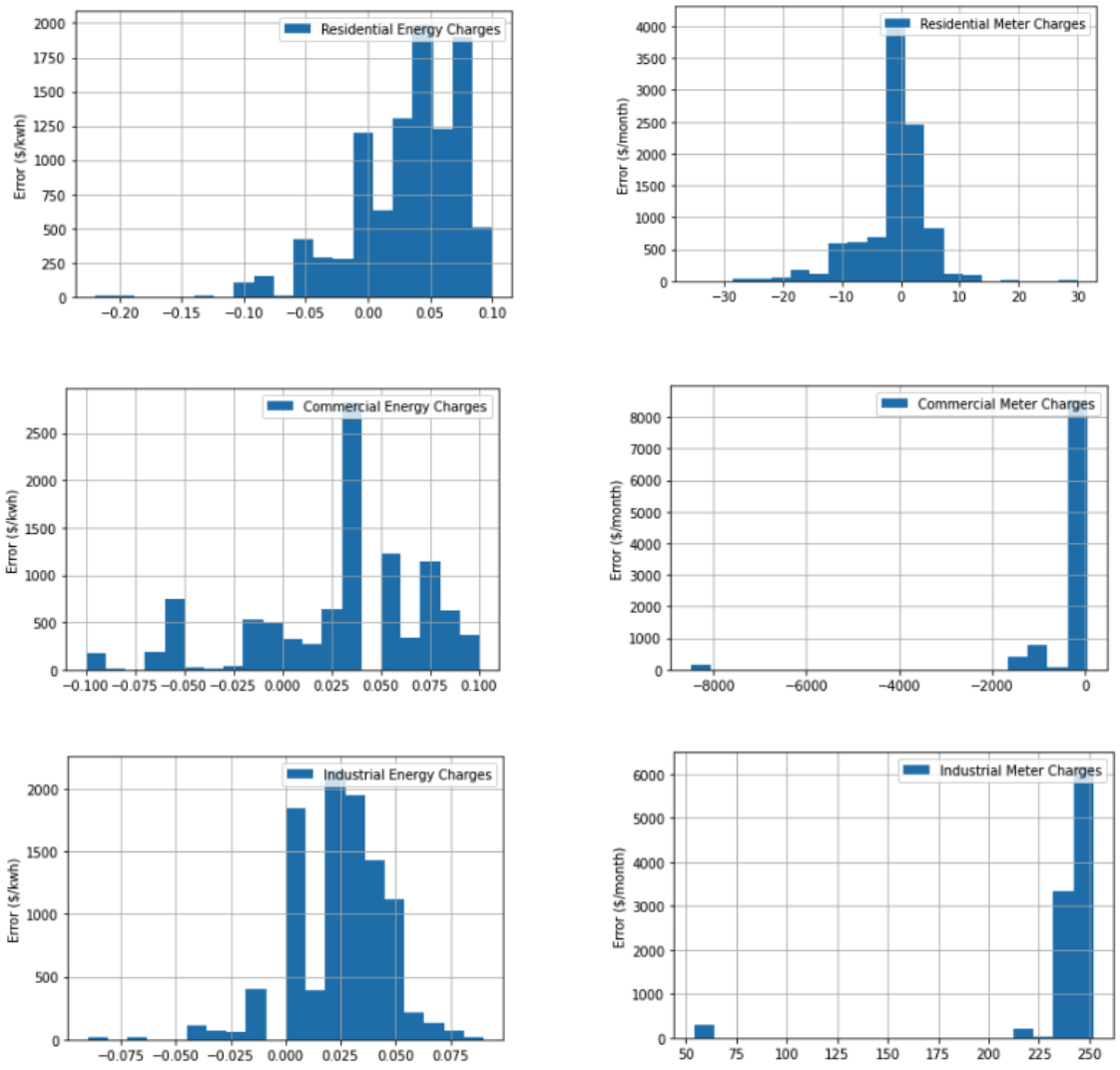

Figure 4-15. Histograms of estimation errors for the SVM model.

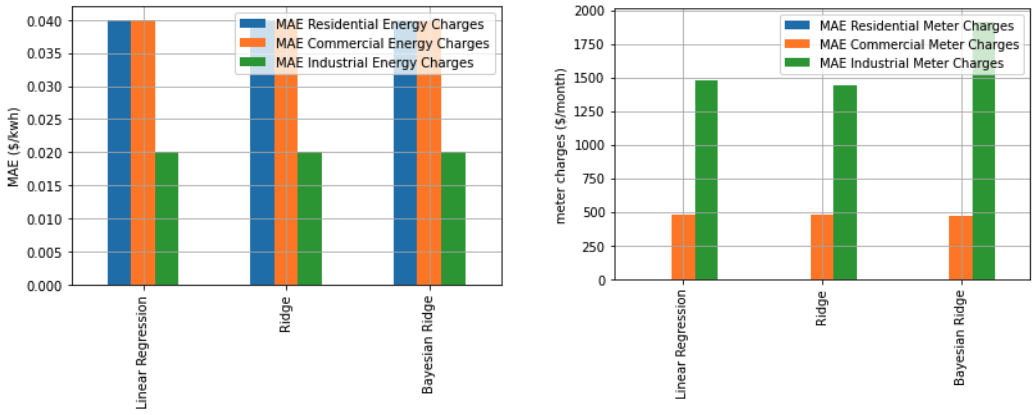

Figure 4-16. MAE values for the linear regression model with different regressors.

Figure 4-17. Results of Variable Importance Analysis for the Linear Regression model.

The error values are presented in Figure 4-18 and Table 4-7; where the error distribution is centered around a mean and median of 0 for all energy charges. For meter charges however, both mean and median divert from the 0 value which is indicative of a biased model.

Table 4-7. Statistical summary of the error for the linear regression model.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	0.00	0.00	0.00	-0.07	-7.43	1531.45
Standard deviation	0.05	0.06	0.02	4.08	1271.40	8816.79
Min	-0.69	-0.36	-0.46	-30.00	-14,918	-3,015
25%	-0.03	-0.03	-0.01	-2.00	-150	240
50%	0.00	0.00	0.00	0.00	103	640
75%	0.03	0.03	0.02	2.00	304	1,130
Max	0.20	0.24	0.09	17.00	3,419	123,673
RMSE	0.06	0.06	0.03	4.06	1,301.32	9,416.48
MAE	0.04	0.04	0.02	2.9	453.64	1,329.20

4.3.5 Artificial Neural Networks (ANN)

155 runs were made by considering conventional neural networks with one, two or three hidden layers. The number of nodes in each layer is incremented from 10 to 170 by 40 to determine the optimal network design. The optimal designs are in Table 4-8.

Figure 4-19 shows the error distribution which does not resemble any of the common distributions (e.g., normal, exponential, etc.) for all charges. This makes it challenging to model the estimation error to understand its impact on the economic analysis.

Table 4-8. Optimal ANN designs with the corresponding MAPE and MAE values.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Design	10,90,90	130,10,130	90,90,170	130,130,170	10,10,130	50,170,90
MAE	0.05	0.07	0.02	4.42	325.52	2,718.84

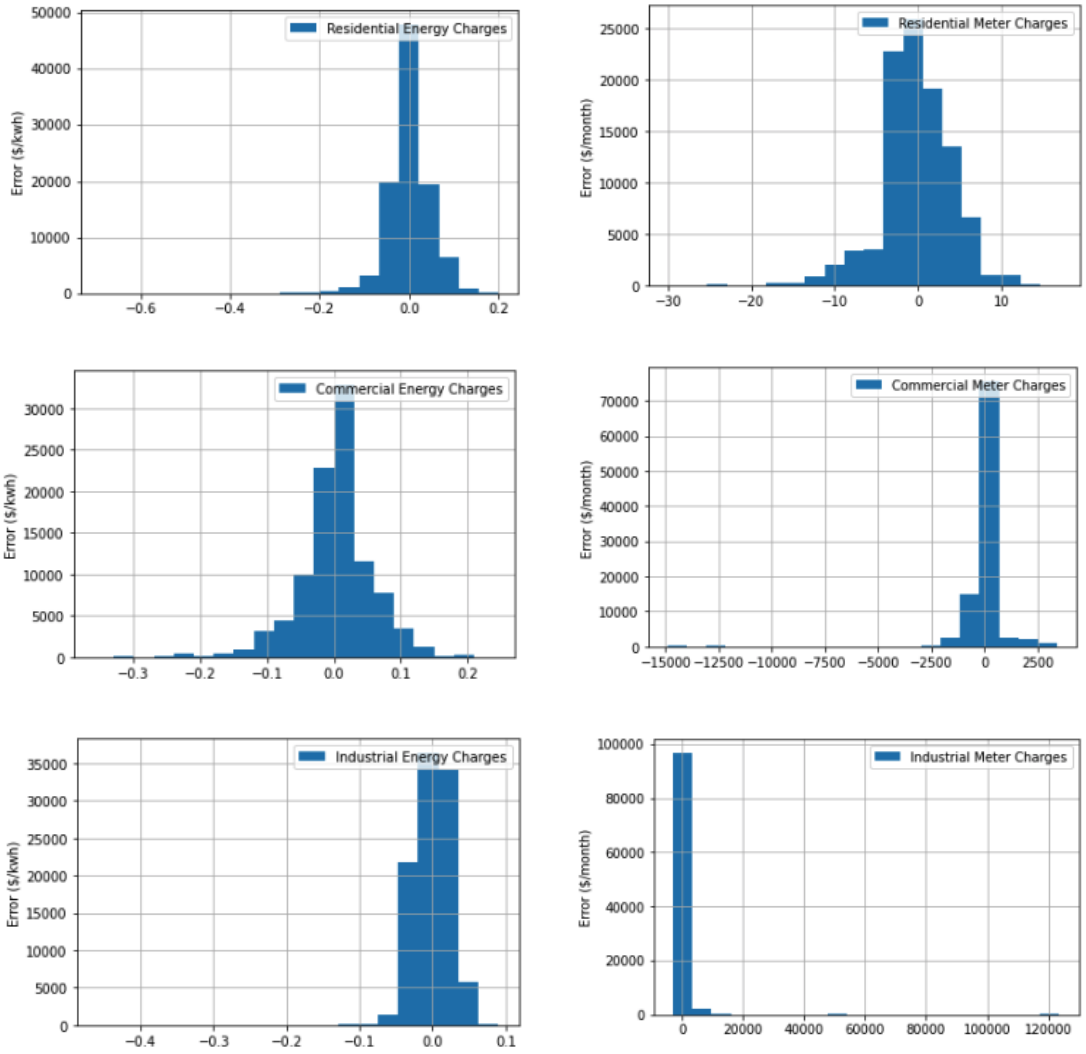

Figure 4-18. Histograms of estimation errors for the linear regression models.

