UNDERSTANDING RESIDENTIAL ELECTRICITY CONSUMPTION PATTERNS AND FORECASTING TRENDS BASED USING MACHINE LEARNING AND OPTIMIZATION

July 2024

A thesis submitted to

The University of Kent

in the subject of Management Science

for the degree

of Doctor of Philosophy

by

Zhifeng Guo

Abstract

Energy is the most important factor driving the progress and development of modern society since Industrial Revolution, and electricity is closely related with human activities. As the global greenhouse gas content continues to increase, reducing carbon emissions has become an important theme in the current global development. Based on this, The Intergovernmental Panel on Climate Change (IPCC) has set different pathways that limit global warming to 1.5 °C and some specific emission standards in different countries. Fossil fuels (coal, oil and gas) are by far the largest contributor to global climate change, while traditional thermal power generation is an important part of fossil fuels, so reducing non-essential electricity consumption can help to reduce fossil fuel emissions from the perspective of demand-side management. To achieve this goal, it is necessary to analyze people's electricity consumption patterns, the social-economical characteristics of electricity end users, and influencing factors that affect electricity consumption. Therefore, this thesis focuses on three aspects of electricity consumption. First, we focus on developing novel machine learning-based approaches for characterizing residential electricity consumption patterns and predicting electricity consumption pattern based on household characteristics. In addition, natural disasters are also important factors that lead to changes in electricity usage, such as earthquake, tsunami and pandemic. COVID-19 has caused a huge impact on people's lives and therefore lead to significant changes in electricity consumption. Second, we study the effect of COVID-19 on electricity consumption, and predict daily consumption over 1-6 weeks based on historical consumption data and population flow indicators in seven America cities by establishing a set of Bayesian structured time series models. Third, we study personalized electricity customer segmentation problem. Identifying different group of customers with both similar electricity usage and social-economical background is key to implement demand-side management programs. We proposed a two-stage constrained clustering framework based on mixed integer linear programming and clustering algorithm, Lagrangian relaxation procedure are proposed to efficiently solve large problem instances. A realistic case studies are analyzed to demonstrate how proposed framework can be used to support personalized customer segmentation problem.

Acknowledgement

As I approach the end of my PhD journey at Kent Business School, University of Kent, I find myself reflecting on the past five years, which have been marked by both unique challenges and profound growth, especially considering the extraordinary circumstances brought on by the COVID-19 pandemic shortly after I began. This unprecedented situation has made my experience both unforgettable and deeply transformative, and I am grateful for the resilience it has instilled in me.

First and foremost, I would like to express my sincere gratitude to my main supervisor, Jesse O'Hanley. His unwavering support and guidance from the very beginning of my application process have been invaluable. His profound knowledge and serious approach to research have inspired me greatly. It has been a privilege to work with him, and I am immensely thankful for the encouragement and assistance he provided throughout my PhD.

I would also like to extend my heartfelt thanks to my secondary supervisor, Dr. Stuart Gibson, for his insightful suggestions and constructive feedback on my papers. His expertise has significantly enriched my work and helped refine my research.

I am grateful to Dr. Zhen Zhu at Kent Business School for serving as my internal examiner and to Professor Huiyu Zhou from the School of Computing and Mathematical Sciences at the University of Leicester for acting as my external examiner. Their thoughtful comments and suggestions on my thesis have greatly contributed to its improvement, and I truly appreciate their time and effort.

Additionally, I express my deep gratitude to the staff and colleagues at Kent Business School whose congratulations and support have been a source of motivation throughout my PhD journey.

Lastly, I want to convey my special thanks to my parents. Their unwavering support and belief in my dreams have been instrumental in my pursuit of this academic endeavor. I am forever grateful for their love and encouragement.

Table of Contents

List of Figures

List of Tables

1. Introduction

This chapter introduces the research background and research topic covered in this thesis, and discusses the specific issues studied in this field, identifies the research questions that have been posed as well as the academic contributions of this thesis, and finally, outlines the structure of this thesis.

1.1 Research background

According to the IPCC, global climate change is currently a major obstacle to human's sustainable development, especially in recent years when extreme weather due to has become increasingly commonplace global climate change, such as heatwaves, heavy precipitation, droughts and tropical cyclones, which pose a great danger to human society (IPCC, 2022). Global carbon emission overload is a major factor contributing to extreme climate change on Earth, with fossil fuels being the most significant source of carbon emissions. 2030 Agenda for Sustainable Development set up 17 Sustainable Development Goals (SDGs) to call for action by all countries to promote prosperity and protecting the planet (Figure 1.1) and three out of the seventeen SDGs are highly related with energy, such as, SDG 7 (affordable and clean energy); SDG 11 (sustainable cities and communities); and SDG 13 (climate action).

Figure 1.1. The Sustainable Development Goals (United Nations)

Since the 1970s, climate change has received more and more attention worldwide, many countries start to take actions to combat climate change, such as the Global Climate Conference held in Geneva in

1979, the establishment of IPCC in 1988 and the Kyoto Protocol in 1997, which set greenhouse gas emission standards for each country. Many countries have also designated corresponding legal regulations, which point the direction for further reducing carbon emissions and provide a policy basis for achieving sustainable development of society.

Figure 1.2 and 1.3 give an overview of the type of primary energy source used for total energy supply and electricity generation worldwide. From viewpoint of energy supply, more than 80% of the global energy usage is based on fossil fuels (Coal with about 27% and oil with 32% come next with natural gas (23%)). In terms of electricity production, Fossil fuel based thermal power station is used to produce almost 64% of world electricity, although the proportion of Hydro and renewables energy is expanding in recent years. Both Hydro and non-hydro renewables only account for about 25% of the global electricity production. In particular, for developing countries, thermal power generation still accounts for a relatively large proportion. Hence the greenhouse gas emissions caused by traditional power generation is still very large, therefore reducing electricity consumption would help to reduce greenhouse gas emissions, which plays a positive role in global sustainable development. Since electricity is most closely related to people's daily lives in modern society, hence reducing electricity consumption without cause inconvenience is a challenging task.

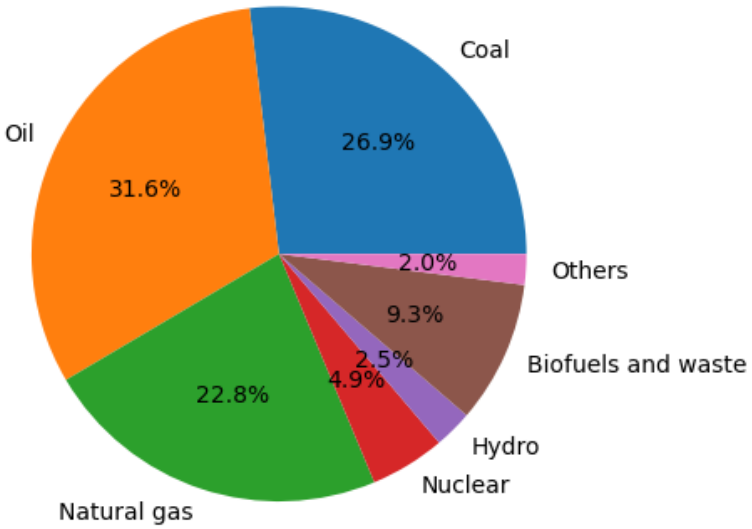


Figure 1.2. World primary energy supply in 2018 (IEA, 2020)

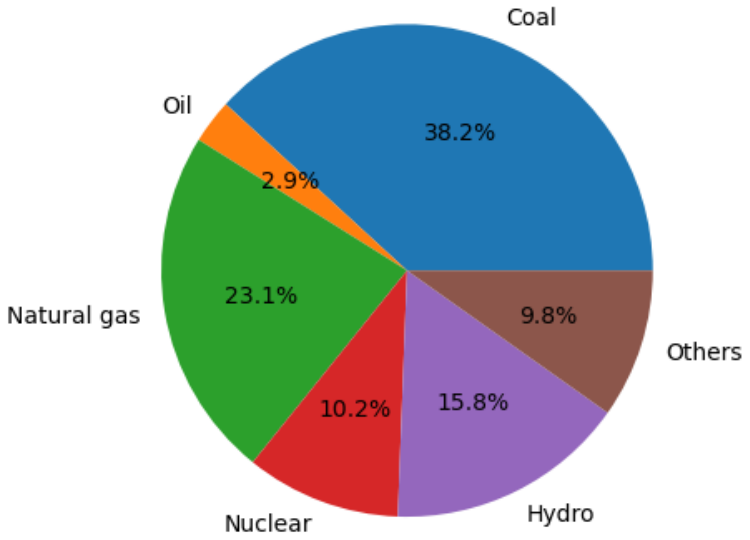


Figure 1.3. Electricity production by primary energy source in 2018 (IEA, 2020)

The electricity market plays a significant role in reducing carbon emissions by promoting the adoption of renewable energy sources (e.g, hydroelectric power, wind power, and solar power), improving grid efficiency, considering environmental externalities (carbon tax), and encouraging consumer behavior change, etc, there are generally two main types of markets in electricity sector, such as wholesale market and retail market. The wholesale market is where electricity is traded in large volumes between producers and sellers (retailers), while the retail market is where retailers sell electricity to end consumers (Hampton *et al*. 2022). These market structures can vary significantly by region and are influenced by factors such as regulatory frameworks, technological advancements, and the integration of renewable energy sources.

There has been extensive research on the relationship between wholesale electricity markets and carbon emissions. This research generally focuses on how market structures, pricing mechanisms, and regulatory frameworks influence the type of electricity generation, investing in renewable energy and low-carbon technologies, which in turn affect carbon emissions. For instance, Wakiyama and Zusman (2021) showed that wholesale reforms, which focus on the power generation and transmission sector, have a more significant impact on reducing CO_2 emissions by increasing energy efficiency and renewable energy use. Green (2008) focused on electricity wholesale markets in the United States and Europe. They concluded that the U.S. standard market design is likely to give better results than the

European models in many areas. Arcos-Vargas *et al*. (2020) studied the effect of the new photovoltaic (PV) facilities on Iberian electricity market in 2017. The installation of new PV facilities could help to decrease CO_2 emissions, and increase economic surpluses, which contribute to the Government's compliance with the commitments in the Paris Agreement. MacIver *et al*. (2021) analyzed the impact of increasing interconnection levels between Great Britain and its European neighbors, their results showed this could reduce emissions in the UK but increase total European emissions due to different carbon pricing across different countries. Andrianesis *et al*. (2021) surveyed the impact of Renewable Energy Sources (RES) on the Italian electricity market based on based on hourly zonal micro-data for 2018, and RES could help to achieve annual savings of carbon emissions nationwide up to 22 Mt CO_2.

However, retail market has often been overshadowed by the more prominent discussions surrounding wholesale market design in the context of electricity supply industry restructuring, despite the advancements in technology and changes in energy policy aimed at promoting competition and efficiency in retail electricity markets. The main reason is that in the power sector, electricity had traditionally been dominated by a monopoly integrated horizontally from production to supply to the end users, and people are accustomed to electricity usage under the monopoly model and lack interaction with electricity retailers. For example, under normal situation, retailers only tell consumers the total amount of electricity usage, rather than the specific load profiles, and consumers cannot review their own electricity consumption behaviors without their load profiles and related price. Hence, this thesis mainly focuses on the research problems in retail market.

In the retail electricity market, there are mainly two types of entities: retailers and customers. Therefore, current research mainly focuses on customer segmentation, pricing strategy, electricity consumption forecasting, etc. Such as Ramos *et al*. (2015) applied several clustering algorithms to obtain the typical load profiles customers' characterization of medium voltage electric consumers, and four distinct typical load profiles were identified. The electricity consumption patterns of microgrid customers were studied by Williams *et al*. (2018), k-means clustering algorithm was applied to cluster 821 customers into 5 five typical load profiles based on daily electricity usage data. Their results could help Microgrid developers optimize power system design given different load profiles. Motlagh *et al*. (2019) applied a feature-based clustering algorithm to group residential electricity customers into 12 clusters, the proposed algorithm deal with the challenge of high dimensionality of time series by feature extraction. Density-

based spatial clustering of applications with noise (DBSCAN) is used to obtain typical load profiles. On this basis, time-of-use (TOU) retail price was formulated as a mixed integer nonlinear programming model (Yang *et al*. 2018). Other clustering algorithm such as self-organizing maps could be found in (Chicco *et al*. 2004). The effect of Time-of-use (TOU) pricing on cost savings for industrial customers in the U.S. by Wang and Li (2015). The result showed cost savings is inconsistent, ranging from −72.0% to +82.6%, when industrial customers changed from flat rates to TOU pricing program. Nezamoddini and Wang (2017) studied the effects of different pricing strategies such as Real-time pricing (RTP) and time-of-use (TOU) pricing on customer's cost savings in the U.S. However, they found the effect is highly program dependent.

In addition, customers could choose preferred retailers based on user experience, and retailers could also develop different electricity products and pricing strategies for their customers. For instance, residential customers should select their preferred energy provider facing different electricity products, 4968 experimental choices from 414 German retail consumers were analyzed by Kaenzig *et al*. (2013), they found price and electricity mix were the most important energy product attributes.

Understanding users' behavior patterns and values can help promote changes in energy-saving behaviors. People may make more reasonable decisions about their energy related activities when immediate information about energy usage data is available. Kopsakangas-Savolainen *et al*. (2017) have indicated that such information can lead to behavioral changes, resulting in electric savings of up to 17%. Albani *et al*. (2017) found understanding customer preferences could help to promote energy-saving behaviors and pay for smart meter services based on more than 1500 respondents in the retail electricity market in Switzerland.

Forecasting is of great significance to electricity retailers, so retailers need to make purchasing plans based on electricity demand forecasts and submit it to the power trading system one day in advance. Therefore, accurate forecasts will increase retailers' profits, while adverse forecasts will lead to losses. In terms of forecasting, many studies focus on predicting the total electricity consumption in a region, such as aggregated electricity consumption from all customers. Such as Bianco *et al*. (2009) developed a linear regression model to forecast a long-term trend in Italy from 1970 to 2007, and historical electricity consumption, gross domestic product (GDP), gross domestic product per capita (GDP per capita) and population were included as explanatory variables. The proposed model could get an error

of ±1% in the best situation and ±11% for the worst situation. Some studies predict the electricity usage of a building or a household. For instance, Chae *et al*. (2016) applied artificial neural network model to forecast electricity usage in commercial buildings and achieved an accuracy of around 90%. Gajowniczek and Ząbkowski (2014) applied SVM and MLP neural network to forecast hourly electricity usage for individual household.

There are also studies that predict different typical load profiles in a region, and then combine them to get the electricity usage forecast for the region. For instance, Andersen *et al*. (2013) forecast aggregated annual electricity consumption by combining projections of different categories of customers in Denmark. Similarly, Imani and Ghassemian (2018) make predictions for each cluster, and then combine them to get the get short-term load forecasting, their results showed the better performance of the proposed method compared to forecasting aggregated load profiles.

The fast development of big data and cloud storage technology in recent years has made it possible to store a large amount of electricity consumption data, which provides an opportunity to study electricity consumption problem more deeply so that more reasonable policies can be specified to reduce electricity consumption and then carbon emissions. The main strategies to adjust electricity consumption in the power sector consist of supply-side management and demand-side management. Supply-side management focuses on actions taken to ensure the generation, transmission and distribution of electricity are conducted efficiently in liberalised energy markets through the reform of the electricity market. It covers every aspect of supply-side management at each stage of the supply chain of electricity, e.g., energy resources used (clean coal technologies, fuel substitution and renewable energy use), generating plants (operational improvements in existing plants, upgrading units and cogeneration), transmission and distribution of electricity (including lines, substations and on-site generation). The main purpose is to stimulate the electricity market, enhance the supply of electricity and stable operation of power system and so on.

Demand-side management, on the other hand, is to provide more flexible paradigm to reduce electricity usage, and has been proved be one of the most successful electricity consumption adjustment strategies. According to US Federal Energy Regulatory Commission, Demand Response (DR) can be defined as "*Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower*

electricity use at times of high wholesale market prices or when system reliability is jeopardized" (Siano, 2014). DSM use a range of programs planned to encourage end-users to adjust their electricity usage, and it can be classified into three groups according to the party that initiates the demand reduction action, such as rate-based or price DR programs, incentive or event-based DR programs and demand reduction bids. Price-based strategy mainly provides users with different prices at different times to adjust their electricity consumption, examples of this kind of scheme are Time of Use rates (TOU), Critical Peak Pricing (CPP) and Real Time Pricing (RTP). Incentive or event-based programs refer to the program there is a certain contractual agreement between the electricity end users and program administrator (e.g., an aggregator, service provider etc.,). In Incentive or event-based programs, the program administrator will provide certain incentives if customer achieves the promised change in electricity consumption. Example of this kind of schemes includes: Direct Load Control (DLC), emergency demand response programs, capacity market programs, interruptible/curtailable program, ancillary services market programs. Demand reduction bids refer to the scheme where electricity customers offer bids to reduce electricity usage when wholesale market is profitable, examples of demand reduction bids are demand bidding and buyback programs.

Both supply-side management and demand-side management play a positive important role in promoting energy conservation and energy efficiency, hence reducing carbon emission. For example, these methods can effectively shift the peak load, reduce the maximum load running time of the power plant, thus reducing the operating cost, and on the other hand, can also guarantee the smooth operation of the grid and reducing the number of corresponding accidents.

Whether it is supply-side management or demand-side management, an effective demand-side management programs often relies on the following aspects:

The first aspect is electricity consumption patterns, whether it is the implementation of a price based or incentive based strategy. Understanding residential electricity consumption pattern would help to specify detailed price-based scheme. For example, typical load profiles provide much information about how electricity usage change over time in each group, this would definitely help to make personalized Time of Use rates and Critical Peak Pricing for different households. Besides, there is a close connection between electricity consumption pattern and various socio-demographic and dwelling characteristics, since electricity consumption patterns mainly depend on people's electricity consumption behavior.

Behaviors are in part governed by household size and socio-economic status. These characteristics and past usage can be used to predict electricity usage patterns of potential new customers, which can enable electricity providers and energy service company to make appropriate price recommendations.

The second aspect is electricity consumption forecasting. Power station and energy management company need to forecast short-term and long-term trend of future power load based on historical usage, accurate forecasting not only has an important guiding role for power plants to dispatch electricity production in energy market (day-ahead market and intraday markets), which is essential to maintain the smooth operation of the power grid, but also has important significance for power service companies to implement demand-side management services. For example, energy service companies can develop different pricing strategies to decrease peak demand, or when the summer heat arrives, power station could make appropriate preparations in advance to ensure peak demand and smooth operation of the power grid if the power station could accurately forecast it in advance. In addition, peak demand and off-peak demand are affected not only by the weather, the climate, the season, the day of the week and other factors but also by disasters and pandemic (such as floods, earthquakes, wildfires and COVID-19). Besides, the availability of renewable energy sources (rooftop solar PV units, wind turbines, biomass generators, fuel cells, etc), smart metering technologies, and storage (battery storage, electric vehicles (EV) chargers) also brings new challenges to forecasting electricity consumption.

The third aspect is electricity customer segmentation, customer segmentation is a classical topic that has been studied for many years. In the energy filed, it has been utilized for different purposes, such as understanding energy consumption pattern and defining tariff designs, since it is impossible to specify a separate price strategy or energy-saving incentives for every electricity consumption customer, so the electricity consumption pattern is very important for the formulation of price strategies, on contrary, we could set the price strategy for a group of customers with the same electricity consumption patterns. Clustering algorithm have been extensively used in different to analyze underlying energy consumption patterns of collected data. For example, consumers could be grouped into high consumption groups, medium consumption groups, and low consumption groups, or they can be divided into finer groups based on energy demand. Energy service company could achieve profit maximization by adopting personalized pricing strategies for different groups. User profile could be obtained by combining corresponding socio-economic features with consumption patterns. However, most clustering

algorithms fail to divide customers into similar groups not only with similar energy consumption but also similar social-economic background.

Based on this, the research of this thesis has important implications for implementation of more accurate demand-side management strategies and reducing carbon emissions. In addition, the findings of this thesis provide useful suggestions for decision making on the impact of pandemic on electricity consumption. A more detailed research topics is described in section 1.2.

1.2 Research topics

This section gives a more detailed description of research topics addressed in this thesis, namely, prediction of residential consumption patterns, electricity demand forecasting in the context of COVID-19 and electricity customer segmentation.

1.2.1 Prediction of residential consumption patterns

Electricity consumption pattern is mainly a typical load profile obtained by clustering on electricity consumption data, In recent years, with the development of artificial intelligence and smart meter technology, it has been possible to cluster and analyze electricity consumption data at different time intervals (e.g., intra-day, day, week, month and year), and many machine learning algorithms could be applied to this field, such as: k-means clustering (MacQueen 1967), hierarchical clustering (Johnson 1967), Gaussian Mixture Models (McLachlan and Basford 1988), spectral clustering (Ng *et al*. 2001), Nonnegative Matrix Factorizations (Paatero and Tapper 1994), self-organizing maps (SOM) (Kohonen 1990), and ensemble clustering (Yu *et al*. 2012). Residential electricity consumption could be affected by many factors, such as household size and socio-economic status (e.g., number of adults and children, level of education, income, and employment status) but also seasonal and climatic factors (e.g., daylight, temperature, and precipitation).

However, existing researches assume that people's electricity consumption patterns are fixed. Since electricity consumption patterns reflect the lifestyle of residents, and people's life patterns change under different seasons, in addition, every holiday season, people's and therefore residents' electricity consumption patterns will change gradually over time. The effect of influencing factors on electricity usage may be non-linear and dynamic, Therefore, these influencing factors can be used to predict the

power consumption pattern, this thesis use different machine learning models to predict the power consumption pattern based on these influencing factors. Energy consumption patterns may vary dynamically through different time intervals (e.g., intra-day, day, week, month, year), so it is necessary to verify the stability of different patterns before predicting consumption pattern. Using stable features (socio-economic) to predict an unstable (intra-day electricity consumption pattern) feature is unreasonable. Intra-day, week, and month patterns are relatively flexible and more likely affected by temperature and daylight. Based on this, we understand short patterns mainly from perspective of, temperature, intra-day, day, week, month, etc. On the contrary to predict long consumption pattern (electricity usage over one year or more) according to household characteristics. These characteristics, including number of occupants, employment status, education level, and social class are strongly linked with different pattern. It is important to have a deep understanding of the role that household characteristics enables better forecasting electricity demand and devise effective demand side management programs.

1.2.2 Electricity demand forecasting in the context of COVID-19

The rapid economic growth has caused an increase in electricity demand, accurate forecast plays an important role in the power station planning, operation, and energy trading, such as unit commitment, economic dispatch, and unit maintenance. For example, for the electricity suppliers, they need to set up production scheduling in advance based on the power demand forecasting results, which implies economic losses if the produced electricity is higher than the actual demand. On the other hand, energy service company also need to submit bids to purchase electricity from energy wholesale market for further sale to end-users, and thus power service company needs to forecast power demand for its area of responsibility.

Depending on the perspective of prediction, it can be broadly categorized into statistical based forecasts, machine learning based forecasts and combination of these two. Statistical-based time series forecasting is mainly to construct a generative model to restore the logic of time series generation as much as possible, such as, Autoregressive Integrated Moving Average (ARIMA), Seasonal and Trend decomposition using Loess (STL) and Exponential Smoothing (ETS), so kind of forecasting techniques is more explanatory. The disadvantage of statistical-based time series forecasting is that it cannot fully utilize the information outside the time series. On the contrary, machine learning beads time series

forecasting technique, including deep learning, mainly converts the time series forecasting problem into a supervised learning model, for example, given a time series $y_1, y_2, y_3, \cdots, y_t$, t is length of time series. If we want to implement a 2 step ahead forecast, the original time series could be turned into a training set $\{(y_1, y_2, y_4), (y_2, y_3, y_5), (y_3, y_4, y_6), \cdots, (y_{t-3}, y_{t-2}, y_t)\}$. The advantage is that its accuracy is usually higher over that of traditional time series forecasting techniques. the disadvantage is that it is not as explanatory as the generative model. In addition, based on the time horizon, forecasting is categorized as short-term, medium-term, and long-term forecasting. In power system, the difficulty of forecasting is mainly from the non-stationarity of the time series, this non-stationarity reflects the uncertainty of power demand, and many factors could lead to the uncertainty. First of all, the heterogeneity of the main body of energy consumption, in each energy sector (residential, commercial, industrial and transportation), and every end-users within each sector also has different electricity consumption pattern, which definitely leads to a certain degree of local fluctuations in the final power load profile. In addition, the seasonal factors and climatic factors have also led to global fluctuations of load profile by influencing people's behaviors related to energy use. Finally, the disaster, such as earthquakes, tsunami and pandemics will lead to a great change in the operation of society, and therefore the electricity usage will naturally change compare to that of in normal situation.

This thesis focuses on electricity demand forecasting in the context of COVID-19. After outbreak of a COVID-19, many countries adopted different lockdown policy to reduce the spread of the virus by reducing population flow, electricity consumption in public places in cities would decrease significantly, on the contrary, electricity consumption in residential sector would rise as most people work from home. Compared with the pre-COVID-19 period, people's lifestyles have changed significantly, and this change poses a greater challenge for electricity demand forecasting. This thesis proposed a Bayesian structural time series (BSTS) based model to comprehensively explore relationships between electricity consumption and interactions among key drivers, including population-related factors (e.g., number of people stay at home, number of people go to work, number of people go to retails), weather-related factors (e.g., temperature) and public health data (e.g., number of positive cases, number of death) in the COVID-19 situation. The BSTS could capture the general trend, seasonal patterns (such as week patterns, month patterns) and other influencing factor related patterns in a time series but also decompose the original electricity time series into multiple personalized components, such as long-term trends, seasonal components, and population flow related factors, which leads to an understanding of

the effect of COVID-19 on electricity consumption. Based on proposed model, this thesis further predicts daily consumption over 1-6 weeks based on historical consumption data and population flow indicators in seven America cities and finally estimate counterfactual electricity usage according to pre-COVID-19 trend, and then provide an analysis of the short-term effect of COVID-19 on electricity usage.

1.2.3 Personalized electricity customer segmentation

Customer segmentation can be defined as the process of dividing customers into groups based on different characteristics or behaviors. Some bases of segmentation that may be used include electricity usage (e.g., hourly, daily, weekly, monthly), household characteristics (e.g., number of occupants, employment status, education level, and social class, etc), psychographic (attitude towards smart meter, renewable energy, and eco-friendly appliances, etc) and energy use behaviors (cooking, showering, space heating, etc). Electric utilities use it to set different types of tariffs for each segment, many clustering methods have been used to segment electricity customers, such as k-means, self-organizing map and Density-based spatial clustering of applications with noise (DBSCAN), etc. Each sample within cluster represents similar property defined by distance metric among them, and clustering algorithms use different distance similarity for various data types including numerical, categorical or mixed type, such as Euclidean distance, Gower distance and Manhattan distance etc. The dataset typically used for electricity customer segmentation consists of two parts (i.e., the historical electricity usage data and the household characteristics). In general, the length of historical usage time series is much longer than the length of household characteristics, which causes an unavoidable problem for clustering algorithm. Machine learning models treat all features equally and therefore it is very challenging to insert the importance or weight of socio-economic factors into traditional clustering algorithm. Because neither machine learning algorithms nor statistically relevant models can distinguish between these two types of features, so when using clustering algorithms for clustering, the clustering results are always inclined to the historical electricity consumption data, and therefore the clustering tends to reflect a similar electricity consumption pattern without similar socio-economics background. Such results make it difficult to implement certain pricing strategies and promotion of new energy devices for this heterogeneous group from the perspective of energy service company.

The purpose of this thesis is to cluster those households that are both similar in electricity consumption

patterns and similar in socio-economics background into same group. This thesis focuses on using mathematical programming approach to overcome aforementioned limits, we proposed a p-median based constrained clustering framework. P-median problem is a very classical facilities location problem and one of the first applications is for data clustering. In the proposed framework, the socio-economic background labels $S_1, S_2, S_3, \cdots, S_N$ are converted into must-link constraints to be added into original p-median problem model, in which the these households from a socio-economic S_i can be set artificially and subjectively not to appear in the same group as the households from label S_j appear in a group. In addition, p-median problem is known to be NP-hard, hence solving large instances of the p-median problem is a difficult task in terms of running time, this thesis proposes a two-stage scheme, in the first stage of solving, k-means is firstly used for the selection of the initial clustering centroids, and in the second stage, the final clustering centroids is determined by the decomposition algorithms, such as lagrangian relaxation.

Based on above research topics, a series of research questions are provided in Sections 1.3

1.3 Research questions

This section summarizes main research questions in different research topics.

Research questions emerging in the prediction of residential consumption patterns are as follows:

(1) How to understand non-linear effect of temperature, day of week and month on electricity usage based on proposed methods, and its advantage over traditional linear regression?

(2) How to understand how people's intra-day electricity consumption patterns (load profile shape) change over time, such as from Monday to Sunday, from summer to winter, and on important holidays?

(3) How to predict household daily consumption pattern over a year based on historical consumption data and household data, and understand user-profile behind each pattern?

Research questions emerging in the electricity demand forecasting in the context of COVID-19 context are as follows:

(1) How to forecast electricity consumption in the context of COVID-19?

(2) How does population mobility cause changes in electricity consumption, and what is the mechanism?

(3) How can the time series of electricity consumption be decomposed in a more personalized way, such as long-term trends, seasonal trend (e.g., week patterns, month patterns and quarterly patterns) and other influencing factor related trend?

(4) How to effectively assess the impact of the COVID-19 on electricity consumption, especially in the context of declining electricity consumption trend in recent years?

Research questions emerging in the personalized electricity customer segmentation context are as follows:

(1) How to cluster the electricity customers so that users in the same group not only have similar power consumption patterns but also have similar socioeconomic backgrounds?

(2) How to build a constrained clustering model based on optimization techniques and be relatively free to add cannot-link and must-link constraints?

(3) How to solve the model in reasonable time given large datasets?

1.4 Research contributions

The aim of this dissertation is to contribute to research on prediction of electricity consumption pattern, electricity demand forecasting, and customer segmentation for Demand-side management (DSM). In particular, novel statistic and optimization models are developed to shed light on aspects that have not been fully considered or neglected in the existing literature.

Research contributions related to the prediction of residential consumption patterns, aimed at answering the research questions that have been detailed in the previous section (Section 1.3), are as follows:

This study attempts to fill these identified gaps. In particularly, using smart meter data from the Republic of Ireland, we try to understand effect of temperature, day of week, and month on electricity usage with the help of a generalized additive model and its advantage over traditional linear regression. Second, after using a novel clustering method to generate load profiles, we examine how intra-day household electricity consumption patterns vary both annually and weekly. Third, using an elastic net model, we predict annual consumption patterns at daily intervals based on household level data and recent prior electricity consumption in an attempt to better understand the user-profile underlying each pattern.

Research contributions related to the electricity demand forecasting in the context of COVID-19, aimed at answering the research questions that have been detailed in the previous section (Section 1.3), are as follows:

This study attempts to fill these identified gaps. In particularly using multiple source datasets, we explore changes in power consumption pattern in both 2019 and 2020 from different time intervals (e.g., day, week, month, year). Second, a Bayesian structural time series model is used provide short and long-term forecasting and to decompose the original power consumption data into multiple personalized components, such as long-term trends, seasonal components, and population flow related factors. Bayesian structural time series (BSTS) has obvious advantages over other classical time series models, such as Autoregressive Integrated Moving Average (ARIMA), Seasonal and Trend decomposition using Loess (STL) and Exponential Smoothing (ETS). First of all, BSTS can perform the decomposition of the time series in an additive way, so it is more interpretable than ARIMA, and it is easier to explain why the time series shows such a change pattern. BSTS could handle any type of seasonality, such as weekly data, not only monthly and quarterly data compared to ETS with STL. In addition. When we decompose a time series into components, such as trend, seasonality and cycles. BSTS could take advantage of external features instead of internal features that rely on differencing, lags and moving averages of a time series, which help to provide feature based decomposition of the model. This personalized decomposition is conducive to understanding the impact of the epidemic on electricity consumption. Third, a causal analysis of electricity consumption and COVID-19-related influencing factors were carried out. We conducted a scenario analysis of the outbreak of COVID-19 and predicted the likely trend of electricity consumption if there were no outbreak based on pre-COVID-19 data, and estimated the true effect of the COVID-19 on electricity consumption.

Research contributions related to the personalized electricity customer segmentation, aimed at answering the research questions that have been detailed in the previous section (Section 1.3), are as follows:

This study attempts to fill these identified gaps, a two-stage constrained clustering algorithm is proposed that combines k-means algorithm and mixed integer programming technique. To be more specific, in the first stage of proposed algorithm, we use the k-means algorithm to select the initial clustering centroids, and then the closest points to these centroids are selected as potential clustering centroids. In

the second stage, the constrained p-median problem built on potential clustering centroids is used for determining final clustering centroids. In order to speeding up solving the problem, a Lagrangian relaxation procedure is proposed to efficiently solve large problem.

1.5 Outline

The remainder of this thesis is organized as follows.

Chapter 2 provides various machine learning algorithms to analyze and understand different aspects of residential electricity consumption, including temporal and climate influences on electricity consumption, daily consumption pattern shifting over time, and prediction of annual load profiles, using smart meter data and household level characteristics from the Republic of Ireland.

Chapter 3 introduces a Bayesian structural time series based forecasting model to forecast normal consumption and usage since the outbreak of COVID-19 in late 2019. It reveals the dynamic relationship between electricity usage and population flow features by providing a personalized decomposition of time series. It also explored the impact of COVID-19 on various electricity consumption patterns, such week patterns and intra-day patterns and estimate counterfactual electricity usage according to pre-COVID-19 trend, and then provide an analysis of the short-term effect of COVID-19 on electricity usage.

Chapter 4 presents a two-stage constrained clustering framework for electricity customer segmentation, which consist of k-means for choosing potential clustering centroids in the first stage, and determined final clusters based on a p-median problem with cannot-link and must-link constraints in the second stage. the model formulation as well as the solution methodologies that have been developed to solve it, a Lagrangian relaxation algorithm are applied to solve proposed model, and provides computational results on real residential electricity dataset in the UK.

Chapter 5 offers some conclusive remarks.

2. Predicting residential electricity consumption patterns based on smart meter and household data: A case study from the Republic of Ireland

2.1 Abstract

We use machine learning algorithms to investigate various aspects of residential electricity consumption for households in the Republic of Ireland. Temperature, day of week, and month of year have an apparent causal effect on consumption. The prevalence of six distinct intra-day load profiles, identified by clustering, changes dramatically between weekdays and weekends as well as seasonally. Key socio-demographic and dwelling characteristics associated with annual load profiles include household makeup and size and occupation of the primary income earner. We further discuss policy and management implications of our findings and propose avenues for future research.

2.2 Introduction

Reducing nonessential energy consumption, including residential electricity, is widely considered one of the most effective and low-cost means of mitigating global climate change and enhancing sustainable development, particularly in developed economies. The residential sector accounts for a significant portion of overall electricity demand – 26.9% worldwide in 2018 (IEA 2020) – with consumption patterns largely reflective of household characteristics (Ürge-Vorsatz *et al*. 2015; Wang *et al*. 2018). With increased population growth and rising living standards, residential electricity consumption has continued to grow over the decades. Gaining a better understanding of household electricity consumption patterns, behaviors, and household characteristics that influence consumption should prove enormously useful to policymakers and electricity generation firms in developing effective strategies to reduce residential electricity consumption.

Recent advances in big data and machine learning algorithms come at an ideal time to help address this challenge. It is now possible to digitally store massive amounts of electricity usage data. Advanced machine learning methods, in turn, provide ideal tools for analyzing these large datasets, opening up new avenues to explore electricity consumption patterns in greater depth. For example, variations in electricity consumption over time can be analyzed at almost any level of granularity, from yearly to

monthly, daily, hourly, or even minute-by-minute (Crone and Kourentzes 2009). As such, it becomes possible to identify how peaks and troughs in demand evolve and begin to delineate typical household energy consumption patterns or load profiles. Here, advanced machine learning techniques are ideally suited for categorizing and predicting household load profiles by identifying key variables driving energy consumption (i.e., feature selection), which can subsequently form the basis for more accurate electricity demand forecasting and support demand side management (Koprinska *et al*. 2015).

Uncovering behavior patterns hidden in typical load profiles is challenging due to the lack of a full mapping between behavior and electricity consumption data. In earlier studies, researchers often investigated behavior through questionnaires and then applied traditional statistical models to try to find any relationships. In the internet era, it is possible to gain fresh insights into different aspects of residential electricity consumption behaviors. In particular, it is now feasible to combine social media network data with energy usage data and then analyze the data using flexible neural network structures to reveal relationships between behavior and residential power demand.

Although there is considerable research on household electricity consumption, there are some apparent deficiencies. For one, most studies fail to comprehensively explore relationships between residential electricity patterns and interactions among key drivers simultaneously, including time-related factors (e.g., season, month, day, hour) (McLoughlin *et al*. 2015) and weather-related factors (e.g., rainfall, temperature, humidity) (Li *et al*. 2019). Another limitation is that researchers often investigate energy consumption patterns and behaviors from a static point of view. Energy consumption patterns may vary dynamically through different time intervals. For example, in temperate regions, daytime during winter is much shorter than it is in summer, which directly impacts heating needs and people's daily activities. A third critique of the existing literature is a general failure to associate electricity usage with household data and historical usage. Both household characteristics and historical electricity usage are likely to influence future usage strongly. Most studies focus solely on characterizing consumption patterns according to dwelling and socio-economic data, while ignoring recent historical usage (e.g., over the past few months).

This study attempts to fill these identified gaps. In particular, using smart meter data from the Republic of Ireland, we examine the effect of temperature, day of week, and month on electricity usage with the help of a generalized additive model. Second, using k-means clustering to extract different intra-day

load profiles, we examine how household electricity consumption patterns vary annually and weekly. Third, after employing k-medoids clustering to generate annual load profiles, we apply an elastic net model to predict annual household consumption patterns based on household-level data and recent past electricity consumption to generate useful user-profile information.

The remainder of this chapter is structured as follows. In the next section, we describe the data used in this study and our methodology for analyzing residential electricity consumption patterns. We then present main results and findings. We conclude with a critical discussion of the implications of our findings and propose avenues for future research.