4.3.6 Areal Interpolation

The estimates of energy charges are set to two known areal features: zip-average and state-average rates of electricity. The MAE values for both features are in Table 4-9. In all sectors, using state-average electricity rates as the areal feature shows better performance than using zip-average electricity rates. All MAE values are within 4 cents per kWh. One limitation of this method is lack of areal features that represent meter

charges. The distributions of the errors are plotted in Figure 4-20 and statistics in Table 4-10.

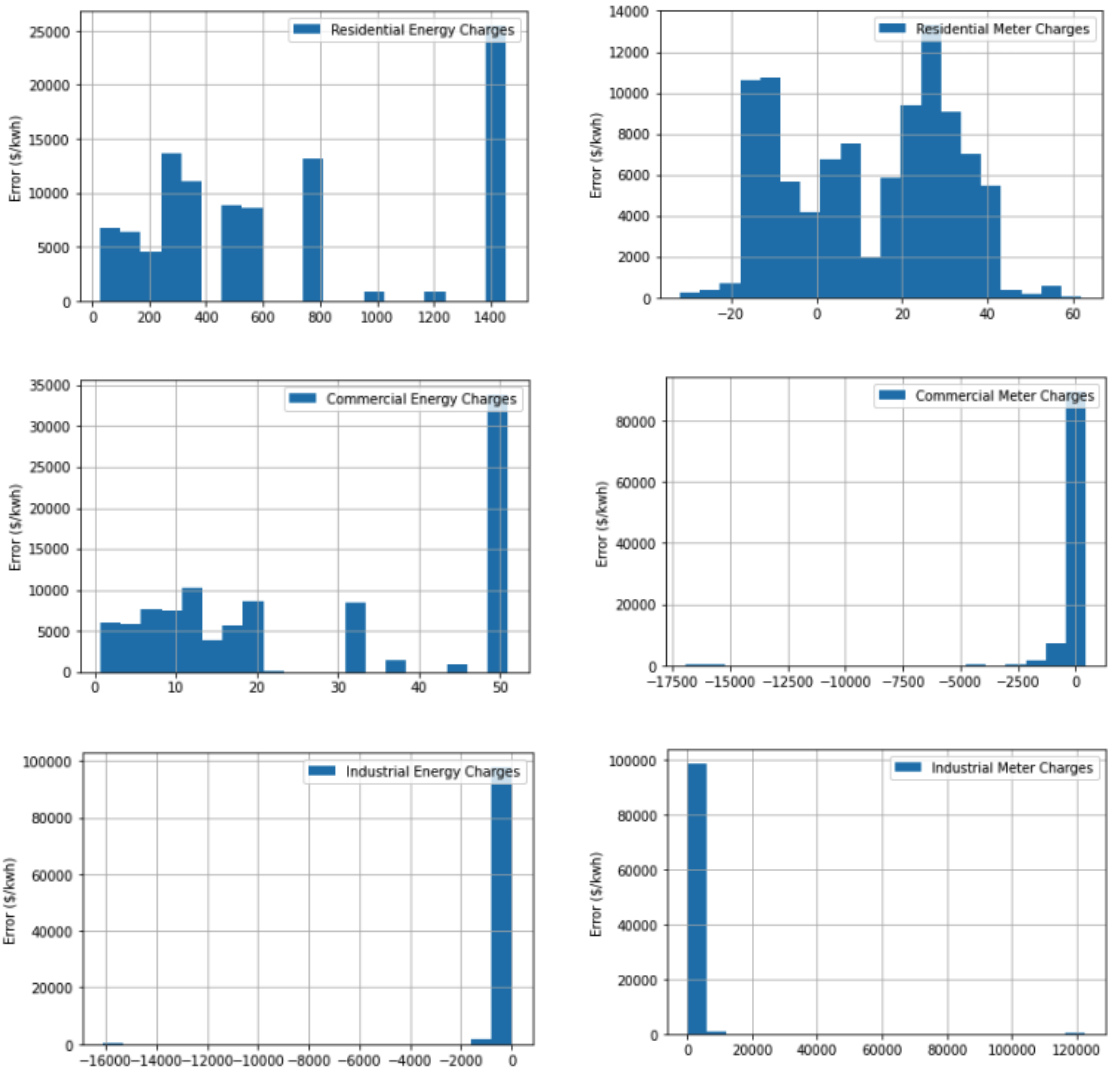

Figure 4-19. Histograms of estimation errors for the artificial neural networks' models.

Table 4-9. Estimation errors of the areal interpolation models.

	Zip-average areal features			State-average areal features		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
MAE	0.04	0.04	0.03	0.03	0.03	0.02

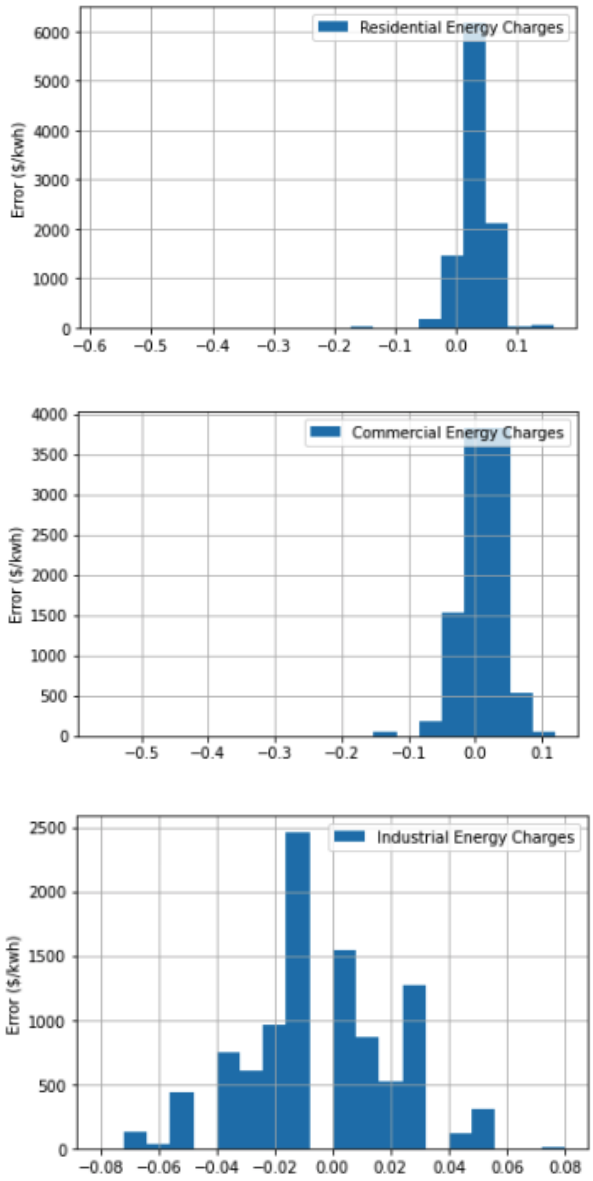

Figure 4-20. Histograms of estimation errors for the areal interpolation model.

Table 4-10. Statistical summary of the error for the areal interpolation model.

	Residential	Commercial	Industrial
Mean	0.03	0.01	0.00
Standard deviation	0.02	0.03	0.03
Min	-0.58	-0.56	-0.08
25%	0.02	-0.01	-0.02
50%	0.04	0.01	-0.01
75%	0.04	0.04	0.01
Max	0.16	0.12	0.08

4.3.7 Point Interpolation

4.3.7.1 Nearest Neighbors

In this model, the estimate is the average of the charges that are applicable in the closest locations. Energy charges remained constant when the number of neighbors changed. This is because neighboring points tend to have the same energy charges. However, for meter charges, the estimates for industrial charges improved as the number of neighbors increased (Figure 4-21). The distribution of the error is in Figure 4-22 and Table 4-11. Like the ANN case, none of the distributions resembled a known distribution and error values were biased in either direction.

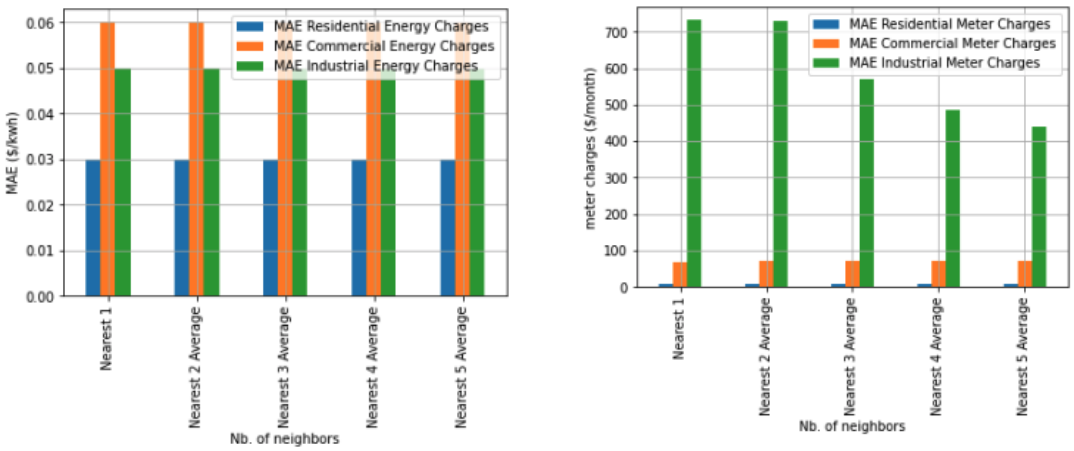

Figure 4-21. MAE values for the point interpolation model.