2.3 Literature Review

Existing studies on this general topic of residential electricity demand profiling can be grouped into several different research themes. The first focuses on obtaining typical load profiles using different levels of granularity for electricity consumption data (i.e., daily, hourly, and minute) and different types of clustering algorithms, such as k-means (Hartigan and Wong 1979), self-organizing maps (SOM) (Kohonen 1990), and ensemble clustering (Yu *et al*. 2012). For example, Räsänen *et al.* (2010) propose using SOM combined with k-means and hierarchical clustering for handling large time-series datasets. Their methodology is applied to year-long hourly electricity consumption data from North Savo, Finland. Experimental results indicate their proposed methodology produced more accurate household load profiles than existing ones used by an electric power company to estimate load.

Meanwhile, Khan *et al.* (2016) present an incremental density-based ensemble clustering method for segmenting factories according to electricity consumption data. Using data from manufacturing factories in Guangdong Province, China, the authors find that their algorithm outperforms several state-of-the-art data stream clustering algorithms. Other relevant studies include those by Benítez *et al*. (2016) and Chévez *et al*. (2017). Benítez *et al*. (2016) present a Hausdorff distance-based dynamic clustering algorithm for identifying and visualizing temporal load profiles. Compared to traditional clustering methods, such as k-means and Fuzzy c-means, the proposed method produces more well-defined and balanced clusters. Chévez *et al*. (2017) use k-means to detect homogeneous areas of residential electricity consumption and associated socio-demographic characteristics.

Three above-mentioned methods are used in the field of data clustering, but they differ in their

algorithms and purposes, k-means is a centroid-based clustering algorithm that partitions the data into k clusters, where each data point belongs to the cluster with the nearest centroids. while SOM is a type of artificial neural network used for clustering and it reduces dimensionality while maintaining topological properties, allowing for the visualization of high-dimensional data, ensemble clustering combines the results of multiple base clustering algorithms that is more robust and reliable than individual clustering algorithm.

A second line of research examines the importance of feature selection prior to carrying out data clustering. Feature selection is critical in the big data era as high-dimensional electricity consumption data become more widely available with the introduction of smart meters. With high-dimensional data, some traditional clustering algorithms are no longer helpful. Hence, feature selection becomes key to tackling challenges associated with analysis of high-dimensional data. An illustrative example is a study by Räsänen and Kolehmainen (2009), who present a feature selection approach for clustering time series based on extracting statistical features within time series. Advantages of their approach include dimensionality reduction of the original time series, increased robustness to missing observations, and the ability to handle time series of different lengths. Motlagh *et al*. (2019) proposed a baseline feature-based clustering algorithm to alleviate the limitations of extreme dimensionality of load time series by converting load time series into map models that can be readily clustered. Choksi *et al*. (2020) proposed a feature-based clustering algorithm aimed at dimensionality reduction, load profile characterization, and probabilistic load variation assessment by combining classical k-means with empirical feature selection.

A third area of research on household load profiles concerns the relationship between household-level data and electricity consumption patterns. For example, McLoughlin *et al.* (2012) use a multiple linear regression model to estimate total electricity consumption, maximum demand, load factor, and time of use of maximum electricity demand based on different dwelling types and occupant socio-economic variables. Conversely, Beckel *et al.* (2014) examine the feasibility of inferring household characteristics, like employment status and number of occupants, from smart meter data. They were able to achieve 70% or more accuracy for each household characteristic using supervised machine learning techniques. More relevant to our work, Singh *et al*. (2019) apply k-means clusters to segment customers based on electricity consumption metrics, socio-demographic/economic and dwelling characteristics, and

georeferenced weather data to improve peak and off-peak load predictions and support targeted demand management programs.

2.4 Methodology

2.4.1 Data

The data used in this study come from 4232 Irish households randomly selected across the country between 14 July 2009 and 31 December 2010. The dataset consists of smart meter electricity consumption data logged at 30 min intervals, along with household surveys carried out in December 2009. The data were collected as part of an electricity customer behavior trial carried out by the Irish Commission for Energy Regulation (CER 2012) and were obtained from the Irish Social Science Data Archive (https://www.ucd.ie/issda). The types of household data collected, along with summary information about dwelling, space heating, water heating, and cooker types, are provided in a series of tables in an appendix (see Table 2.10-2.11 and Table A.1-A.4). Key things to point out are that most surveyed homes (72%) are heated by oil or gas (only 6% are heated by electricity), electricity represents the single largest category for water heating (35%) and cooking (70%), and properties tend to be rather large (less than 2% are apartments and only 14% are terraced houses). After data cleaning, a sample of 3639 households remained to analyze and pattern shifting over time (see §2.4.3), and aggregated electricity consumption from these households are used to analyze time and climate effects on electricity consumption (see §2.4.2). A subset of 2874 households could be matched with household survey data to carry out a household segmentation analysis and prediction of electricity consumption patterns (see §2.4.4).

Daily temperature data from 47 weather stations across Ireland (see Figure A.1) for the same time interval as the electricity monitoring data were sourced from the Irish Meteorological Service (https://www.met.ie). Due to a lack of knowledge about the location of households, hourly temperatures were averaged across all stations and used as a common temperature time series for all households. Though not ideal, average heating degree days at each station showed low variability (see Figure A.1), suggesting that most households in Ireland experience similar temperatures regardless of location.

2.4.2 Time and climate effects on electricity usage

It is hypothesized that Irish residential electricity consumption behavior is mainly affected by seasonally varying factors like temperature and month (a proxy for daylight). In addition, electricity consumption often has clear day-of-week patterns. The effect of such influencing factors on electricity usage may be non-linear; hence we applied a generalized additive model (GAM) (Hastie and Tibshirani 1990) to study the relationship between average electricity consumption and influencing factors.

GAM is a non-parametric extension of a generalized linear model (GLM) in which non-linear smooth functions express relationships between the response and predictor variables to capture non-linearities within the data. For our analysis, we considered three predictor variables for household electricity usage (temperature, day of week, and month), which resulted in the following GAM:

$$Elect_t = \beta_0 + g_1(Temp_t) + g_2(Day_t) + g_3(Month_t) + \varepsilon_t \quad t = 1, 2, \dots, T \tag{2.1}$$

where $Elect_t$ is daily usage on the t^{th} day, $Temp_t$ is average the daily temperature on day t, Day_t is the day of week on day t (values 1 to 7), $Month_t$ is the month of year on day t (values 1 to 12), each $g_s(\cdot)$ is a smooth function of the corresponding covariate with thin plate regression splines as a smoothing basis, and ε_t is an error item with normal distribution. We used the penalized iteratively re-weighted least squares (PIRLS) algorithm to solve model (2.1).

We also consider the following linear regression model as a benchmark for model (2.2).

$$Elect_t = \beta_0 + \beta_1 * Temp_t + \beta_2 * Day_t + \beta_3 * Month_t + \varepsilon_t \quad t = 1, 2, \dots, T \tag{2.2}$$

where $Elect_t$, $Temp_t$, Day_t, $Month_t$ and ε_t are the same as defined previously and β_0, β_1, β_2 and β_3 are linear regression coefficients. Model (2.2) was solved via least squares (Fahrmeir *et al.*, 2007).

2.4.3 Electricity consumption pattern shifting

Residential daily electricity consumption patterns mainly depend on people's electricity consumption behavior. Behaviors are in part governed by household size and socio-economic status (e.g., number of adults and children, level of education, income, and employment status), but also seasonal and climatic factors (e.g., daylight, temperature, and precipitation). Accordingly, daily electricity consumption patterns do not remain fixed over time. To better understand this, we combine data re-aggregation with the k-means algorithm (Hartigan and Wong, 1979) to analyze changes in household intra-day electricity consumption, referred to as electricity consumption pattern shifting.

We first extracted total intra-day electricity consumption patterns for all households by taking intra-day usage time series data, re-aggregating it to a daily basis, and then applying the k-means algorithm on the re-aggregated data. The result is an assignment of each household ℓ on day d to pattern p. The number of each pattern on a particular day can then be tallied to determine the distribution of patterns and analyze how the distribution varies gradually (or not) through time to help identify underlying drivers. The proposed method is explained in more detail below.

Data re-aggregation: Before applying the k-means algorithm to extract intra-day electricity consumption patterns, it was necessary to reformat the original dataset from wide to long format, as shown in Table 2.1. Here, each row indicates a whole day's record of electricity usage for household ℓ ($\ell = 1,2,\dots,n$). The first column is household index ℓ or ID, the second column denotes the day i ($i = 1,2,\dots,T$) electricity usage was recorded, and the remaining columns indicate electricity usage c_{ℓ,i,t_j} for household ℓ on day i during a 30-minute time interval t_j ($j = 1,2,\dots,48$).

Table 2.1. Aggregated intra-day electricity usage.

ID	Day	t_1	t_2	...	t_{48}
1	1	$c_{1,1,t_1}$	$c_{1,1,t_2}$	...	$c_{1,1,t_{48}}$
1	2	$c_{1,2,t_1}$	$c_{1,2,t_2}$	...	$c_{1,2,t_{48}}$
⋮	⋮	⋮	⋮	⋱	⋮
1	T	c_{1,T,t_1}	c_{1,T,t_2}	...	$c_{1,T,t_{48}}$
2	1	$c_{2,1,t_1}$	$c_{2,1,t_2}$	...	$c_{2,1,t_{48}}$
2	2	$c_{2,2,t_1}$	$c_{2,2,t_2}$	...	$c_{2,2,t_{48}}$
⋮	⋮	⋮	⋮	⋱	⋮
2	T	c_{2,T,t_1}	c_{2,T,t_2}	...	$c_{2,T,t_{48}}$
⋮	⋮	⋮	⋮	⋱	⋮
n	1	$c_{n,1,t_1}$	$c_{n,1,t_2}$	...	$c_{n,1,t_{48}}$
n	2	$c_{n,2,t_1}$	$c_{n,2,t_2}$	...	$c_{n,2,t_{48}}$
⋮	⋮	⋮	⋮	⋱	⋮
n	T	c_{n,T,t_1}	c_{n,T,t_2}	...	$c_{n,T,t_{48}}$

Pattern extraction and assignment: In the pattern extraction phase, k-means is applied on columns t_1 to t_{48} in Table 2.1. The outputs consist of intra-daily patterns, namely the mean value of each cluster and the household-date-pattern assignment matrix (Table 2.2). From this, we derive the pattern-date-proportion matrix (Table 2.3). In Table 2.3, value $f_{v,d}$ represents the proportion of each pattern p_v on day d as calculated from Table 2.2.

Table 2.2. Example household-date-pattern assignment matrix.

Day / ID	1	2	3	...	T

1	p_1	p_1	p_4	...	p_5
2	p_3	p_1	p_7	...	p_2
⋮	⋮	⋮		⋱	⋮
n	p_5	p_k	p_k	...	p_6

Table 2.3. Pattern-date-proportion matrix.

Day / Pattern	1	2	...	T
p_1	$f_{1,1}$	$f_{1,2}$	...	$f_{1,T}$
p_2	$f_{2,1}$	$f_{2,2}$	...	$f_{2,T}$
⋮	⋮	⋮	⋱	⋮
p_k	$f_{k,1}$	$f_{k,2}$	...	$f_{k,T}$

Table 2.3 lets us easily track how each pattern changes over time. To more easily visualize this, we further define a pattern shifting matrix (Table 2.4) and utilize it to analyze electricity consumption pattern shifting between any two days d and $d^{'}$.

Table 2.4. Pattern shifting matrix.

d \ $d^{'}$	$P_1^{'}$	$P_2^{'}$	⋯	$P_k^{'}$
P_1	S_{11}	S_{12}	⋯	S_{1k}
P_2	S_{21}	S_{22}	⋯	S_2
⋮	⋮	⋮	⋱	⋮
P_k	S_{k1}	S_{k2}	⋯	S_{kk}

In Table 2.4, value $S_{vw} \in [0,1]$ represents the shifting proportion, defined as:

$$S_{vw} = \begin{cases} \dfrac{|P_v \cap P_w^{'}|}{|P_v|}, & |P_v| > 0 \\ 0, & |P_v| = 0 \end{cases} \tag{2.3}$$

where P_v represents the set of pattern v households on day d and $P_w^{'}$ represents the set of pattern w households on day $d^{'}$.

Our proposed method has two main advantages over a more traditional approach of directly clustering on household electricity usage wide formatted time series data. First, it avoids clustering on high-dimensional data. Second, it is more effective at revealing changes in daily electricity consumption patterns over time. Without data re-aggregation, it is more difficult to detect meaningful intra-day patterns in long time series possessing a distinct daily cycle. Typically, only long-term cycles within the data can be extracted without re-aggregation. An alternative is to cluster on a day-by-day basis, but this presents issues regarding the large number of times clustering needs to be performed. Moreover, the patterns obtained by clustering on each day separately are unlikely to be universal. Our proposed way of clustering avoids both of these problems.

2.4.4 Household segmentation

We further investigated how household characteristics and past usage can be used to forecast annual electricity usage patterns on a daily interval. An important application of this analysis is the segmentation of potential new customers, enabling electricity providers and energy comparison websites to make appropriate price plan recommendations.

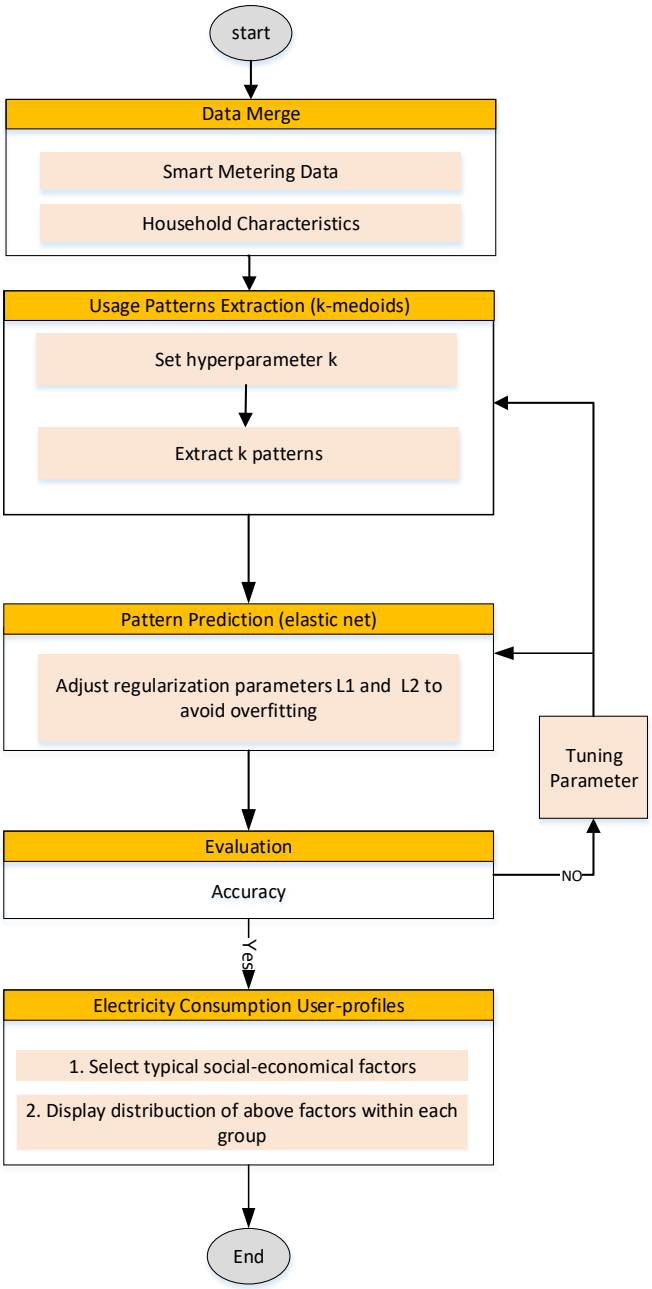


Figure 2.1. Overview of how to predict annual household electricity consumption patterns.

As mentioned previously, daily electricity consumption patterns often vary over time. On the contrary, household socio-economic descriptors tend to be relatively stable. This creates an inherent challenge whereby stable (socio-economic) features are used to predict an unstable (electricity consumption pattern) feature. To overcome this, we applied k-mediods (Reynolds *et al.*, 2006) on household characteristics and past electricity consumption information to first cluster households around different typical yearly load profiles. Unlike k-means, k-medoids chooses actual data points as centers, which has the advantage of providing greater interpretability of the links between underlying household characteristics and consumption patterns. Having extracted a set of typical load profile patterns, we subsequently applied an elastic net model (Zou and Hastie 2005) to predict the likelihood of a household having any given load profile. Finally, we generated user-profile information for each load profile. Figure 2.1 shows an overview of our methodology. The methodology consists of four distinct steps: data merging, load profile pattern extraction, load profile prediction, and generation of user-profile information. Each of these steps is discussed in more detail below.

Data merge

Household characteristics are detailed in two tables provided in an appendix (see Tables A.9-A.10). This includes type of residence, number of occupants, employment status, level of education, and social class. As part data processing, we first removed any households with missing data. Next, we split the smart meter data into time intervals: 2009-07-14 to 2009-12-31 and 2010-01-01 and 2010-12-31. Data from 2009 were treated as "historical" usage; from this, average daily usage is computed. The three categories of data – household information, 2009 historical usage, and 2010 contemporary usage – were then used in the next step for load profile pattern extraction. The first two data types were later used for load profile prediction.

Load profile pattern extraction

In the second step, household characteristics and historical and contemporary electricity usage data were used to segment consumers using the Partitioning Around Medoids (PAM) algorithm. Importantly, PAM can accept mixed-type distance similarity. Indeed, several household characteristics are categorical, like residence type and education level. For hyperparameter k (i.e., number of clusters), values between 3 and 6 were considered. The silhouette coefficient can be used to assess the quality of

clusters based on the degree of similarity and dissimilarity among them. However, the cluster validity index tends to result in choosing a smaller number of clusters, which is not ideal for customer segmentation. The output from this step was a set of load profiles with similar household information, historical electricity usage, and contemporary electricity consumption patterns.

Load profile prediction

The load profile prediction step aims to classify households according to load profile patterns based on household and historical electricity usage data. For this, we used an elastic net model (Zou and Hastie 2005), which combines lasso (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970). Elastic net overcomes some of the limitations of the lasso model. For example, in the "large p, small n" case (i.e., high-dimensional data with few samples), lasso selects at most n variables before it saturates. If there is a group of highly correlated variables, then the lasso model tends to select one variable from the group and ignore the others. In order to overcome this, the elastic net adds a quadratic part to the penalty, also known as Tikhonov regularization, which comes from ridge regression.

We used coordinate descent to solve for the regression parameters and parameter grid search to optimize the model's hyperparameters, namely, the number of clusters, elastic net penalty, and regularization parameter. With parameter grid search, parameter values can be determined to optimize the model's accuracy. We also implemented a deep neural network (LeCun *et al*. 1998), random forest (Breiman 2001), gradient boosting machine (Friedman 2001), and support vector machine (Cortes and Vapnik 1995) as benchmarking models.

User-profiles

In the final step of our analysis, socio-demographic and dwelling information were analyzed for each load profile pattern. Household characteristics, including number of occupants, employment status, education level, and social class, were anticipated to be strongly linked with different load profiles. A better understanding of the linkages between household characteristics and consumption behaviors will likely help in devising more effective demand-side management programs in the electricity sector.

2.5 Results

Model implementation and computational experiments were carried out using Python and R on a PC

with an Intel Core i5-8250U CPU running at 1.60GHz with 8.0GB RAM.

2.5.1 Influence of time and climate on electricity usage

We implemented the GAM model (2) in the R package mgcv (Wood 2017), where each smooth function $g_s(\cdot)$ is estimated by a penalized regression spline. Results of the GAM and benchmark linear regression models are displayed in Tables 2.5 and 2.6. Estimated temperature, day of week, and month curves are plotted in Figure 2.2, with the shaded area representing twice the standard error bands.

Table 2.5. Linear regression model results.

Parameter	Est.	Std.	t value	p-value
Intercept***	26.938	0.265	101.466	$< 2\times10^{-16}$
Temperature***	-0.610	0.014	-43.828	$< 2\times10^{-16}$
Day***	0.221	0.036	6.016	3.32×10^{-9}
Month***	0.263	0.022	11.916	$< 2\times10^{-16}$

*** significant at the 0.1% level.

Table 2.6. GAM model results.

Parameter	Est.	Std.	t value	Eff. DF	F value	p-value
Intercept***	23.935	0.051	469.3			$< 2\times10^{-16}$
s(Temperature)***				4.667	62.78	$< 2\times10^{-16}$
s(Day)***				4.483	23.76	$< 2\times10^{-16}$
s(Month)***				9.299	69.34	$< 2\times10^{-16}$

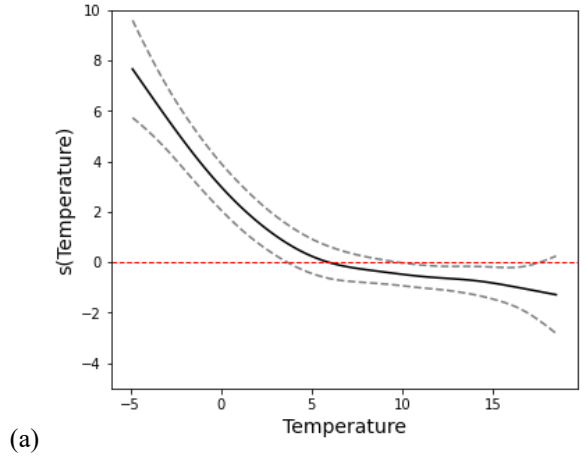
*** significant at the 0.1% level.

For both models, we observe that all predictor variables are statistically significant at the 0.1% level. As one would expect, there is an inverse relationship between electricity demand and temperature, as indicated by the negative sign for the corresponding regression coefficient estimate. For the GAM model, the near zero p values for s(Temperature), s(Day), and s(Month) suggests that these smoothing terms each have a powerful influence on electricity usage.

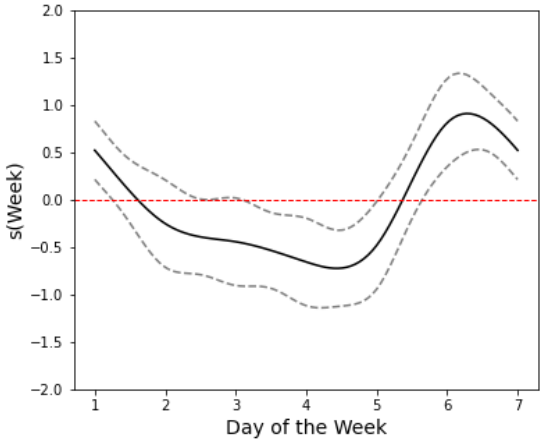
Table 2.7. Comparison of GAM and linear regression model accuracy.

Model	Adj. R^2	AIC	RMSE	MAPE
Regression	0.79	2095.48	1.69	5.14%
GAM	0.90	1720.30	1.16	3.04%

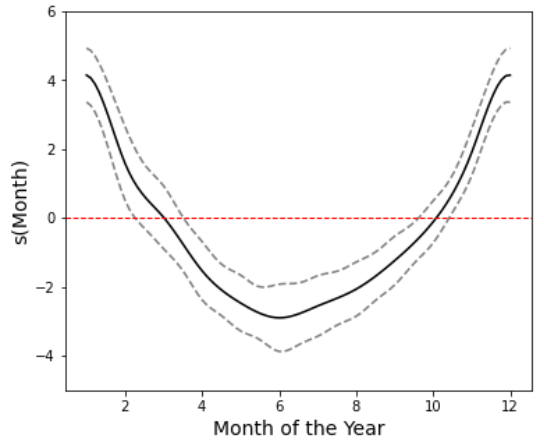
The performance of each model is reported in Table 2.7. We observe that GAM produces a superior fit compared to a multiple linear regression, which is evident by the higher adjusted R^2 value (90% of variation explained by the three predictors) and lower Akaike information criterion (AIC), root mean square error (RMSE), mean absolute percentage error (MAPE) values.



(a)



(b)



(c)

Figure 2.2. Effect of temperature (a), day of week (b), and month of year (c) on residential electricity demand.

Figure 2.2 shows nonlinear effect of temperature (a), day of week (b), and month of year (c) on residential electricity usage, s(Temperature), s(Day), and s(Month) are regression splines estimated from GAM. Black dashed lines indicate the 95% confidence intervals for the regression spline, and red dash line across at zero. Inspection of Figure 2.2 reveals a clear monthly pattern to electricity usage,

with above average usage in the late autumn and winter (November to February) and below average usage in the spring, summer, and early autumn (March to October). Intuitively, this would appear to be driven by behavioral responses to seasonal variation, namely reduced time at home and use of appliances in warmer months with more daylight (March to October) and greater use of appliances and increased secondary electric heating and water heating in colder months with less daylight (November to February). There is also a clear weekly pattern for residential electricity usage. Demand is noticeably higher at the weekend (Saturday and Sunday) when people are typically at home than during weekdays when they are typically at work. One interesting observation is that demand steadily declines from Monday to Friday. Why this should be the case is not clear.

Compared to a basic regression model, the GAM model is able to capture non-linearities in the temperature response function. Moreover, while month of year will partly capture temperature-driven electricity usage patterns, the use of temperature as a separate variable better accounts for daily variation in the need for water heating, secondary electric heating, and other activities. Indeed, household electricity consumption shows a near linear increase as temperature drops from 20 to 5°C and then rapidly increases as temperatures drop below 5°C (see Figure 2.2a).

2.5.2 Shifting electricity consumption patterns

We identified six typical intra-day load profiles (see Table 2.8 and Figure 2.3). Patterns 1-3 are all similar in that they show a bump in electricity usage in the morning starting around 06:00, increased electricity usage beginning at 16:00 that eventually peaks in the evening at around 19:00, presumably when dinner is being prepared, followed by a noticeable decline starting at 22:00 when people go to bed. Pattern 1 stands out from these and the rest in terms of the very low amount of electricity used in a 24-hour period and low temporal variability. Pattern 6 is similar to patterns 1-3, but peaks much earlier in the late afternoon/very early evening around 16:30, presumably corresponding to an early dinner time. Patterns 4 and 5 differ considerably from the others. Pattern 4 has a prominent peak in the middle of the day at noon that then drops off but remains high in the evening before declining at 22:00. Pattern 5, meanwhile, has the highest electricity usage overall, which begins steadily rising at 05:00 in the morning until 17:30 in the early evening and then rapidly decreases.

Table 2.8. Summary statistics for patterns 1-6 (n = 3639).

Pattern	1	2	3	4	5	6
Proportion	34.4%	9.9%	32.6%	9.8%	3.4%	9.9%
Avg. daily usage (kWh)	9.88	38.19	22.21	36.49	71.32	35.33
Peak avg. usage (kWh/30 min)	0.42	2.35	1.12	2.25	5.36	1.87
Max avg. ramp rate (kWh/min)	0.08	0.59	0.24	0.56	1.80	0.33

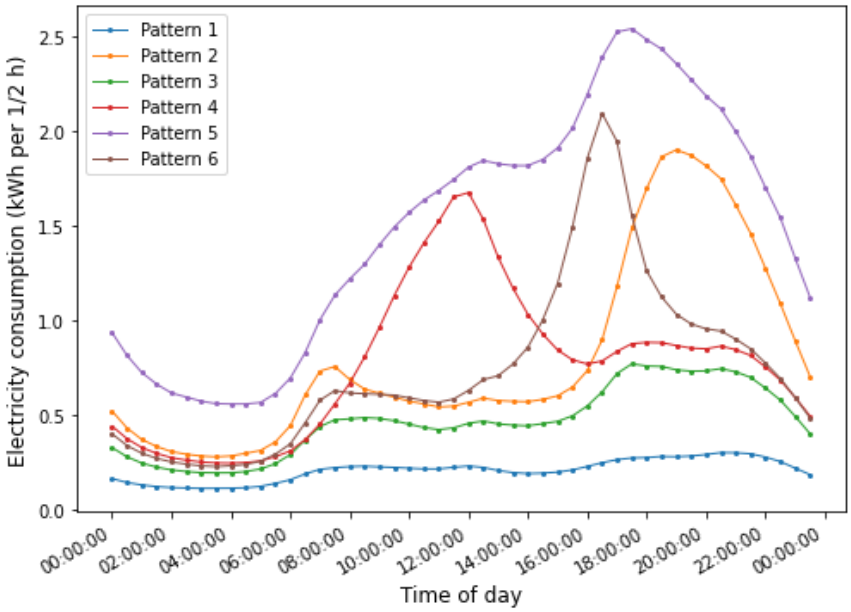


Figure 2.3. Typical intra-day consumption patterns (kWh per ½h).

It is worth noting that mean annual electricity consumption for the different load patterns is relatively high (8730.3 kWh), nearly double what one might expect (~4500 kWh). We suspect that this is mainly because few small dwellings (apartments and terraced houses) are included in the household survey dataset, which is probably not fully representative of Irish households more generally.

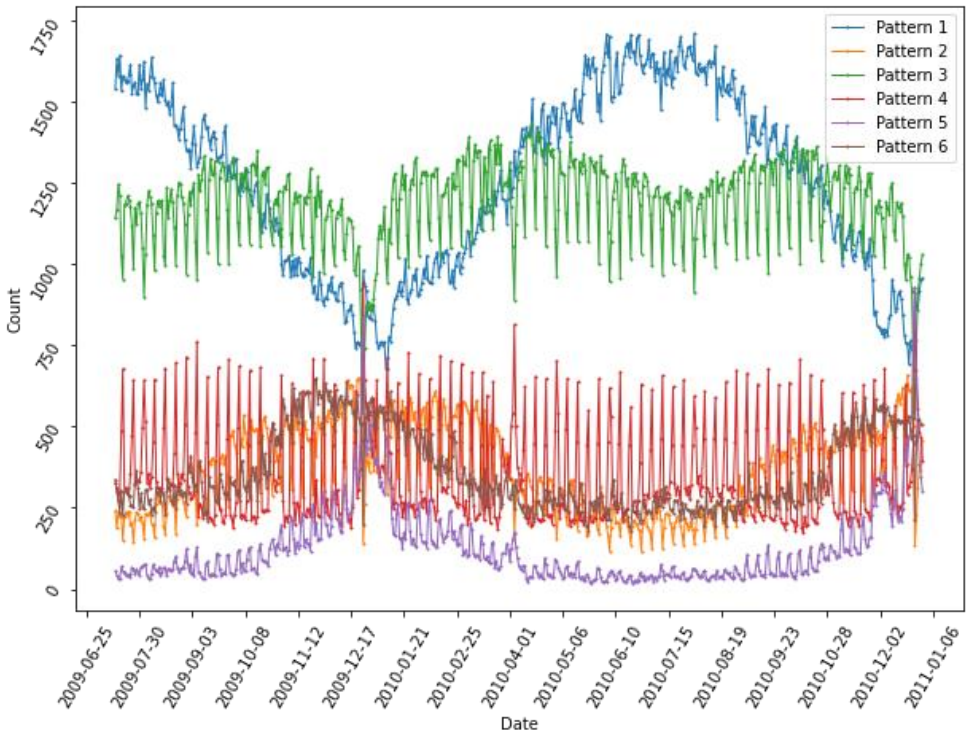


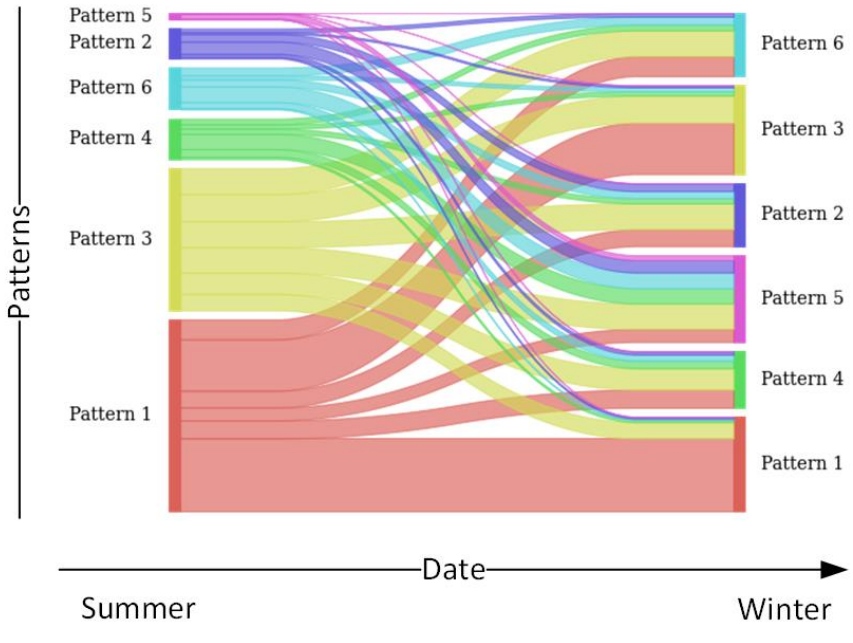
Figure 2.4. Frequency of each intra-day consumption pattern over time.

A graphical representation of the pattern-date-proportion matrix showing the number of households exhibiting each daily load profile pattern for one and a half years is displayed in Figure 2.4. As can be seen, pattern 1 and pattern 3 represent the dominant load profiles among Irish households, accounting on average for 34% and 33%, respectively, over the study period. Pattern 1 shows apparent seasonal variation characterized by a summer peak in July and a winter nadir in December. In contrast, patterns 2, 5, and 6 show opposite seasonal variation, with highs in December and lows in July. More specially, we observe that the proportion of pattern 1 households reduces from 42-44% in July to as low as 20-21% in December, while patterns 2, 5, and 6 households collectively account for 15-17% in July to as much as 47-49% in December (see Table A.5). This kind of pattern oscillation demonstrates the extent that households modify their behavior and electricity usage in response to long-term changes in temperature, amount of daylight, and other seasonal factors. Patterns 3 and 4, meanwhile, form a more constant proportion of households, albeit with high weekly and yearly volatility.

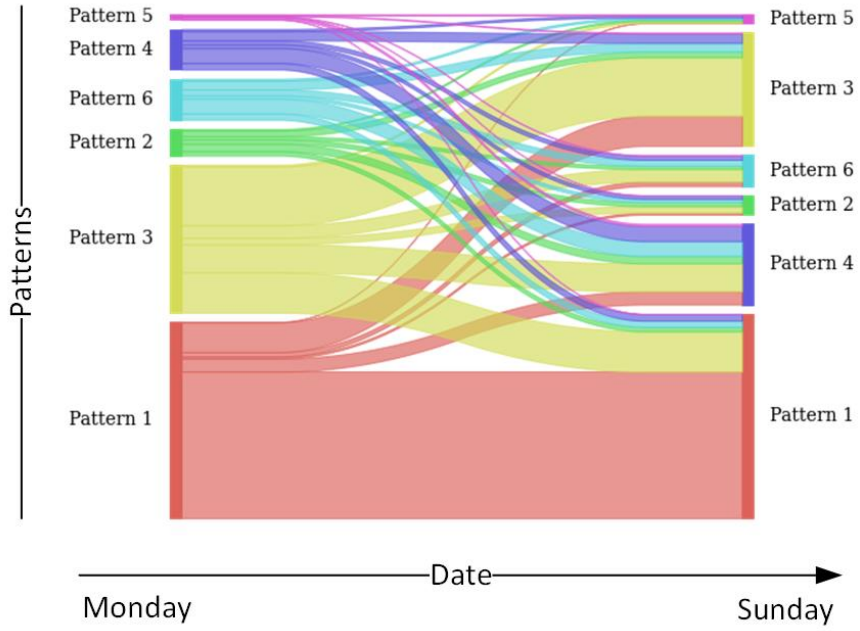
Importantly, there is a noticeable weekday versus weekend relationship to the proportion of pattern 3 and pattern 4 households, with pattern 3 being more frequent on weekdays and pattern 4 more frequent on weekends. Other patterns also have a strong weekly cycle, with patterns 2 and 6 being more prevalent on weekdays and pattern 5 on weekends. We also note a very distinctive load profile distribution on

Christmas Day, with patterns 4 and 5 usually representing a small proportion of households (12% over the study period) but forming a majority or clear plurality (49-51%) on Christmas Day.

An investigation of seasonal transitions (Figure 2.5a and Table A.6) reveals several perhaps unexpected insights. Moving from summer (14 July 2009) to winter (24 December 2009), one might expect, as a result of pattern 1 being counter-cyclical with patterns 2, 5, and 6, that the decline of pattern 1 and increase of patterns 2, 5, and 6 would be simply down to pattern 1 households transitioning to those other patterns. The story is more complex than that. While a significant proportion (26%) of pattern 1 households transition to patterns 2, 5, and 6, more (27%) transition to the other dominant pattern 3. Meanwhile, the bulk of pattern 3 households (54%) shift to patterns 2, 5, and 6 between summer and winter. Thus, the significant growth in these less common patterns during the summer is driven more by pattern 3 than pattern 1 shifting. Patterns 2, 4, and 6, like patterns 1 and 3, also show a significant shift to other patterns. Only pattern 5 seems relatively stable, with 73% of households maintaining this pattern between summer and winter. The result is that winter shows a much more uniform distribution among patterns 1-6 in winter compared to summer.



(a)



(b)

Figure 2.5. Sankey diagrams of summer (14 July 2009) to winter (24 December 2009) to winter pattern shifting (a) and weekday (27 July 2009) to weekend (2 August 2009) pattern shifting (b).

Looking at weekday-to-weekend pattern shifting (see Figure 2.5b and Table A.7), we find that pattern 1 is by far the most stable, with only around 25% of households having a pattern 1 on a weekday transitioning to some other pattern on the weekend. In contrast, patterns 2, 5, and 6 show the most flux, with 75-85% of households having one of these patterns on a weekday and moving to a different pattern on the weekend. Patterns 3 are 4, meanwhile, are somewhat more stable between weekdays and weekends, with 60-63% of households transitioning to a different pattern on the weekend.

In terms of overall makeup (see Figure 2.5b and Table A.6), pattern 1 accounts for a roughly equal share of households on weekdays (43%) and weekends (45%). Pattern 4, meanwhile, increases substantially on weekends, going from 9% of households on a weekday to 18% on the weekend. Conversely, pattern 3 shows a noticeable decrease, going from 32% of households on weekdays to 25% on weekends; patterns 2, 5, and 6 show more modest decreases or increases on weekends than weekdays.

2.5.3 Household segmentation

Load profile patterns

We identified a total of 5 distinct annual electricity consumption patterns (see Figure 2.6 and Table 2.9). All five have a similar usage profile, typified by higher (lower) usage in the winter (summer), albeit

with very different amounts of average daily usage. Pattern 2 households consume the most, with average daily usage of 39.92 kWh, while pattern 5 households consume the least, with average daily usage of just 9.98 kWh. The most common usage profile is pattern 1, which accounts for nearly a third (29%) of all Irish households. Pattern 1 households consume an intermediate amount of electricity, with daily usage at 26.02 kWh.