Table 4-11. Estimation errors of the point interpolation models with the number of neighbors set to 5.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	-0.03	-0.06	-0.04	-5.24	-2.2	359.29
Standard deviation	0.04	0.05	0.05	9.05	107.18	3,795.23
Min	-0.17	-0.14	-0.12	-33	-232	-253
25%	-0.06	-0.11	-0.08	-11	-45	-41
50%	0	-0.04	-0.04	-3	-10	-11
75%	0	-0.01	-0.01	1	33.25	112
Max	0.06	0.06	0.05	15	423	59,536

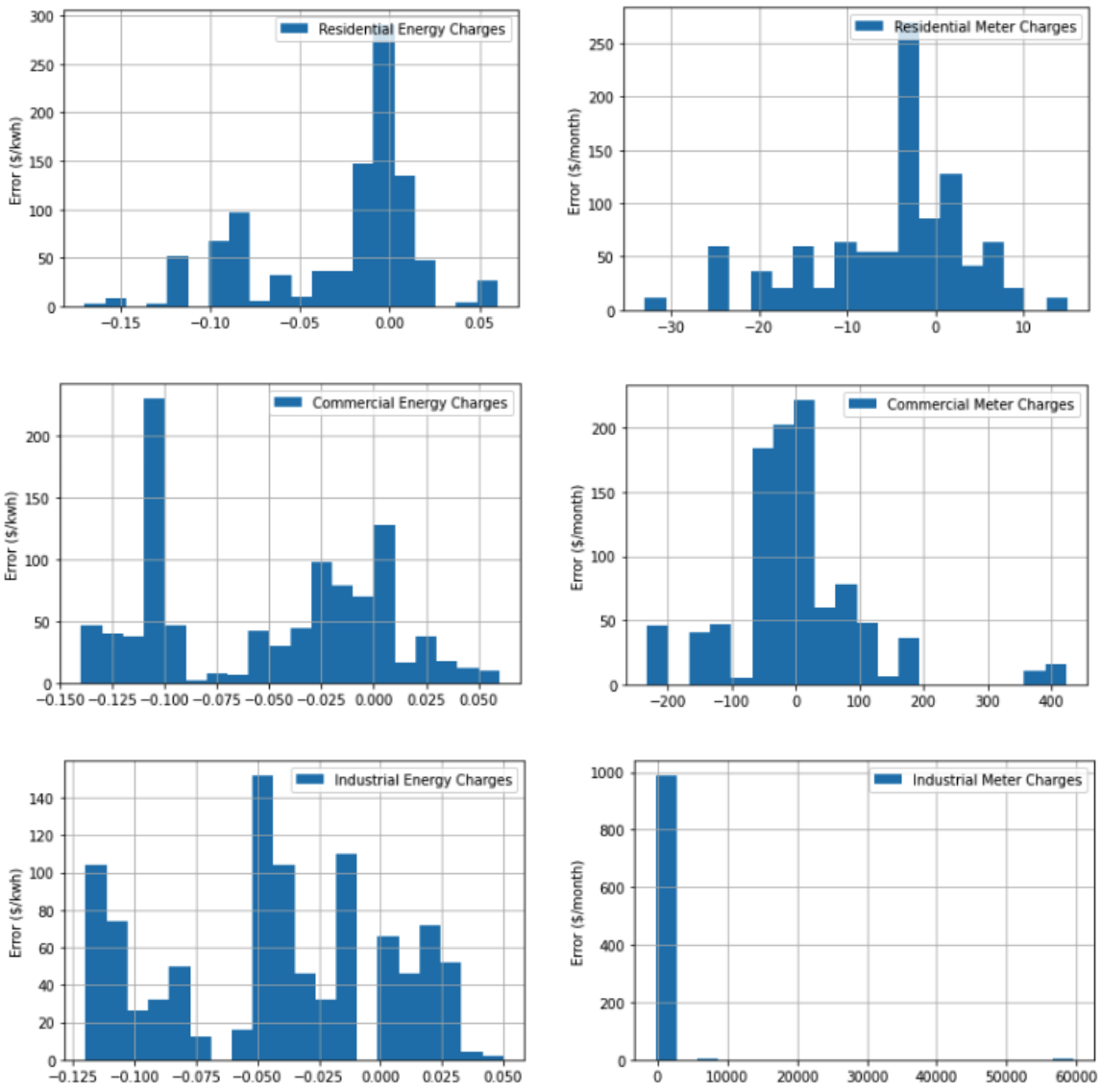

Figure 4-22. Error distribution for the point interpolation model with the number of neighbors set to 5.

4.3.7.2 Nearest Neighbors within a Radius

In this model, the estimate is the average of the charges that are applicable in all the locations that are within a certain radius. Three values are considered for the radius: 10, 15 and 20 miles. The results show an optimum radius of 15 miles. Lower or higher values lead to higher error values (Figure 4-23).

Error distribution figures are in Figurer 4-24 and Table 4-12. In all six cases, the distributions were skewed in either direction.

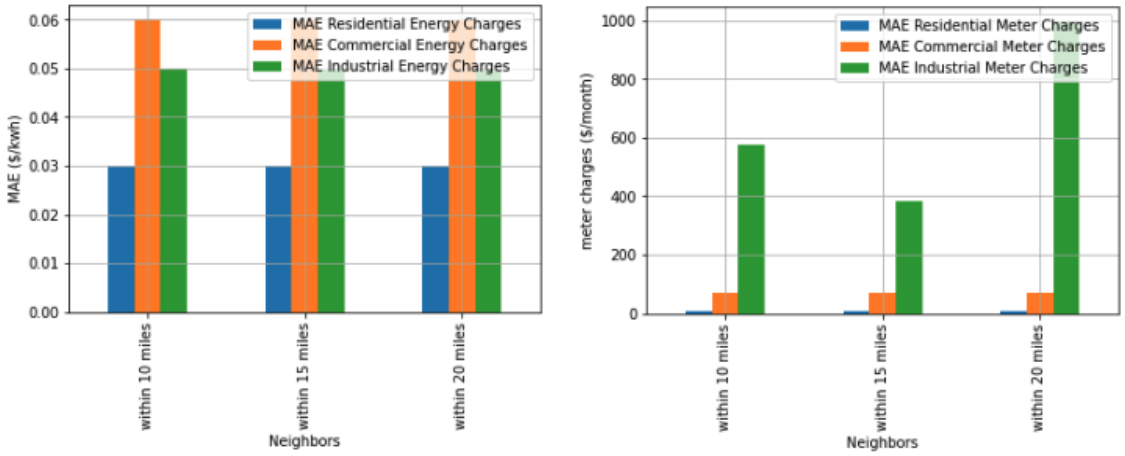

Figure 4-23. MAE values for the point interpolation model.

Table 4-12. Estimation errors of the point interpolation models with the radius of 15.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	-0.03	-0.06	-0.04	-5.24	-2.2	359.29
Std. dev.	0.04	0.05	0.05	9.05	107.18	3,795.23
Min	-0.17	-0.14	-0.12	-33	-232	-253
25%	-0.06	-0.11	-0.08	-11	-45	-41
50%	0	-0.04	-0.04	-3	-10	-11
75%	0	-0.01	-0.01	1	33.25	112
Max	0.06	0.06	0.05	15	423	59,536

4.3.7.3 Inverse Distance Weighting (IDW)

Four values of the parameter rho were considered: 1, 1.5, 2 and 2.5. For each run, the MAE values are calculated (Figure 4-25). For energy charges, all parameters led to

close error figures. For the case of industrial meter charges however, as the value of rho increased, the error values decreased.

Energy charges were all estimated within a 0.07 $/kwh range with negative mean values which is symptomatic of a bias like shown in Figure 4-26 and Table 4-13 where most points fall below 0, which means that the model tends to underestimate the energy charges.

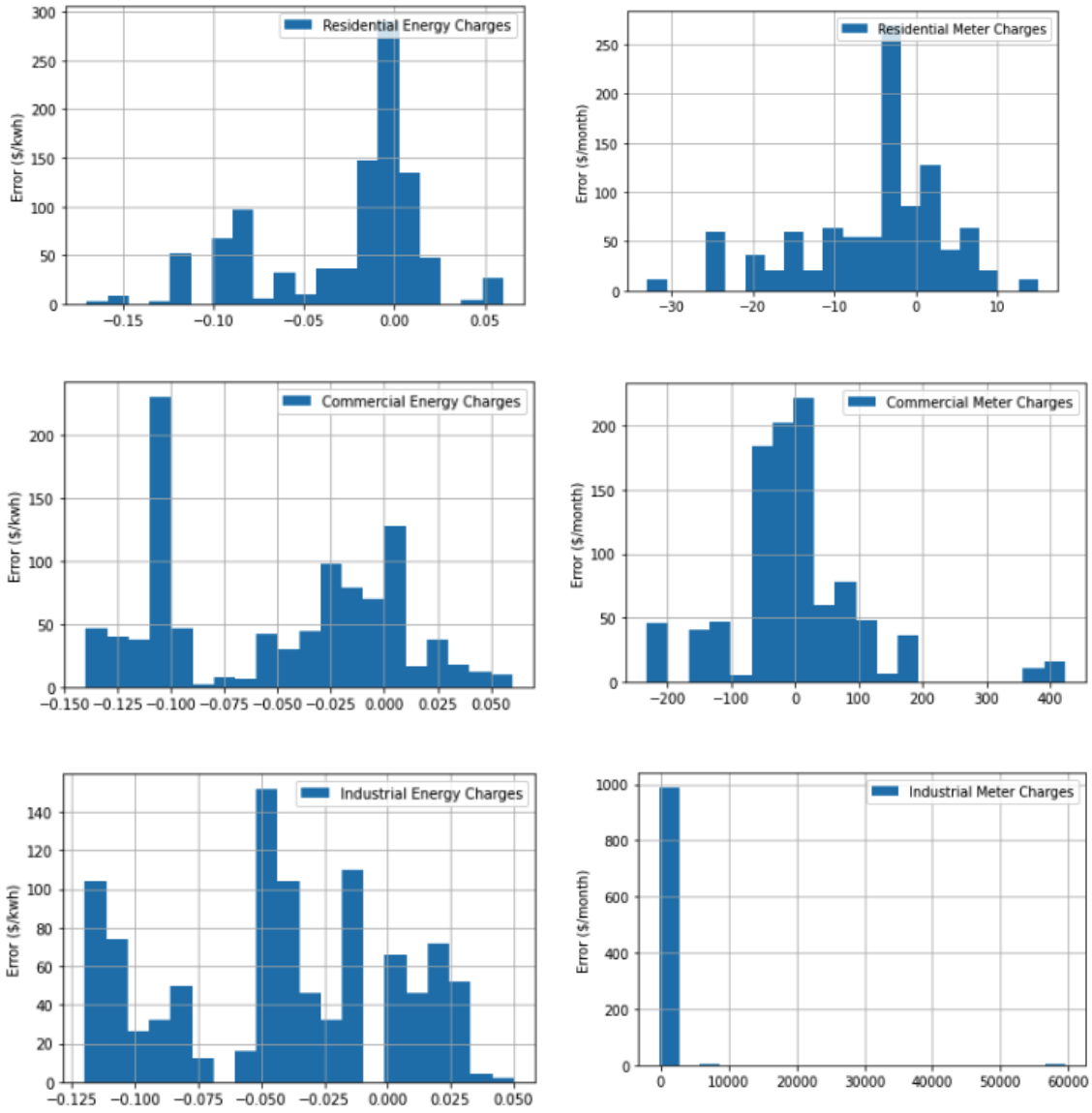

Figure 4-24. Error distribution for the point interpolation model with a radius of 15.

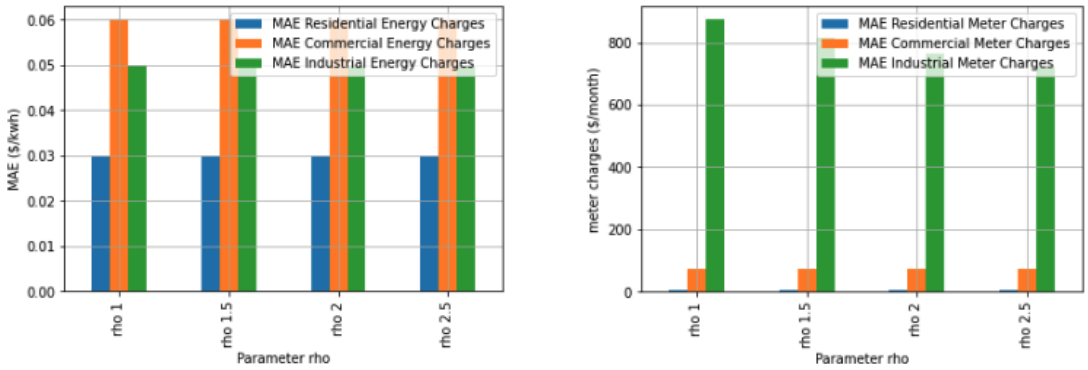

Figure 4-25. MAE values for the IDW model with different parameter values.

Table 4-13. Estimation errors of the IDW model with a parameter rho of 2.5.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Mean	-0.03	-0.06	-0.04	-5.29	-1.7	643.04
Standard deviation	0.04	0.05	0.05	9	108.38	5,065.74
Min	-0.17	-0.14	-0.12	-33	-232	-253
25%	-0.06	-0.11	-0.08	-11	-45	-42.75
50%	0	-0.04	-0.04	-3	-10	-11
75%	0	-0.01	-0.01	1	31	113
Max	0.07	0.07	0.05	14	423	70,093

4.3.8 Ordinary Kriging

For each independent variable, five linear and non-linear variograms are considered: linear, power, gaussian, exponential, spherical. These are the assumed shapes of the function that describes the relationship between the distance of point-pairs within the same vicinity and difference in charges at the two points. The variograms that led to the lowest MAE values are in Table 4-14.

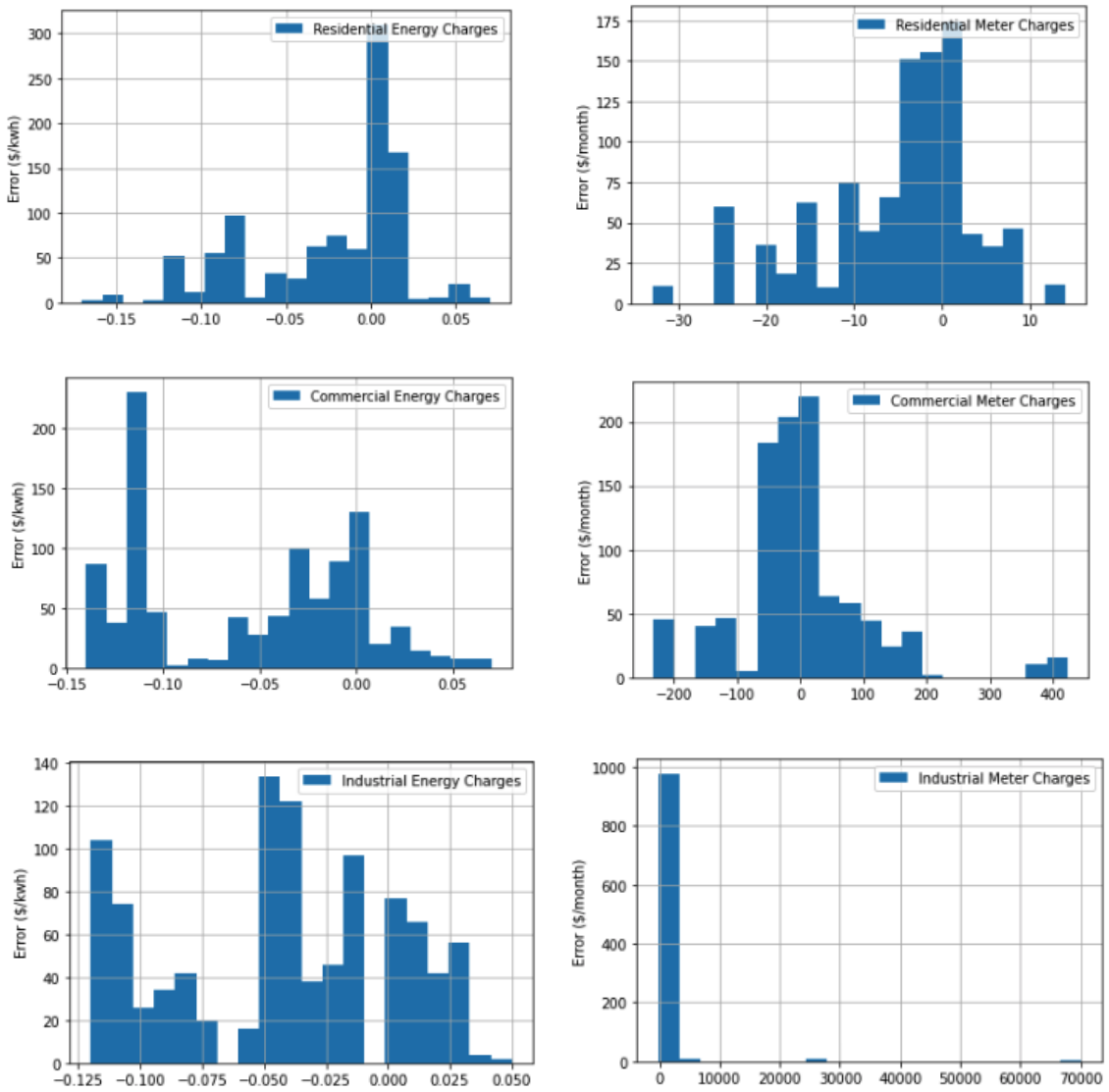

Figure 4-26. Error distribution values for the IDW model with a parameter rho of 2.5.

Table 4-14. Optimal variograms with the corresponding error values.

	Energy charges			Meter charges		
	Residential	Commercial	Industrial	Residential	Commercial	Industrial
Variogram	Linear	Spherical	Exponential	Power	Exponential	Exponential
MAE	0.01	0.04	0.03	2.26	53.84	1,627.79
RMSE	0.02	0.07	0.04	4.31	254.41	2,864.24

4.3.9 Analyzing Model Performance Results

4.3.9.1 Residential Energy Charges

The MAE values are between 0.01 $/kwh for the ordinary kriging model and 0.05 for the SVM and ANN models. The KNN and decision tree models also performed well with 0.02 $/kwh. However, in terms of error range, the kriging model has a wider error range than KNN which in turn has a wider range than the decision tree model. It is worth noting that with the decision tree model, it is easier to plot the tree and understand the decision process while KNN and ordinary kriging do not offer those facilities. It is also worth noting that the median in the ordinary kriging is larger than the mean which is indicative of a biased estimator. On the other hand, it has a lower RMSE value of 0.02.

The distribution of the error is close to a Generalized Normal Distribution (GND) with the parameter beta of 0.1133 (see Figure 4-19 and Table 4-15.)

Table 4-15. Estimation errors for residential energy charges.