We further investigated the proportions of different intra-day electricity consumption patterns associated with each annual load profile (Tables A.11-A.12). We observe that the five annual patterns are linked to distinct combinations of intra-day patterns that vary between summer and winter. Specifically, annual load profile 1 is majority (0.56) daily pattern 3 in summer, but formed mostly of patterns 4 and 5 (0.63) in winter. Load profile 2, on the other hand, has roughly equal proportions of patterns 2, 4, and 6 (0.18-0.21 each) and a slightly higher proportion of pattern 3 (0.28) in summer, but is majority pattern 5 (0.62) in winter. Annual load profiles 3 and 5 have in summer very high proportions of pattern 1 (0.87-0.91), a small proportion of pattern 3 (0.07-0.10), and little of the other patterns (0.02-0.03). In winter, daily pattern 1 (0.47-0.67) remains the dominant pattern, but with higher proportions of patterns 3 and 4 (0.27-0.41) and increases in the other patterns (0.06-0.13). Finally, annual load profile 4 is composed mainly of daily patterns 1 and 3 (0.90) in summer, but has similar shares of patterns 1, 3, and 5 (0.16-0.21 each) and a plurality of pattern 4 (0.33) in winter.

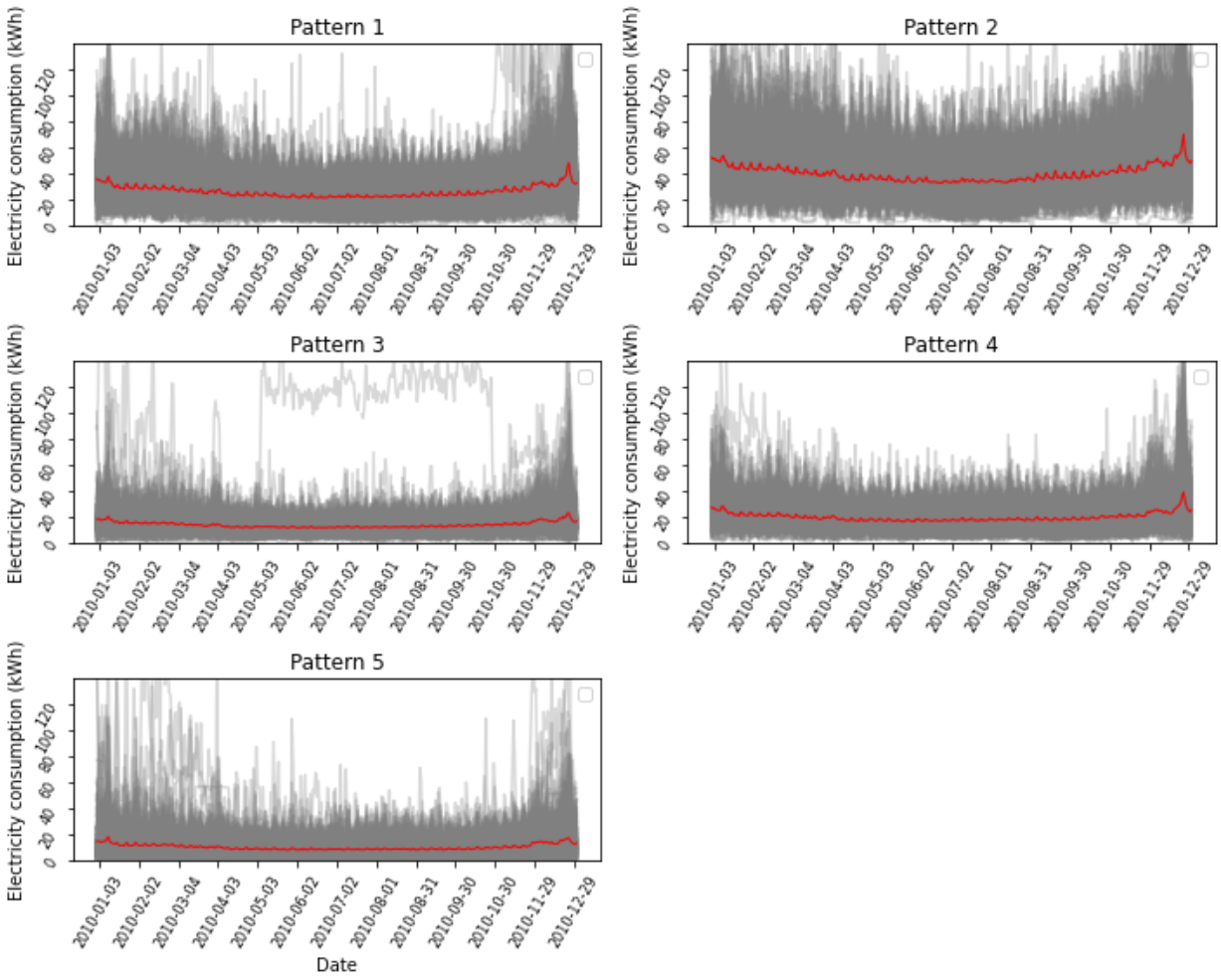


Figure 2.6. Annual electricity consumption patterns on a daily basis.

Table 2.9. Annual electricity consumption pattern summary statistics (n = 2874).

Pattern	1	2	3	4	5
Proportion (%)	29.37	24.36	13.99	18.86	13.43
Avg. daily usage (kWh)	26.02	39.92	13.73	19.75	9.98

Load profile prediction

Figure 2.7 displays the training and test accuracy of the five machine learning models (deep neural network (DNN), random forest (RF), gradient boosting machine (GBM), elastic net, and support vector machine (SVM)) used to classify households into one of the five annual load profiles based on household and historical electricity usage information. Parameter settings for the five machine learning models are provided in Table A.11. The elastic net model performs the best overall, with a test accuracy of nearly 85%, thus demonstrating the effectiveness of combining the L_1 and L_2 regularizations, closely followed by the random forest (RF) model with an 83% test accuracy. Perhaps due to limited training data and complex structure of deep neural networks, the performance of deep neural networks is not

outstanding. Despite having the best training accuracy (96%), support vector machine had the lowest test accuracy (70%), indicating it overfitted the training data substantially.

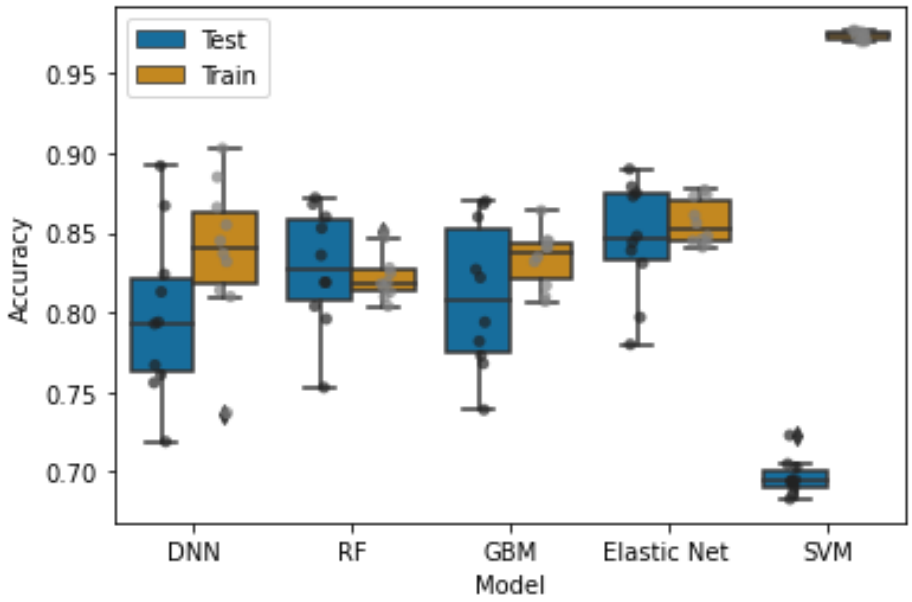


Figure 2.7. Boxplot of training accuracy and test accuracy for the five machine learning models. Accuracy based on the fraction of samples (annual load profile patterns) correctly classified.

User-profiles

As part of our analysis, we try to elucidate the connection between a household load profile and various socio-demographic and dwelling characteristics, namely household type (live alone vs only adults vs adults and children), occupation (aka social grade) and employment status of the chief income earner, number of household appliances, number of residents, and number of bedrooms (see Table 2.10-2.11). Results are summarized in Table 2.12 and further discussed below.

Table 2.10. Partial list of household characteristics in the CER dataset.

Characteristic	Description	Categories
Education level	Level of education of chief income	Primary level Secondary level Third level
Employment status	Employment status of chief income earner	Employee Self-employed Unemployed Retired
Occupation	National Readership Survey (NRS) social grade of chief income earner	AB – upper and middle class C1 – lower middle class C2 – skilled working class DE – working class and nonworking
Internet access	Internet access availability	Yes No

Household type	Number/ages of people living with	Live alone All over 15 years old Adults and children under 15 years old
House type	Type of property	Apartment Semi-detached Detached Terraced Bungalow Refused
Homeownership	Own or rent property	Rent (from a private landlord) Rent (from a local authority) Own outright (not mortgaged) Own with mortgage Other
No. bedrooms	Number of bedrooms	Range 1 to 6
No. people	Number of residents	Range 1 to 6
No. household appliances	Number of household appliances	Range 11 to 29

Table 2.11. NRS Social grade classification scheme.

Social Grade	Social Class	Description
A	Upper middle class	Higher managerial role (administrative or professional)
B	Middle middle class	Intermediate managerial role (administrative or professional)
C1	Lower middle class	Supervisory or clerical and junior managerial role (administrative or professional)
C2	Skilled working class	Skilled manual worker
D	Working class	Semi-skilled and unskilled manual worker
DE	Nonworking	Pensioners, casual and lowest grade workers, unemployed with state benefits only

User Profile 1: This user profile represents the largest subgroup of Irish households (29%) with moderate electricity usage (26 kWh daily average). User profile 1 can generally be classified as middle-income, working adult families and more affluent retirees with moderate energy usage. Household type is predominately made up of adults only (62%), some with children (33%), and very few living alone (5%). In terms of social grade, a majority are either lower middle class (29% C1) or nonworking (34% DE), with most being employees (53%) or retirees (29%). Profile 1 users have an intermediate number of bedrooms (3-4) and intermediate number of household appliances (15-18).

User Profile 2: This subgroup of households has the highest overall electricity usage of any user profile (40 kWh daily average), including the highest Christmas Day usage peak (see Figure 2.6). This user profile can be generally classed as affluent families with children and high energy usage. A majority of households have children under 15 years old (50%) and are mainly employed in higher/intermediate

(26% AB) or junior (35% C1) administrative and professional roles. Unsurprisingly, this user profile is also characterized by the highest percentage of self-employed (19%), the lowest percentage of retirees (13%), the most number of appliances (15-19), and the most number of bedrooms (3-5).

User Profile 3: This subgroup of households can be mainly described as lower-income adult families and less affluent retirees with low energy usage (14 kWh daily average). Most households are composed of adults (58%) only or single adults (33%). A majority are retired (51%). Among the employed and self-employed, most households are lower middle class (20% C1) in junior administrative and professional roles. This user subgroup has the fewest household appliances (13-16) and lives in smaller-sized houses (2-4 bedrooms).User Profile 4: Households in this subgroup are similar to user profile 3, but with higher electricity usage (20kWh daily average). In comparison to user profile 3, households are slightly less affluent, with more lower-middle class (22% C1) and working class (18% C2) main income earners, have fewer adults living alone (20%), more families with children (15%), and more appliances (15-18), and live in larger houses (3-4 bedrooms).

User Profile 5: The final and smallest subgroup of Irish households (13%) has the lowest energy usage (10 kWh daily average) of any user profile. This user profile is best described as middle-income single adults and more affluent retirees with low energy usage. A clear majority live alone (58%), the highest of any user profile; very few have children (12%). Nearly half are employed in higher/intermediate (13% AB) or junior (36% C1) administrative and professional roles. Households with this user profile also have relatively few appliances (14-17) and live in smaller-sized houses (2-4 bedrooms).

In summary, household type, occupation, and employment status are all highly correlated with household electricity usage of Irish households. Each of these factors is likely to strongly influence lifestyle, including work and leisure patterns, which, in turn, result in different electricity usage profiles. Our procedure for extracting typical annual load profiles and linking this to socio-demographic and dwelling data to generate user profiles offers a simple, transparent, and effective approach to a challenging, cross-domain matching problem that combines massive smart meter data with household data to extract meaningful market segmentation information.

Table 2.12. Primary socio-demographic and dwelling characteristics associated with each annual load profile.

User profile (proportion)	Avg. daily consumption (kWh)	Household type	Occupation (proportion)	Employment status (proportion)	No. household appliances	No. people	No. bedrooms
1 (29.4%)	26.0 kWh	Live alone (5%) Only adults (62%) Adults and children (33%)	AB (18%) C1 (29%) C2 (19%) DE (34%)	Employee (53%) Self-employed (12%) Unemployed (6%) Retired (29%)	15-18	2-4	3-4
2 (24.4%)	39.9 kWh	Live alone (3%) Only adults (47%) Adults and children (50%)	AB (26%) C1 (35%) C2 (18%) DE (20%)	Employee (61%) Self-employed (19%) Unemployed (6%) Retired (14%)	15-19	2-6	3-5
3 (14.0%)	13.7 kWh	Live alone (33%) Only adults (58%) Adults and children (9%)	AB (11%) C1 (21%) C2 (12%) DE (56%)	Employee (35%) Self-employed (7%) Unemployed (7%) Retired (51%)	13-16	2-3	2-4
4 (18.9%)	19.8 kWh	Live alone (20%) Only adults (65%) Adults and children (15%)	AB (8%) C1 (22%) C2 (18%) DE (52%)	Employee (35%) Self-employed (7%) Unemployed (7%) Retired (51%)	15-18	2-3	3-4
5 (13.4%)	10.0 kWh	Live alone (58%) Only adults (30%) Adults and children (12%)	AB (13%) C1 (36%) C2 (17%) DE (34%)	Employee (57%) Self-employed (8%) Unemployed (8%) Retired (27%)	14-17	2	2-4

2.6 Conclusion and discussion

This study considers various aspects of electricity consumption based on an analysis of smart meter electricity data and household level characteristics, namely temporal and climate influences on electricity consumption, daily consumption pattern shifting over time, and prediction of annual load profiles. For the first of these, we established an apparent causal effect of temperature, day of week, and month on household electricity consumption. We found that the GAM model that uses smoothing functions for independent variables produced a noticeably better fit than standard linear regression (+0.11 adjusted R^2 value). More importantly, we observed that 5°C represents a critical threshold for electricity consumption – for temperatures below 5°C, electricity demand rapidly increases, whereas at 5°C and above, there is a small negative effect of increasing temperature on electricity demand. This monotonically decreasing temperature sensitivity curve is notably different from the "U" shaped curve of other countries, like the US and China. Season also has a powerful influence on electricity consumption. In spring to early autumn (March and October), electricity demand is less than in late autumn and winter (November to February). Day of week also affects electricity usage, with Mondays and weekends seeing noticeably higher demand than other weekdays. It is likely that this is driven by increased leisure time at home (weekends) and increased cleaning and food preparation on the first day of the week (Monday).

As part of our electricity consumption pattern shifting analysis, we identified six distinct intra-day patterns between July 2009 and December 2010. Moving from summer to winter, household intra-day load profiles changed dramatically in shape and volume. For example, the most common pattern reduced from 42-44% of households in July to 21-25% in December, indicating the extent to which daily electricity consumption behaviors change over a year (i.e., higher energy use associated with secondary electric heating, water heating, and a more in-door lifestyle during colder months with less daylight versus more outdoor time and less energy use in warmer months with more daylight). In addition, we observed a clear weekday versus weekend cycle and a very noticeable spike in demand on Christmas day. We note that the data collection period corresponds to just after the 2007-08 Financial Crisis when the Irish economy was in a state of flux. Indeed, GDP in Ireland fell 5.1% in 2009 compared to 2008 (European

Commission 2013). This context may have influenced observed household electricity consumption behaviors.

Finally, we assessed the performance of various machine learning models to predict annual electricity consumption patterns based on household characteristics and historical usage. The best fitting model, the elastic net model, achieved nearly 85% test accuracy for a 5-class classification problem. Further analysis of socio-demographic and dwelling characteristics associated with each load profile revealed some interesting findings. Besides household makeup and size, occupation of the main income earner (social class indicator) had a strong influence on electricity usage. It was found that those employed in higher/intermediate and junior administrative and professional roles had distinctly higher energy demand than semi-skilled and unskilled workers and the unemployed, even when children lived in the household.

These results have meaningful implications for electricity suppliers/operators and policymakers. Firstly, given the strong influence of temperature, day of week, and season on electricity usage, power suppliers and power grid operators can predict electricity demand based on climate forecasts and proactively take measures to balance supply and demand. Secondly, given the extent of daily electricity consumption pattern shifting seasonally, over a week, and on national holidays (Christmas), the use of time-of-use price schemes may go a long way to smoothing out electricity demand over a day. Finally, our market segmentation of user profiles could support more tailored energy plans for new customers based on simple household information.

In terms of potential future lines of research, extensions of our methodological approach might focus on formulating and implementing a real-time demand response system. Our current work did not examine the extent to which supply-side mechanisms household electricity consumption, which future work might address. In addition, a more robust analysis could factor in other weather variables besides just temperature, like precipitation, cloudiness, and humidity, which may affect electricity demand as well as the potential influence of government policy. Finally, our general approach could be extended to other utility sectors, like water and natural gas usage, and the commercial user sector. In short, there is ample opportunity for follow-up work.

3. Influence of population mobility on electricity consumption in seven U.S. cities during the COVID-19 pandemic

3.1 Abstract

We examine COVID-19 pandemic impacts on electricity consumption in seven US cities. A high-level analysis reveals reductions in electricity consumption were mostly short-term, mainly when lockdowns were first introduced. To gain a more in-depth understanding of COVID-19's impact, Bayesian structural time series modeling was used to decompose electricity consumption into multiple tailored components. We find that models incorporating population mobility achieved high accuracy rates using pre-COVID data and even better rates using post-COVID data. Interestingly, while electricity usage dropped during the first 6 weeks of the pandemic in most cities, in one it rose above the long-term trend.

3.2 Introduction

Following an outbreak in late 2019, the COVID-19 virus quickly became a global pandemic, causing significant loss of life and negatively impacting health and daily life (Gautam and Hens. 2020; Meyer *et al*. 2022; Maital and Barzani 2020). As the virus continued to spread, most countries around the world took active measures to contain the spread of the disease, including lockdowns and restrictions on travel. In this context, new daily patterns of behavior began to emerge as a result of social distancing measures. With these changes in human activity, energy consumption patterns changed as well.

There is already a large and growing literature on the impacts of COVID-19 on energy consumption and demand. This includes COVID-19 induced changes in total energy usage (Kang *et al*. 2021), energy consumption patterns (Carvalho *et al*. 2021), and electricity consumption forecasting (Obst *et al*. 2021). For example, Narajewski and Ziel (2020) analyzed the impact of power demand shifts in European countries due to COVID-19 lockdown. Results showed that COVID-19 lockdowns had a significant impact on the level of electricity demand in Europe and altered weekly usage patterns. Chen *et al*. (2020) provided decision makers with a framework for using high-frequency indicators (such as electricity consumption) to assess the

economic impact of COVID-19 in near real-time. They found that during the pandemic, electricity consumption in Europe dropped by 10-15%. Gulati *et al*. (2021) tested the impact analysis of electricity consumption in Haryana, India using machine learning (ML) algorithms and artificial neural networks to provide electricity load forecasting to help the power station to understand the electricity consumption in the region in advance. Results revealed that artificial neural networks were more accurate than other types of ML models.

A closely related research theme has examined the nexus between COVID-19, energy production, and the environment, in particular how reduced power demand led to reductions in greenhouse gas (GHG) emissions and improved air-quality. Han *et al*. (2021), for example, used national and provincial GDP data in combination with the China Emission Accounts and Datasets to estimate carbon emission reductions in the first quarter (Q1) of 2020. They found a reduction of 257.7 Mt CO_2 (11.0%) compared with Q1 2019. Similarly, Hilares *et al*. (2020) examined the impact of social distancing measures adopted during the COVID-19 pandemic on GHG emissions from the Peruvian Interconnected Power System. They estimated reduction of 1.6 MtCO2e, equivalent to -60% compared to a reference scenario. Meanwhile, Lahcen *et al*. (2020) showed that the COVID-19 pandemic negatively affected the economy, but that the accompanying reduction in GHG emissions was disproportionate. Well-designed public policies, they argue, can simultaneously achieve economic growth and emissions abatement.

Impacts of COVID-19 on air quality have been examined by various authors in relation to energy production and transportation. Park *et al*. (2021) examined differences in CO and NO2 concentrations in Seoul between February and March 2020 compared to 2019. They found significant reductions in both indicators, up to –11.9% and –41.7%, respectively. Filonchyk *et al*. (2020) focused on the East China region, trying to estimate the impact of COVID-19 on SO_2, NO_2, and CO concentrations and aerosol optical depth levels. Results showed that the strict COVID-19 lockdown in China improved air quality initially, but as power station and oil refineries were brough back online, pollution levels quickly returned to previous levels. Jain and Sharma (2020) evaluated temporal and spatial changes of five standard air pollutants in Indian megacities over equivalent periods in 2019 and 2020. With the exception of ozone, pollutants in all major cities significantly decreased. A study by Peters *et al*. (2020)

demonstrates that improved air quality has the additional benefit of boosting solar power generation, observing that insolation increased in Delhi by 8.3% following institution of a national lockdown. A review by Eroğlu (2021) discusses a broad overview of COVID-19 impacts on the environment and renewable energy sector.

In a limited number of cases, researchers have looked at improving electricity consumption forecasting by factoring in COVID-19 government policies designed to combat the pandemic. For example, Werth *et al*. (2021) used the Oxford Coronavirus Government Response Tracker stringency index to investigate the impact of COVID restrictions on electricity usage in 16 European countries. They found four restriction measures, namely "stay at home", "school closing", "restriction on internal movements", and "workplace closing", were significantly correlated with load reductions during the pandemic. They suggested that such restriction measures could be accounted for in energy consumption forecast models to obtain more accurate predictions in the event of future emergencies.

Buechler *et al*. (2022) explored changes in electricity consumption in 58 different countries and regions around the world from January to October 2020. Their results show that stricter and more wide spread government restrictions on mobility, in particular shopping and recreation, lead to obviously decreases in electricity consumption during the initial phase of the pandemic. In addition, the impact of restrictions on electricity consumption evolved over time, with impacts being felt more at the beginning of the pandemic and then gradually decreasing.

Among the few studies to estimate counterfactual electricity usage is Prol and Sungmin (2020), who used dynamic harmonic regression to estimate baseline daily electricity consumption assuming COVID-19 never occurred in nine of the most impacted European countries and US states. Their proposed method was able to achieve test accuracy rates of between 2.7% and 4.6% based on mean average percentage error. They found a non-linear relationship between a COVID-19 response stringency index and declines in electricity consumption.

At present, there has been a lot of research on the interplay between COVID-19 and energy consumption. However, there are three main deficiencies. First, most studies fail to investigate changes in energy consumption patterns at varying time scales (e.g., intra-day, daily, weekly

cycles), though there are exceptions (Burleyson *et al*. 2021). Any reductions in electricity consumption as a result of COVID-19 will invariably be accompanied by a change in load profile vis-à-vis changes to behavioral patterns. Exploring changes in electricity consumption patterns is generally more helpful in understanding changes in people's behavior and vice versa.

A second limitation is that studies have typically investigated changes in energy consumption due to COVID-19 from a static point of view. There is a lack of comprehensively exploring relationships between electricity consumption and key drivers, including population-related factors (e.g., number of people staying at home versus going to work or retail stores), weather-related factors (e.g., temperature), and public health data (e.g., number of positive cases, number of fatalities), all of which vary over time. A nice counter example to this generalization is Li *et al*. (2022), who investigate the link between electricity demand and the effective reproductive number of COVID-19 in Germany and 5 US states.

Third, the existing literature has mostly failed to unpick the causal factors behind shifts in energy consumption. Because electricity consumption in 2020 was generally less than 2019, most studies infer that this decline in power consumption was caused entirely by the pandemic. In fact, in both 2019 and 2020, electric power consumption in the US fell slightly (EIA 2023a), ostensibly due to more mild weather (EIA 2023b). Therefore, it would be interesting to identify how much any decrease was caused directly by the pandemic versus other factors.

This study attempts to fill these gaps. More specifically, using an integrated time series combining electricity demand with weather records, COVID-19 caseloads, and mobile device location information, we first explore how electricity consumption patterns changed between 2019 and 2020 over varying time scales (year, month, week, day) in seven US cities. Bayesian structural time series (BSTS) models are subsequently used to decompose the electricity consumption data into multiple tailored components, including long-term trend, seasonality, and population mobility related factors. From this, we infer the importance of different factors in influencing electricity demand as well as predict net changes in energy consumption that can be attributed specifically to COVID-19.

BSTS has obvious advantages over other classical time series models, such as Autoregressive

Integrated Moving Average (ARIMA), Seasonal and Trend decomposition using Loess (STL), and Exponential Smoothing (ETS) variants. First of all, BSTS can perform the decomposition of the time series in an additive way, so it is more interpretable than ARIMA and is easier to explain why the time series exhibits a particular change in pattern. BSTS can also simultaneously handle multiple types of seasonality (e.g., weekly, monthly, quarterly), unlike standard STL and ETS. Perhaps most important of all, BSTS can further take advantage of external features (e.g., population mobility variables), not just internal features generated by differencing, lags, and moving averages of a time series. Including internal features can alter the structure of the underlying time series and produce models that are more focused on local effects. BSTS, on the other hand, is able to decompose the entire time series and provide a more holistic feature based decomposition. The end result is that BSTS can both high prediction accuracy and strong explanatory ability.

The advantages imparted by BSTS allow us to perform a bespoke decomposition to better understand the impact of the pandemic on electricity consumption in relation to other dependent factors, including population mobility. As part of this, we conduct a scenario analysis of the COVID-19 outbreak, predicting the likely trend of electricity consumption had no outbreak occurred based on pre-COVID-19 data compared to what did occur in order to estimate the "true" effect of COVID-19 on electricity consumption.

The remainder of this chapter is as follows. In the next section, we give a detailed description of the data used in this study and our methodology for forecasting electricity consumption based on causal inference. This is followed by a presentation of main results and findings. We conclude with a discussion of the implications of our findings and avenues for future research.

3.3 Methodology

3.3.1 Data

We use a cross-domain open access data center COVID-EMDA+ (Coronavirus Disease and Electricity Market Data Aggregation+) developed by Ruan *et al*. (2020) to build our model and measure the impact of COVID-19 on the US power sector. The data center integrates information from the electricity market with heterogeneous data sources, such as COVID-19

public health data, weather, mobile device location information, and satellite image data. A summary of the variables in the COVID-EMDA+ dataset is provided in Table 3.1.

Table 3.1. Variable information for the COVID-EMDA+ dataset.

Variable	Category	Description	Period*
Electricity	Electricity Market	Electricity usage (30 minutes)	a
Temperature	Weather	Daily average air temperature	a
Cases	Public Health	Daily newly confirmed COVID-19 cases	b
Deaths	Public Health	Daily recorded COVID-19 deaths	b
Home	Mobile Device Location	Daily total number of devices that stay at home for 24 hours	b
Work	Mobile Device Location	Daily total number of devices that go to places of work for more than 6 hours	b
Restaurant and Recreation	Mobile Device Location	Daily total number of visits to restaurants and recreation venues	b
Retail	Mobile Device Location	Daily total number of visits to retail establishments	b

* a = 1 Jan 2017 to 16 Apr 2021; b = 1 Jan 2019 to 16 Apr 2021

3.3.2 Effect of COVID-19 on electricity consumption patterns

Figure 3.1 displays the number of recorded COVID-19 cases in Los Angeles, Houston, Boston, New York City, Chicago, Philadelphia, and Kansas City (hereafter the seven US cities) from the start of 2020 to early 2021. Apart from Los Angeles, Houston, and Kansas City, the number of cases in these cities showed a large uptick starting in Mar 2020 (2020-03), reaching an initial peak in May 2020 (2020-05), followed by a slight decline and then a second, much larger peak between Nov 2020 and Jan 2021 (2020-11 to 2021-01). The initial peaks in Los Angeles, Houston, and Kansas City started months later than the other cities, around Jun 2020 (2020-06).

In most places across the US, daily routines were significantly affected by the COVID-19 pandemic from Mar 2020 onward. Significant proportions of people were required or voluntarily opted to work remotely from home and avoid public spaces, thus significantly reducing travel to normal places of work, as well as visits to retail establishments, restaurants, and recreation venues. In the first part of our investigation, we carry out a descriptive analysis of overall net change in electricity consumption during the pandemic as well as changes in electricity consumption patterns at varying sub-yearly time scales (month, week, day). As a reference point, we cross compare electricity consumption during equivalent periods in 2019.

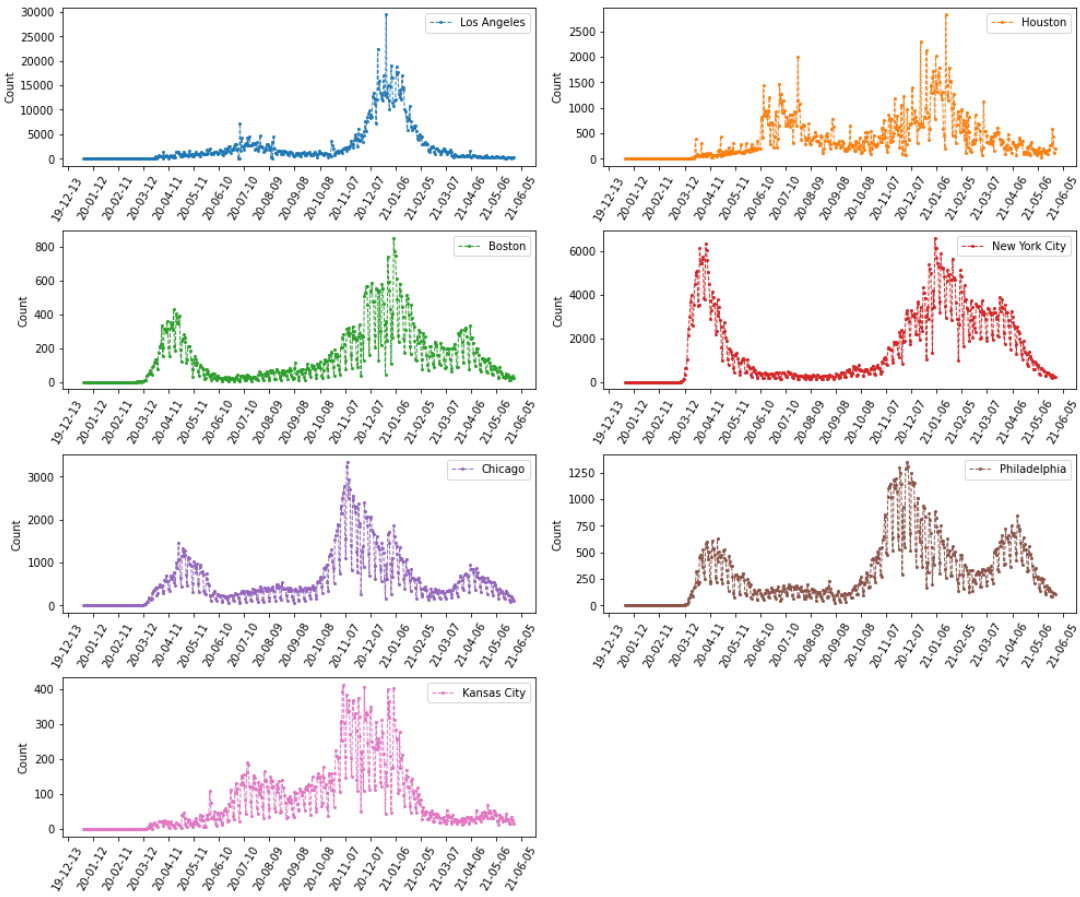


Figure 3.1. Daily reported COVID-19 cases in Los Angeles, Houston, Boston, New York City, Chicago, Philadelphia and Kansas City from 1 Jan 2020 to 16 Apr 2021.

3.3.3 Daily load forecasting in the context of COVID-19

The second part of our analysis makes use of BSTS to capture the dynamic relationship between electricity consumption and various influencing factors, including mobile device location data which we use as a proxy for degree of population mobility. The relative importance of each influencing factor is examined as well the accuracy of short- and long-term forecasts produced by our BSTS models. Lastly, we use our BSTS models to predict daily electricity consumption in the absence of COVID-19 in an attempt to estimate changes in electricity usage caused by the pandemic. The various steps involved in our analysis (Figure 3.2) are discussed in more detail in the following subsections.

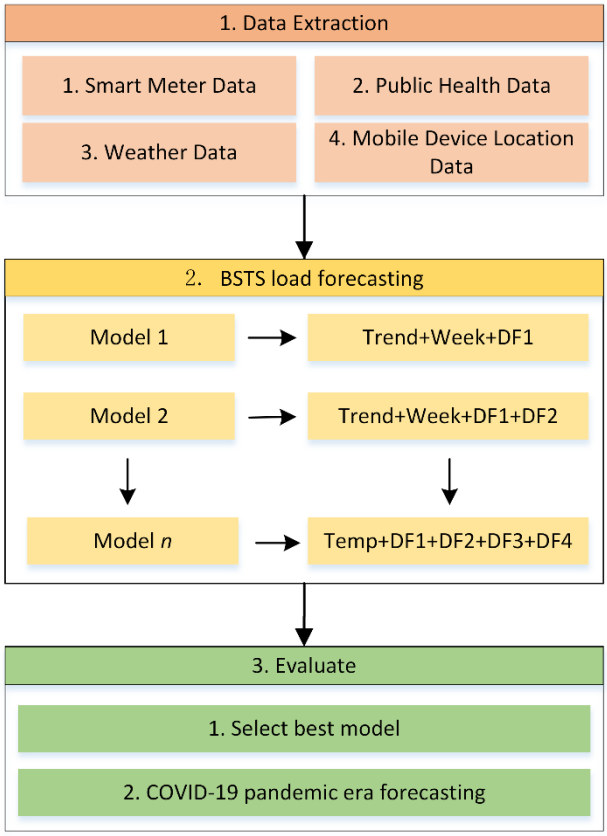


Figure 3.2. Overview of how to predict city electricity consumption and causal inference.

Data extraction

In the data extraction stage, we created multiple datasets containing different features contained in the open data hub COVID-EMDA+. This included different levels of aggregation of electricity usage over time, time interval, air temperature, and number of people categorized as being at home, work, restaurants and recreation venues, and retail establishments (Table 3.1).

Bayesian structural time series load forecasting

In the load profile prediction step, our goal is to forecast electricity usage based on population mobility metrics and other variables. For our purposes, we use a BSTS model (Brodersen *et al.* 2015). A BSTS model can be described by a pair of equations relating observation y_t of a real-valued time series to a vector of latent state variables α_t, such that:

$$y_t = \boldsymbol{Z}_t^T \alpha_t + \epsilon_t \qquad \epsilon_t \sim N(0, H_t) \qquad (3.1)$$

$$\alpha_{t+1} = \boldsymbol{T}_t \alpha_t + \boldsymbol{R}_t \eta_t \qquad \eta_t \sim N(0, Q_t) \qquad (3.2)$$

Equation (3.1) is called the observation equation because it links the observed data y_t with the unobserved latent state α_t. Equation (3.2) is called the transition equation because it defines

how the latent state evolves over time. Matrices $\boldsymbol{Z}_t$, $\boldsymbol{T}_t$ and $\boldsymbol{R}_t$ typically contain a mix of known values and unknown parameters. Model (3.1)-(3.2) is said to be a state space model.

BSTS provides considerable flexibility with which to decompose a time series based on trend, seasonality, regressors, and potentially other state variables. In our application, we use a typical BSTS model obtained by adding trend, seasonality, and dynamic regression components. Such a model can be written as:

$$y_t = \mu_t + \tau_t + \boldsymbol{\beta}_t^T \boldsymbol{x}_t + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{3.3}$$

$$\mu_t = \mu_{t-1} + \delta_{t-1} + u_t \qquad u_t \sim N(0, \sigma_u^2) \tag{3.4}$$

$$\delta_t = \delta_{t-1} + v_t \qquad v_t \sim N(0, \sigma_v^2) \tag{3.5}$$

$$\tau_t = -\sum_{s=1}^{S-1} \tau_{t-s} + w_t \qquad w_t \sim N(0, \sigma_w^2) \tag{3.6}$$

where μ_t in (3.3)-(3.4) is the local linear trend, δ_t in (3.4)-(3.5) is the slope of the trend generated by a random walk, and τ_t in (3.3) and (3.6) is a seasonality factor that can vary over time by adding w_t. Terms ϵ_t, u_t, v_t, and w_t are all independent Gaussian random noise with mean 0 and variances σ_ϵ^2, σ_u^2, σ_v^2 and σ_w^2, respectively. Vector $\boldsymbol{\beta}_t$ contains the dynamic regression coefficients for explanatory variables $i = 1, \dots, n$ specific to period t. Each individual coefficient β_{it} is assumed to follow a random walk of the form $\beta_{i(t+1)} = \beta_{it} + \eta_{it}$, such that $\eta_{it} \sim N\left(0, \sigma_{\beta_i}^2\right)$. In BSTS, a noninformative spike and slab prior distribution is typically used for the dynamic regression coefficients $\boldsymbol{\beta}_t$, which in effect assumes that most elements of $\boldsymbol{\beta}_t$ are zero. Equation (3.3) indicates the observed value y_t for a time series can be decomposed into terms for trend μ_t, seasonality τ_t, and explanatory variables $\boldsymbol{x}_t$ in an additive way. Model (3.3)-(3.6) can be solved using a Monte Carlo Markov Chain (MCMC) algorithm. Further details about BSTS are provided in Durbin and Koopman (2012) and Brodersen *et al*. (2015).