	Models with ancillary data						Models with no ancillary data			
	KNN	SVM	DT	LR	AI	ANN	NGN	IDW	NNR	OK
Mean	0.00	0.03	0.00	0.00	0.03	678.90	-0.03	-0.03	-0.03	0
Standard deviation	0.04	0.04	0.04	0.05	0.02	500.21	0.04	0.04	0.04	0.06
Min	-0.72	-0.22	-0.72	-0.69	-0.58	27.17	-0.17	-0.17	-0.17	-0.72
25%	-0.01	0.01	-0.01	-0.03	0.02	286.69	-0.06	-0.06	-0.06	-0.03
50%	0.00	0.04	0.00	0.00	0.04	525.57	0	0	0	0.01
75%	0.01	0.06	0.01	0.03	0.04	1,452.75	0	0	0	0.03
Max	0.65	0.1	0.42	0.20	0.16	1,453.26	0.06	0.07	0.06	0.79
Range	1.37	0.132	1.14	0.89	0.74	1,426.09	0.23	0.23	0.23	1.51
RMSE	0.04	0.05	0.04	0.06	0.04	0.08	0.05	0.05	0.05	0.02
MAE	0.02	0.05	0.02	0.04	0.03	0.05	0.03	0.03	0.03	0.01

4.3.9.2 Commercial Energy Charges

The MAE values are between 0.03 for the KNN, decision tree and aerial interpolation model and 0.06 for the IDW and both point interpolation models. Among the three models with the lowest MAE, the decision tree model has the lowest range of 0.64 but a larger RMSE than both KNN and areal interpolation. Additionally, the areal interpolation method is centered around a mean and median of 0.01 which is indicative of a biased estimator, although with a smaller RMSE. The error range is between 0.20 for the models that rely on calculating the charge by averaging the nearest 5 points where this is a commercial charge, or the average of all commercial charges applied within 20 miles (see Table 4-16).

The distribution of the error is close to a Double Weibull Distribution with the parameter tail length of 2.4244.

Table 4-16. Estimation errors for commercial energy charges.

	Models with ancillary data						Models with no ancillary data			
	KNN	SVM	DT	LR	AI	ANN	NGN	IDW	NNR	OK
Mean	0.00	0.03	0.00	0.00	0.01	27.07	-0.06	-0.06	-0.06	0.00
Std. dev.	0.04	0.04	0.04	0.06	0.03	18.87	0.05	0.05	0.05	0.04
Min	-0.39	-0.1	-0.36	-0.36	-0.56	0.72	-0.14	-0.14	-0.14	-0.18
25%	-0.02	0	-0.02	-0.03	-0.01	10.28	-0.11	-0.11	-0.11	-0.01
50%	0.00	0.03	0.00	0.00	0.01	18.37	-0.04	-0.04	-0.04	0.01
75%	0.02	0.05	0.02	0.03	0.04	50.67	-0.01	-0.01	-0.01	0.02
Max	0.40	0.1	0.29	0.24	0.12	50.97	0.06	0.07	0.06	0.18
Range	0.79	0.2	0.64	0.60	0.68	50.25	0.20	0.21	0.20	0.36
RMSE	0.04	0.05	0.05	0.06	0.03	0.07	0.08	0.08	0.08	0.07
MAE	0.03	0.05	0.03	0.04	0.03	0.05	0.06	0.06	0.06	0.04

4.3.9.3 Industrial Energy Charges

Based on the MAE values, the methods with ancillary data performed consistently better than geo-statistics methods with no ancillary data. The MAE values were between 0.01 for the KNN and decision tree models and 0.05 for the IDW and point interpolation models. The KNN has a lower RMSE than the decision tree model but a slightly wider error range of 0.96. The error ranges were between 0.16 for the areal interpolation model and 0.96 for the K-Nearest Neighbors model (Table 4-17). The areal interpolation model estimates the commercial rate as the state's average rate for the commercial sector. This data is published yearly by the EIA. The distribution of the error is close to the Asymmetric Laplace distribution with the parameter kappa of 0.8704.

Table 4-17. Estimation errors for industrial energy charges.

	Models with ancillary data						Models with no ancillary data			
	KNN	SVM	DT	LR	AI	ANN	NGN	IDW	NNR	OK
Mean	0.00	0.02	0.00	0.00	0.00	-311.59	-0.04	-0.04	-0.04	0.00
Std. dev.	0.02	0.02	0.02	0.02	0.03	912.51	0.05	0.05	0.05	0.04
Min	-0.48	-0.09	-0.47	-0.46	-0.08	-16,111	-0.12	-0.12	-0.12	-0.17
25%	0.00	0	-0.01	-0.01	-0.02	-397.24	-0.08	-0.08	-0.08	-0.02
50%	0.00	0.02	0.00	0.00	-0.01	-164.67	-0.04	-0.04	-0.04	0
75%	0.00	0.04	0.01	0.02	0.01	-90.13	-0.01	-0.01	-0.01	0.02
Max	0.48	0.09	0.37	0.09	0.08	13.47	0.05	0.05	0.05	0.15
Range	0.96	0.18	0.84	0.54	0.16	16,124	0.17	0.17	0.17	0.32
RMSE	0.02	0.03	0.03	0.03	0.03	0.03	0.06	0.06	0.06	0.04
MAE	0.01	0.03	0.01	0.02	0.02	0.02	0.05	0.05	0.05	0.03

4.3.9.4 Residential Meter Charges

The MAE values are between 1.78 for the decision tree model and 7.40 for the point interpolation model that averages the charges from the five nearest neighbors. The decision

tree also has the narrowest error range of 39. The error range for residential meter charges is between 39 for the decision tree model and 63 for the K-Nearest Neighbors model (Table 4-18). Like residential energy charges, the closest distribution to the error is the Generalized Normal Distribution with the parameter beta equal to 0.1798.

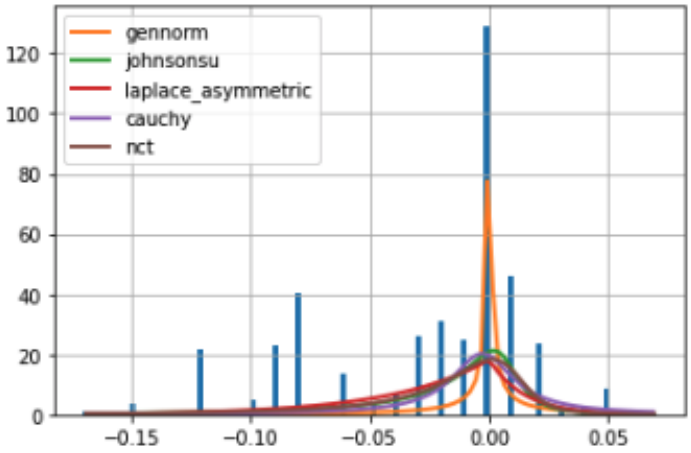

Figure 4-27. Statistical distributions that are close to the distribution of the error in estimating energy charges with the 5 nearest geographic neighbors (NGN).

Table 4-18. Estimation errors for residential meter charges.

	Models with ancillary data					Models with no ancillary data			
	KNN	SVM	DT	LR	ANN	NGN	IDW	NNR	OK
Mean	-0.09	-0.97	-0.11	-0.07	12.21	-5.24	-5.29	-5.24	0.43
Std. dev.	3.25	5.81	3.13	4.08	19.38	9.05	9	9.05	16.39
Min	-32.00	-35	-38.00	-30.00	-32.00	-33	-33	-33	-96.00
25%	0.00	-1	-1.00	-2.00	-6.00	-11	-11	-11	-9.00
50%	0.00	0	0.00	0.00	16.00	-3	-3	-3	4.00
75%	0.00	2	1.00	2.00	29.00	1	1	1	6.00
Max	31.00	30	1	17.00	62.00	15	14	15	104.00
Range	63	65	39	47	94.00	48	47	48	200
RMSE	3.27	5.86	3.80	4.06	5.76	10.45	10.41	10.45	4.31
MAE	2.10	3.43	1.78	2.9	4.42	7.40	7.35	7.31	2.26

4.3.9.5 Commercial Meter Charges

The MAE values are between 70 for the point interpolation method that averages the five nearest charges and 453.64 for the linear regression model. The IDW has a slightly higher MAE of 41.11 than the linear interpolation method.

Table 4-19. Estimation errors for commercial meter charges.

	Models with ancillary data					Models with no ancillary data			
	KNN	SVM	DT	LR	ANN	NGN	IDW	NNR	OK
Mean	-3.99	-316.27	12.65	-7.43	-92	-2.2	-1.7	-2.2	4.65
Std Dev	1,349.22	1092.4	1204.76	1,271.40	1,437	107.18	108.38	107.18	182.86
Min	-17203	-8497	-17,126	-14,918	-16,954	-232	-232	-232	-6,987.00
25%	-37.00	-212	-31.00	-150	-7	-45	-45	-45	14.00
50%	0.00	0	9.00	103	128	-10	-10	-10	19.00
75%	72.00	19	116.00	304	266	33.25	31	33.25	37.00
Max	17,203.00	47	16,561	3,419	441	423	423	423	780.00
Range	34,406.00	1144	33,687	18,337	17,395	655	655	655	7,767.00
RMSE	1,338.86	1,137.24	1,546.61	1,301.32	1390.4	106.9	108.4	106.9	254.41
MAE	361.48	336.66	379.76	453.64	325.52	70	71.11	69.93	220

Table 4-20.Estimation errors for industrial meter charges.

	Models with ancillary data					Models with no ancillary data			
	KNN	SVM	DT	LR	ANN	NGN	IDW	NNR	OK
Mean	1547.15	240.29	1,544.91	1,531.45	1,511	359.29	643.04	359.29	-1.87
Std. dev.	12,507.39	31.88	11,006	8,816	7,915	3,795	5,065	3,795	2,864
Min	1.00	54	1.00	-3,015	4	-253	-253	-253	-19,181
25%	156.00	239	175.00	240	595	-41	-42.75	-41	-525.00
50%	354.00	250	383.00	640	879	-11	-11	-11	0
75%	930.00	250	988.00	1,130	1,218	112	113	112	840.00
Max	251,528	252	249,722	123,673	122,640	59,536	70,093	59,536	11,854
Range	251,529.00	306	249,723	126,688	122,636	59,789	70,346	59,789	31,035.00
RMSE	8,821.67	1,555.19	9,995.00	9,416.48	9,387.18	4,585.10	5,104	3,810.32	2,864.24
MAE	1,156.58	597.66	877.32	1,329.20	1,315.75	438	721.44	381.21	1,627.79

The error range for commercial meter charges is between 655 for the IDW and point interpolation models that do not require ancillary data to more than 34,000 dollars for the KNN model (Table 4-19). The closest distribution to the error is the Johnson SU's distribution with the shape parameters a and b equal to -0.1687 and 0.7909 respectively (Figure 4-28).

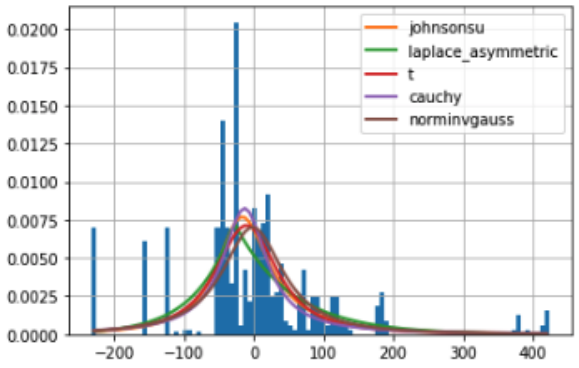

Figure 4-28. Statistical distributions that are close to the distribution of the error in estimating commercial meter charges with the 5 nearest geographic neighbors (NGN).

4.3.9.6 Industrial Meter Charges

The interpolation method that averages meter charges within 15 miles has the lowest MAE of 381.21 which is significantly lower than all other methods. The error range for industrial meter charges is between 59,789 for the model that averages the nearest five industrial meter charges to more than 250,000 for the KNN model (Table 4-20).

Although the SVM model has a higher MAE, it has a much lower range of 306 instead of nearly 60,000. It also has the lowest MRSE value of 1,555.19.

The closest distribution to the error is the Generalized Normal Distribution with the parameter beta equal to 0.1798.

4.4 Summary

In this chapter, several models were run with different configurations to minimize their error rates. For each model, the error values were retrieved along with the statistical distributions of the errors. In the next chapter, the research questions are answered in light of the results reported in this chapter.

Chapter 5 — Discussion and Conclusions

5.1 Discussion

Only 10% of retail tariffs in the database were complete and up to date. An assessment of the frequency of updates has shown a decline in terms of how often rates are being updated. When the complete tariffs were examined, many unusual values were observed. Two approaches were proposed to eliminate the outliers: a univariate method that assumes a normal distribution of all charges, and a multivariate distribution. The multivariate method, which is based on the Mahalanobis distance only removed between 5% to 10% of the tariffs. After manually checking the results, the multivariate method proved to be far more reliable.

The problem of estimating tariffs was broken down into estimating six different variables: residential energy and meter charges, commercial energy and meter charges, industrial energy, and meter charges. Ten types of models were constructed, and their configuration optimized to minimize their error rates. The models fall broadly in two categories: those that require ancillary data such as state power data, and those that do not.

For residential energy charges, the ordinary kriging with a linear variogram led the lowest MAE figures of 1 cent per kWh. This is followed by the KNN and decision trees with MAE values of 0.02 each.

For commercial energy charges, the lowest MAE values were 0.03 which were achieved with the KNN, decision tree and aerial interpolation models. It is worth noting that this last method consisted of using the state-average values of commercial electricity rates published by the Energy Information Administration.

For industrial energy charges, MAE values of 0.01 were achieved by using the KNN and decision tree models, with the aerial interpolation model recording a MAE value of 0.02. These results resemble what is observed with the estimation of commercial energy charges.

For residential meter charges, the decision tree model returned the lowest MAE of 1.78 followed by KNN with a MAE of 2.10 and then ordinary kriging with 2.26. The decision tree model showed the lowest error range of 39.

For commercial meter charges, the geo-statistics methods IDW and linear interpolation performed substantially better than all other methods with a MAE ranging between 69.93 and 71.11 and an error range in the interval 348 to 355.

For industrial meter charges, although the IDW and linear interpolation methods led to the lowest MAE values like it was observed in the estimation of commercial meter charges, the SVM model showed a slightly higher MAE of 597.66 (instead of 381.11), but a much lower error range of 306 (instead of 3,810) and mean of 240.29 (instead of 359.29).

On the research question of whether tariffs can be estimated anywhere in the US, it was found that in regions where there is data for regions within 15 miles, those charges can be estimated by using geo-statistics methods that do not require any ancillary data. In regions where there is no data available in nearby locations, ancillary data becomes necessary to build machine learning models. For the case of commercial and industrial energy charges, the state's average rate is a good estimate anywhere in the US.

These results on which class of algorithms perform best are different from what has been reported in the literature on different classes of problems. For instance, Kim et al. (2022) considered machine learning and geo-statistics methods to estimate real-estate

pricings in Seoul, South Korea and found that Artificial Neural Networks performed consistently better than Kriging and Inverse Distance Weighting.

On the question of how to model the error rates, it was shown that estimation errors can be approximated to known statistical distributions with known parameters. These statistical distributions can be used to factor uncertainty associated with tariff estimation in the overall project economics.

5.2 Contributions to Body of Knowledge

This research examined a broad range of statistical inference models, which have been used extensively to estimate unknown parameters in the fields of environmental studies, real-estate, social studies, etc. The methods explored demonstrated how those same techniques can be adapted to estimate tariff charges anywhere in the US. It highlights how the optimal modeling technique can be selected based on the type of charge to be predicted and the availability of data in the vicinity of the points of interest.

One key contribution of this study is the literature review which addressed a gap in the literature on what factors affect energy and meter charges in the US. One facet of this contribution is that the impact of market deregulation on prices in the United States remains inconclusive.

Finally, this study determined the statistical distribution of the estimation errors associated with different estimation algorithms. These statistical distributions will enable the economic analysis of clean energy projects with estimated tariffs.

5.3 Recommendations for Future Research

Researchers at the Department of Energy published a large dataset of load-profile data and real load data for residential and commercial buildings in locations that cover all

climate regions in the United States (Frick et al., 2019). Combining those datasets with the tariff estimation models proposed in this praxis opens countless research opportunities in the area of large-scale deployment of clean-energy projects.

This research provides insights into estimating retail tariffs anywhere in the United States. This opens the door to several research opportunities that require retail tariff data in large geographies where manual data collection is challenging. One such example would be assessing the impact of specific clean-energy interventions across the US (e.g., where are home batteries most profitable).

Lin et al. formulated an electric vehicle routing problem under time-dependent electricity prices (Lin et al., 2021). The study did not discuss how those prices are obtained. The findings in this research offer a mechanism to estimate those prices and run this type of routing problems in a more realistic setting.

This work looked at meter and energy charges only. Expanding this work to understand the factors that affect demand charges and determining which models perform best for their estimation will be greatly beneficial to accommodate the need for these charges. This is particularly important for demand-reduction applications (e.g., energy storage, power factor correction, etc.)

On the topic of outlier detection, only two methods were considered. Further investigation into removing outliers from the URDB database and understanding their influence on the performance of the various methods will further enhance their performance.

Finally, the scope of this research was the United States. Expanding the methods proposed here to other geographies in the world can help address fill the gaps in tariff data globally.

References

Aggarwal, C. (2017). An introduction to outlier analysis. Outlier analysis. Springer.

Alaboz, P., Dengiz, O., Demir, S., & Şenol, H. (2021). Digital mapping of soil erodibility factors based on decision tree using geostatistical approaches in terrestrial ecosystem. Catena, 207, 105634.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Anderson, J. E. (1981). Ridge estimation of house value determinants. Journal of Urban Economics, 9(3), 286-297.

ASEM (American Society for Engineering Management), About ASEM. May 5, 2022. Retrieved from https://www.asem.org/about

Aszemi, N. M., & Dominic, P. D. D. (2019). Hyperparameter optimization in convolutional neural network using genetic algorithms. International Journal of Advanced Computer Science and Applications, 10(6).

Balk, B., & Elder, K. (2000). Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. Water Resources Research, 36(1), 13-26.

Bekele, A., et al. "Comparative evaluation of spatial prediction methods in a field experiment for mapping soil potassium." Soil Science 168.1 (2003): 15-28.

Bernal, B., Molero, J. C., & De Gracia, F. P. (2019). Impact of fossil fuel prices on electricity prices in Mexico. Journal of Economic Studies.

Besag J. (1986). On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society: Series B (Methodological), Volume 48, Issue 3, Pages 259-279.

Bradshaw, J., & Chang, S. (2013). Past performance as an indicator of future performanc e: Selecting an industry partner to maximize the probability of program success. Defense Acquisition Research Journal, 20(1), 4980. Retrieved from http://dau.dodlive.mil/files/2014/11/ARJ_65-Bradshaw.pdf

Chaplain, C. T. (2011). Space acquisitions: DoD delivering new generations of satellites, but space system acquisition challenges remain. Hampton Roads International Sec urity Quarterly, 52.

Chen, F. W., & Liu, C. W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. Paddy and Water Environment, 10(3), 209-222.

Chen, N. (2022). House Price Prediction Model of Zhaoqing City Based on Correlation Analysis and Multiple Linear Regression Analysis. Wireless Communications and Mobile Computing, 2022.

Cheng, K. S., Lin, Y. C., & Liou, J. J. (2008). Rain-gauge network evaluation and augmentation using geostatistics. Hydrological Processes: An International Journal, 22(14), 2554-2564.

Comber A.; Zeng W. (2019). Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geography Compass, Volume 13, Issue 10.

Deason, J., & Schwartz, L. (2016). Higher Fixed Charges. Recovery of Utility Fixed Costs: Utility, Consumer, Environmental and Economist Perspectives, 565.

Denholm, P.; R.M. Margolis; S. Ong; B. Roberts. (2009). Break-Even Cost for Residential Photovoltaics in the United States: Key Drivers and Sensitivities. NREL TP-6A2-46909. Golden, CO: National Renewable Energy Laboratory.

Dorji, U. J., Plangprasopchok, A., Surasvadi, N., & Siripanpornchana, C. (2019, November). A machine learning approach to estimate median income levels of sub-districts in Thailand using satellite and geospatial data. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery (pp. 11-14).

Faruqui, A., & Leyshon, K. (2017). Fixed charges in electric rate design: A survey. The Electricity Journal, 30(10), 32-43.