Since BSTS does not give a unique decomposition of time series, we constructed and tested multiple models (Table B.1). The main logic we used for constructing models was that electricity consumption had an underlying trend and this trend could be extracted in one of two ways. The first relied on a moving average or other smoothing methods to characterize the local

linear trend. The second was to use temperature, which is closely related to electricity consumption. For both methods, we consider whether to include a day of week (seasonality component) or not. Finally, we considered the influence of population mobility variables (dynamic components) on electricity consumption, specifically numbers of people at home, traveling to work, visiting restaurants and recreation venues, and shopping at retail establishments.

Given potential interactive effects among the population mobility variables, we tried all combinations containing at least one population mobility factor. In addition, we created two null models with trend and seasonality, but no population mobility predictor variables. This resulted in a total of 62 models (Table B.1). For example, model 1, decomposes the electricity consumption data into a local linear trend (Local_Trend), day of week (Week), number of people traveling to work (Work), number of people staying at home (Home), number of people visiting restaurants and recreation venues (Restaurant_Recreation), and number of people shopping at retail establishments (Retail) as follows.

$$\begin{aligned} \text{Load}_t = \text{Local\_Trend}_t + \text{Week}_t + \beta_{1t} \times \text{Work}_t + \beta_{2t} \times \text{Home}_t \\ + \beta_{3t} \times \text{Restaurant\_Recreation}_t + \beta_{4t} \times \text{Retail}_t + \epsilon_t \end{aligned} \tag{3.7}$$

In (3.7), Load_t, Local_Trend_t, and Week_t correspond, respectively, to y_t, μ_t, and τ_t in equation (3.3). Parameters β_{1t}, $\beta_{2,t}$, β_{3t}, and β_{4t} denote the dynamic coefficients for the four explanatory variables (i.e., Work, Home, Restaurant_Recreation, Retail). Note that since coefficients vary with time, the relationship between any of the explanatory variables and Load is allowed to be non-linear.

In cases where temperature (Temp) is used to describe the long-term trend, we add an additional coefficient to the model. Model 31, for example, is expressed as:

$$\begin{aligned} \text{Load}_t = \beta_{0t} \times \text{Temp}_t + \text{Week}_t + \beta_{1t} \times \text{Work}_t + \beta_{2t} \times \text{Home}_t \\ + \beta_{3t} \times \text{Restaurant\_Recreation}_t + \beta_{4t} \times \text{Retail}_t + \epsilon_t \end{aligned} \tag{3.8}$$

Again, since coefficient β_{0t} are dynamic, non-linear relationships between load and temperature can be modeled. For this reason, it was not necessary to use heating and or cooling degree days as explanatory variables.

We made the explicit choice of not including the number of recorded COVID-19 cases to avoid bias. Many people, especially during the early phases of the pandemic when cheap and rapid antigen tests were still not widely available, may have been infected but not recorded as positive. Recorded cases, therefore, may not be reflective of the true number of cases. Besides, population mobility variables partially account for recorded COVID-19 cases. We also did not include number of deaths given there can be a 2 to 8 week lag between contracting the disease and death (WHO 2020).

Model evaluation

To evaluate performance, we tested models over pre- and post-COVID-19 timeframes (Figure 3.3). For the pre-COVID-19 timeframe, we trained and tested the models on data obtained strictly prior to the first peak of the pandemic. For the post-COVID-19 timeframe, we used a longer time series containing the pandemic's first peak and then validated model performance during its second peak. Since the timing of the second peak varied among the seven US cities, we designed tailored training and test datasets.

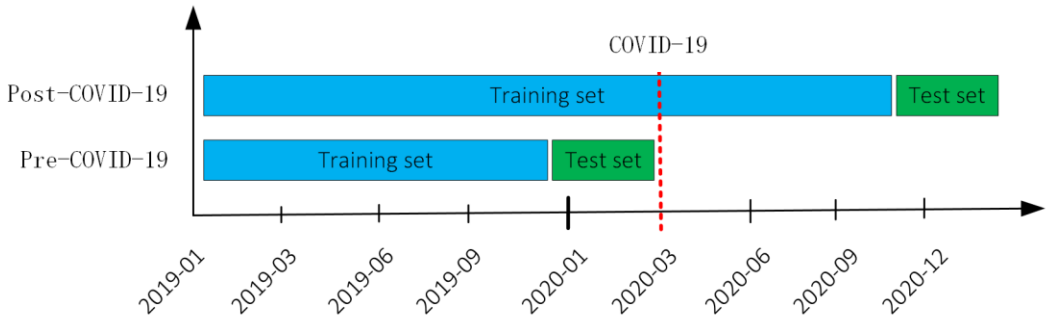


Figure 3.3. Pre- and post-COVID-19 training and testing.

In the final step of our analysis, we use the best fitting pre-COVID null model to estimate the short-term impact of COVID-19 on electricity consumption in each city. A natural assumption would be that following the initial outbreak of COVID-19 and ensuing lockdowns, city wide electricity usage would reduce. However, at the time of the outbreak, electricity consumption in each city was already declining as temperatures gradually began to rise in March. Attributing any reduction in electricity usage to COVID-19, therefore, is not straightforward. To help unravel this, we used the pre-COVID-19 null model to forecast electricity consumption during the first 6 weeks of the pandemic. We took the difference between predicted and observed

values to estimate the net decrease in electricity consumption caused by COVID-19. A null model, which excludes population mobility as a predictor, was used to estimate counterfactual electricity usage, rather than say the overall best fitting pre-COVID model, due to the fact that COVID-19 and the ensuing lockdowns themselves strongly influenced population mobility patterns during the early part of the pandemic.

3.3.4 Implementation

Model development and experimentation were performed using a combination of Python and R on a standard PC with an Intel Core TM i5-8250U CPU running at 1.60GHz with 8.0GB RAM. BSTS models were solved using the R bsts package. For the MCMC algorithm, the number of draws was set to 1000.

3.4. Results

3.4.1 Comparison of electricity consumption patterns in 2019 and 2020

Total US electricity consumption in 2020 was about 4% lower than in 2019 (Table 3.2), with big decreases in the commercial (–5.4%) and industrial (–4.3%) sectors. The largest relative decrease was observed in the public transportation sector (–14.2%), mainly as result of reduced population mobility associated with COVID-19 lockdowns. Conversely, residential electricity consumption increased slightly (+1.7%) in 2020 due to significant numbers of people spending more time at home.

Table 3.2. Electricity consumption by sector in 2019 and 2020. Source: https://www.eia.gov/totalenergy/data/browser/index.php.

		Electricity Consumption*		
Sector	Proportion (%)	2019	2020	Change (%)
Residential	39.4%	1440.3	1464.6	+1.7%
Commercial	34.6%	1360.9	1287.4	–5.4%
Industrial	25.8%	1002.4	959.1	–4.3%
Transportation	0.2%	7.6	6.5	–14.2%

*Billion kWh

At monthly time scales, total electricity varied considerably in from 2019 to 2020 in the seven US cities that are the focus of this study (Figure 3.4). Of note, electricity consumption was considerably lower in Jan 2020 compared to Jan 2019 for all but one city. This aligns with the

national picture, which saw electricity usage decrease year-on-year in 2019 and 2020 (EIA 2023a). The outlier was Houston, which had slightly higher electricity consumption in Jan 2020.

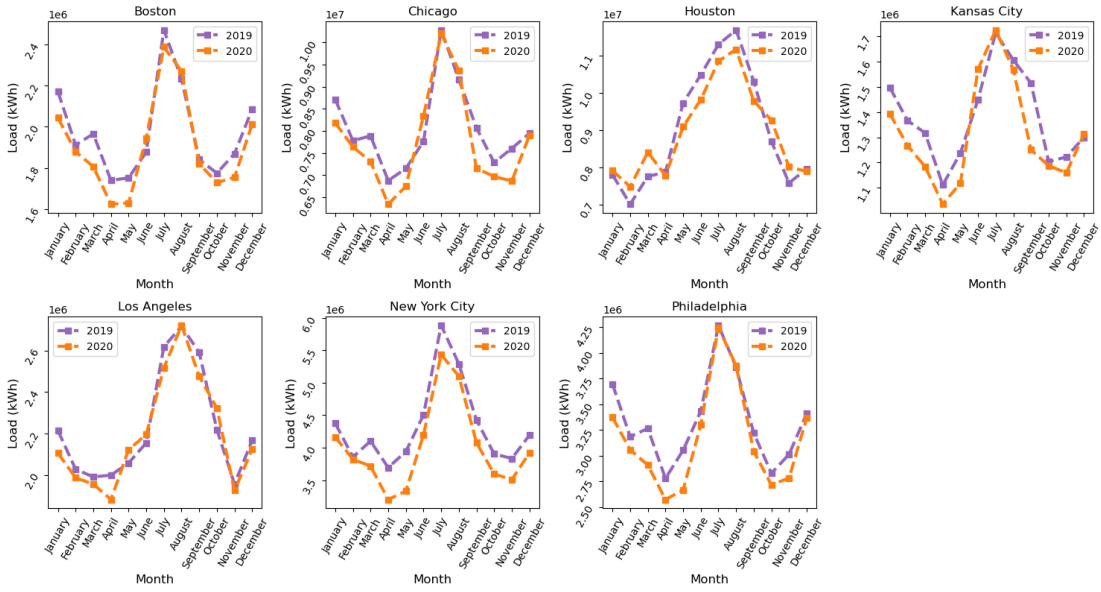


Figure 3.4. Monthly electricity usage in 2019 and 2020 in seven US cities.

In five out of seven cities (i.e., Boston, Chicago, Kansas City, New York, Philadelphia), the largest reductions in electricity consumption between 2019 and 2020 occurred during the months of Mar, Apr, and May 2020 when lockdowns were first introduced. In the same five cities, differences in monthly electricity consumption compared to 2019 were much smaller between Jun and Aug, albeit with some exceptions, as the number of COVID-19 cases were declining. In one or more months from Sep to Nov 2020 as cases began to ramp up again, the five cities experienced another significant drop in electricity usage relative to the previous year, though smaller in magnitude compared to Mar-May. By Dec, usage in 2020 was similar to 2019. Overall, the effect of COVID-19 on electricity consumption in those cities was more significant during the pandemic's first peak compared to the second, in spite of the number of cases often times being far higher in Dec 2020 than in Mar-Apr 2020. The two outliers were Los Angeles and Houston. For Los Angeles, May electricity usage was slightly higher in 2020 and there was no second drop in electricity usage in the autumn. In Houston, on the other hand, electricity consumption only started reducing in May 2020 and continued all the way to Sep, but then exceeded 2019 usage in Oct and Nov 2020.

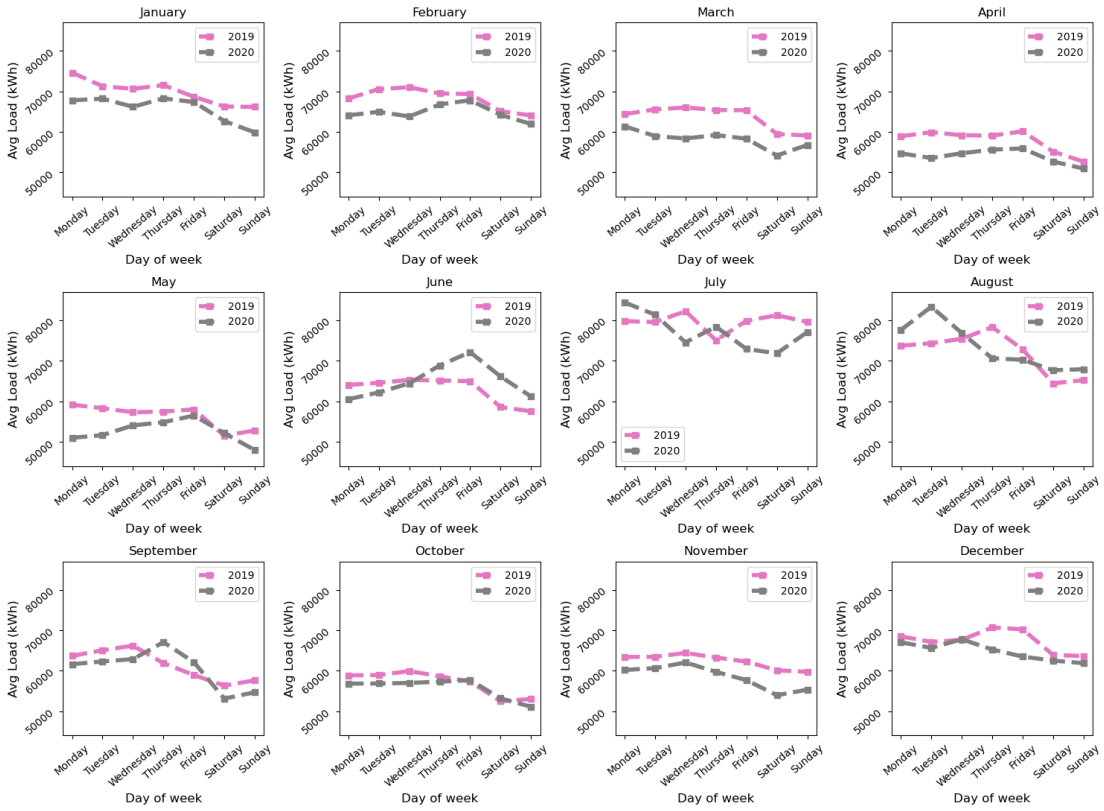


Figure 3.5. Average electricity consumption Monday to Sunday each month in 2019 and 2020 for Boston.

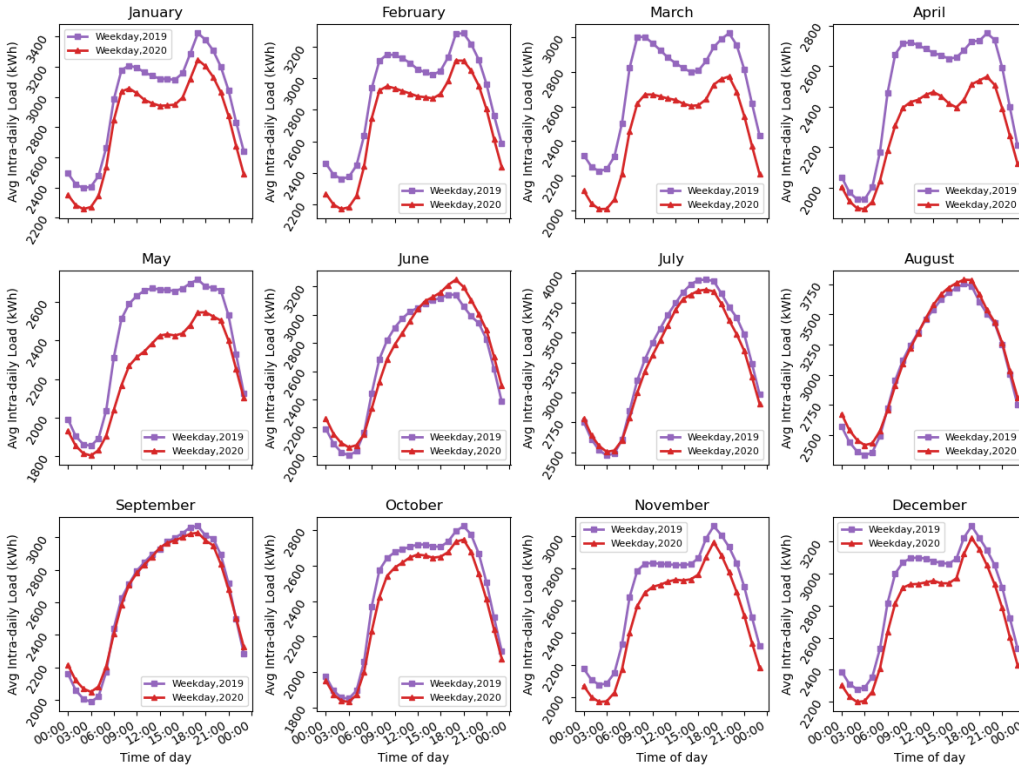
We further investigated weekly electricity consumption patterns (Figure 3.5, Figures B.1-B.6). For Boston, we can see that before and during the early part of the pandemic, weekly usage patterns in 2019 and 2020 were similar in profile, but with daily consumption lower in 2020. Weekly patterns were characterized by having higher consumption on weekdays and lower consumption on weekends. From May to Sep, however, electricity consumption changed considerably as the economy recovered from COVID-19. Specifically, weekly patterns show more volatility, with 2020 electricity usage occasionally exceeding that of 2019 on certain days. From Oct to Dec 2020, weekly patterns began to realign with those in 2019. Dynamic changes in weekly electricity consumption were similar in the other six US cities (Figures B.1-B.6).

A final analysis of electricity consumption patterns was undertaken at the intra-daily level (Figure 3.6, Fig B.7.-B.12). As a general rule, electricity usage peaks in the early evening (18:00) from late spring (May/Jun) to early autumn (Sep/Oct) in each of the seven US cities. Intra-day electricity usage is noticeably different, however, in the colder months (typically Oct/Nov to Mar/Apr), with a double peak occurring in the morning (9:00) and evening (18:00). In all seven

cities, intra-day electricity usage in July is unique in that weekdays and weekends patterns are virtually undisguisable, whereas in other months weekday usage usually follows a similar shape but is significantly higher during daytime hours.

Comparing 2020 and 2019, intra-day load profiles were generally similar in shape in any given month, but shifted downward, particularly during Mar to May when the first peak of the pandemic hit. Other than that, there does not seem to be a consistent difference in intra-day profiles before or during the pandemic, except for the odd month here and there of 2020 in some cities (e.g., Apr-Jun in Los Angeles, Mar and Nov in Houston, Jun in Chicago, and Oct in Kansas City), which is inconsistent with the observation of variability for weekly electricity usage patterns between May and Sep 2020. Indeed, and there was no obvious change in intra-day load profiles during the later part of 2020 when cases were surging, indicating that it was primarily the first COVID-19 peak that had the most perceptible impact on electricity consumption.

(a)



(b)

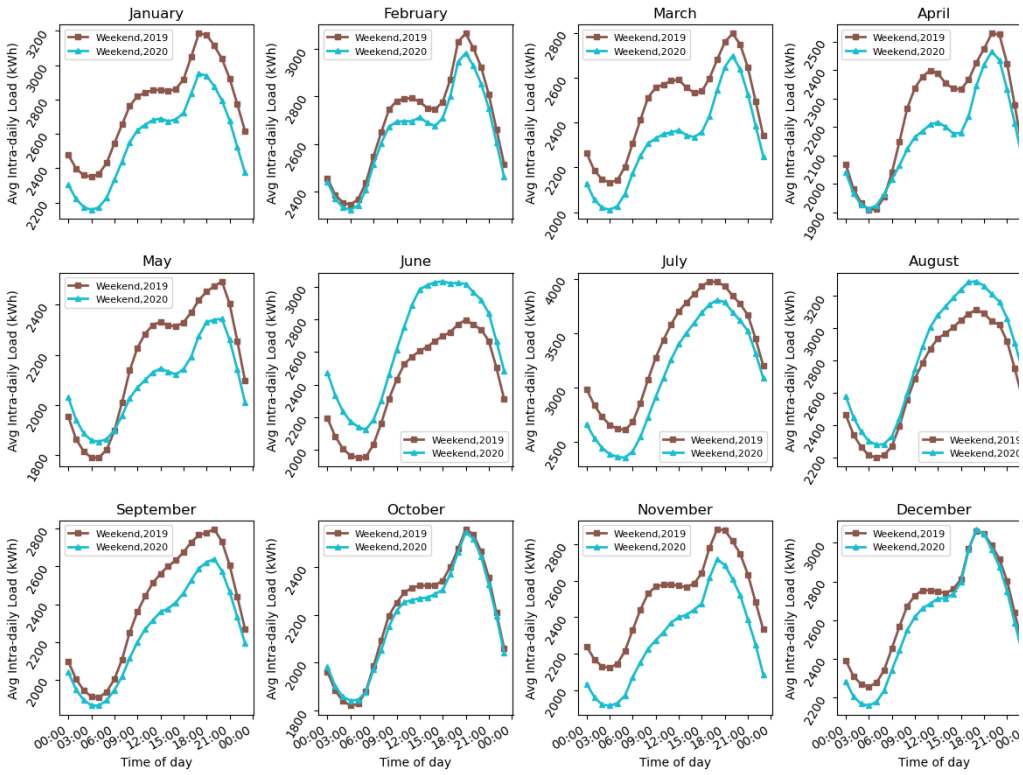


Figure 3.6. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Boston.

3.4.2 BSTS model performance

Here, we analyze forecasting performance of the 60 BSTS models developed for each city before and after the outbreak of COVID-19 that contain at least one dynamic (population mobility) factor. Models were numbered in such a way that they accounted for additional components, with models 1-15 including local linear trend (hereafter local trend) and day of week seasonality, models 16-30 including local trend only, models 31-45 including temperature trend plus weekly seasonality, and models 46-60 including only a temperature trend. For each model, we assessed out-of-sample mean absolute percentage error (MAPE) over 1-6 weeks starting in the second half of January 2020 (pre-COVID, Tables B.2-B.8) and over 1-6 weeks starting in December 2020 (post-COVID, Tables B.9-B.15).

We find that in nearly all cases, models with a temperature trend had higher, often times significantly higher average errors than those with a local trend. This is mainly down to the fact that the local trend for daily load is typically much smoother and less volatile than temperature. Take pre-COVID models for Los Angeles as an example. Models 1-15 (local trend plus weekly

seasonality) and models 16-30 (local trend only) had average errors of 4.89% and 2.88%, respectively, while models 31-45 (temperature trend plus weekly seasonality) and models 46-60 (temperature trend only) had average errors of 7.01% and 8.43%, respectively. Kansas City was the sole outlier in which post-COVID models with a local trend did not completely outrank temperature trended models, though post-COVID models with a local trend still had the lowest overall average error. Our interpretation of this is that although there is a close relationship between temperature and electricity consumption, this relationship is mainly an extreme value relationship (Di Lascio *et al*. 2020), meaning that very high or low temperatures are associated with significant increases in electricity consumption. Interestingly, average MAPE for models with local trend and weekly seasonality (models 1-15) were in nearly all cases very similar to models local trend only (models 16-30), suggesting that population mobility factors at least partially account for weekly seasonality patterns.

Unsurprisingly, the accuracy of short-term forecasts was generally higher than that of long-term forecasts for both pre- and post-COVID models. Average MAPE for Los Angeles pre-COVID with a local trend only (models 16-30), for example, increases from 2.31% to 3.63% moving from a 1 to 6 week timespan. More surprising is the fact that for all cities, post-COVID models had slightly better overall fit than pre-COVID models. The biggest difference occurred for Kansas city with pre-COVID models having an average MAPE of 14.42%, while post-COVID models had an average MAPE of 9.62% (almost 5 points lower).

Table 3.3. Best performing pre- and post-COVID models.

	Pre-COVID			Post-COVID		
City	Model	MAPE	Model components	Model	MAPE	Model components
Los Angeles	16	2.16%	Local_Trend, Work, Home, Restaurant_Recreation, Retail	16	2.01%	Local_Trend, Work, Home, Restaurant_Recreation, Retail
Houston	3	2.72%	Local_Trend, Week, Home	4	3.49%	Local_Trend, Week, Restaurant_Recreation
Boston	14	4.16%	Local_Trend, Week, Work, Restaurant_Recreation, Retail	31	4.45%	Temp, Week, Work, Home, Restaurant_Recreation, Retail
New York	29	2.84%	Local_Trend, Work, Restaurant_Recreation, Retail	14	2.86%	Local_Trend, Week, Work, Restaurant_Recreation, Retail
Chicago	30	2.85%	Local_Trend, Home, Restaurant_Recreation, Retail	13	4.46%	Local_Trend, Week, Work, Home, Retail
Philadelphia	6	5.41%	Local_Trend, Week, Work, Home	8	4.48%	Local_Trend, Week, Restaurant_Recreation, Retail
Kansas City	6	8.01%	Local_Trend, Week, Work, Home	46	5.36%	Temp, Work, Home, Restaurant_Recreation, Retail

The single best fitting pre- and post-COVID models based on average error across all forecasting windows are reported in Table 3.3. Errors vary from a low of 2.01% (post-COVID Los Angeles model) to a high of 8.01% (pre-COVID Kansas City model). A clear majority of best fitting models (12 out of 14) used a local trend as opposed to a temperature trend as well as included a weekly seasonality component (9 out of 14). We note that except for Houston, all best fitting models further include 2-4 population mobility factors, with staying at home and or visiting restaurants and recreation venues included in all models.

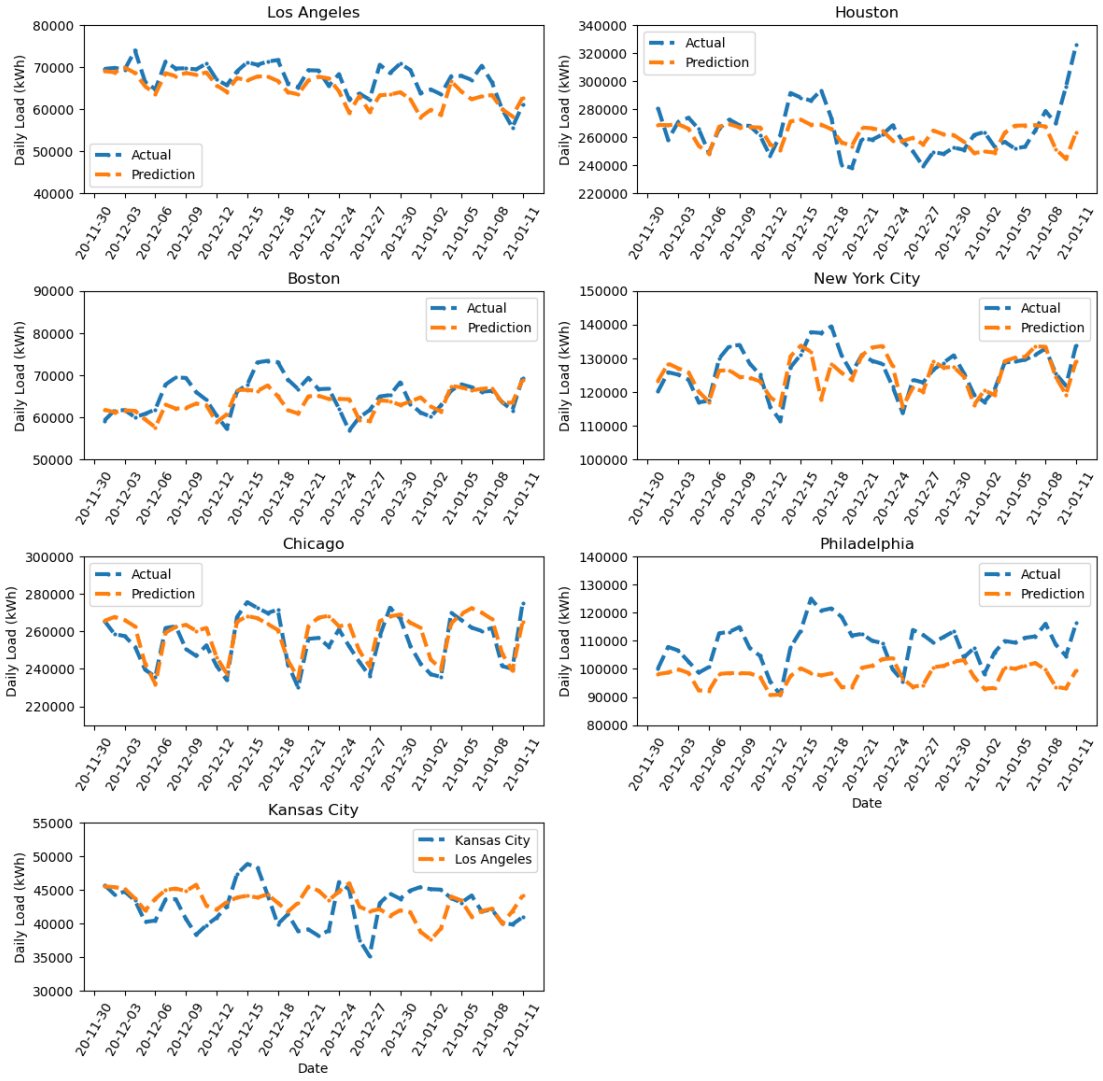


Figure 3.7. Comparison of observed versus predicted daily electricity in seven US cities from Nov 2020 to Jan 2021 based on best fitting post-COVID models.

Particularly noteworthy is the fact that most of the best fitting models had MAPEs less than 5%, indicating very good forecasting accuracy (Montaño Moreno *et al*. 2013). Indeed, BSTS looks

to be very competitive with other state-of-the-art statistical and ML methods reported in the literature. Prol and Sungmin (2020), for example, used 4 different forecasting methods (dynamic harmonic ARIMA regression, neural network autocorrelation, trigonometric seasonality with Box-Cox transformation, and STL) to predict electricity consumption and were able to achieve an average MAPE of 3.6% for the 9 geographic regions they examined. We had average MAPEs of 4.0% and 3.9% for pre- and post-COVID models, respectively. Excluding Kansas City (the city with the worst forecasting accuracy), average MAPE drops to 3.4% and 3.6%, respectively, for pre- and post-COVID models, as good or better than Prol and Sungmin (2020).

Plots of predicted daily electricity usage from the best fitting post-COVID models generally show a good fit with observed values (Figure 3.7), in particular for Chicago, Los Angeles and New York City. Unsurprising Kansas City and Philadelphia, given their much higher MAPEs, have less robust fits to actual daily electricity usage. More specifically, the overall trend of predicted and observed values is relatively consistent for Philadelphia, though the forecasting error is larger than other cities. In Kansas City, both the predicted trend and error are significantly worse than other cities, indicating that the relationship between population mobility factors and electricity consumption may be more complex and in need of further study.

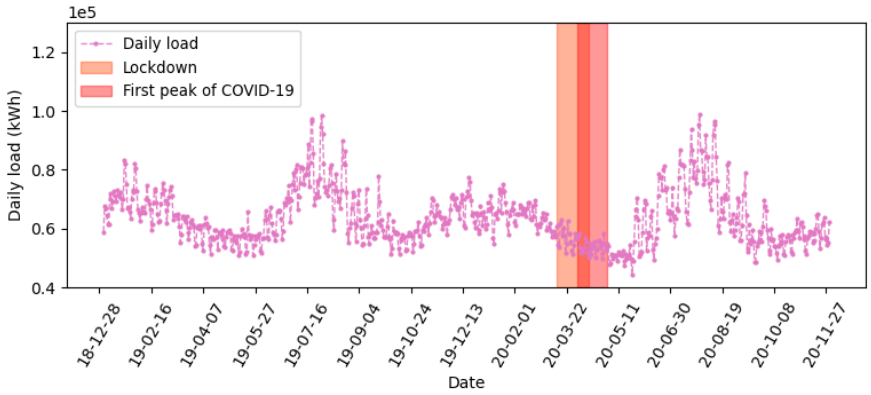
3.4.3 Effect of population mobility on electricity usage

A decomposition of the Boston electricity usage time series reveals there are clear long-term trend, weekly seasonality, and population mobility influences on electricity usage (Figure 3.8). Weekly seasonality (Figure 3.8a), for example, is responsible for the large up/down swings that can be seen in daily electricity usage through time (Figure 3.8d). Interestingly, the weekday consumption component becomes more variable and more similar to weekend usage after the first lockdown (Figure 3.8d), which aligns with previous observations regarding changes in weekly patterns, especially from May to Sep 2020 (see §3.4.1).

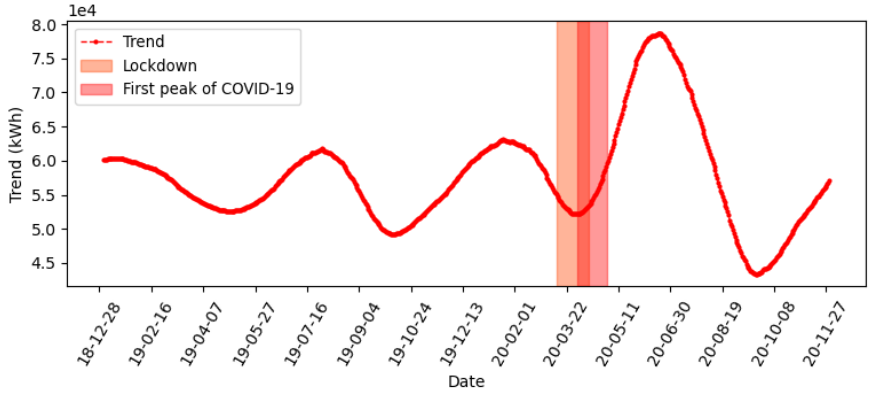
Unsurprising, trend accounts for the major component of overall electricity usage as indicated by the close orientation of the trend and daily load curves, both in terms of shape and magnitude (Figure 3.8b). It is worthwhile pointing out, however, that while the trend profiles for 2019 and

2020 are similar, the trend in 2020 shows much larger fluctuations and a shift in the timing of usage peaks and troughs during and following the initial lockdown period. In particular, there is, relative to the same period in 2019, a very sharp dip in the trend between Feb and mid-Mar 2020, followed immediately by a very steep rise that continues through May. Counterintuitively, the time interval of this trend rise corresponds precisely with the first COVID-19 peak and accompanying lockdown during which electricity consumption fell relative to 2019 (Figure 3.8a). As is clearly evident, the observed fall in daily electricity usage is almost entirely driven by the negatively trending dynamic effect from mid-Mar to May (Figure 3.8d), indicating that reduced population mobility was indeed the primary cause for the pronounced reduction in total electricity consumption. This is further supported by the fact that precisely as population mobility increases after May 2020, observed electricity consumption rises quickly and eventually even exceeds 2019 levels.

(a)



(b)



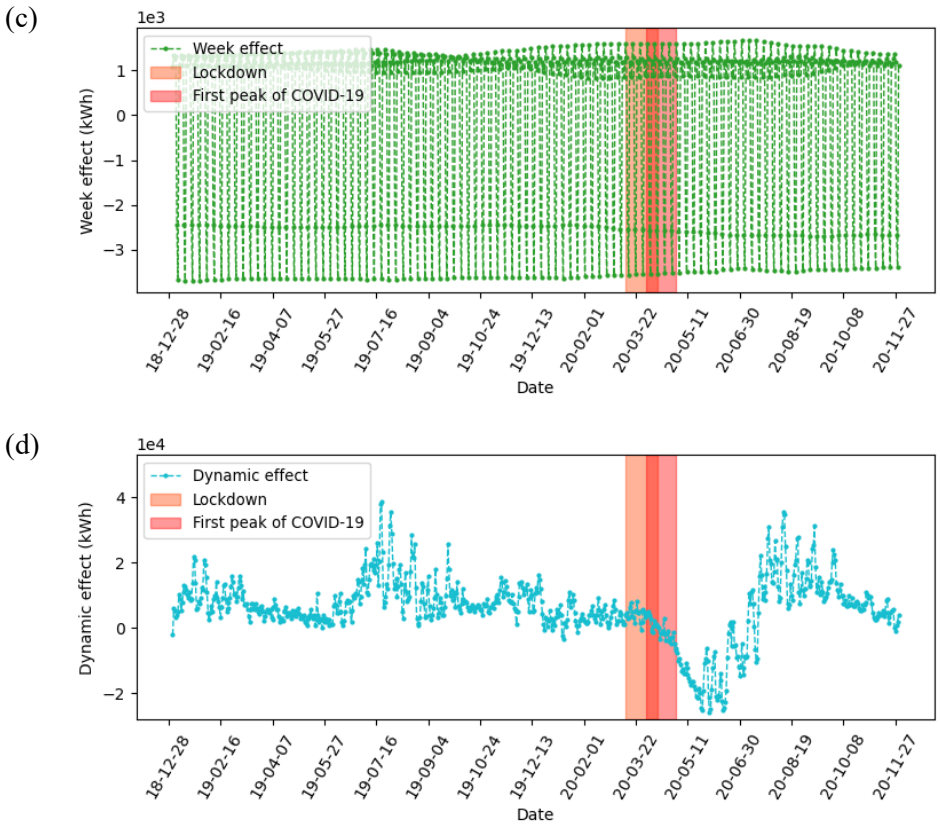


Figure 3.8. Daily load (a), trend (b), weekly seasonality (c), and population mobility (d) post-COVID model components for Boston from Dec 2018 to Dec 2020.

3.4.4 Causal impact of COVID-19 on electricity consumption

Plots of predicted daily electricity usage from the best fitting pre-COVID null models against observed electricity usage indicate a variable impact of COVID-19 on overall electricity usage (Figure 3.9). In at least six cities (Los Angeles, Boston, New York, Chicago, Philadelphia, Kansas City), electricity usage would have been expected to exhibit a slight to moderate downward trend between mid-Mar and late Apr 2020, even in the absence of the pandemic. Factoring in this downtrend, it is apparent that five out of seven cities (Los Angeles, Boston, New York, Chicago, Philadelphia) have an identifiable reduction in electricity consumption that can be safely attributed to the pandemic. Two cities (Houston, Kansas City) show a mixed or even positive impact of COVID-19 on aggregate electricity usage.