Foysal, M., Ahmed, F., Sultana, N., Rimi, T. A., & Rifat, M. H. (2021). Convolutional Neural Network Hyper-Parameter Optimization Using Particle Swarm Optimization. In Emerging Technologies in Data Mining and Information Security (pp. 363-373). Springer, Singapore.

Fox, J. R. (2012). Defense acquisition reform, 1960-2009: An elusive goal (Vol. 51). Retrieved from https://www.fas.org/sgp/crs/natsec/R43566.pdf

Frick, N. M., Wilson, E., Reyna, J., Parker, A., Present, E., Kim, J., ... & Eckman, T. (2019). End-Use Load Profiles for the US Building Stock: Market Needs, Use Cases, and Data Gaps.

Geron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Gholami, V., Khaleghi, M. R., Pirasteh, S., & Booij, M. J. (2022). Comparison of self-organizing map, artificial neural network, and co-active neuro-fuzzy inference system methods in simulating groundwater quality: geospatial artificial intelligence. Water Resources Management, 36(2), 451-469.

Gholami, V., and Sahour, H. (2022). Prediction of groundwater drawdown using artificial neural networks. Environmental Science and Pollution Research, 1-14.

Guigues, V., Sagastizábal, C., & Zubelli, J. P. (2014). Robust management and pricing of liquefied natural gas contracts with cancelation options. Journal of Optimization Theory and Applications, 161(1), 179-198. doi:10.1007/s10957-013-0309-5.

Goulard, M., & Voltz, M. (1993). Geostatistical interpolation of curves: a case study in soil science. In Geostatistics Tróia'92 (pp. 805-816). Springer, Dordrecht.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Ho, W. K., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48-70.

Jalali, V., Khasheei, S. A., & Homaei, M. (2013). Comparing Geostatistics Techniques and Nonparametric K-Nearest Neighbor Technique For Predicting Soil Saturated Hydraulic Conductivity. Journal of Water and Soil Conservation, 20(5), 147-162.

Jasinska, E. Preweda, E. (2021). Statistical Modelling of the Market Value of Dwellings, on the Example of the City of Kraków. Sustainability 2021, 13, 9339.

Jing, C., Zhou, W., Qian, Y., & Yan, J. (2020). Mapping the urban population in residential neighborhoods by integrating remote sensing and crowdsourcing data. Remote Sensing, 12(19), 3235.

Kanevski, M., Gilardi, N., Mayoraz, E., & Maignan, M. (1999). Environmental spatial data classification with Support Vector Machines (No. REP_WORK). IDIAP.

Kim, J., Lee, Y., Lee, M. H., & Hong, S. Y. (2022). A Comparative Study of Machine Learning and Spatial Interpolation Methods for Predicting House Prices. Sustainability, 14(15), 9056.

Kuzyakova, I. F., Romanenkov, V. A., & Kuzyakov, Y. V. (2001). Geostatistics in soil agrochemical studies. Eurasian Soil Science C/c of Pochvovedenie, 34(9), 1011-1017.

Ktena, A., Panagakis, G., & Hivziefendic, J. (2019). A study of the retail electricity prices increasing trend in European retail electricity markets. In 2019 IEEE 60th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON) (pp. 1-6).

Langlois-Bertrand S.; Pineau P. (2018). Pricing the transition: Empirical evidence on the evolution of electricity rate structures in North America. Energy Policy, Volume 117, Pages 184-197.

Li J.; Heap A.; Potter A.; Daniell J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, Volume 26, Issue 12, Pages 1647-1659.

Lin, B., Ghaddar, B., & Nathwani, J. (2021). Electric vehicle routing with charging/discharging under time-variant electricity prices. Transportation Research Part C: Emerging Technologies, 130, 103285.

Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. Scientific Programming, 2022.

Lu, G. Y., & Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. Computers & geosciences, 34(9), 1044-1055.

Lu, Q., & Jiang, T. (2001). Pixon-based image denoising with Markov random fields. Pattern Recognition, 34(10), 2029-2039.

MacKay, A., & Mercadal, I. (2022). Deregulation, Market Power, and Prices: Evidence from the Electricity Sector. Available at SSRN 3793305.

Mohammadi, H. (2009). Electricity prices and fuel costs: Long-run relations and short-run dynamics. Energy Economics, 31(3), 503-509.

Moreno, B., López, A. J., & García-Álvarez, M. T. (2012). The electricity prices in the European Union. The role of renewable energies and regulatory electric market reforms. Energy, 48(1), 307-313.

Necoechea-Porras, P. D., López, A., & Salazar-Elena, J. C. (2021). Deregulation in the energy sector and its economic effects on the power sector: A literature review. Sustainability, 13(6), 3429.

Ong, S. Denholm, P. Doris, E. (2010). The Impacts of Commercial Electric Utility Rate Structure Elements on the Economics of Photovoltaic Systems. NREL TP-6A2-46782. Golden, CO: National Renewable Energy Laboratory.

Ong, S. McKeel R. (2012, May 13-17). National Utility Rate Database. World Renewable Energy Forum. Denver, Colorado, United States.

Oosthuizen, A. M., Inglesi-Lotz, R., & Thopil, G. A. (2022). The relationship between renewable energy and retail electricity prices: Panel evidence from OECD countries. Energy, 238, 121790.

Pazzalia, J. (2022). Sparking Debate: Private or Public? The Effect of Ownership on Electric Utility Performance.

Pedregosa, F., Varoquaux, Gael, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Ping, J. L., Green, C. J., Zartman, R. E., & Bronson, K. F. (2004). Exploring spatial dependence of cotton yield using global and local autocorrelation statistics. Field Crops Research, 89(2-3), 219-236.

Ranjit, M. P., Ganapathy, G., Sridhar, K., & Arumugham, V. (2019, July). Efficient deep learning hyperparameter tuning using cloud infrastructure: Intelligent distributed hyperparameter tuning with bayesian optimization in the cloud. In 2019 IEEE 12th international conference on cloud computing (CLOUD) (pp. 520-522). IEEE.

Rehman, S., & Ghori, S. G. (2000). Spatial estimation of global solar radiation using geostatistics. Renewable Energy, 21(3-4), 583-605.

Razeghi, G., Shaffer, B., & Samuelsen, S. (2017). Impact of electricity deregulation in the state of California. Energy Policy, 103, 105-115.

Rintamäki, T., Siddiqui, A. S., & Salo, A. (2017). Does renewable energy generation decrease the volatility of electricity prices? An analysis of Denmark and Germany. Energy Economics, 62, 270-282.

Sekulić, Aleksandar, Milan Kilibarda, Gerard Heuvelink, Mladen Nikolić, and Branislav Bajat. "Random forest spatial interpolation." Remote Sensing 12, no. 10 (2020): 1687.

Shahneh, M., Oymak, S., Magdy, A. (2021). A-GWR: Fast and Accurate Geospatial Inference via Augmented Geographically Weighted Regression. In Proceedings of the 29th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '21). Association for Computing Machinery, New York, NY, USA, 564–575.

Shi, Q., Abdel-Aty, M., & Lee, J. (2016). A Bayesian ridge regression analysis of congestion's impact on urban expressway safety. Accident Analysis & Prevention, 88, 124-137.

Sitharam, T. G., Samui, P., & Anbazhagan, P. (2008). Spatial variability of rock depth in Bangalore using geostatistical, neural network and support vector machine models. Geotechnical and Geological Engineering, 26(5), 503-517.

Suchenwirth, L., Stümer, W., Schmidt, T., Förster, M., & Kleinschmit, B. (2014). Large-scale mapping of carbon stocks in riparian forests with self-organizing maps and the k-nearest-neighbor algorithm. Forests, 5(7), 1635-1652.

Taber, J. T., Chapman, D., & Mount, T. D. (2005). Examining the effects of Deregulation on retail electricity prices.

Tchuente, D., & Nyawa, S. (2021). Real estate price estimation in French cities using geocoding and machine learning. Annals of Operations Research, 1-38.

Trujillo-Baute, E., del Río, P., & Mir-Artigues, P. (2018). Analysing the impact of renewable energy regulation on retail electricity prices. Energy Policy, 114, 153-164.

Tuba, E., Bacanin, N., Strumberger, I., & Tuba, M. (2021). Convolutional neural networks hyperparameters tuning. In Artificial intelligence: theory and applications (pp. 65-84). Springer, Cham.

US Department of Energy, Energy Information Administration, Today in Energy (2019, August 15). Investor-owned utilities served 72% of US electricity customers in 2017. Retrieved from https://www.eia.gov/todayinenergy/detail.php?id=40913

US Environmental Protection Agency (2022). Inventory of US Greenhouse Gas Emissions and Sinks: 1990-2020 (Report No. 430-R-22-003). US Environmental Protection Agency. https://www.epa.gov/system/files/documents/2022-04/us-ghg-inventory-2022-main-text.pdf

Wackernagel, H. (2003). Ordinary kriging. In Multivariate geostatistics (pp. 79-88). Springer, Berlin, Heidelberg.

Webster, R., & Oliver, M. A. (2007). Geostatistics for environmental scientists. John Wiley & Sons.

Willis, H. L., & Philipson, L. (2018). Understanding electric utilities and de-regulation. CRC Press.

Wohlberg, B., Tartakovsky, D. M., & Guadagnini, A. (2005). Subsurface characterization with support vector machines. IEEE transactions on geoscience and remote sensing, 44(1), 47-57.

Zarnikau, J., & Whitworth, D. (2006). Has electric utility restructuring led to lower electricity prices for residential consumers in Texas?. Energy Policy, 34(15), 2191-2200.

Appendix A

Table A-1. Mix of electricity generation in all US states in 2020.