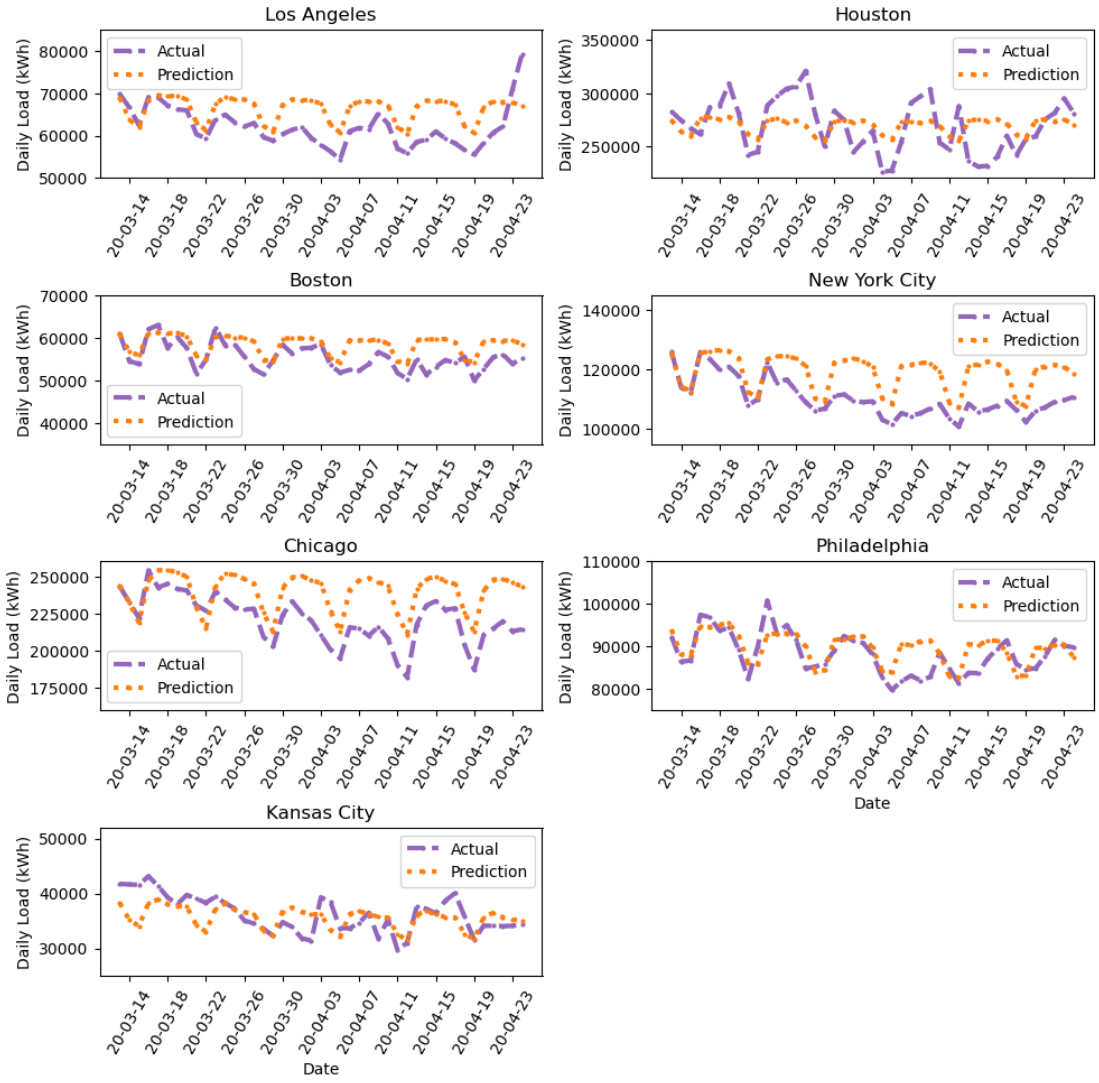


Figure 3.9. Predicted versus actual electricity usage from mid Mar to late Apr 2020 in seven US cities based on best fitting pre-COVID null models.

More specifically, Chicago showed the largest reduction in electricity usage as a result of COVID-19 (Table B.16), –2.4% after the first week of lockdown and –8.2% after the sixth week. This was followed closely by New York, which saw a fairly steady decrease in electricity, culminating in a drop of 7.3% after 6 weeks. The impact of COVID-19 was marginally less in Los Angeles and Boston, ending –6.4% and –5.3% after 6 weeks, respectively. Philadelphia had a consistent but only slight reduction in electricity usage during the COVID-19 lockdown period, going form 0.2% lower in week 1 to 1.5% lower in week 6. Meanwhile, Houston showed no consistent change in electricity usage, going from positive to negative over the weeks and ending only slightly below (–0.1%) what would have been expected by week 6. Kansas City

was the only city with a consistently positive change in electricity usage, starting off with much higher electricity usage (+10%) than would have been predicted after 1 week, then dropping but still remaining positive (+1.7%) after 6 weeks.

3.5. Discussion and conclusion

In this study, we examined how electricity consumption changed between 2019 and 2020 in seven US cities. This was done first through a descriptive analysis of electricity consumption patterns at varying time scales and second by applying BSTS forecasting to integrated time series with the aim of decomposing electricity consumption into multiple tailored components, including long-term trend, seasonality, and various population mobility related factors. From this, we were able to infer the importance of different influencing factors on electricity demand as well as predict net changes in energy consumption that could be attributed specifically to COVID-19.

Our findings reveal that observed reductions in electricity consumption during the pandemic were mostly short-term, mainly during the period when lockdowns were initially introduced from Mar to Apr 2020. Usage quickly rebounded and by Dec 2020 was similar to 2019 for most cities we looked at, in spite of the fact that the number of COVID-19 cases was often times far higher at this time than in Mar-Apr 2020.

We also find that daily electricity consumption is highly correlated with population mobility factors (e.g., number of people staying at home, going to work, and visiting restaurants, recreation venues, and retail establishments), suggesting that traditional methods of forecasting electricity usage may need to be revised to include such features. Indeed, our BSTS models based on population mobility indicators were able to achieve accuracy rates of 2.2-8.0% using pre-COVID data and even better accuracy rates of 2.0-5.4% using post-COVID data. Using these models, we were able to discern that in five out of seven cities COVID-19 caused an appreciable reduction of 1.5-8.2% in electricity usage over the first 6 weeks when lockdowns were introduced. In the other two cities, electricity usage was virtually the same (–0.1%) or actually higher (+1.7%) than would have been expected from the long-term trend.

The methodologies employed in this chapter the conclusions that we draw from our analysis

helps further our understanding of pandemic induced changes in electricity consumption. Key policy implications are that 1) the impacts of a pandemic on electricity usage are likely to be short-lived (i.e. spanning several weeks) and 2) any impacts are likely to be highly variable, ranging from negative through to neutral and to even positive effects once underlying long-term trends and population mobility are taken into account.

4. A p-median based approach to constrained clustering

4.1 Abstract

Constrained clustering is a semi-supervised method for dividing multi-attribute datasets into groups of similar elements that incorporates restrictions how data points should and or should not be clustered together. In this study, we investigate the application of the p-median model coupled with instance-level constraints, specifically must-link and cannot-link constraint, for performing constrained clustering. A Lagrangian relaxation procedure is proposed to efficiently solve large problem instances. To further speed up solution times, a simple but highly effective variable reduction technique is devised for identifying a small set of potential cluster centers. We test our modeling framework and solution approach on set of benchmark datasets and several real-world household electricity consumption datasets. We find that high-quality solutions with small optimality gaps can obtained with moderate computational effort. In addition, analysis of the largest household electricity consumption dataset for reveals that constrained p-median clusters have less extreme variability with respect to average energy usage, a more even spread of cluster sizes, and much less overlap between high- and low-income households within the same clusters compared to classic k-means clustering.

4.2 Introduction

Clustering analysis is a widely employed unsupervised data mining technique for dividing multi-attribute datasets into groups of similar elements (Rokach and Maimon 2005; Xu and Wunsch 2005). It has application in a wide array of fields, such as image segmentation (Coleman and Andrews 1979), molecular biology (Nugent and Meila 2010; Oyelade *et al*. 2016), customer relationship management (Ngai *et al*. 2009), recommendation systems (Phanich *et al*. 2010), social network analysis (Pham *et al*. 2011), and information retrieval (Leuski 2001; McCallum *et al*. 2000). The main clustering algorithms commonly in use are: k-means clustering (MacQueen 1967), hierarchical clustering (Johnson 1967), the Gaussian Mixture Model (McLachlan and Basford 1988), spectral clustering (Ng *et al*. 2001), self-organizing maps (Kohonen 2012), and Nonnegative Matrix Factorization (Paatero and Tapper 1994).

Clustering algorithms such as these can work either directly on the original set of features or new features generated through feature engineering (e.g., dimensionality reduction and feature aggregation).

Adding human intuition or subjectivity to a clustering algorithm can enhance its validity and the interpretability of clusters. In some cases, for example, an analyst may have prior domain knowledge regarding how elements should or should not be clustered together. Such information can come from experts, physical reasoning, or logical deduction. Semi-supervised clustering (aka constrained clustering) is a process of incorporating such information into a clustering algorithm with the aim of generating more logical clusters. Constrained clustering has been used in several domains and applications, such as text mining (Basu *et al*. 2004), video object classification (Yan *et al*. 2004), and RNA sequencing data (Tian *et al*. 2021).

In this chapter, we investigate the use of the p-median problem coupled with instance-level constraints for performing constrained clustering. Instance-level constraints place restrictions on cluster membership of data point pairs and include "must-link" and "cannot-link" constraints (Wagstaff *et al*. 2001). Respectively, these constraints require a given pair of points to be in or not be in the same cluster. An example where cannot-link constraints might be advantageous is to prevent high- and low-income customers from appearing in the same clusters when doing a customer segmentation analysis. Although consumption levels of the two groups may be similar, formulating rational customer sales and or relationship management strategies may necessitate these two customer groups be treated separately. Conversely, in the context of image segmentation, it may be desirable to have nearby pixels grouped together, in which case must-link constraints should be included.

A p-median problem with instance-level constraints can be readily formulated as a mixed-integer linear program (MILP). To efficiently solve large problem instances with verifiable optimality bounds, we propose a Lagrangian relaxation algorithm. We further investigate an efficient variable reduction technique involving the use of standard k-means to identify a subset of potential cluster centers that significantly reduces computational times, while still producing demonstrably good solutions.

Research on constrained clustering can be broadly divided into machine learning and mathematical programing based approaches (González-Almagro *et al*. 2023), with the vast majority of studies falling into the former category. Among the more well-known machine learning algorithms for incorporating instance-level constraints are COP-KMeans (Wagstaff *et al*. 2001), PCKMeans (Basu *et al*. 2004a), HMRF-KMeans (Basu *et al*. 2004b) and CMWK-Means (de Amorim 2012). Each of these extends the classic k-means model and has been shown to produce dramatic increases in cluster accuracy on both artificially constructed and real-world datasets. However, being in essence construction and local search heuristics, their degree of optimality is unknown and they need to be run repeatedly to ensure some level of confidence in solution quality. Additionally, while computational overhead may be low, not all guarantee that clustering constrains will be satisfied (e.g., PCKMeans and HMRF-KMeans).

Incorporation of constraints, other than instance-level constraints, has also been investigated in the machine learning literature. In Ng (2000), the k-means algorithm is modified to ensure that clusters all have a fixed number of elements. Bradley *et al*. (2020), meanwhile, add constraints to k-means clustering to avoid empty clusters or clusters with few points. Ge *et al*. (2007) extend this by further by imposing minimum variance of clusters in addition to a minimum number of points in each cluster in order to find clusters which are balanced in terms of cardinality and variance.

Mathematical programing is another powerful tool for both unconstrained and constrained clustering. For unconstrained clustering, a number of approaches have been adopted, including MILP solved by a variety of exact and heuristic methods (Brusco 2003, Xia and Peng 2005, Sağlam *et al*. 2006), dynamic programming (Os and Meulman 2004), and semi-definite programming (Peng and Wei 2007, Aloise and Hansen 2009). Constrained clustering has been addressed by the likes of Xia (2009) and Babaki *et al*. (2014). Xia (2009) proposed an adaptation of Tuy's cutting plane algorithm to find optimal solutions to k-means clustering with must-link and cannot-link constraints. Numerical experiments show the algorithm is not only able to find better solutions than COP-KMeans, but also efficiently solve medium to large problem instances involving many thousands of data points, usually in fractions of a second and only several minutes in the worst case. Babaki *et al*. (2014), meanwhile, propose the use of

column generation to solve minimum sum-of-squares clustering with must- and cannot-link constraints as well as anti-monotone constraints other than cannot-link (e.g., constraints limiting the maximum cluster size or overlap with a subset of points). Although an exact method with guaranteed optimality bounds, the solution approach of Babaki *et al*. (2014) is computationally intensive and only tested on small instances involving less than 200 data points with at most 35 dimensions and 4 labels.

Mulvey and Crowder (1979) were the first to recognize that the p-median problem could be used in unconstrained clustering. The p-median problem is a classic facility location model that seeks to locate p facilities in order to minimize the average distance between a set of customer points and their nearest assigned facility (Daskin and Maass 2015). In the context of clustering, p-median distance corresponds to a dissimilarity measure between data points. The model is equivalent to the k-medoids model well-known in the field of machine learning. The suitability of the p-median problem for unconstrained clustering has been investigated by various authors, including Klastorin (1985), Brusco and Köhn (2008), and Benati and García (2014).

To our knowledge, use of the p-median problem for constrained clustering has only been examined by Randel *et al*. (2019). The authors introduce an MILP formulation of the p-median problem with instance-level constraints and propose an efficient variable neighborhood search (VNS) heuristic. The heuristic is tested on a set of benchmark datasets generated by Xia (2009) with solution quality assessed by comparing against branch and bound (CPLEX). They find that the VNS heuristics is able to find optimal to near optimal solutions in a matter of seconds to minutes for instances with 600 or more data points.

In summary, different semi-supervised algorithms have their own advantages and disadvantages. For machine learning based approaches, algorithms are typically based on traditional clustering paradigms and adopt various strategies to incorporate instance-level constraints. Their solution quality is difficult to assess as no optimality bounds are available and they have rarely been compared against global optimization methods. In addition, machine learning algorithms are usually highly specialized in that they are designed to only handle must-link and cannot-link constraints, but cannot be easily generalized incorporate other types of constraints (e.g., anti-monotone constraints). Compared with machine learning algorithms, mathematical programing

overcomes many of these disadvantages, but are generally have a much higher computational overhead and, therefore, may be limited to tackling small problem sizes. Additional research is clearly needed on the development of alternative mathematical programing approaches capable of producing verifiable near-optimal solutions to large, constrained clustering problems. This includes decomposition techniques like Lagrangian relaxation.

The remainder of this chapter is structured as follows. In the next section, we provide a mathematical formulation of our problem along with our proposed solution methodology. This is followed by a presentation of results of numerical experiments on a set of benchmark datasets and several real-world household electricity consumption datasets. We conclude with a critical discussion of the implications of our findings and propose avenues for future research.

4.3 Methodology

4.3.1 Constrained p-Median Model

Consider the following notation. Let N, indexed by i, j, and k, be a set of data points. The distance metric d_{ij} represents the dissimilarity between points i and j. The number of cluster centers (i.e., medians) is denoted by p. Set $\mathcal{ML}$ defines the pairs of points (i, j) such that i and j must be in the same cluster, while set $\mathcal{CL}$ defines the pairs of points (i, j) such that i and j must be assigned to different clusters. The decision variables of the problem are given below.

$$y_j = \begin{cases} 1 & \text{if data point } j \text{ is selected as a cluster center} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{ij} = \begin{cases} 1 & \text{if data point } i \text{ is assigned to the cluster centered on center } j \\ 0 & \text{otherwise} \end{cases}$$

With this in place, an MILP formulation of the p-median problem with instance-level constraints is given as follows.

$$\min \sum_{i \in N} \sum_{j \in N} d_{ij}\, x_{ij} \tag{4.1}$$

$$s.t.$$

$$\sum_{j \in N} y_j = p \tag{4.2}$$

$$\sum_{j \in I} x_{ij} = 1 \qquad \forall i \in N \qquad (4.3)$$

$$x_{ij} \leq y_j \qquad \forall i, j \in N \qquad (4.4)$$

$$x_{ij} - x_{kj} = 0 \qquad \forall (i,k) \in \mathcal{ML}, j \in N \qquad (4.5)$$

$$x_{ij} + x_{kj} \leq 1 \qquad \forall (i,k) \in \mathcal{CL}, j \in N \qquad (4.6)$$

$$x_{ij} \in \{0,1\} \qquad \forall i, j \in N \qquad (4.7)$$

$$y_j \in \{0,1\} \qquad \forall j \in N \qquad (4.8)$$

The objective (4.1) minimizes the sum of distances between all points and their assigned clusters. Constraint (4.2) requires that exactly p points be selected as cluster center. Equalities (4.3) ensure that each data point is assigned to exactly one cluster, while inequalities (4.4) require that data points only be assigned to selected cluster center. Constraints (4.5) and (4.6) define the must-link and cannot-link constraints, respectively. More specifically, (4.5) states for points i and k that must belong to the same cluster, if $x_{ij} = 1$, then $x_{kj} = 1$ and vice versa. For points i and k that cannot be in the same cluster, (4.6) specifies if $x_{ij} = 1$, then $x_{kj} = 0$ and vice versa. Finally, constraints (4.7) and (4.8) impose binary restrictions on the decision variables. Note that without inequalities (4.5)-(4.6), the problem reduces to a standard p-median model.

Compared with k-means, a p-median type model has obvious advantages for constrained clustering. First and foremost, it can be solved directly to find a global optimum. As pointed out by Steinley (2015), other advantages include the fact that 1) centers are selected from the points in a dataset (as opposed to being the central point of a cluster), thereby allowing one to identify the most representative element of each cluster; 2) because dissimilarity between points is usually taken as Euclidean distance (instead of quadratic distance), it is typically more robust to outliers within a dataset; and 3) the flexibility of having a more generic pre-defined distance metric means that distances do not need to be constantly recomputed each time centers are updated and there is no imposition of having triangle inequality or distance symmetry violations.

On the downside, however, as the number of data points increases, the number of variables and constraints in a p-median model grow dramatically. This can quickly make even moderately sized datasets computationally intractable with an exact approach. To partially overcome the challenge of solving large problem instances, we present below a Lagrangian relaxation

algorithm together with a simple variable reduction technique to help reduce the number of potential center points and further speed up the Lagrangian procedure.

4.3.2 Lagrangian Relaxation

Lower Bound Problem

The Lagrangian relaxation is a classic decomposition method in which complicating constraints are relaxed and included in the objective function as penalties. The relaxed problem becomes easier to solve, while still providing useful information such as error bound on optimality.

In our implementation, we chose to relax equality constraints (4.3) with Lagrangian multipliers $\mu_i, \forall i \in N$. The result is the following dual problem:

$$z_{LR}(\boldsymbol{\mu}) = \max_{\mu} \min_{\boldsymbol{x},\boldsymbol{y}} \sum_{i \in N} \sum_{j \in N} (d_{ij} - \mu_i) x_{ij} + \sum_{i \in N} \mu_i \tag{4.9}$$

subject to constraints (4.2) and (4.4)-(4.8). For fixed values of the Lagrangian multipliers μ_i, the resulting minimization problem over variables x_{ij} and y_j can be solved using a branch and bound solver like CPLEX. This yields a lower bound to the original problem (4.1)-(4.8).

It is worth noting that a more natural choice would be to relax inequalities (4.4), which would decouple the original problem into two simpler subproblems involving the selection of cluster center variables y_j and cluster assignment variables x_{ij} separately. However, previous studies (Daskin and Maass 2015) and preliminary tests on our part showed that relaxing (4.3) produced much tighter lower bounds, albeit at the expense of solving a much more difficult relaxed problem at each iteration

Upper Bound Problem

To try and produce a feasible solution and valid upper bound to the original problem, we consider two approaches. The first is to fix the centers $y_j = 1$ selected in the relaxed dual problem and then use a branch and bound solver (e.g., CPLEX) to determine the x_{ij} assignment variables. Although guaranteed to find the best possible assignment of points to centers, the obvious disadvantage of this is the potentially long run times, especially for large problem

instances.

A second way of finding a feasible upper bound is to use a heuristic for assigning data points to centers, with centers still fixed based on the solution to the relax dual problem. For this, we implemented a modification of the approach proposed by Randel *et al*. (2019). The heuristic is divided into two main steps.

> **Step 1.** Create super points from must-link constraints. Temporarily ignoring cannot-link constraints, assign each regular or super point to its nearest available center.
>
> **Step 2.** Examine the list of point pairs with cannot-link constraint violations and try to restore feasibility by reassigning one of the two points to its nearest center without a constraint violation. After feasibility is restored, see if a reassigned point can be moved to a closer center to try to improve the solution further.

In Step 1, sample points involved in must-link constraints, which need to be assigned to the same cluster center, are merged into a super point and the distances between the super point and all other points updated. For example, assuming points p_1 and p_9 have a must-link constraint, a new super point s_1 is created and the distance between s_1 and every other point p_k is changed to $d_{1k} = d_{13} + d_{93}$. Note that the choice of specifying point p_1 or p_9 as the super point root is arbitrary. For further details, please refer to Randel *et al*. (2019). Once super points have been created, Step 1 ends by assigning each point to its nearest cluster center, while temporarily ignoring cannot-link constraints.

Step 2 involves recovering feasibility for the solution found in the first step. The procedure for doing this is outlined in Algorithm GR (GR for greedy reassignment). The notation for Algorithm GS is defined as follows. Let $Y(i) = \left\{ j \in J' \middle| x_{ij} = 1 \right\}$ denote the cluster center of point i and let $O(i) = \arg\min_{j \in J'} \left\{ d_{ij} | \nexists k \in N, (i,k) \in \mathcal{CL}, x_{kj} = 1 \right\}$ be the nearest "open" center to point i that does not violate any cannot-link constraints, with $J' = \left\{ j \middle| y_j = 1 \right\}$ being the set of p currently selected centers. The net cost of reassigning point i from $Y(i)$ to $O(i)$ is given by $\Delta d(i) = d_{iO(i)} - d_{iY(i)}$ if $O(i)$ exists, $\Delta d(i) = \infty$ otherwise.

Algorithm GR. Greedy reassignment procedure for restoring problem feasibility.

1. $V \leftarrow \emptyset, P \leftarrow \emptyset$
2. **for** each cannot-link constraint $(i, k) \in \mathcal{CL}$
3. **if** (i, k) violated **then**
4. $V \leftarrow (i, k)$
5. **end if**
6. **end for**
7. **for** each $(i, k) \in V$
8. $net\_cost_i = \Delta d(i)$
9. $net\_cost_k = \Delta d(k)$
10. **if** $net\_cost_i = \infty$ **and** $net\_cost_k = \infty$ **then**
11. **stop** feasibility cannot be restored
12. **else if** $net\_cost_i < net\_cost_k$ **then**
13. reassign i from $Y(i)$ to $O(i)$
14. $P \leftarrow i$
15. **else**
16. reassign k from $Y(k)$ to $O(k)$
17. $P \leftarrow k$
18. **end if**
19. **end for**
20. **for** each $i \in P$
21. **if** $\Delta d(i) < 0$
22. reassign i from $Y(i)$ to $O(i)$
23. **end if**
24. **end for**

In lines 1-6 of Algorithm GS, the set V of violated cannot-link constraints is populated with point pairs (i, k). In lines 7-19, an attempt is made to iteratively remove violated constraints in V. This is done by first computing separately the net cost of reassigning either point i or point k to its nearest open center (lines 8-9). If neither point can be reassigned (line 10 true), the algorithm stops (line 11). Otherwise, the reassignment with the lower cost is selected (lines 12-18). Note that both $O(\cdot)$ and $Y(\cdot)$ need to be updated any time a point is reassigned. Finally, in steps 20-24, a check is made to see if any point that was previously reassigned (set P) can be moved to a closer open center. The rationale behind this that a point may have been allocated

to cluster center further away as part of feasibility restoration because a closer center was unavailable for assignment due to a cannot-link constraint violation.

Our algorithm for restoring feasibility differs considerably from Randel *et al*. (2019). Whereas Randel *et al*. iteratively move a point violating at least one cannot-link constraint from its current center to its nearest open center until feasibility is restored (or not), in our implementation, we look at point pairs involved in violated constraints to make a greedy reassignment of one of the points and stop after one complete pass. This has the advantage of potentially having fewer reassignments to restore feasibility and, therefore, shorter run times compared to Randel *et al*.

It is also worth noting that our restore feasibility routine, like Randel *et al*. (2019), is not guaranteed to produce a feasible solution. In the event this occurs, no feasible upper bound will be produced, but the Lagrangian relaxation algorithm can still continue using the best known upper bound to compute updated Lagrangian multipliers (see below).

Subgradient Optimization

To find the best lower bound possible, we used subgradient optimization to update the Lagrangian multipliers in an iterative fashion. Let z_{UB}^{*} and $z_{LR}(\boldsymbol{\mu}^{t})$ denote the values of the best known upper bound and the current lower bound at iteration t, respectively, and let x_{ij}^{t} be the partial solution to the relaxed problem at iteration t (i.e., the values of the cluster assignment variables). Multipliers were updated at every iteration according to the rule:

$$\mu_{i}^{t+1} = \mu_{i}^{t} - \alpha^{t}\left(\sum_{j\in J} x_{ij}^{t} - 1\right) \tag{4.10}$$

$$\alpha^{t} = \frac{\theta^{t}\left(z_{UB}^{*} - z_{LR}(\boldsymbol{\mu}^{t})\right)}{\sum_{i\in N}\left(\sum_{j\in N} x_{ij}^{t} - 1\right)^{2}} \tag{4.11}$$

where μ_{i}^{t} is the multiplier value at iteration t, α^{t} it step size at iteration t, and θ^{t} is a user-defined constant, which, in the usual way, is initially set to 2 and divided by 2 if the lower bound fails to improve after a set number of iterations, in our case 3 iterations.

Starting values for the Lagrangian multipliers were all set to a fixed value:

$$\bar{d}(p) = \frac{1}{p}\sum_{i\in N}\sum_{j\in N}\frac{d_{ij}}{|N|^2} \tag{4.12}$$

This is simply the average distance between each pair of points divided by the number of cluster centers. Lastly, stopping conditions for the Lagrangian procedure included whether: (i) a maximum number of iterations t_{max} was reached; (ii) the relative difference between the upper and lower bounds was below some threshold ε; and (iii) the θ^t tuning parameter was less than some threshold θ_{min}. In our experimentation, we set $t_{max} = 200$, $\varepsilon = 0.0001$ (i.e., 0.01%), and $\theta_{min} = 0.0001$.

4.3.3 Variable Reduction Technique

To reduce the computational burden of having to solve very large MILPs for datasets of medium/large size, we adopted a two-stage solution approach (see Figure 4.1). In the first stage, unconstrained cluster centers are identified using k-means or some other similarly fast algorithm (e.g. k-medoids). The points closest to each k-means cluster center (or the actual centroids for k-medoids) are then taken as the set of candidate medians M for a constrained p-median model, with $j \in M$ replacing $j \in N$ in (4.1)-(4.8). This reduced model is then solved in the second stage to find the final set of p cluster centers. The Lagrangian relaxation algorithm remains exactly the same, except for the denominator in equation (4.12), which changes from $|N|^2$ to $|N||M|$.

There is a lot of flexibility in how to apply k-means for selecting candidate centers. One approach is to simply use a single fixed value of $k > p$. For example, given a dataset with 1000 points, 250 must-link constraints, and 250 cannot-link constraints, if one wished to find $p = 10$ clusters and set $k = 30$, the number of variables and constraints would both decrease by two orders of magnitude, specifically going from 1,001,000 to 30,030 variables and from 1,501,001 to 46,001 constraints, which is a considerable reduction.

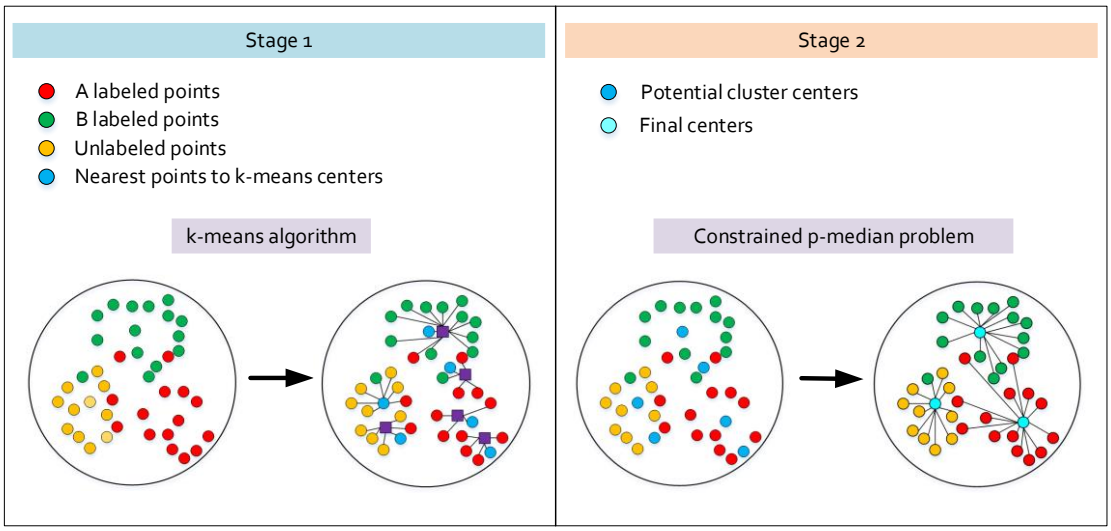


Figure 4.1. Two-stage variable reduction technique for constrained clustering. Data points labeled A and B are assumed to have cannot-link constraints. Values for k and p are equal to 5 and 3, respectively.

While this would undoubtedly make the constrained p-median problem much easier to solve, there are obvious concerns about the quality of the resulting solutions. One might expect, at least for some clusters, that the chosen cluster centers for a constrained p-median model would be nearby to the cluster centers selected using unconstrained k-means for $k = p$. If $k > p$, however, it may be that none of the cluster centers for k-means are close in any sense to the $k = p$ centers, especially if k is only marginally larger than p.

To address this, a more elaborate procedure is to use k equal to a multiple of p or a series of increasing k values, starting with $k = p$, whereby each set of centers found using k-means for a specific value of k is iteratively added to the candidate set of constrained cluster centers. Different series of k values can be chosen, such as increments or multiples of p (e.g., $k = (p, p+1, p+2, \dots)$, ., $k = (p, p + \lfloor p/2 \rfloor, p + 2 \cdot \lfloor p/2 \rfloor, \dots)$ and $k = (p, 2p, 3p, \dots)$).

4.4 Experimental Results

The optimization model and solution methods were implemented in Python using the docplex package and CPLEX callable libraries version 22.1.1. Runs of the k-mean algorithm were executed with the sklearn package for Python. All experiments were performed on the same multi-core Lenovo ThinkBook laptop (13th Gen Intel i7-13700H processor, 2.40 GHz per core)

with 32GB of RAM and running Windows 10 64-bit operating system.

4.4.1 Datasets

We tested the performance of the model, the variable reduction technique, and Lagrangian relaxation algorithm on set of 8 small-sized benchmark datasets (see Table 4.1), as well as several medium-sized and one large-sized UK household electricity consumption datasets. Benchmark datasets are the same ones used by Randel *et al*. (2019), thus allowing a direct comparison with their VNS heuristic. Details of the benchmark datasets are provided in Xia (2009).

Table 4.1. Summary of benchmark datasets. Respectively, parameters n, f, and p refer to the number of data points, the number of features, and the number of classes/clusters.

				Configuration 1*		Configuration 2*	
Instance	n	f	p	$\lvert\mathcal{ML}\rvert$	$\lvert\mathcal{CL}\rvert$	$\lvert\mathcal{ML}\rvert$	$\lvert\mathcal{CL}\rvert$
Soybean	47	35	4	2	12	4	2
Protein	116	20	6	9	6	13	9
Iris	150	4	3	6	6	8	4
Wine	178	13	3	22	13	36	22
Ionosphere	351	34	2	26	18	61	32
Control	600	60	6	30	15	45	30
Balance	625	4	3	78	47	109	63
Yeast	1484	8†	10	148	89	260	148

* Note: the number of must- and cannot-link constraints reported in Xia (2009) and Randel *et al*. (2019) refer to the number of points involved, not pairwise restrictions, which are given here.

† Correction to Xia (2009).

The household electricity consumption dataset derives from the UK Power Networks' Low Carbon London project (AECOM Building Engineering 2018). The project, which ran from late 2007 to the end of 2010, measured electricity consumption at 30-minute intervals for over 14,000 households across the UK. The dataset contains energy consumption time series (kWh per 30 min), unique household identifier, date and time, and Acorn household segmentation type (CACI 2024). Acorn provides a geo-demographic breakdown of the UK's population

(Table C.1). It divides postcodes into 62 types based on detailed social-economical characteristics then aggregates these into 18 intermediate groups and 6 top level categories (HM Land Registry 2023). Acorn is designed to segment households based on lifestyle characteristics as opposed to property characteristic.

For our analysis, we created a number of datasets from the original UK Power Networks (UKPN) dataset. The large dataset contained household electricity consumption data for 8102 households with a full six months' worth of electricity consumption readings (24 July 2009 to 20 December 2009). This dataset (denoted UKPN-L) has a total of 150 features that includes Acorn types in the ranges 1-17 and 19-56. Note, there are only two type 55 households and one type 56 household. We attached one of three labels (aka classes) to each household based on their Acorn type. Households of type 1 were designated high-income and given a label "H", while households of type 52-57 were considered low-income and assigned a label "L". In all, there were 296 H-labeled and 187 L-labeled data points. All other households ($n = 7609$) were given a label "O" for other.

Must-link constraints were included to force a subset of high-income (alternatively low-income) households to appear in the same cluster. In all, we added 327 H-to-H and 130 L-to-L must-link constraints by selecting at random, with replacement, two H- and two-L labeled points, respectively, for a total of 457 must-link constraints. Specified numbers of must-link constraints are based on the formula $\lfloor 0.0075 \times n(n-1)/2 + 0.5 \rfloor$, which is simply 0.0075 times the total number of pairwise combinations rounded to the nearest whole number, n being the number of data points for a particular label (i.e., 296 or 187). In a similar manner, 1384 random H-to-L cannot-link constraints were generated to prevent high- and low-income households from appearing in the same clusters. The number of cannot-link constraints is based on the formula $\lfloor 0.025 \times n_H \times n_L + 0.5 \rfloor$, where n_H and n_L are number of H- and L-labeled points (i.e., 296 and 187, respectively).

From the UKPN-L dataset, we subsequently generated 5 medium sized datasets (denoted UKPN-M1 to UKPN-M5) with 4000 randomly selected points each (see Table 4.2). For the medium-sized datasets, cannot-link and must-link constraints were added for any pair of points that had such constraints in the large dataset.

Table 4.2. Summary of medium-sized UK Power Network datasets. All datasets have 4000 points, 150 features, and 3 classes.

Instance	$\lvert\mathcal{ML}\rvert$	$\lvert\mathcal{CL}\rvert$
UKPN-M1	101	309
UKPN-M2	106	329
UKPN-M3	110	342
UKPN-M4	118	335
UKPN-M5	101	286

4.4.2 Benchmark Dataset Results

Variable Reduction Technique

For the variable reduction technique, we evaluated a variety of rules for setting parameter k. These included using different single values for k (denoted SV1-SV3), specifically $k = 2p$ (SV1), $k = 5p$ (SV2), and $k = 10p$ (SV3), plus the following three ways of combining multiple values of k (denoted MV1-MV3).

- First four 1-step increments: $k = (p, p+1, p+2, p+3, p+4)$ (MV1)
- First three half p-step increments: $k = (p, p + \lfloor p/2 \rfloor, p + 2\lfloor p/2 \rfloor)$ (MV2)
- First three multiplicative terms: $k = (p, 2p, 3p)$ (MV3)

Computational results confirm the overall effectiveness of the variable reduction technique (see Table 4.3 and Tables C.2-C.3). Although there was some variation, rule MV1, in general, achieved the best results in terms solution quality for 12 out of 14 cases. In only two cases each were rules MV2 and SV3 best.

Irrespectively, variable reduction, in most cases, resulted in little or no degradation in solution quality. For five datasets (Soybean, Protein, Iris, Ionosphere, and Control), an optimal solution was still found. For the other three datasets (Wine, Balance, and Yeast), optimality gaps were generally small, ranging from 0.3 to 0.8%. More impressively, solution times were radically reduced, especially for the four larger datasets. On Balance, for example, times were reduced from over 200 seconds to 10 seconds or less. On Yeast, times were reduced from 38-247 minutes down to around 1 second, representing a time savings of 3 to 4 orders of magnitude.

Table 4.3. Best performing variable reduction technique (VR) rule on benchmark datasets.

		Without VR	With VR		
Instance	Config	Time (s)	Best Rule	Time (s)	Gap (%)
Soybean	1	0.23	MV1	0.10	0.00
	2	0.14	MV1	0.06	0.00
Protein	1	0.70	MV1	0.17	0.00
	2	0.54	MV1	0.11	0.00
Iris	1	2.54	MV1	0.10	0.00
	2	2.37	MV1	0.10	0.00
Wine	1	3.62	SV3	0.27	0.31
	2	5.66	SV3	0.27	0.27
Ionosphere	1	17.46	MV1	0.22	0.00
	2	15.96	MV1	0.22	0.00
Control	1	36.76	MV2	1.19	0.00
	2	33.59	MV2	0.96	0.00
Balance	1	225.97	MV1	10.06	0.57
	2	206.73	MV1	8.32	0.73
Yeast	1	14,835.61	MV1	1.19	0.76
	2	2,281.82	MV1	0.96	0.81

Lagrangian Relaxation

Overall, the Lagrangian relaxation solution method performed well on the benchmark datasets (see Table 4.3). Unsurprisingly, better optimality gaps were achieved by using CPLEX to compute a feasible upper bound. For CPLEX, gaps ranged from just 0.01 to 0.1% on the six smaller datasets (Soybean, Protein, Iris, Wine, Ionosphere, Control) to a high of 0.5-0.9% for the largest two datasets (Balance and Yeast). Run times ranged from several seconds to several minutes.

Feasible upper bounds found using the heuristic resulted in optimality gaps that were mostly the same as CPLEX for the six smaller datasets, but noticeably higher (1.2-2.7%) for the largest two datasets. On the other hand, run times using the heuristic were around 9% less (on average) than CPLEX, except for the Iris dataset, which were three to four times more.

Table 4.4. Performance of Lagrangian relaxation on benchmark datasets for the best variable reduction rule (see Table 4.3) using either CPLEX or an heuristic to compute a feasible upper bound (UB) solution. Columns under VNS are results reported for Randel *et al*. (2019).

		CPLEX UB			Heuristic UB			VNS	
Instance	Config	Gap (%)	Itr	Time (s)	Gap (%)	Itr	Time (s)	Gap (%)	Time (s)
Soybean	1	0.01	39	2.46	0.01	32	2.48	0.00	0.00
	2	0.01	41	2.67	0.01	53	2.46	0.00	0.00
Protein	1	0.11	119	18.60	0.11	108	11.92	0.00	0.01
	2	0.01	143	22.31	0.01	183	22.89	0.00	0.00
Iris	1	0.01	38	4.59	0.08	200	19.18	0.00	0.01
	2	0.01	52	6.17	0.07	200	17.43	0.00	0.00
Wine	1	0.01	110	33.53	0.01	110	25.94	0.00	0.01
	2	0.01	141	47.16	0.02	200	55.79	0.00	7.08
Ionosphere	1	0.01	65	15.84	0.01	66	15.62	0.00	5.13
	2	0.01	34	9.23	0.01	34	10.33	0.02	65.04
Control	1	0.01	61	33.44	0.01	61	23.81	0.00	0.13
	2	0.01	45	14.88	0.01	45	19.49	0.00	0.12
Balance	1	0.51	116	95.86	2.73	85	55.56	0.002	110.42
	2	0.85	89	76.80	2.56	87	63.24	0.00	91.51
Yeast	1	0.81	95	254.09	2.51	101	198.76	0.00	97.78
	2	0.93	93	285.92	1.15	113	256.79	0.004	124.44

Compared to the VNS heuristic developed by Randel *et al*. (2019), our Lagrangian method produces only marginally higher optimality gaps for the six smaller datasets (Soybean, Protein, Iris, Wine, Ionosphere, Control). Solution times for the Lagrangian are, in most cases, longer for these datasets, but there is one notable exception (e.g., Ionosphere confirmation no. 2). For the two largest datasets, solution times for the Lagrangian compare favorably with VNS, but with distinctly higher optimality gaps.

4.4.3 Medium UK Power Networks Dataset Results

Results for the medium UKPN datasets merely confirm the overall effectiveness of the Lagrangian relaxation algorithm (Table 4.5). Optimality gaps were mostly less than 1%. In only two cases did the gap exceed 1%. In the majority of cases, optimality gap increased as the number of clusters p increases. In two cases, however, there was increase in gap followed by a drop as p becomes larger.

Table 4.5. Performance of Lagrangian relaxation on medium-sized UKPN datasets using variable reduction rule MV1 and the heuristic method for computing a feasible upper bound.

Instance	p	Best LB	Best UB	Gap (%)	Itr	Time (s)
UKPN-M1	6	197,299.2	197,670.8	0.19	127	430.73
	9	187,697.1	188,186.4	0.26	121	448.17
	12	181,011.0	182,178.6	0.65	200	877.49
UKPN-M2	6	195,521.1	196,844.1	0.68	113	384.80
	9	183,283.3	185,685.2	1.31	200	792.89
	12	180,269.6	180,696.4	0.24	155	688.03
UKPN-M3	6	194,403.4	194,853.8	0.23	123	399.47
	9	185,364.7	186,329.0	0.52	179	606.42
	12	178,622.8	179,972.0	0.76	130	568.83
UKPN-M4	6	193,276.3	193,847.5	0.30	200	688.35
	9	184,859.9	185,521.5	0.36	200	747.38
	12	177,285.4	180,177.7	1.63	200	841.26
UKPN-M5	6	192,518.7	193,528.6	0.52	100	351.81
	9	182,763.6	184,082.2	0.72	200	711.73
	12	177,839.1	178,863.2	0.58	105	472.52

Run times ranged from 5.9 to 14.6 minutes. The average was 10 minutes. In 6 out of 15 cases, the maximum number of iterations was reached. In the other 9 cases, a total of 128 iterations were required on average. Both run time and iterations generally show a positive relationship with the number of clusters p.

4.4.4 Large UK Power Networks Dataset Results

k-Means Clustering

Here we report detailed results comparing k-means clustering with our proposed constrained p-median clustering method on the large UK Power Networks dataset UKPN-L. We begin by computing silhouette scores to determine an 'optimal' number of clusters p (Rousseeuw, 1987) for annual load profiles. We find (Figure 4.2) that the preferred number of clusters is in the range 15-17. For the convenience of visualization, we set $p = 15$ for both k-means and

constrained clustering.

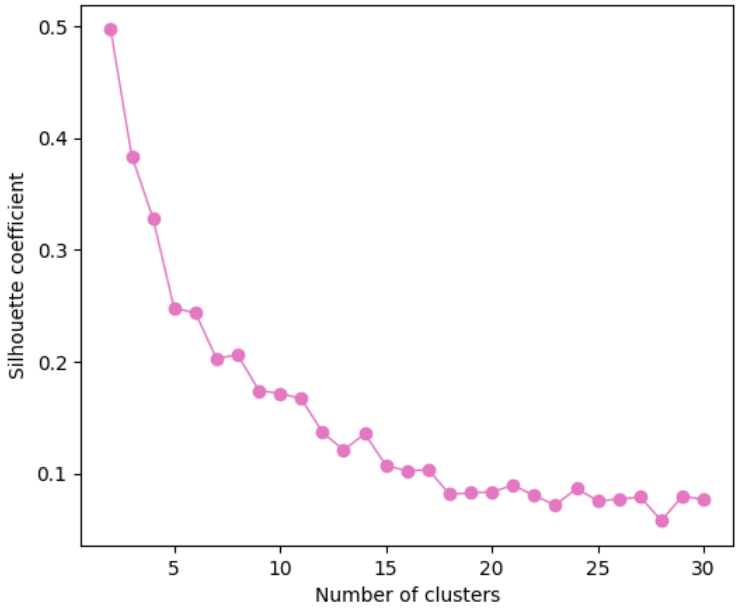


Figure 4.2. Average silhouette coefficient versus number of clusters for the UKPN-L dataset.

Table 4.6. Summary statistics for k-means clusters 1-15.

Pattern	1	2	3	4	5
Count	1473	1464	1215	1074	904
Proportion (%)	18.18	18.07	15.00	13.26	11.16
Average usage (kWh/day)	8.91	6.63	11.52	4.23	14.45
Pattern	6	7	8	9	10
Count	544	392	307	227	204
Proportion (%)	6.71	4.84	3.79	2.80	2.52
Average usage (kWh/day)	17.83	22.76	18.91	11.56	28.09
Pattern	11	12	13	14	15
Count	136	70	68	18	6
Proportion (%)	1.68	0.86	0.84	0.22	0.07
Average usage (kWh/day)	21.76	29.59	38.15	45.98	67.85

Results for k-means clustering (see Table 4.6 and Figure C.1) reveal a number of distinct clusters that vary considerably in terms of the number of data points contained within them and daily average electricity consumption (measured in kWh/day). The two smallest clusters (no. 14-15), for example, contain just 18 and 6 households, respectively, which amounts to just 0.22% and 0.07% of the full sample. The largest cluster (no. 1), by contrast, contains 1473 data points,

equivalent to 18.2% of the total.

Average daily usage for the five larger clusters (no. 1-5) is on the low side, ranging from 4.2 to 14.5 kWh/day. Meanwhile, the five medium-sized clusters (no. 6-10) have moderate average daily usage of 11.6 to 28.1 kWh/day. Finally, the five smaller clusters (no. 11-16) have moderate to very high average daily usage between 21.8 and 67.9 kWh/day.

Table 4.7. Number of H-, L-, and O-labeled households in each k-means cluster.

Pattern	H	L	O	Q_H	Q_L	w	D
1	46	36	1391	0.561	0.439	0.170	0.122
2	23	42	1399	0.354	0.646	0.135	0.292
3	45	12	1158	0.789	0.211	0.118	0.579
4	13	62	999	0.173	0.827	0.155	0.653
5	35	11	858	0.761	0.239	0.095	0.522
6	30	3	511	0.909	0.091	0.068	0.818
7	30	3	359	0.909	0.091	0.068	0.818
8	22	2	283	0.917	0.083	0.050	0.833
9	8	4	215	0.667	0.333	0.025	0.333
10	18	8	178	0.692	0.308	0.054	0.385
11	4	2	130	0.667	0.333	0.012	0.333
12	4	0	66	1.000	0.000	0.008	1.000
13	12	0	56	1.000	0.000	0.025	1.000
14	4	2	12	0.667	0.333	0.012	0.333
15	2	0	4	1.000	0.000	0.004	1.000
						wD	0.510

The results also reveal that in spite of there being a fairly large number of clusters, there is a significant overlap between high- and low-income households within most clusters (see Table 4.7 and Figure C.1). To measure the degree of overlap within clusters, we computed a dispersion index for each cluster j as follows.

$$D_j = \left|Q_{Hj} - 0.5\right| + \left|Q_{Lj} - 0.5\right| \tag{13}$$

In (13), Q_{Hj} and Q_{Lj} denote, respectively, the fractions of H- and L-labeled points in cluster j relative to the total number of H- and L-labeled points in the cluster. By assumption, if there

are no H- or L-labeled points in cluster j, then $Q_{Hj} = Q_{Lj} = 0$. Metric D_j ranges from 1, indicating no overlap of H- and L-labeled points within the same cluster, down to 0, indicating an even 50-50 split. The following weighted sum of the D_j's produces an overall measure of label overlap for a full clustering solution:

$$wD = \sum_{j=1}^{p} w_j D_j \tag{14}$$

where w_j is the fraction of H- and L-labeled points in cluster j relative to the total number of H- and L-labeled points in the whole dataset. Like D_j, total weighted dispersion wD approaches 1 if there is no overlap among any cluster.

According to our dispersion index, only clusters 12, 13, and 15 show no overlap (D = 1). The others either have very high overlap with D < 0.385 (no. 1, 2, 9-11, 14) or moderate to low overlap with D in the range 0.522-0.833 (no. 3-8). Clusters 10 and 14 are noteworthy for the fact they represent high average daily electricity consumption households with a significant proportion of low-income households.

Constrained p-Median Clustering

Even for the large UKPN dataset UKPN-L, the Lagrangian relaxation algorithm manages to produce good quality results in a reasonable amount of computing time (see Figure 4.3 and Table C.4). Convergence of the best lower and upper bounds occurs quickly. Within 30 iterations, the optimality gap decreases from over 2000% initially to 2.35%. After 43 iterations, the gap is below 1% and after 82 iterations the best feasible upper bound is achieved, finally stopping after 106 iterations. The optimality gap at this point is 0.60%. In terms of run time, each iteration takes around 11 seconds on average. Total run time is under 19 minutes. For comparison, k-means produces solutions in a matter of seconds.

Compared to k-means, constrained p-median clusters are qualitatively similar in that larger clusters (no. 1-11) tend to have low to moderate average energy usage of 4.1 to 23.9 kWh/day, while smaller clusters (no. 12-15) have moderate to high energy usage of 22.4 to 44.6 kWh/day (see Table 4.8 and Figure C.2). What stands out, however, is that constrained p-median clusters have: (i) less extreme variability in terms of average energy usage; (ii) a slightly more even

spread of cluster sizes (i.e., fewer clusters with extremely few data points); and (iii) far less overlap between high- and low-income households within most clusters (compare Tables 4.6-4.7 with Tables 4.8-4.9 and Figure C.1 with Figure C.2).

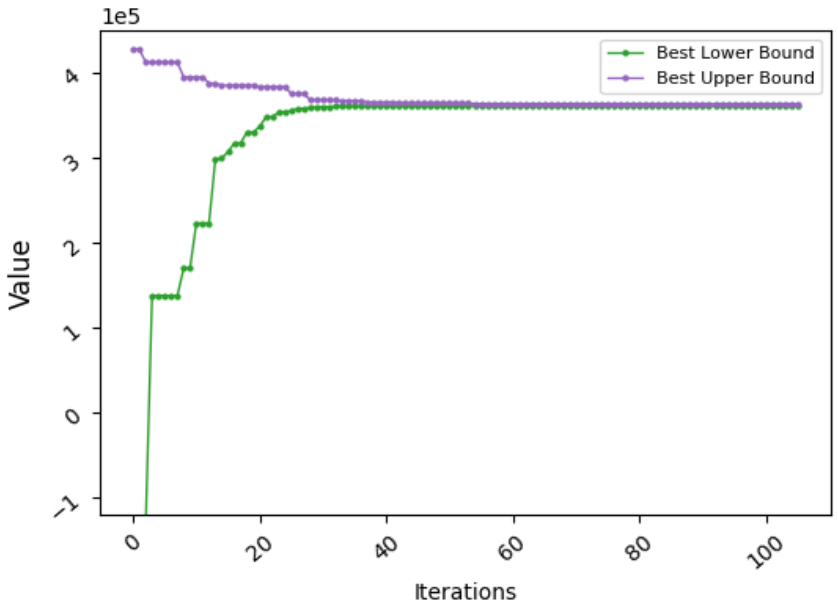


Figure 4.3. Best lower and upper bound for Lagrangian relaxation of dataset UKPN-L.

Table 4.8. Summary statistics for constrained p-median clusters 1-15.

Pattern	1	2	3	4	5
Count	1340	1140	988	886	886
Proportion (%)	16.54	14.07	12.19	10.94	10.94
Average usage (kWh/day)	6.43	14.94	8.03	9.66	11.68
Pattern	6	7	8	9	10
Count	881	569	339	261	247
Proportion (%)	10.87	7.02	4.18	3.22	3.05
Average usage (kWh/day)	4.11	18.16	23.89	19.74	11.89
Pattern	11	12	13	14	15
Count	183	135	130	70	47
Proportion (%)	2.26	1.67	1.6	0.86	0.58
Average usage (kWh/day)	14.87	22.40	29.92	31.43	44.55

Regarding the first two points, daily average energy consumption for k-means ranges from 4.2 to 67.9 kWh/day. For the constrained p-median solution, that range is reduced down

considerably to 4.1 to 44.6 kWh/day. Additionally, whereas the largest and smallest clusters for k-means contain 1473 points (18.2%) and a mere 6 points (0.07%), respectively, for constrained p-median these values are 1340 (16.5%) and 47 (0.6%), respectively.

The most noticeable difference from k-means (point three above) is the much lower overlap of different label types within clusters (see Table 4.9 and Figure C.2). With constrained clustering, the number of clusters with no overlap between H- and L-labeled points increases from 3 to 9 (i.e., $D = 1$ for cluster no. 1, 2, 9-15). Of the remaining clusters, two have fairly low overlap with $D > 0.579$ (no. 3, 6), and only 4 have significant overlap with $D < 0.2$ (no. 4, 5, 7, 8). Overall, dispersion (wD) has increased by 73% from 0.51 for the k-means solution to 0.88 for the constrained p-median solution.

Table 4.9. Number of H-, L-, and O-labeled households in each constrained p-median cluster.

Pattern	H	L	O	Q_H	Q_L	w	D
1	0	114	1226	0.000	1.000	0.236	1.000
2	258	0	882	1.000	0.000	0.534	1.000
3	4	15	969	0.211	0.789	0.039	0.579
4	12	10	864	0.545	0.455	0.046	0.091
5	8	9	869	0.471	0.529	0.035	0.059
6	1	24	856	0.040	0.960	0.052	0.920
7	3	2	564	0.600	0.400	0.010	0.200
8	4	4	331	0.500	0.500	0.017	0.000
9	2	0	259	1.000	0.000	0.004	1.000
10	0	8	239	0.000	1.000	0.017	1.000
11	0	0	183	0.000	0.000	0.000	1.000
12	0	1	134	0.000	1.000	0.002	1.000
13	1	0	129	1.000	0.000	0.002	1.000
14	1	0	69	1.000	0.000	0.002	1.000
15	2	0	45	1.000	0.000	0.004	1.000
						wD	0.880

4.5 Conclusion

This chapter investigates the application of the p-median model combined with must-link and cannot-link constraints to the problem of constrained clustering. Unlike machine learning

algorithms, which have traditionally been used in constrained clustering, our modeling approach has the benefit of producing solutions with verifiable optimality bounds. It can also be generalized to incorporate other types of instance-level constraints, such imposing a minimum or maximum number of points in each cluster and soft cannot-link constraints limiting the number of pairs of points with cannot-link constraints allowed in the same cluster.

A Lagrangian relaxation procedure is proposed to efficiently solve medium to large problem instances. Subgradient optimization is used in the usual way to update Lagrangian multipliers in the dual problem and a greedy based heuristic is used to find feasible upper bound solutions in each iteration. To speed up solution times, a simple but highly effective variable reduction technique is incorporated, which applies k-means (other similarly fast algorithm) to identify a candidate set of cluster centers.

We test our modeling framework and solution approach on a bank of benchmark datasets as well as several much larger real-world household electricity consumption datasets. We find that high-quality solutions with small optimality gaps, typically under 1%, can obtained in second to minutes even for problem instances with 8000 plus data points.

In terms of potential lines of future research, a particularly useful one would be to build and test new heuristic procedures for generating feasible upper bound solutions as part of a Lagrangian relaxation procedure. Our heuristic (or at least our implementation), though adequate, is not especially fast and sometimes fails to find the best assignment of data points to cluster centers. Another direction of research would be to evaluate entirely different solution approaches to solve a constrained p-median model, such as metaheuristics (other than VNS) or possibly matheuristics. Finally, alternative model formulations of the underlying p-median problem could be considered, such as the COBRA (Church 2003), BEAMR (Church 2008), integer pseudo-Boolean (Goldengorin and Krushinsky 2011), and radius constraint (García *et al*. 2011) models.

5. Conclusion

This chapter summarizes the main contributions of this dissertation and offers some further research directions.

5.1 Research summary

The thesis focuses on three important aspects of demand-side management, namely, prediction of residential consumption patterns, electricity demand forecasting, and electricity customer segmentation.

In the first study, we firstly analyze the residential electricity consumption pattern in Ireland using clustering algorithm, and secondly analyze the various factors affecting electricity consumption, namely, the effect of temperature, week and month on electricity, electricity consumption pattern shifting, and the prediction of electricity consumption pattern respectively. Finally, we provide a machine learning based framework for predicting yearly consumption pattern based on household characters and historical usage. Our proposed framework could achieve 84.56% accuracy for 5-classification problem on test dataset. Based on results, many interesting conclusions were found: (1) 5℃ is the threshold value, that less than 5℃ will have reduce electricity demand, while 5℃ or above will have negative effect on electricity demand, The effect between 5℃ and 20℃ on demand is very small and this implies that hot temperature effect is positive but small. (2) Our proposed method can effectively identify pattern shifting phenomenon. More specially, six typical intra-day patterns are identified between July, 2009 and December, 2010. From summer to winter, people's intra-day load profiles change from the perspective of both shape and volume. (3) In terms of occupation of chief income earner, Higher and intermediate managerial, administrative, professional occupation or supervisory, clerical & junior managerial, administrative, professional occupations trend to use more than these who has semi-skilled & unskilled led manual occupations, unemployed and lowest grade occupations, although they both live with children.

In the second study, this thesis analysis the electricity consumption pattern both in the context of COVID-19 and normal situation. The second study constructs a series models for forecasting

electricity consumption in the context of COVID-19, which selects factors that related with epidemic such as the number of people working at home, the number of people shopping in supermarkets, etc., as independent variables, so as to better reflect the relationship between electricity consumption and population mobility factors in the context of the COVID-19. The results of this study show that the proposed model can not only make short-term and long-term forecasts, but also evaluate the impact of lockdown policy on electricity consumption. The results suggest: (1) The drop in electricity consumption due to the pandemic is short-term, mainly in March and April, and is caused not only by lockdown policy but also electricity consumption down trend from 2019 to 2020. (2) The daily electricity consumption is also highly correlated with population flow factors such as the number of people stay at home, number of people go to workplace, number of people go to retails and so on, indicating that the traditional model of forecasting in the electricity sector will now need to be augmented with such features. (3) Our proposed prediction model based on population flow indicators can generally achieve an accuracy rate of 2.16%-8.01% before COVID -19, and with accuracy of 2.01%-5.36% in post-COVID-19 data in 7 different cities in US. (4) within six weeks of lockdown, four out of seven cities did COVID-19 cause an appreciable reduction of 3.6-7.6% in electricity usage over the first 6 weeks when lockdowns were introduced. In the other three cities, electricity usage actually increased slightly by 0.53-3.4%. This could help policymakers to better understand the impact of COVID-19 from different aspects, better explore the impact mechanism, and would provide guidance for the recovery of the economic development and production after the pandemic.

In the final study, we propose a two-stage constrained clustering framework based on optimization techniques for electricity customer segmentation. The proposed clustering algorithm is formed by adding must-link and cannot-link constraints on the basis of the original p-median problem. In order to speed up the solution of the model, we developed the Lagrangian relaxation method for solving the model. The results of our method show that: (1) In terms of clustering effect, compared with k-means, the proposed method could achieve more personalized clustering, so that high-income and low-income household are not appear in the same group as much as possible by adding cannot-link constraints. (2) Compared with the

traditional p-median algorithm, the Lagrangian relaxation can greatly reduce the running time, mainly because in the first stage we select potential candidate clusters medoids.

5.2 Further research directions

This section focuses briefly on new research questions that the authors have identified in the process of completing doctoral dissertation.

5.2.1 Integrating renewables in electricity markets and households

The topic of climate change has received increasing attention in recent years, and hence reducing greenhouse gas emissions is of great importance for the world to achieve sustainable development. Replacing traditional energy sources with renewable energy sources has become an urgent requirement today due to the fact the main energy source driving society' development is still fossil fuels. In recent years, the share of renewable energy in the electricity system has been increasing gradually, so it is important to better integrate it into electricity market and household to achieve a reduction in carbon emissions. Besides, from the perspective of economic, access to renewable energy will lead to a reduction in the cost of electricity, further reducing the price of electricity and thus benefiting consumers. However, the integration of renewable energy sources also brings greater challenge to energy market, the electricity market is fundamentally different from the traditional economic market, the electric power system needs to maintain a balance between power supply and demand at all times, in order to better balance the supply and demand, the electricity market introduced the day-ahead market and the intra-day market respectively. The day-ahead market refers to the power supply and demand side must submit their power supply and demand bids in the energy market 24 hours in advance based on electricity demand forecast, which means electricity is exchanged before the physical delivery begins. The system will automatically match offers from both sides, in this way, the market clearing price is determined. The intra-day market is designed to react to unexpected changes, such as power plant outages and maintenance or unexpected changes in demand occurs due to weather, accidents, and other influencing factors on the demand side in day-ahead market. Weather-dependent renewables like wind and solar are more volatile and unstable, significant deviations may arise between the forecasted demand and the final demand as

proportions of renewable energy gets higher, hence the integration of renewable energy into the energy market and households will pose a greater challenge to the power system. Besides, the renewables energy source has strong seasonal characteristics, although residential electricity consumption also show seasonal patterns. Yet they are not synchronized, so future research needs to integrate several unsynchronized systems into a unified intelligent energy management system using cutting-edge techniques, such as deep learning and reinforcement learning.

5.2.2 Analysis of environmental policy in the power sector

The frequent occurrence of climate extremes in recent years, such as drought, hill fire, rising global temperatures, melting glaciers, sea-level rise, extreme heat, etc., pose considerable threats to our environment and sustainable development, indicates that greenhouse gases emission has reached the limit compared to the pre-industrial revolution. Electric power generates the second largest (25%) share of greenhouse gas emissions and includes emissions from electricity production in the U.S. in 2021, hence reducing GHG emissions from electric power production would definitely contribute to sustainability. However, power plants in most OECD countries minimize their production cost or maximize their profit without taking into account the social welfare and environmental impacts of greenhouse gases, thus leading to market failures, so incorporating environmental externalities into market mechanisms is feasible way for plant owner to make more reasonable production decisions or investment in renewable energy technologies. In order to reduce pollution, many countries have started to take various forms of regulation to limit greenhouse gas emissions, such as technical technology standards, emission standards, and market-based instruments (e.g., emission tax and cap-and-trade). These environmental policies provide a promising pathway to mitigating climate change in the future. However, there is still a fundamental issue that needs to be addressed, for example, power companies aim to maximize profits, while policymakers hope to enhance social welfare, and these two objectives are in conflict with each other, Hence future research should focus on building game-theoretic modelling framework to handle this kind of conflicts, such as bi-level programming and equilibrium methods.

5.3.4 Clustering high-dimensional time series data

Energy big data is currently gaining more and more attention due to the continuous development of smart grid, and data generated by electric power systems have increased dramatically. Households' energy consumption can be continuously recorded and transferred to data center. From the perspective of energy consumption pattern mining, this brings more opportunities, because more data often means more information can be mined. However, from the perspective of algorithm, more data requires better computing power and more efficient algorithms. For example, in energy consumption pattern mining, the typical load profile is generated by clustering the time series of energy usage. With the popularity of smart meters, electricity usage can be recorded every 15 minutes, which leads to a very high dimensional time series given one year, thus it will take long time to get the results for the clustering algorithm. Another aspect is that the power consumption data behind often represents human behavior, and this human behavior is often cyclical, such as during weekdays, each day has a similar part of the power consumption pattern, while at weekends the power consumption shows another pattern. Besides, there are similar patterns in every season. Therefore, electricity consumption data is a high-dimensional and periodic time series, and the processing of this high-dimensional time series generally requires transformation processing, such as feature extraction, for example, using principal component analysis for feature extraction, but these feature extraction methods often do not take into account the temporal characteristics of the time series, so future research aims to seek a feature extraction method that can both reduce the dimension of the time series and retain the time-series characteristics is the future.

6. Research contributions

This section reports the papers that have been published during my PhD study.

6.1 Papers

Guo, Z., O'Hanley, J. R., & Gibson, S. (2022). Predicting residential electricity consumption patterns based on smart meter and household data: A case study from the Republic of Ireland. *Utilities Policy*, **79**, 101446. (**Chapter 2**)

Guo, Z., O'Hanley, J. R., & Gibson, S. (2024). Influence of population mobility on electricity consumption in seven US cities during the COVID-19 pandemic. *Utilities Policy*, **90**, 101804. (**Chapter 3**)

References

AECOM Building Engineering (2018) Energy Demand Research Project: Early Smart Meter Trials, 2007-2010. UK Data Service. SN: 7591. DOI: 10.5255/UKDA-SN-7591-1.

Albani, A., Domigall, Y., & Winter, R. (2017). Implications of customer value perceptions for the design of electricity efficiency services in times of smart metering. *Information Systems and e-Business Management*, **15**, 825-844.

Aloise, D., Hansen, P. (2009) A branch-and-cut SDP-based algorithm for minimum sum-of-squares clustering. *Pesquisa Operacional*, **29**, 503-516.

Andersen, F. M., Larsen, H. V., & Boomsma, T. K. (2013). Long-term forecasting of hourly electricity load: Identification of consumption profiles and segmentation of customers. *Energy conversion and Management*, **68**, 244-252.

Andrianesis, P., Biskas, P., & Liberopoulos, G. (2021). Evaluating the cost of emissions in a pool-based electricity market. *Applied Energy*, **298**, 117253.

Arcos-Vargas, A., Nuñez, F., & Román-Collado, R. (2020). Short-term effects of PV integration on global welfare and CO2 emissions. An application to the Iberian electricity market. *Energy*, **200**, 117504.

Avella, P., Boccia, M., Salerno, S., & Vasilyev, I. (2012). An aggregation heuristic for large scale p-median problem. *Computers & Operations Research*, **39**(7), 1625-1632.

Babaki, B., Guns, T., Nijssen, S. (2014) Constrained clustering using column generation. Pp. 438-454 in: *Integration of AI and OR Techniques in Constraint Programming*, 11th International Conference, CPAIOR, Cork, Ireland, 19-23 May 2014, Proceedings. Springer.

Basu, S., Banerjee, A., Mooney, R.J. (2004a) Active semi-supervision for pairwise constrained clustering. Pp. 333-344 in: *Proceedings of the 2004 SIAM International Conference on Data Mining*, Lake Buena Vista (FL), USA, 22-24 April 2004. Society for Industrial and Applied Mathematics.

Basu, S., Bilenko, M., Mooney, R.J. (2004b) A probabilistic framework for semi-supervised clustering. Pp. 59-68 in: *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle (WA), USA, 22-25 August 2004. ACM.

Beckel, C., Sadamori, L., Staake, T., Santini, S. (2014) Revealing household characteristics from smart meter data. *Energy* **78**: 397-410.

Benati, S., García, S. (2014) A mixed integer linear model for clustering with variable selection. *Computers and Operations Research*, **43**, 280-285.

Benítez, I., Díez, J.L., Quijano, A., Delgado, I. (2016) Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance. *Electric Power*

Systems Research **140**: 517-526.

Bianco, V., Manca, O., & Nardini, S. (2009). Electricity consumption forecasting in Italy using linear regression models. *Energy*, **34**(9), 1413-1421.

Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000). Constrained k-means clustering. *Microsoft Research,* Redmond, 20(0), 0.

Breiman, L. (2001). Random forests. *Machine Learning* **45**(1): 5-32.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, **9**(1), 247-274.

Brusco, M. J. (2003). An enhanced branch-and-bound algorithm for a partitioning problem. *British Journal of Mathematical and Statistical Psychology*, **56**(1), 83-92.

Brusco, M. J., & Köhn, H. F. (2008). Optimal partitioning of a data set based on the p-median model. *Psychometrika*, **73**(1), 89.

Bryan, E., Ringler, C., Okoba, B., Roncoli, C., Silvestri, S., Herrero, M. (2013). Adapting agriculture to climate change in Kenya: household strategies and determinants. *Journal of Environmental Management* **114**: 26-35.

Buechler, E., Powell, S., Sun, T., Astier, N., Zanocco, C., Bolorinos, J., Flora, J., Boudet, H., & Rajagopal, R. (2022). Global changes in electricity consumption during COVID-19. *iScience*, **25**(1). 103568.

Burleyson, C. D., Rahman, A., Rice, J. S., Smith, A. D., & Voisin, N. (2021) Multiscale effects masked the impact of the COVID-19 pandemic on electricity demand in the United States. *Applied Energy*, **304**, 117711.

CACI (2024) Acorn: Powerful Geodemographics, Customer Insight and Resource Targeting. CACI Limited. Available at: https://www.caci.co.uk/datasets/acorn/ [Last accessed: 15 Mar 2024].

Carvalho, M., Bandeira de Mello Delgado, D., de Lima, K. M., de Camargo Cancela, M., dos Siqueira, C. A., & de Souza, D. L. B. (2021). Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns. *International Journal of Energy Research*, **45**(2), 3358-3364.

CER (Commission for Energy Regulation). (2012) CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010 [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00. Available at: https://www.ucd.ie/issda/data/commissionforenergyregulationcer/ [Last accessed: 6 May 2021].

Chae, Y. T., Horesh, R., Hwang, Y., & Lee, Y. M. (2016). Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings. *Energy and Buildings*, **111**, 184-194.

Chen, S., Igan, D. O., Pierri, N., Presbitero, A. F., Soledad, M., & Peria, M. (2020). Tracking the economic impact of COVID-19 and mitigation policies in Europe and the United States. *IMF Working Papers*, **2020**(125).

Cherry, C., Scott, K., Barrett, J., Pidgeon, N. (2018) Public acceptance of resource-efficiency strategies to mitigate climate change. *Nature Climate Change* **8**(11): 1007-1012.

Chévez, P., Barbero, D., Martini, I., Discoli, C. (2017) Application of the k-means clustering method for the detection and analysis of areas of homogeneous residential electricity consumption at the Great La Plata region, Buenos Aires, Argentina. *Sustainable Cities and Society* **32**: 115-129.

Chicco, G., Napoli, R., Piglione, F., Postolache, P., Scutariu, M., & Toader, C. (2004). Load pattern-based classification of electricity customers. *IEEE Transactions on Power Systems*, 19(2), 1232-1239.

Choksi, K.A., Jain, S., Pindoriya, N.M. (2020) Feature based clustering technique for investigation of domestic load profiles and probabilistic variation assessment: Smart meter dataset. *Sustainable Energy, Grids and Networks* **22**: 100346.

Church, R.L. (2003) COBRA: a new formulation of the classic p-median location problem. *Annals of Operations Research*, **122**(1-4), 103-120.

Church, R.L. (2008) BEAMR: An exact and approximate model for the p-median problem. *Computers and Operations Research*, **35**(2), 417-426.

Coleman, G. B., & Andrews, H. C. (1979). Image segmentation by clustering. *Proceedings of the IEEE*, **67**(5), 773-785.

Cortes, C., Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**(3): 273-297.

Crone, S.F., Kourentzes, N. (2009) Input-variable specification for neural networks-an analysis of forecasting low and high time series frequency. *2009 International Joint Conference on Neural Networks*, pp. 619-626.

Daskin, M.S., Maass, K.L. (2015) The p-median problem. Pp. 21-45 in: *Location Science*. Cham: Springer International Publishing.

De Amorim, R.C. (2012) Constrained clustering with Minkowski weighted k-means. Pp. 13-17 in: *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, Budapest, Hungary, 20-22 November 2012. IEEE.

Di Lascio, F. M. L., Menapace, A., & Righetti, M. (2020). Joint and conditional dependence modelling of peak district heating demand and outdoor temperature: a copula-based approach. *Statistical Methods & Applications*, **29**(2), 373-395.

Durbin, J., & Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, **89**(3), 603-616.

Durbin, J., & Koopman, S. J. (2012). Time series analysis by state space methods (Vol. 38). OUP Oxford.

EIA (2023a). Electricity. U.S. Energy Information Administration (EIA). Available at: https://www.eia.gov/electricity/data.php [Last accessed: 8 Sep 2023].

EIA (2023b). Electricity explained. U.S. Energy Information Administration (EIA). Available at: https://www.eia.gov/energyexplained/electricity/use-of-electricity.php [Last accessed: 9 Jan 2024].

Eroğlu, H. (2021). Effects of COVID-19 outbreak on environment and renewable energy sector. *Environment, Development and Sustainability*, **23**(4), 4782-4790.

European Commission, Directorate-General for Economic and Financial Affairs, European economic forecast : Spring 2013, Publications Office, 2013, https://data.europa.eu/doi/10.2765/46506.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B. (2007) Regression. Springer-Verlag, Berlin/Heidelberg.

Filonchyk, M., Hurynovich, V., Yan, H., Gusev, A., & Shpilevskaya, N. (2020). Impact assessment of COVID-19 on variations of SO_2, NO_2, CO and AOD over East China. *Aerosol and Air Quality Research*, **20**(7), 1530-1540.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* **29**(5): 1189-1232.

Gajowniczek, K., & Ząbkowski, T. (2014). Short term electricity forecasting using individual smart meter data. *Procedia Computer Science*, **35**, 589-597.

García, S., Labbé, M., Marín, A. (2011) Solving large p-median problems with a radius formulation. *INFORMS Journal on Computing*, **23**(4), 546-556.

Gautam, S. & Hens, L. (2020). COVID-19: Impact by and on the environment, health and economy." *Environment, Development and Sustainability* **22**, 4953-4954.

Ge, R., Ester, M., Jin, W., Davidson, I. (2007) Constraint-driven clustering. Pp. 320-329 in: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose (CA), USA, 12-15 August 2007. ACM.

Gharehgozli, O., Nayebvali, P., Gharehgozli, A., & Zamanian, Z. (2020). Impact of COVID-19 on the Economic Output of the US Outbreak's Epicenter. *Economics of Disasters and Climate Change*, **4**(3), 561-573.

Goldengorin, B., Krushinsky, D. (2011) A computational study of the pseudo-Boolean approach to the p-median problem applied to cell formation. Pp. 503-516 in: *Network Optimization*, 5th International Conference on Network Optimization, Hamburg, Germany, 13-16 June 2011, Proceedings. Springer.

González-Almagro, G., Peralta, D., De Poorter, E., Cano, J.-R., García, S. (2023) Semi-supervised constrained clustering: An in-depth overview, ranked taxonomy and future research directions. *arXiv*, 2303.00522.

Green, R. (2008). Electricity wholesale markets: designs now and in a low-carbon future. The

Energy Journal, **29**(2_suppl), 95-124.

Gulati, P., Kumar, A., & Bhardwaj, R. (2021). Impact of COVID-19 on electricity load in Haryana (India). *International Journal of Energy Research*, **45**(2), 3397-3409.

Hampton, H., Foley, A., Del Rio, D. F., Smyth, B., Laverty, D., & Caulfield, B. (2022). Customer engagement strategies in retail electricity markets: A comprehensive and comparative review. *Energy Research & Social Science*, **90**, 102611.

Han, P., Cai, Q., Oda, T., Zeng, N., Shan, Y., Lin, X., & Liu, D. (2021). Assessing the recent impact of COVID-19 on carbon emissions from China using domestic economic data. *Science of the Total Environment*, **750**, 141688.

Hansen, P., Brimberg, J., Urošević, D., & Mladenović, N. (2009). Solving large p-median clustering problems by primal–dual variable neighborhood search. *Data Mining and Knowledge Discovery*, **19**, 351-375.

Hartigan, J.A., Wong, M.A. (1979) Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **28**(1): 100-108.

Hastie, T.J., Tibshirani, R.J. (1990) Generalized additive models. Monographs on Statistics and Applied Probability 43, Chapman and Hall/CRC, New York/Boca Raton.

He, P., Liang, J., Qiu, Y., Li, Q, Xing, B. (2020) Increase in domestic electricity consumption from particulate air pollution. *Nature Energy*, **5**: 985-995.

Hilares, K., Vargas, R., & Gastelo-Roque, J. A. (2020, September). Impact of COVID-19 on the GHG emissions of the Peruvian Interconnected Electrical System. *In 2020 IEEE XXVII International Conference on Electronics, Electrical Engineering and Computing* (INTERCON) (pp. 1-4). IEEE.

HM Land Registry (2023) Acorn consumer classification (CACI). Official Statistics, Updated 22 Dec 2023. Available at: https://www.gov.uk/government/statistics/quality-assurance-of-administrative-data-in-the-uk-house-price-index/acorn-consumer-classification-caci [Last accessed 15 Mar 2024].

Hoerl, A.E., Kennard, R.W. (1970) Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**(1): 69-82.

IEA (2020) Key world energy statistics 2020. International Energy Agency (IEA). Available at: https://www.iea.org/reports/key-world-energy-statistics-2020/final-consumption [Last accessed 7 Sep 2022]

Imani, M., & Ghassemian, H. (2018). Electrical load forecasting using customers clustering and smart meters in Internet of Things. In 2018 9th International Symposium on Telecommunications (IST) (pp. 113-117). IEEE.

Jain, S., & Sharma, T. (2020). Social and travel lockdown impact considering coronavirus disease (COVID-19) on air quality in megacities of India: present benefits, future challenges and way forward. *Aerosol and Air Quality Research*, **20**(6), 1222-1236.

Johnson S C. (1967). Hierarchical clustering schemes. *Psychometrika*, **32**(3): 241-254.

Kaenzig, J., Heinzle, S. L., & Wüstenhagen, R. (2013). Whatever the customer wants, the customer gets? Exploring the gap between consumer preferences and default electricity products in Germany. *Energy Policy*, **53**, 311-322.

Kang, H., An, J., Kim, H., Ji, C., Hong, T., & Lee, S. (2021). Changes in energy consumption according to building use type under COVID-19 pandemic in South Korea. *Renewable and Sustainable Energy Reviews*, **148**, 111294.

Khan, I., Huang, J.Z., Ivanov, K. (2016) Incremental density-based ensemble clustering over evolving data streams. *Neurocomputing* **191**: 34-43.

Klastorin, T. D. (1985). The p-median problem for cluster analysis: A comparative test using the mixture model approach. *Management Science*, **31**(1), 84-95.