State	Nuclear	Coal	Gas	Petroleum	Hydropower	Geothermal	Solar	Wind	Other
AK	32	16	40.5	0	8.8	0	0.3	0	2.5
AL	0	12.6	37.6	16	30.5	0	0	2.8	0.6
AR	28.8	12.5	46.4	0	5.9	0	5.5	0.6	0.2
AZ	28.5	29.1	31.8	0.1	8.1	0	0.5	0	1.9
CA	8.4	0.1	48.1	0	11	6.1	15.7	7	3.4
CO	0	35.9	34	0	3.3	0	2.8	23.5	0.4
CT	38.2	0	57	0	1	0	0.6	0	3.3
DC	0	2	95.2	0.3	0	0	1.2	0.1	1.2
DE	0	0	64.8	0	0	0	8.8	0	26.4
FL	11.8	6.7	75.3	0.7	0.1	0	2.6	0	2.8
GA	27.6	11.7	49.1	0.2	3.1	0	3.3	0	5
HI	0	12.5	0	66.1	1	2.2	5.7	6.2	6.3
IA	0	0.1	20.7	0	58.7	0.5	2.9	14.2	2.9
ID	57.8	17.9	14	0	0.1	0	0.1	9.9	0.4
IL	0	53.2	37.6	0.1	0.3	0	0.5	7.3	1.1
IN	4.9	23.8	11.7	0.2	1.5	0	0	57.5	0.3
KS	19.5	31.2	5.6	0.2	0.1	0	0.1	43.3	0.1
KY	0	68.8	23.1	0.1	7.3	0	0.1	0	0.7
LA	16.6	3.8	72.2	3.4	1.4	0	0.1	0	2.5
MA	0	0.6	16.9	0.4	34.4	0	0.3	24	23.4
MD	41.8	9.3	38.9	0.2	4.7	0	1.7	1.5	1.8
ME	0	0	76.1	0.2	3.2	0	8.5	1.5	10.5
MI	28.9	26.7	33.8	0.9	0.9	0	0.2	6.4	2.2
MN	26	24.9	19.7	0.1	2	0	3.1	21.6	2.7
MO	9.8	7	80.4	0	0	0	0.7	0	2.1

MS	10.5	70.5	10.9	0.1	3	0	0.1	4.7	0.2
MT	0	36	1.7	1.9	46.6	0	0.1	12.6	1.1
NC	16.8	51.2	3.9	0	4	0	0.2	23.6	0.2
ND	0	4.8	66.1	0	4.8	10.2	13	0.8	0.2
NE	59	0.8	21.8	0.2	8.9	0	0	3.1	6.1
NH	43.5	1.5	50.3	0.1	-0.2	0	2.6	0	2.2
NJ	0	37.2	35.7	0.5	0.5	0.2	4.9	20.9	0.1
NM	29.1	0.1	40.1	0.2	23.6	0	0.8	3.8	2.3
NV	34.1	16.8	33.7	0.1	5.3	0	7.2	0.4	2.4
NY	0	57.3	3.6	0.1	8.1	0	0	30.8	0.1
OH	15	37.2	44	0.9	0.3	0	0.2	1.9	0.6
OK	0	7.1	52.4	0	4.6	0	0.1	35.4	0.4
OR	0	2.5	28.9	0	51.9	0.2	1.7	13.2	1.6
PA	33.1	10.3	52.3	0	1.3	0	0.1	1.7	1.2
RI	0	0	91.8	0.1	0	0	2.6	2.9	2.6
SC	55.8	12.7	24.8	0.1	2.5	0	1.8	0	2.4
SD	0	9.7	6.9	0.1	50.5	0	0	32.9	0
TN	47.3	18.4	20.2	0.1	12.4	0	0.4	0.1	1
TX	8.7	16.6	52.7	0	0.4	0	1.7	19.6	0.4
UT	0	61.5	25.4	0.1	2.6	1	6.7	2.2	0.6
VA	0	0	0.1	0	57.8	0	8	16.2	17.8
VT	29.5	3.7	60.7	0.2	0.5	0	1.4	0	3.9
WA	8.3	4.5	12.4	0	66.1	0	0	7.3	1.4
WI	0	88.4	5	0.3	3.1	0	0	3.3	0
WV	16	39	34.8	0.2	4.7	0	0.2	2.9	2.2
WY	0	80	4.3	0.1	2.6	0	0.4	12.3	0.2

Table A-2. The number of electricity customers in all US states in 2020.

State	Residential customers	Commercial customers	Industrial customers
AK	315,208	61,993	1,129
AL	2,280,741	371,888	7,240
AR	1,413,490	197,869	35,978
AZ	2,896,339	331,229	7,595
CA	13,834,719	1,725,533	148,130
CO	2,400,357	384,518	15,209
CT	1,521,112	154,894	4,130
DC	290,466	26,672	1
DE	446,276	56,764	878
FL	9,731,237	1,256,569	22,587
GA	4,487,431	592,220	23,822
HI	442,002	59,734	816
IA	1,403,386	243,762	9,507
ID	782,559	114,707	28,759
IL	5,339,610	628,868	5,561
IN	2,920,266	363,465	19,383
KS	1,282,532	236,430	23,979
KY	2,013,910	312,014	5,982
LA	2,112,928	296,222	19,276
MA	2,817,549	411,448	10,877
MD	2,376,983	256,738	8,966
ME	717,559	97,575	2,681
MI	4,423,595	546,115	5,580
MN	2,464,753	303,702	9,042
MO	2,833,918	387,872	10,108
MS	1,308,149	237,370	10,343
MT	522,382	110,977	11,414

NC	4,695,096	710,220	9,822
ND	387,506	76,834	8,933
NE	864,842	155,282	62,716
NH	633,234	109,068	3,180
NJ	3,618,587	526,725	11,629
NM	905,885	145,459	9,436
NV	1,226,566	169,743	3,316
NY	7,239,162	1,143,347	7,535
OH	5,014,959	636,519	19,746
OK	1,795,629	290,192	20,468
OR	1,785,131	239,645	26,353
PA	5,448,109	720,919	15,462
RI	441,573	60,057	1,692
SC	2,377,020	395,288	3,714
SD	407,532	74,236	4,109
TN	2,930,482	501,589	1,015
TX	11,515,333	1,539,553	264,126
UT	1,143,136	135,113	11,001
VA	3,506,844	437,477	3,693
VT	316,948	59,554	252
WA	3,168,238	389,802	26,741
WI	2,742,424	360,316	5,621
WV	862,279	146,769	11,444
WY	276,029	57,958	11,334

Table A-3. Average price of electricity to customers for all US states in $/kwh.

State	Residential	Commercial	Industrial
AK	0.23	0.20	0.16
AL	0.13	0.12	0.06
AR	0.10	0.09	0.06
AZ	0.12	0.10	0.06
CA	0.20	0.18	0.14
CO	0.12	0.10	0.07
CT	0.23	0.17	0.13
DC	0.13	0.12	0.08
DE	0.13	0.09	0.07
FL	0.11	0.09	0.07
GA	0.12	0.10	0.06
HI	0.30	0.28	0.24
IA	0.12	0.10	0.06
ID	0.10	0.08	0.06
IL	0.13	0.09	0.07
IN	0.13	0.11	0.07
KS	0.13	0.10	0.07
KY	0.11	0.10	0.05
LA	0.10	0.09	0.05
MA	0.22	0.16	0.15
MD	0.13	0.10	0.08
ME	0.17	0.13	0.09
MI	0.16	0.12	0.07
MN	0.13	0.10	0.08
MO	0.11	0.09	0.07
MS	0.11	0.10	0.06
MT	0.11	0.11	0.05

NC	0.11	0.09	0.06
ND	0.10	0.09	0.07
NE	0.11	0.09	0.07
NH	0.19	0.15	0.13
NJ	0.16	0.12	0.10
NM	0.13	0.10	0.06
NV	0.11	0.07	0.06
NY	0.18	0.15	0.06
OH	0.12	0.10	0.06
OK	0.10	0.08	0.05
OR	0.11	0.09	0.06
PA	0.14	0.09	0.06
RI	0.22	0.16	0.16
SC	0.13	0.10	0.06
SD	0.12	0.10	0.08
TN	0.11	0.11	0.05
TX	0.12	0.08	0.05
UT	0.10	0.08	0.06
VA	0.12	0.08	0.06
VT	0.20	0.16	0.11
WA	0.10	0.09	0.05
WI	0.14	0.11	0.07
WV	0.12	0.09	0.06
WY	0.11	0.10	0.07

Table A-4. Average electricity revenue per customer for all US states in US dollars.

State	Residential	Commercial	Industrial
AK	1,496	7,970	183,531
AL	1,727	6,616	249,520
AR	1,324	4,835	27,446
AZ	1,640	8,889	112,791
CA	1,403	10,870	45,887
CO	1,055	5,364	75,935
CT	1,939	11,933	90,517
DC	1,067	30,285	14,881,000
DE	1,405	6,604	156,929
FL	1,544	6,511	52,496
GA	1,559	7,542	74,611
HI	1,952	12,765	977,868
IA	1,293	4,741	165,396
ID	1,140	4,263	19,887
IL	1,128	6,616	486,074
IN	1,444	6,782	152,181
KS	1,362	6,532	33,625
KY	1,399	5,985	246,928
LA	1,393	6,693	91,847
MA	1,586	9,009	82,959
MD	1,494	10,019	29,457
ME	1,149	4,913	86,822
MI	1,318	7,611	332,999
MN	1,225	7,391	166,069
MO	1,384	6,433	86,803
MS	1,537	5,767	83,361
MT	1,158	4,455	20,412

NC	1,421	5,617	165,970
ND	1,359	7,796	82,368
NE	1,313	5,203	13,613
NH	1,440	5,694	77,249
NJ	1,315	8,279	57,962
NM	1,040	5,939	53,735
NV	1,324	5,258	201,837
NY	1,326	8,786	122,224
OH	1,288	6,467	146,059
OK	1,309	5,042	45,874
OR	1,228	5,917	33,756
PA	1,379	4,170	193,658
RI	1,569	9,427	59,109
SC	1,658	5,456	403,679
SD	1,461	6,106	55,539
TN	1,508	7,048	1,073,847
TX	1,591	7,165	24,009
UT	963	6,976	51,904
VA	1,581	9,339	296,980
VT	1,329	4,969	608,325
WA	1,149	6,250	42,640
WI	1,193	6,697	287,432
WV	1,489	4,456	75,843
WY	1,159	5,530	55,418