Köhn, H. F., Steinley, D., & Brusco, M. J. (2010). The p-median model as a tool for clustering psychological data. *Psychological methods*, **15**(1), 87.

Kohonen, T. (1990) The self-organizing map. *Proceedings of the IEEE* **78**(9): 1464-1480.

Koprinska, I., Rana, M., Agelidis, V.G. (2015) Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems* **82**: 29-40.

Kopsakangas-Savolainen, M., Mattinen, M. K., Manninen, K., & Nissinen, A. (2017). Hourly-based greenhouse gas emissions of electricity–cases demonstrating possibilities for households and companies to decrease their emissions. *Journal of Cleaner Production*, **153**, 384-396.

Lahcen, B., Brusselaers, J., Vrancken, K., Dams, Y., Paes, C. D. S., Eyckmans, J., & Rousseau, S. (2020). Green recovery policies for the COVID-19 crisis: modelling the Impact on the economy and greenhouse gas emissions. *Environmental and Resource Economics*, **76**(4), 731-750.

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R. (2012) Efficient backprop. Pp 9-48 in: Neural networks: Tricks of the Trade. Springer, Berlin/Heidelberg.

Leuski, A. (2001) Evaluating document clustering for interactive information retrieval. Pp. 33-40 in: *Proceedings of the 10th International Conference on Information and Knowledge Management*, Atlanta (GA), USA 5-10 October 2001. ACM.

Li, Z., Ye, H., Liao, N., Wang, R., Qiu, Y., & Wang, Y. (2022). Impact of COVID-19 on electricity energy consumption: A quantitative analysis on electricity. *International Journal of Electrical Power and Energy Systems*, **140**, 108084.

MacIver, C., Bukhsh, W., & Bell, K. R. (2021). The impact of interconnectors on the GB electricity sector and European carbon emissions. *Energy Policy*, **151**, 112170.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (pp. 281-297).

Maital, S., & Barzani, E. (2020). The global economic impact of COVID-19: A summary of

research. *Samuel Neaman Institute for National Policy Research*, **2020**, 1-12.

McCallum A, Nigam K, Ungar L H. (2000) Efficient clustering of high-dimensional data sets with application to reference matching. Pp.169-178 in: *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston (MA), USA 20-23 August 2000. ACM.

McLoughlin, F., Duffy, A., Conlon, M. (2012) Characterising domestic electricity consumption patterns by dwelling and occupant socio-economic variables: An Irish case study. *Energy and Buildings* **48**: 240-248.

McLoughlin, F., Duffy, A., Conlon, M. (2015) A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied Energy* **141**: 190-199.

McLachlan, G. J., & Basford, K. E. (1988) Mixture models: Inference and applications to clustering. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **38**(2), 384-385.

Meier, E., Moy, C. (2004) Social grading and the census. *International Journal of Market Research* **46**(2): 141-170.

Meyer, B. H., Prescott, B., & Sheng, X. S. (2022). The impact of the COVID-19 pandemic on business expectations. *International Journal of Forecasting*, **38**(2), 529-544.

Mirasgedis, S., Sarafidis, Y., Georgopoulou, E., Kotroni, V., Lagouvardos, K., Lalas, D. P. (2007) Modeling framework for estimating impacts of climate change on electricity demand at regional level: Case of Greece. *Energy Conversion and Management* **48**(5): 1737-1750.

Montaño Moreno, J. J., Palmer Pol, A. L., Sesé Abad, A. J., & Cajal Blasco, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, **25**(4), 500-506.

Motlagh, O., Berry, A., O'Neil, L. (2019) Clustering of residential electricity customers using load time series. *Applied Energy* **237**: 11-24.

Mulvey, J. M., & Crowder, H. P. (1979). Cluster analysis: An application of Lagrangian relaxation. *Management Science*, **25**(4), 329-340.

Narajewski, M., & Ziel, F. (2020). Changes in electricity demand pattern in Europe due to COVID-19 shutdowns. *In IAEE Energy Forum*.(Special issue) (pp. 44-47).

Nezamoddini, N., & Wang, Y. (2017). Real-time electricity pricing for industrial customers: Survey and case studies in the United States. *Applied energy*, **195**, 1023-1037.

Ng, A., Jordan, M., Weiss, Y. (2001) On spectral clustering: Analysis and an algorithm. Pp. 849-856 in: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Vancouver, Canada, 3-8 December 2001. MIT Press.

Ng, M. K. (2000). A note on constrained k-means algorithms. *Pattern Recognition*, **33**(3), 515-519.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, **36**(2), 2592-2602.

Nugent, R., Meila, M. (2010) An overview of clustering applied to molecular biology. Pp. 369-404 in: *Statistical Methods in Molecular Biology*, Methods in Molecular Biology, **620**. Humana Press.

Obst, D., De Vilmarest, J., & Goude, Y. (2021). Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France. *IEEE Transactions on Power Systems*, **36**(5), 4754-4763.

Os, B., Meulman, J. (2004) Improving dynamic programming strategies for partitioning. *Journal of Classification*, **21**(2), 207-230.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., Adebiyi, E. (2016) Clustering algorithms: Their application to gene expression data. *Bioinformatics and Biology Insights*, **10**, 237-253.

Paatero P, Tapper U. (1994). Positive matrix factorization: A non‐negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**(2): 111-126.

Park, H., Jeong, S., Koo, J. H., Sim, S., Bae, Y., Kim, Y., ... & Bang, J. (2020). Lessons from COVID-19 and Seoul: Effects of reduced human activity from social distancing on urban CO2 concentration and air quality. *Aerosol and Air Quality Research*, **21**(1), 200376.

Peng, J., Wei, Y. (2007). Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, **18**(1), 186-205.

Peters, I. M., Brabec, C., Buonassisi, T., Hauch, J., & Nobre, A. M. (2020). The impact of COVID-19-related measures on the solar resource in areas with high levels of air pollution. *Joule*, **4**(8), 1681-1687.

Pham, M. C., Cao, Y., Klamma, R., Jarke, M. (2011) A clustering approach for collaborative filtering recommendation using social network analysis. *Journal of Universal Computer Science*, **17**(4), 583-604.

Phanich M, Pholkul P, Phimoltares S. (2010) Food recommendation system using clustering analysis for diabetic patients. Pp. 1-8 in: *Proceedings of 2010 International Conference on Information Science and Applications*, Seoul, Republic of Korea, 21-23 April 2010. IEEE.

Prol, J. L., & Sungmin, O. (2020). Impact of COVID-19 measures on short-term electricity consumption in the most affected EU countries and USA states. *iScience*, **23**(10), 101639.

Ramos, S., Duarte, J. M., Duarte, F. J., & Vale, Z. (2015). A data-mining-based methodology to support MV electricity customers' characterization. *Energy and Buildings*, **91**, 16-25.

Räsänen, T., Voukantsis, D., Niska, H., Karatzas, K., Kolehmainen, M. (2010) Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. *Applied Energy* **87**(11): 3538-3545.

Räsänen, T., Kolehmainen, M. (2009) Feature-based clustering for electricity use time series data. *International Conference on Adaptive and Natural Computing Algorithms* **5495**: 401-412.

Randel, R., Aloise, D., Mladenovíc, N., Hansen, P. (2019) On the k-medoids model for semi-supervised clustering. Pp 13-27 in: *Variable Neighborhood Search*, 6th International Conference on Variable Neighborhood Search, Sithonia, Greece, 4-7 October 2018, Revised Selected Papers. Springer.

Reynolds, A.P., Richards, G., de la Iglesia, B., & Rayward-Smith, V.J. (2006) Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**(4): 475-504.

Rokach, L., Maimon, O. (2005) Clustering methods. Pp. 321-352 in: *Data Mining and Knowledge Discovery Handbook*. Springer.

Rousseeuw, P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65.

Ruan, G., Wu, D., Zheng, X., Zhong, H., Kang, C., Dahleh, M.A., Sivaranjani, S., & Xie, L. (2020) A cross-domain approach to analyzing the short-run impact of COVID-19 on the U.S. electricity sector. *Joule*, **4**(11), 1-16.

Sağlam, B., Salman, F. S., Sayın, S., & Türkay, M. (2006). A mixed-integer programming approach to the clustering problem with an application in customer segmentation. *European Journal of Operational Research*, **173**(3), 866-879.

Singh, S., Yassine, A., Benlamri, R. (2019) Consumer segmentation: Improving energy demand management through households socio-analytics. Pp 1038-1045 in: IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE.

Steinley, D. (2015) K-medoids and other criteria for crisp clustering. Pp. 55-68 in: *Handbook of Cluster Analysis*, Chapman and Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Tian, T., Zhang, J., Lin, X., Wei, Z., Hakonarson, H. (2021). Model-based deep embedding for constrained clustering analysis of single cell RNA-seq data. *Nature communications*, **12**(1), 1873.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1): 267-288.

Ürge-Vorsatz, D., Cabeza, L.F., Serrano, S., Barreneche, C., Petrichenko, K. (2015) Heating and cooling energy trends and drivers in buildings. *Renewable and Sustainable Energy Reviews* **41**: 85-98.

van Vliet, M.T.H., Yearsley, J.R., Ludwig, F., Vögele, S., Lettenmaier, D.P., Kabat, P. (2012) Vulnerability of US and European electricity supply to climate change. *Nature Climate*

Change **2**(9): 676-681.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S. (2001) Constrained k-means clustering with background knowledge. Pp. 577-584 in: *Proceedings of the 18th International Conference on Machine Learning*, Williamstown (MA), USA, 28 June - 1 July 2001. Morgan Kaufmann Publishers Inc.

Wang, Y., & Li, L. (2015). Time-of-use electricity pricing for industrial customers: A survey of US utilities. *Applied Energy*, **149**, 89-103.

Wakiyama, T., Zusman, E. (2021). The impact of electricity market reform and subnational climate policy on carbon dioxide emissions across the United States: A path analysis. *Renewable and Sustainable Energy Reviews*, **149**, 111337.

Wang, F., Li, K., Duić, N., Mi, Z., Hodge, B.M., Shafie-khah, M., Catalão, J.P. (2018) Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. *Energy Conversion and Management* **171**: 839-854.

Werth, A., Gravino, P., Prevedello, G. (2021). Impact analysis of COVID-19 responses on energy grid dynamics in Europe. *Applied Energy*, **281**, 116045.

WHO (2020). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). World Health Organization (WHO).

Williams, N. J., Jaramillo, P., Campbell, K., Musanga, B., & Lyons-Galante, I. (2018). Electricity consumption and load profile segmentation analysis for rural micro grid customers in Tanzania. In 2018 IEEE PES/IAS PowerAfrica (pp. 360-365). IEEE.

Wood, S.N. (2017) Generalized additive models: An introduction with R, 2nd Edition. Chapman and Hall/CRC, New York/Boca Raton.

Xia, Y. (2009) A global optimization method for semi-supervised clustering. *Data Mining and Knowledge Discovery*, **18**, 214-256.

Xia, Y., Peng, J. (2005) A cutting algorithm for the minimum sum-of-squared error clustering. Pp. 150-160 in: *Proceedings of the 2005 SIAM International Conference on Data Mining*, Newport Beach (CA), USA, 21-23 April 2005. SIAM.

Xu R, Wunsch D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, **16**(3): 645-678.

Yan, R., Zhang, J., Yang, J., Hauptmann, A. (2004) A discriminative learning framework with pairwise constraints for video object classification. Pp 1-8 in: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington (DC), USA, 2004. IEEE.

Yang, J., Zhao, J., Wen, F., & Dong, Z. (2018). A model of customizing electricity retail prices based on load profile clustering analysis. *IEEE Transactions on Smart Grid*, **10**(3), 3374-3386.

Yu, Z., You, J., Wong, H.-S., Han, G. (2012) From cluster ensemble to structure ensemble. *Information Sciences*, **198**: 81-99.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2): 301-320.

APPENDICES

A. Supplementary Figures and Tables for Chapter 2

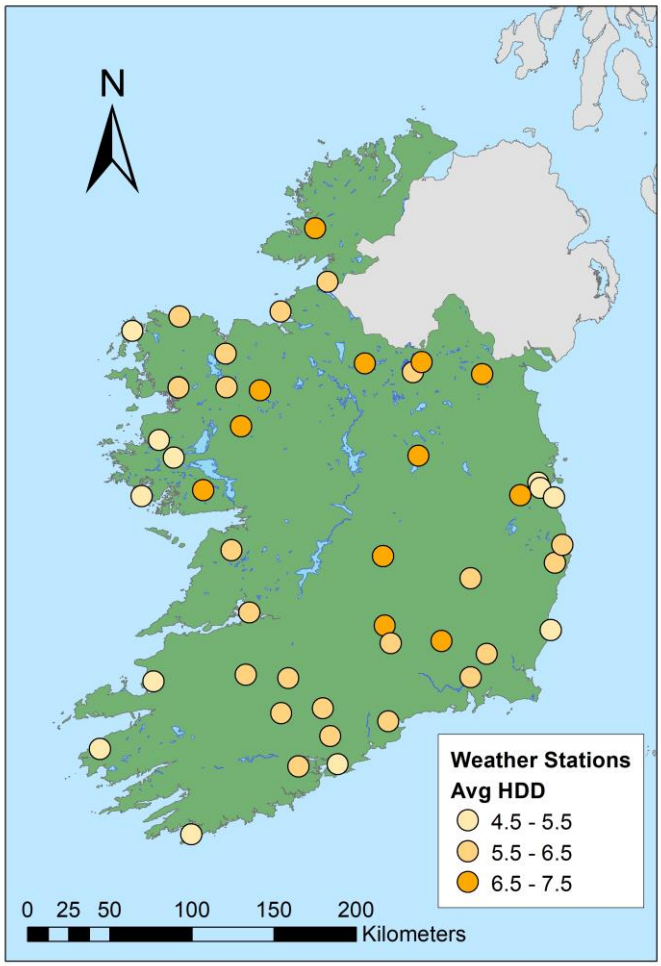


Figure A.1. Locations and associated average heating degree days (HDD) of weather stations in the Republic of Ireland.

Table A.1. Distribution of house type in the CER dataset.

House Type	Percentage
Apartment	1.66
Semi-detached house	30.05
Detached house	27.06
Terraced house	14.25
Bungalow	26.78
Refused	0.20

Table A.2. Distribution of spacing heating type in the CER dataset.

Heating Type	Percentage
Electricity (electric central heating or storage heating)	3.35
Electricity (plug-in heaters)	2.76
Gas	24.59
Oil	47.18
Solid fuel	21.17
Renewable (e.g., solar)	0.43
Other	0.52

Table A.3. Distribution of water heating type in the CER dataset.

Heating for Water	Percentage
Central heating system	7.95
Electric (immersion)	33.84
Electric (instantaneous heater)	0.95
Gas	14.33
Oil	24.26
Solid fuel boiler	9.69
Renewable (e.g., solar)	0.97
Other	8.01

Table A.4. Distribution of cooker type in the CER dataset.

Cook Type	Percentage
Electric cooker	69.70
Gas cooker	25.66
Oil fired cooker	2.38
Solid fuel cooker (stove aga)	2.26

Table A.5. Proportion of each pattern on selected days of the year.

Pattern / Date	P_1	P_2	P_3	P_4	P_5	P_6
2009-07-14	0.423	0.066	0.314	0.090	0.015	0.093
2009-12-24	0.208	0.139	0.196	0.125	0.191	0.140
2009-12-25	0.269	0.038	0.148	0.259	0.230	0.055
2009-12-26	0.256	0.072	0.204	0.177	0.145	0.145
2010-07-14	0.440	0.064	0.326	0.082	0.012	0.074
2010-12-24	0.199	0.138	0.187	0.123	0.223	0.130
2010-12-25	0.254	0.037	0.145	0.251	0.255	0.058
2010-12-26	0.239	0.062	0.210	0.185	0.178	0.126

Table A.6. Typical summer (14 July 2009) to winter (24 December 2009) pattern shifting matrix.

2009-12-24 / 2009-07-14	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0.379	0.088	0.268	0.094	0.068	0.104
P_2	0.050	0.254	0.063	0.100	0.413	0.121
P_3	0.115	0.182	0.191	0.151	0.179	0.182
P_4	0.058	0.129	0.107	0.187	0.377	0.141
P_5	0.036	0.055	0.036	0.091	0.727	0.055
P_6	0.030	0.169	0.092	0.145	0.374	0.190

Table A.7. Typical weekday (27 July 2009) to weekend (2 August 2009) pattern shifting matrix.

Sunday / Monday	P_1	P_2	P_3	P_4	P_5	P_6
P_1	0.746	0.010	0.153	0.067	0.002	0.022
P_2	0.166	0.156	0.213	0.280	0.062	0.123
P_3	0.273	0.044	0.397	0.189	0.012	0.085
P_4	0.161	0.084	0.230	0.373	0.053	0.099
P_5	0.064	0.043	0.170	0.362	0.255	0.106
P_6	0.153	0.077	0.212	0.362	0.049	0.147

Table A.8. Proportion of each intra-day pattern over a selected week.

Pattern / Date	P_1	P_2	P_3	P_4	P_5	P_6
2009-07-27 (Monday)	0.428	0.058	0.323	0.088	0.013	0.090
2009-07-28 (Tuesday)	0.419	0.065	0.308	0.091	0.017	0.100
2009-07-29 (Wednesday)	0.419	0.064	0.325	0.092	0.014	0.086
2009-07-30 (Thursday)	0.443	0.059	0.327	0.091	0.012	0.068
2009-07-31 (Friday)	0.423	0.061	0.317	0.099	0.019	0.081
2009-08-01 (Saturday)	0.432	0.056	0.288	0.136	0.019	0.069
2009-08-02 (Sunday)	0.446	0.042	0.247	0.176	0.021	0.068

Table A.9. Proportion of the six intra-day patterns associated with the five annual load profiles on 14 July 2010 (summer weekday)

Intra-day pattern / Annual profile	P_1	P_2	P_3	P_4	P_5	P_6
1	0.210	0.057	0.560	0.073	0.000	0.100
2	0.070	0.214	0.283	0.201	0.048	0.184
3	0.873	0.003	0.098	0.017	0.000	0.009
4	0.481	0.017	0.416	0.043	0.002	0.041
5	0.912	0.003	0.070	0.006	0.000	0.009

Table A.10. Proportion of the six intra-day patterns associated with the five annual load profiles on 25 December 2010 (Christmas day).

Intra-day pattern / Annual profile	P_1	P_2	P_3	P_4	P_5	P_6
1	0.108	0.043	0.123	0.370	0.263	0.093
2	0.019	0.049	0.048	0.231	0.620	0.033
3	0.466	0.023	0.270	0.138	0.057	0.046
4	0.181	0.051	0.205	0.329	0.164	0.070
5	0.670	0.012	0.132	0.134	0.028	0.024

Table A.11. Machine learning model parameters.

Model	Parameter	Value
DNN	Number of net layers	4
	Number of hidden layers	2
	Nodes per layer	[22, 200, 200, 5]
	L1	0.005
	L2	0.003
	Maximum number of iterations	500
	Activation function	Rectifier
	Learning rate	0.1
RF	Number of trees	2000
	Maximum tree depth	20
	Early stopping based on stopping_metric convergence	2
	Relative tolerance of metric-based stopping criterion	0.001
GBM	Number of trees	2000
	Maximum tree depth	5
	Early stopping based on stopping_metric convergence	2
	Relative tolerance of the metric-based stopping criterion	0.01
	Learning rate	0.001
Elastic Net	Link function	Multinomial
	Solver	L_BFGS
	Regularization factor between L1 and L2	0.9
SVM	Cost	7
	Gamma	0.1
	Kernel	Radial basis

B. Supplementary Figures and Tables for Chapter 3

Table B.1. List of tested models.

Model	Trend Component	Seasonal Component	Dynamic Factor 1	Dynamic Factor 2	Dynamic Factor 3	Dynamic Factor 4
1	Local_Trend	Week	Work	Home	Restaurant_Recreation	Retail
2	Local_Trend	Week	Work	.	.	.
3	Local_Trend	Week	.	Home	.	.
4	Local_Trend	Week	.	.	Restaurant_Recreation	.
5	Local_Trend	Week	.	.	.	Retail
6	Local_Trend	Week	Work	Home	.	.
7	Local_Trend	Week	.	Home	Restaurant_Recreation	.
8	Local_Trend	Week	.	.	Restaurant_Recreation	Retail
9	Local_Trend	Week	Work	.	.	Retail
10	Local_Trend	Week	Work	.	Restaurant_Recreation	.
11	Local_Trend	Week	.	Home	.	Retail
12	Local_Trend	Week	Work	Home	Restaurant_Recreation	.
13	Local_Trend	Week	Work	Home	.	Retail
14	Local_Trend	Week	Work	.	Restaurant_Recreation	Retail
15	Local_Trend	Week	.	Home	Restaurant_Recreation	Retail
16	Local_Trend	.	Work	Home	Restaurant_Recreation	Retail
17	Local_Trend	.	Work	.	.	.
18	Local_Trend	.	.	Home	.	.
19	Local_Trend	.	.	.	Restaurant_Recreation	.
20	Local_Trend	.	.	.	.	Retail
21	Local_Trend	.	Work	Home	.	.
22	Local_Trend	.	.	Home	Restaurant_Recreation	.
23	Local_Trend	.	.	.	Restaurant_Recreation	Retail
24	Local_Trend	.	Work	.	.	Retail
25	Local_Trend	.	Work	.	Restaurant_Recreation	.
26	Local_Trend	.	.	Home	.	Retail
27	Local_Trend	.	Work	Home	Restaurant_Recreation	.
28	Local_Trend	.	Work	Home		Retail
29	Local_Trend	.	Work	.	Restaurant_Recreation	Retail
30	Local_Trend	.	.	Home	Restaurant_Recreation	Retail
31	Temp	Week	Work	Home	Restaurant_Recreation	Retail
32	Temp	Week	Work	.	.	.
33	Temp	Week	.	Home	.	.
34	Temp	Week	.	.	Restaurant_Recreation	.
35	Temp	Week	.	.	.	Retail
36	Temp	Week	Work	Home	.	.
37	Temp	Week	.	Home	Restaurant_Recreation	.
38	Temp	Week	.	.	Restaurant_Recreation	Retail
39	Temp	Week	Work	.	.	Retail
40	Temp	Week	Work	.	Restaurant_Recreation	.

41	Temp	Week	.	Home	.	Retail
42	Temp	Week	Work	Home	Restaurant_Recreation	.
43	Temp	Week	Work	Home		Retail
44	Temp	Week	Work	.	Restaurant_Recreation	Retail
45	Temp	Week	.	Home	Restaurant_Recreation	Retail
46	Temp	.	Work	Home	Restaurant_Recreation	Retail
47	Temp	.	Work	.	.	.
48	Temp	.	.	Home	.	.
49	Temp	.	.	.	Restaurant_Recreation	.
50	Temp	.	.	.	.	Retail
51	Temp	.	Work	Home	.	.
52	Temp	.	.	Home	Restaurant_Recreation	.
53	Temp	.	.	.	Restaurant_Recreation	Retail
54	Temp	.	Work	.	.	Retail
55	Temp	.	Work	.	Restaurant_Recreation	.
56	Temp	.	.	Home	.	Retail
57	Temp	.	Work	Home	Restaurant_Recreation	.
58	Temp	.	Work	Home	.	Retail
59	Temp	.	Work	.	Restaurant_Recreation	Retail
60	Temp	.	.	Home	Restaurant_Recreation	Retail
61	Local_Trend	Week	.	.	.	.
62	Temp	Week	.	.	.	.

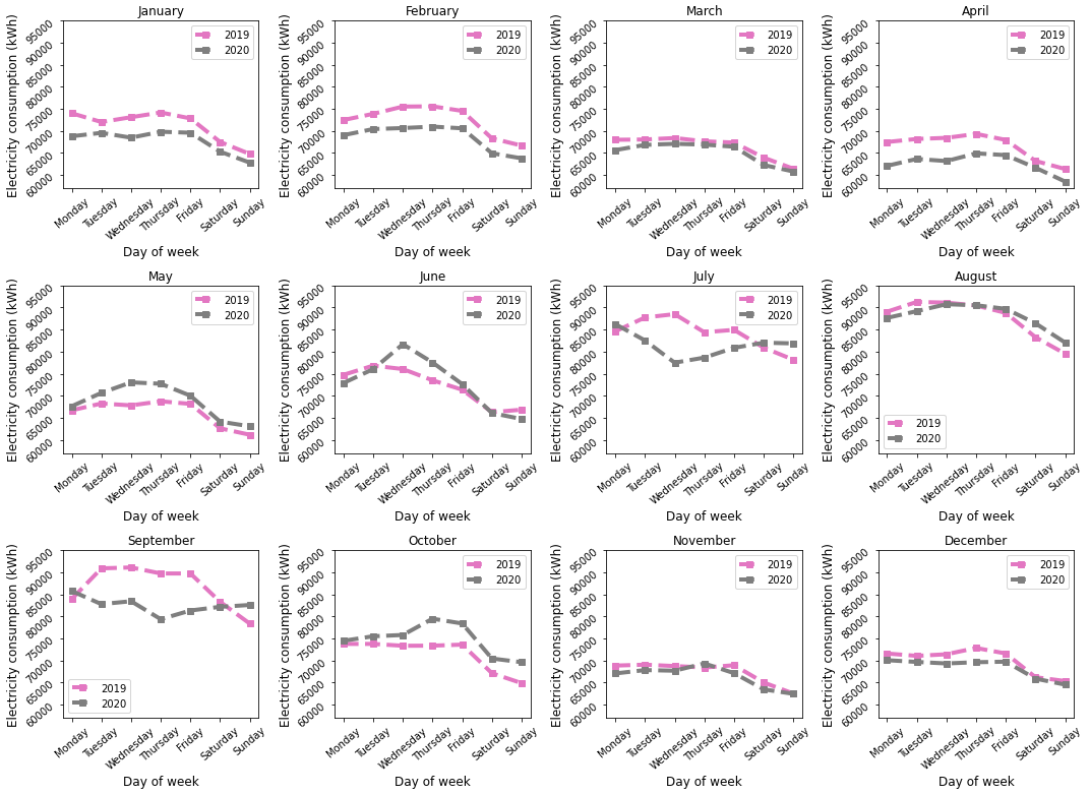


Fig B.1. Weekly electricity consumption in Los Angeles from Jan to Dec in 2019 and 2020.

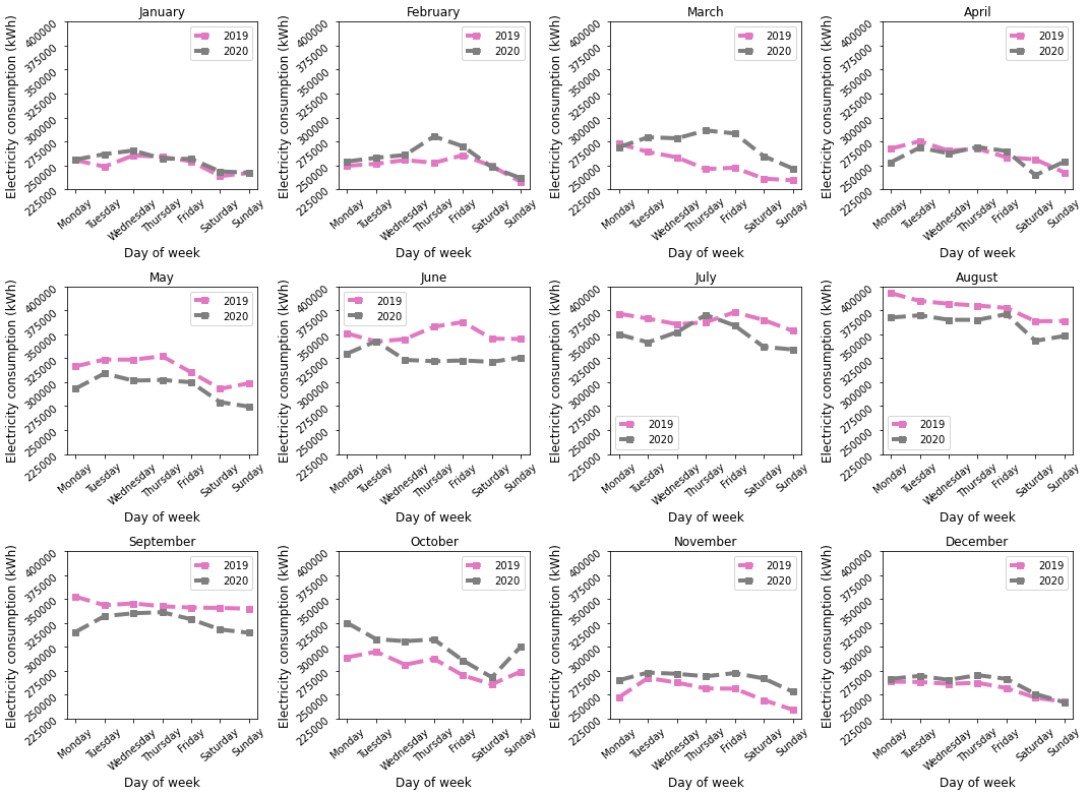


Fig B.2. Weekly electricity consumption in Houston from Jan to Dec in 2019 and 2020.

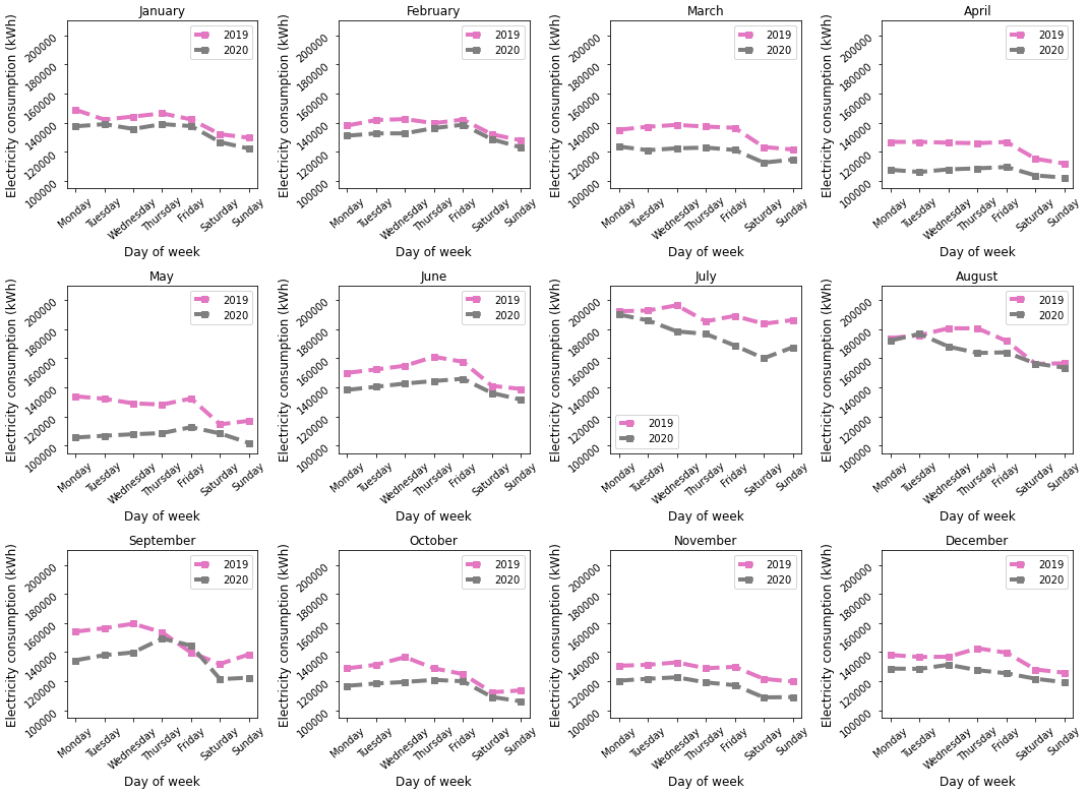


Fig B.3. Weekly electricity consumption in New York from Jan to Dec in 2019 and 2020.

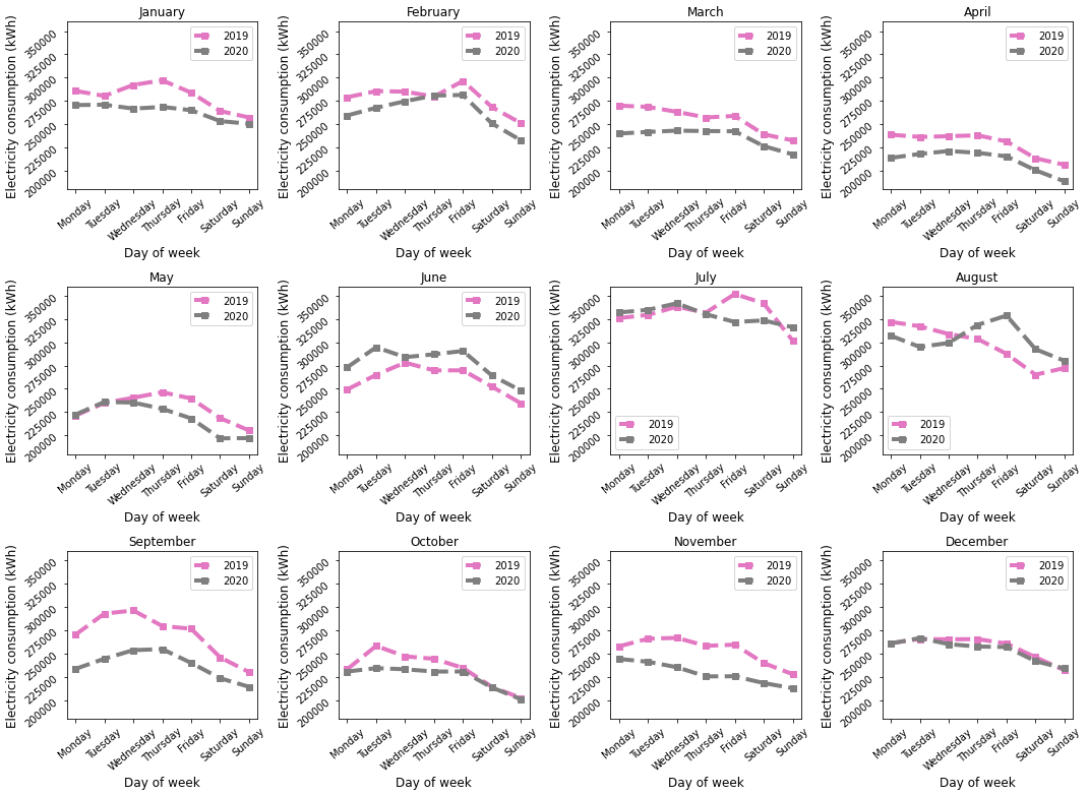


Fig B.4. Weekly electricity consumption in Chicago from Jan to Dec in 2019 and 2020.

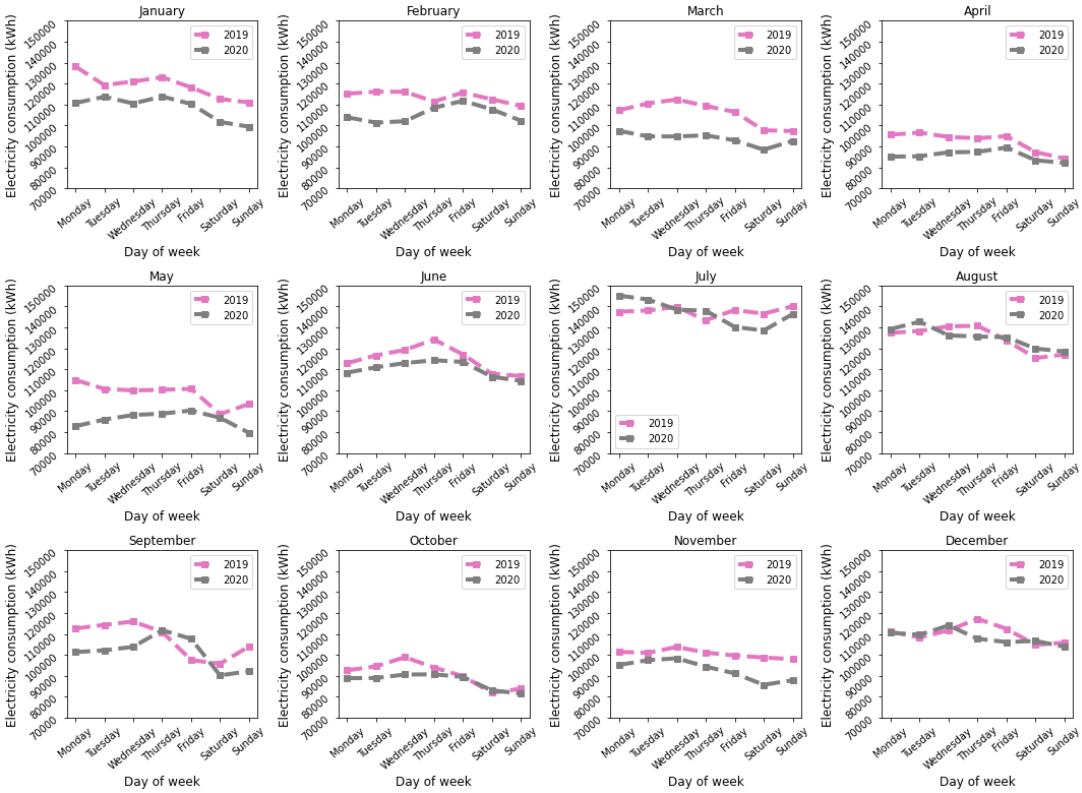


Fig B.5. Weekly electricity consumption in Philadelphia from Jan to Dec in 2019 and 2020.

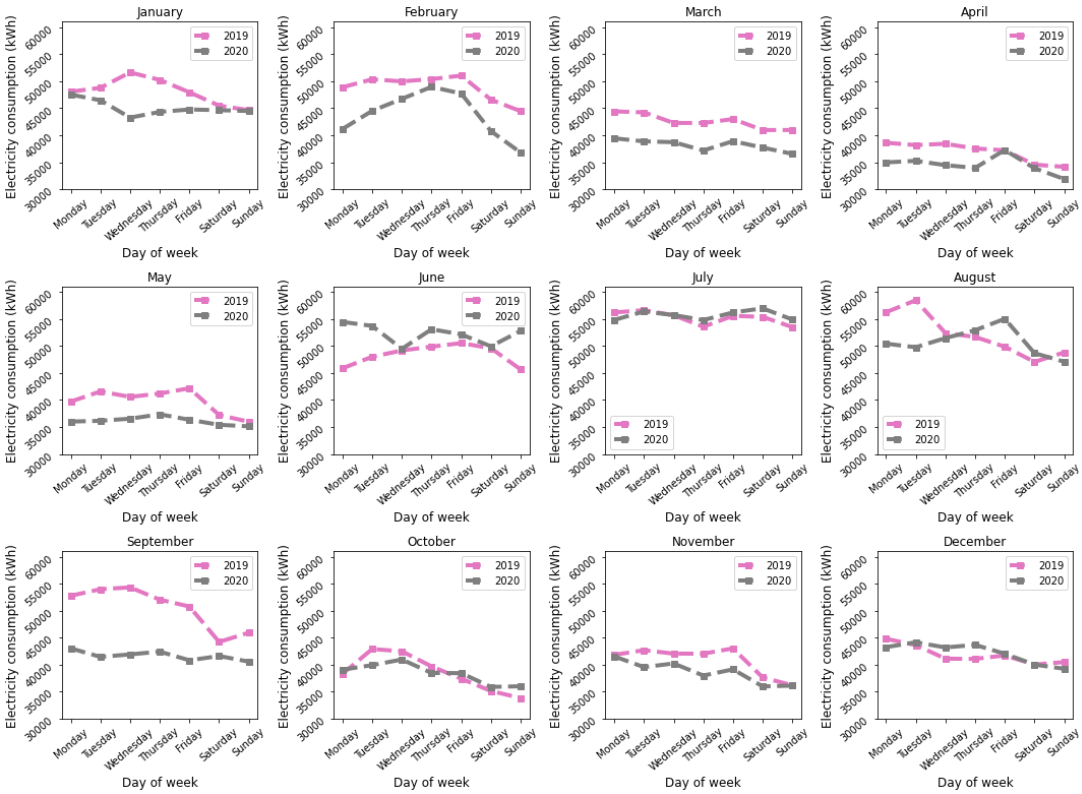


Fig B.6. Weekly electricity consumption in Kansas City from Jan to Dec in 2019 and 2020.

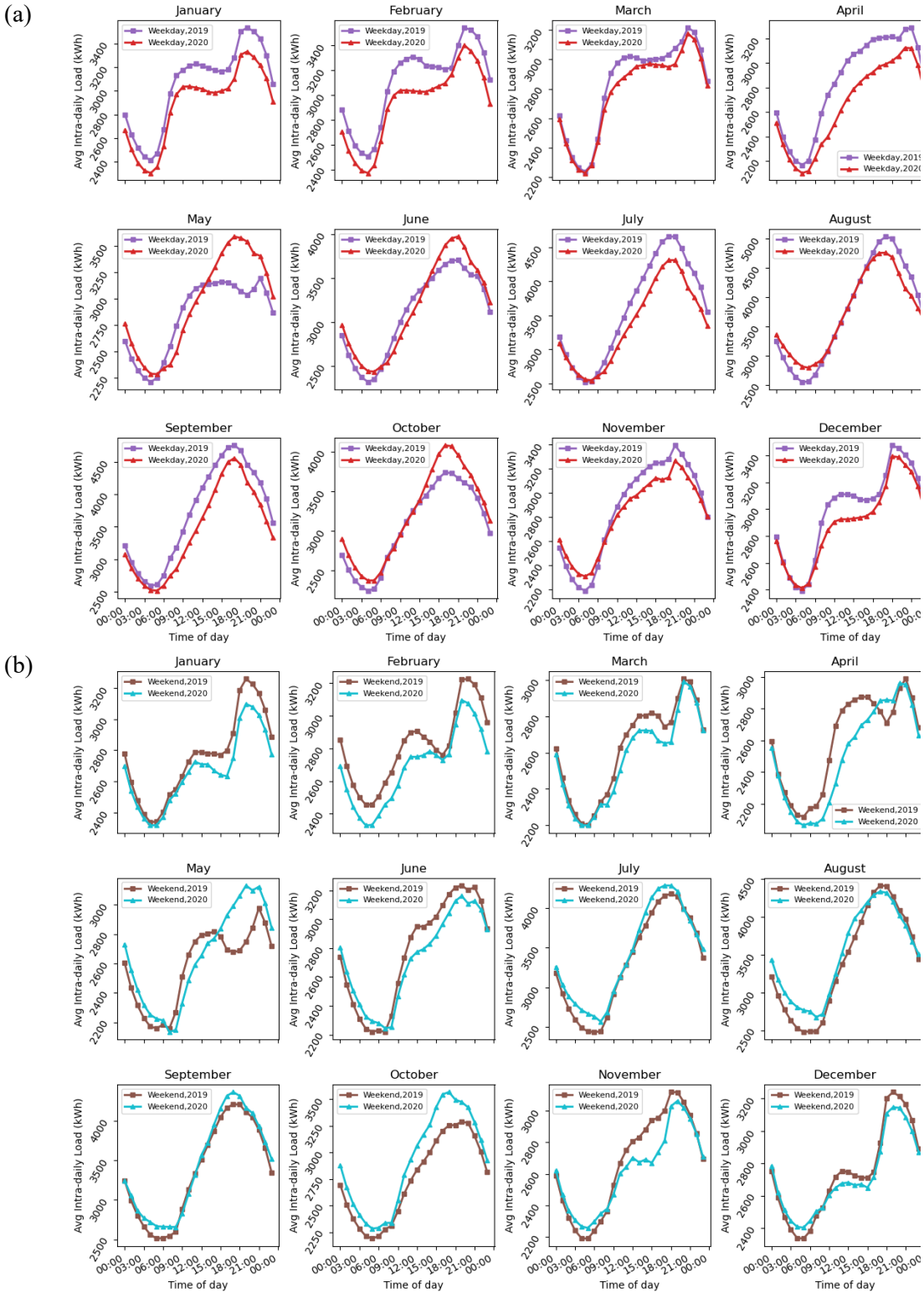


Fig B.7. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Los Angeles.

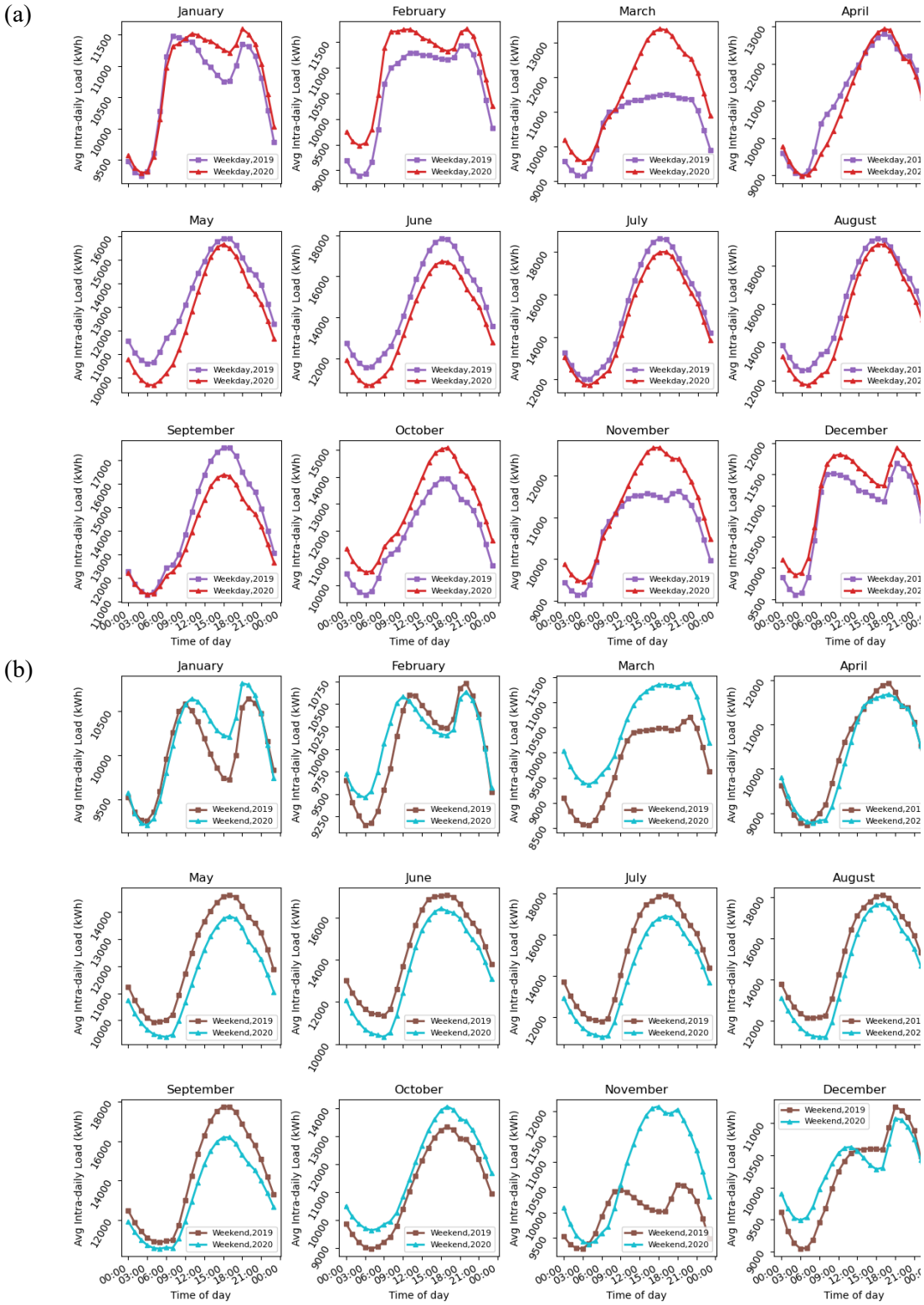


Fig B.8. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Houston.

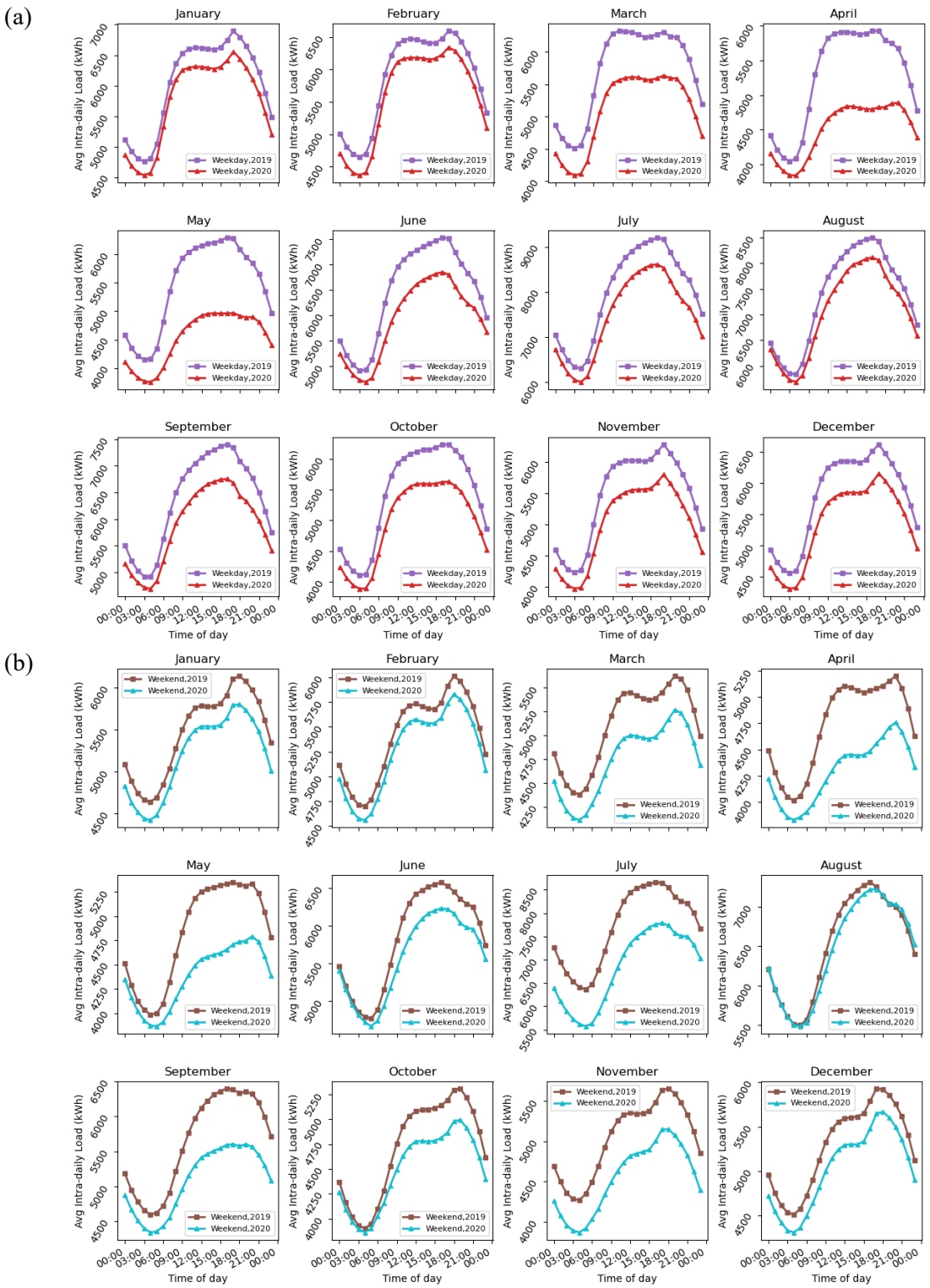


Fig B.9. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for New York.

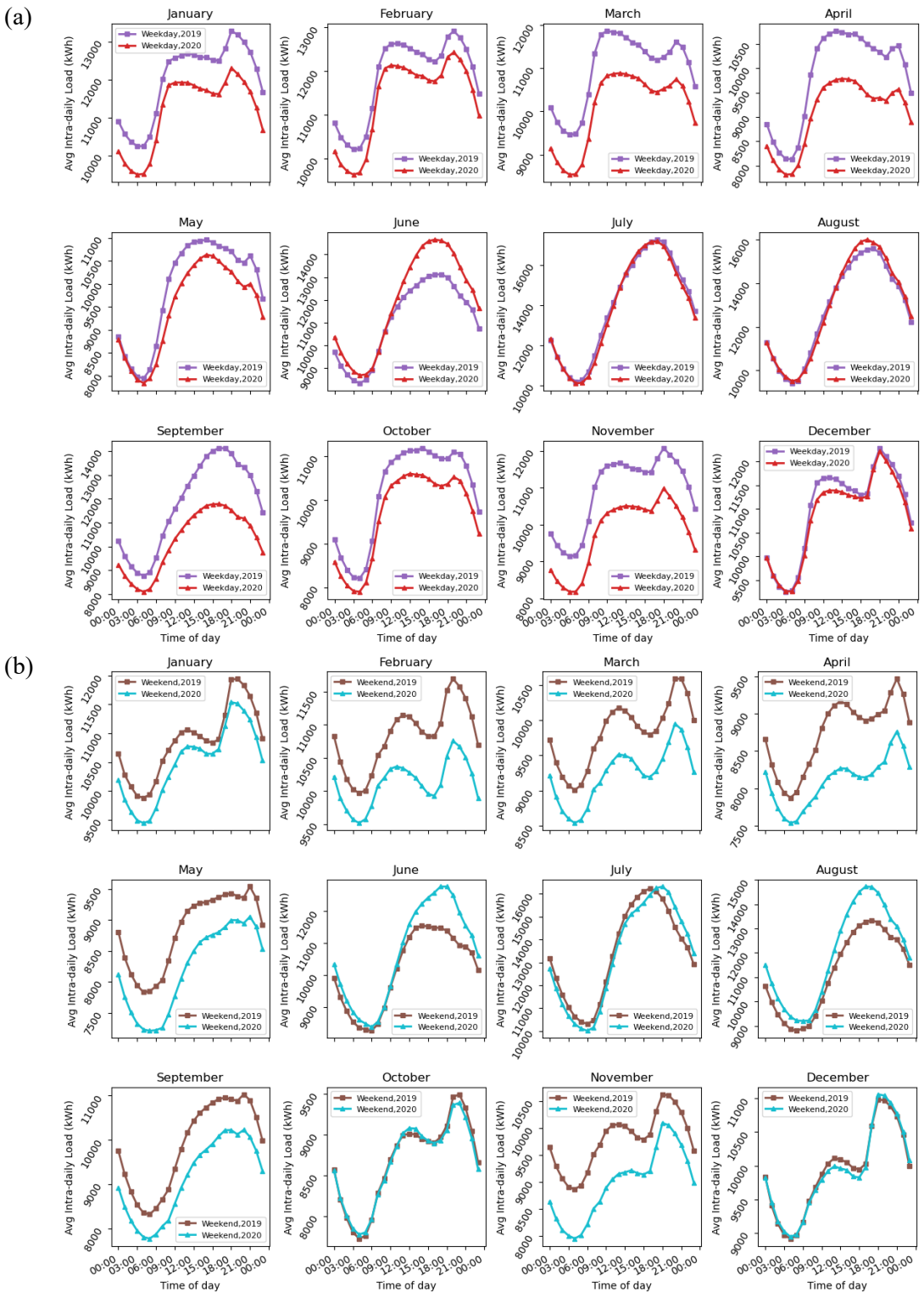


Fig B.10. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Chicago.

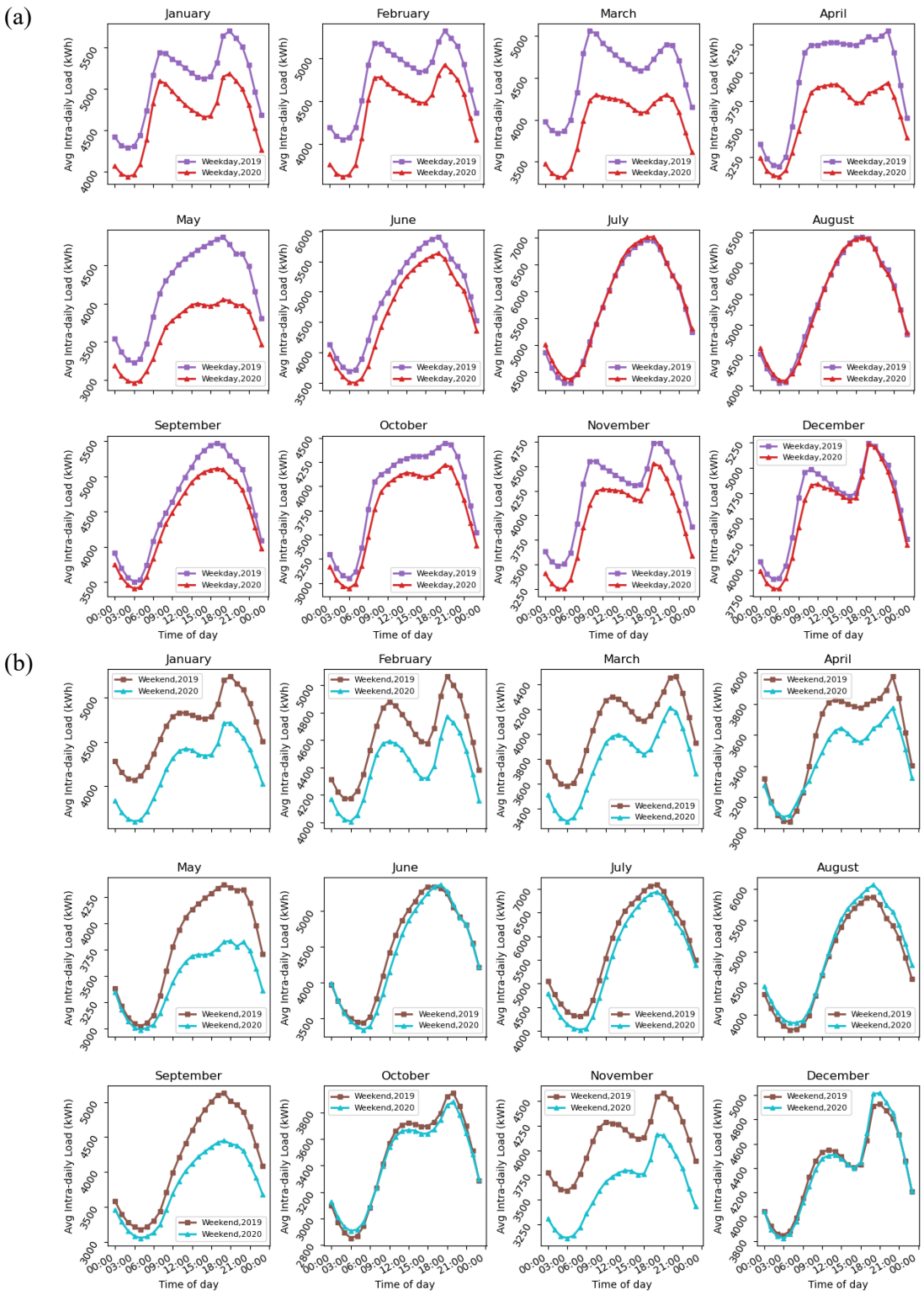


Fig B.11. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Philadelphia.

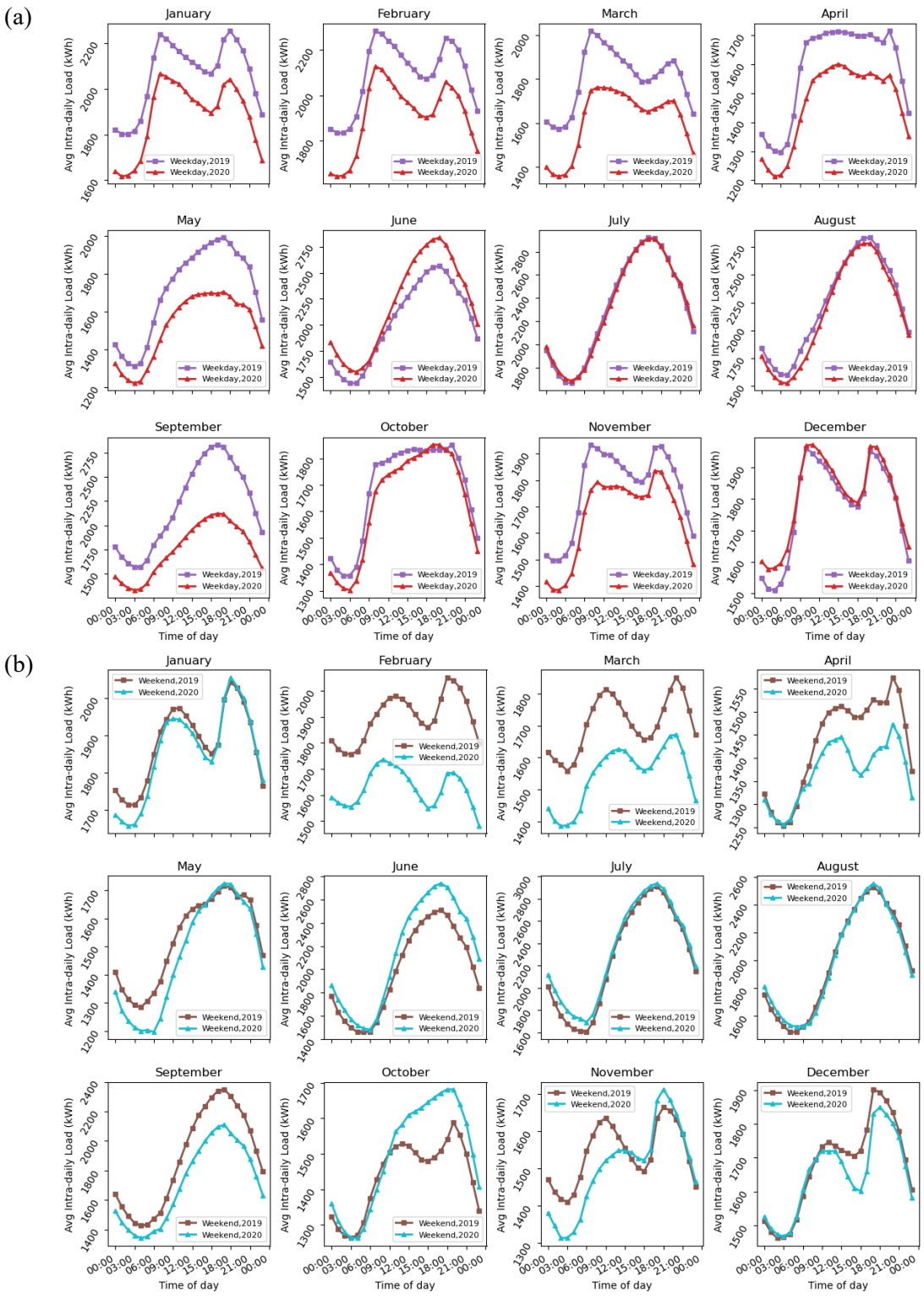


Fig B.12. Average intra-day electricity consumption (00:00 to 24:00) on weekdays (a) and weekends (b) each month in 2019 and 2020 for Kansas City.

Table B.2. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Los Angeles over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	1.40%	2.08%	3.36%	4.06%	4.65%	5.11%	3.44%
2	2.16%	1.81%	2.22%	2.52%	2.89%	2.42%	2.34%
3	3.57%	4.43%	4.95%	6.05%	7.41%	6.82%	5.54%
4	4.80%	6.81%	8.59%	11.31%	14.69%	16.88%	10.51%
5	3.42%	4.74%	5.82%	7.76%	10.20%	11.72%	7.28%
6	2.43%	2.45%	2.23%	2.92%	2.54%	2.87%	2.57%
7	4.53%	6.05%	7.63%	8.48%	10.33%	10.76%	7.96%
8	2.23%	2.99%	3.59%	3.84%	4.76%	4.80%	3.70%
9	1.81%	2.39%	3.68%	4.07%	5.32%	5.46%	3.79%
10	1.95%	2.67%	4.09%	4.67%	6.22%	6.57%	4.36%
11	3.37%	4.56%	5.77%	6.16%	7.43%	7.38%	5.78%
12	2.08%	2.52%	4.10%	4.50%	5.59%	6.81%	4.26%
13	1.83%	2.41%	3.98%	4.36%	5.16%	6.52%	4.04%
14	1.62%	1.83%	2.87%	2.87%	3.09%	3.91%	2.70%
15	2.68%	3.61%	5.44%	5.70%	6.23%	7.14%	5.13%
Models 1-15	**2.66%**	**3.42%**	**4.55%**	**5.28%**	**6.43%**	**7.01%**	**4.89%**
16	2.04%	2.38%	1.92%	1.92%	2.43%	2.18%	2.14%
17	4.35%	4.54%	3.42%	3.54%	3.47%	2.87%	3.70%
18	1.30%	1.48%	2.10%	1.77%	2.42%	4.17%	2.21%
19	2.13%	3.30%	4.73%	4.63%	5.15%	5.00%	4.15%
20	1.87%	2.87%	3.83%	3.75%	3.75%	3.98%	3.34%
21	1.32%	1.85%	2.59%	1.76%	2.34%	3.14%	2.17%
22	2.07%	1.87%	3.17%	3.74%	6.00%	7.10%	3.99%
23	2.44%	2.61%	2.76%	2.68%	3.16%	3.28%	2.82%
24	2.72%	2.92%	2.70%	2.37%	1.98%	2.35%	2.51%
25	3.00%	3.07%	2.86%	2.46%	2.25%	2.34%	2.66%
26	1.95%	1.96%	3.03%	3.52%	5.28%	6.38%	3.68%
27	2.43%	2.34%	2.12%	2.22%	1.99%	2.71%	2.30%
28	2.26%	2.12%	2.10%	2.14%	1.91%	2.49%	2.17%
29	2.23%	2.19%	2.12%	2.28%	2.28%	2.42%	2.25%
30	2.61%	2.49%	2.60%	3.30%	3.46%	4.09%	3.09%
Models 16-30	**2.31%**	**2.53%**	**2.80%**	**2.80%**	**3.19%**	**3.63%**	**2.88%**
31	3.43%	6.28%	8.58%	9.75%	10.92%	11.12%	8.35%
32	14.13%	11.02%	10.54%	13.33%	14.46%	13.89%	12.90%
33	7.63%	10.11%	11.47%	11.24%	11.38%	11.68%	10.59%
34	4.91%	6.64%	7.36%	6.97%	8.01%	8.42%	7.05%
35	5.11%	7.12%	7.53%	6.80%	8.25%	8.36%	7.19%
36	7.54%	6.36%	6.10%	5.45%	4.95%	4.67%	5.84%
37	5.28%	7.82%	10.47%	11.53%	13.16%	13.08%	10.22%
38	4.90%	6.85%	7.26%	6.36%	7.93%	8.08%	6.89%
39	5.23%	7.11%	7.38%	6.50%	8.24%	8.42%	7.15%
40	4.89%	6.54%	7.12%	6.36%	7.79%	8.14%	6.81%
41	3.95%	7.39%	10.29%	11.33%	12.84%	12.82%	9.77%

42	3.90%	6.27%	8.22%	9.58%	10.93%	11.58%	8.41%
43	2.84%	5.98%	8.21%	9.56%	10.74%	11.42%	8.13%
44	5.10%	7.10%	7.28%	6.48%	8.08%	8.40%	7.07%
45	4.59%	7.61%	10.04%	10.99%	12.94%	14.03%	10.03%
Models 31-45	**5.56%**	**7.35%**	**8.52%**	**8.82%**	**10.04%**	**10.27%**	**8.43%**
46	3.44%	3.94%	5.39%	5.84%	6.47%	6.60%	5.28%
47	11.97%	8.77%	8.58%	10.35%	12.18%	12.50%	10.73%
48	7.24%	9.17%	10.69%	9.78%	11.13%	10.62%	9.77%
49	3.65%	5.88%	7.54%	6.42%	8.50%	8.57%	6.76%
50	4.89%	6.44%	7.44%	6.45%	8.50%	8.60%	7.05%
51	8.53%	6.54%	5.53%	5.47%	5.38%	5.19%	6.11%
52	3.47%	5.55%	7.64%	8.15%	9.12%	9.47%	7.23%
53	7.57%	8.09%	8.49%	7.55%	9.28%	9.41%	8.40%
54	6.31%	5.65%	5.70%	5.84%	7.71%	7.54%	6.46%
55	5.39%	4.75%	5.09%	5.21%	7.14%	6.85%	5.74%
56	2.78%	5.06%	7.26%	8.38%	9.18%	9.54%	7.03%
57	3.72%	4.17%	4.57%	4.61%	4.88%	4.68%	4.44%
58	3.26%	3.68%	4.68%	4.78%	5.34%	5.24%	4.50%
59	6.85%	6.11%	6.33%	6.01%	8.01%	7.72%	6.84%
60	2.71%	6.29%	9.11%	10.63%	11.93%	11.84%	8.75%
Models 46-60	**5.45%**	**6.01%**	**6.94%**	**7.03%**	**8.32%**	**8.29%**	**7.01%**
61	2.86%	3.28%	3.80%	3.85%	4.16%	4.26%	3.70%
62	12.55%	10.17%	10.45%	10.47%	13.59%	13.15%	11.73%

Table B.3. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Houston over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	2.54%	2.50%	2.70%	2.99%	2.93%	3.49%	2.86%
2	4.09%	3.25%	3.44%	3.28%	2.99%	3.71%	3.46%
3	1.59%	2.36%	2.48%	2.94%	3.39%	3.55%	2.72%
4	1.55%	3.40%	3.85%	5.66%	6.78%	5.90%	4.52%
5	2.25%	4.39%	4.08%	7.16%	7.30%	8.72%	5.65%
6	2.45%	2.26%	3.59%	3.42%	3.95%	8.32%	4.00%
7	1.60%	5.33%	3.52%	3.17%	4.61%	4.07%	3.72%
8	2.27%	6.76%	5.44%	5.29%	7.51%	5.16%	5.41%
9	2.44%	3.61%	2.98%	3.04%	3.54%	4.34%	3.33%
10	2.43%	3.00%	2.98%	3.01%	3.48%	4.54%	3.24%
11	1.84%	5.45%	3.51%	3.25%	4.55%	4.11%	3.78%
12	2.10%	2.90%	3.14%	3.69%	2.79%	3.05%	2.94%
13	2.08%	2.77%	2.99%	3.56%	2.76%	3.08%	2.87%
14	2.30%	3.34%	2.81%	3.36%	3.06%	4.07%	3.16%
15	1.69%	2.53%	2.51%	3.01%	3.88%	6.09%	3.29%
Models 1-15	**2.21%**	**3.59%**	**3.33%**	**3.79%**	**4.23%**	**4.81%**	**3.66%**
16	2.58%	2.66%	2.70%	2.90%	2.80%	3.00%	2.77%
17	5.46%	5.01%	4.37%	4.84%	4.16%	4.50%	4.72%
18	2.29%	2.24%	2.96%	3.91%	3.39%	3.70%	3.08%

19	3.15%	3.40%	4.25%	4.25%	4.73%	5.06%	4.14%
20	2.56%	3.14%	4.15%	4.14%	4.74%	5.14%	3.98%
21	3.21%	2.99%	3.98%	3.51%	3.15%	3.91%	3.46%
22	1.53%	1.98%	2.54%	2.57%	3.54%	5.56%	2.95%
23	3.88%	5.59%	6.35%	7.68%	8.82%	9.80%	7.02%
24	2.86%	2.80%	3.40%	3.16%	3.62%	4.17%	3.33%
25	3.62%	3.06%	3.90%	3.30%	3.47%	3.98%	3.55%
26	1.74%	2.15%	2.57%	3.09%	3.90%	4.92%	3.06%
27	2.81%	3.00%	3.05%	3.76%	3.57%	4.21%	3.40%
28	2.74%	2.97%	2.99%	3.54%	3.49%	3.93%	3.28%
29	2.20%	2.50%	2.84%	2.89%	3.43%	3.12%	2.83%
30	2.48%	3.99%	4.25%	3.82%	5.32%	5.36%	4.20%
Models 16-30	**2.87%**	**3.17%**	**3.62%**	**3.82%**	**4.14%**	**4.69%**	**3.72%**
31	6.23%	5.18%	5.42%	5.72%	6.15%	6.01%	5.78%
32	30.89%	42.79%	42.11%	43.63%	43.18%	43.54%	41.02%
33	10.30%	11.26%	14.70%	17.33%	17.37%	16.31%	14.54%
34	6.23%	8.19%	9.54%	7.97%	8.30%	8.87%	8.19%
35	7.09%	9.71%	11.29%	9.78%	10.03%	10.69%	9.76%
36	18.19%	21.55%	19.29%	19.31%	20.46%	22.08%	20.15%
37	2.95%	6.52%	8.04%	9.61%	10.48%	10.18%	7.96%
38	6.91%	9.13%	10.98%	9.55%	9.42%	10.12%	9.35%
39	11.43%	16.07%	16.65%	15.39%	13.64%	13.92%	14.52%
40	10.45%	14.63%	15.31%	14.12%	12.70%	12.96%	13.36%
41	3.54%	7.15%	9.64%	11.30%	12.32%	11.83%	9.30%
42	7.56%	6.04%	5.72%	5.20%	6.00%	5.81%	6.05%
43	6.05%	5.08%	5.33%	5.70%	6.54%	6.33%	5.84%
44	10.47%	14.65%	16.23%	14.97%	13.00%	12.88%	13.70%
45	3.03%	6.80%	8.73%	10.21%	11.68%	11.84%	8.71%
Models 31-45	**9.42%**	**12.32%**	**13.26%**	**13.32%**	**13.42%**	**13.56%**	**12.55%**
46	7.13%	5.83%	5.87%	6.01%	6.68%	6.40%	6.32%
47	25.31%	37.22%	37.10%	34.96%	33.81%	33.25%	33.61%
48	20.79%	14.35%	14.31%	15.09%	17.95%	18.23%	16.79%
49	16.15%	18.02%	21.63%	19.23%	17.67%	18.23%	18.49%
50	12.56%	15.28%	18.28%	15.85%	14.91%	15.62%	15.42%
51	16.75%	17.17%	13.88%	13.49%	14.29%	15.04%	15.10%
52	9.06%	9.38%	10.29%	10.64%	11.37%	11.83%	10.43%
53	10.20%	12.74%	15.03%	12.62%	12.29%	13.15%	12.67%
54	14.63%	18.29%	19.14%	16.36%	15.29%	14.50%	16.37%
55	16.36%	20.38%	20.82%	18.70%	17.52%	16.76%	18.42%
56	6.59%	9.00%	10.75%	12.15%	12.83%	13.08%	10.73%
57	9.90%	8.06%	7.11%	6.07%	6.70%	6.75%	7.43%
58	7.26%	5.92%	5.94%	6.03%	7.01%	6.38%	6.42%
59	13.97%	17.83%	18.47%	16.40%	15.30%	15.23%	16.20%
60	4.36%	9.06%	12.63%	14.99%	16.63%	15.84%	12.25%
Models 46-60	**12.73%**	**14.57%**	**15.42%**	**14.57%**	**14.68%**	**14.69%**	**14.44%**
61	2.80%	2.97%	3.73%	3.77%	3.75%	4.09%	3.52%
62	8.31%	18.45%	23.62%	19.98%	21.54%	22.33%	19.04%

Table B.4. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Boston over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	2.96%	3.17%	5.82%	6.37%	5.26%	5.12%	4.78%
2	2.72%	3.37%	5.66%	6.05%	6.96%	6.30%	5.17%
3	3.16%	3.95%	5.07%	4.84%	4.62%	4.55%	4.37%
4	4.37%	4.33%	6.62%	8.69%	9.35%	9.47%	7.14%
5	4.38%	4.34%	6.68%	8.21%	9.02%	9.22%	6.97%
6	3.83%	4.55%	5.56%	5.34%	5.20%	5.73%	5.04%
7	2.73%	3.30%	5.55%	5.06%	4.50%	4.42%	4.26%
8	4.93%	7.40%	8.69%	12.73%	11.06%	10.46%	9.22%
9	3.99%	6.50%	6.75%	9.59%	7.27%	7.39%	6.92%
10	3.86%	6.17%	6.51%	9.34%	6.94%	7.04%	6.64%
11	2.62%	3.49%	5.93%	5.69%	5.18%	5.10%	4.67%
12	2.86%	4.12%	5.64%	4.50%	4.79%	4.92%	4.47%
13	2.61%	3.93%	5.61%	4.57%	4.78%	5.05%	4.43%
14	2.67%	3.48%	5.36%	4.42%	4.41%	4.61%	4.16%
15	3.22%	3.64%	6.35%	5.84%	5.90%	6.70%	5.27%
Models 1-15	**3.40%**	**4.38%**	**6.12%**	**6.75%**	**6.35%**	**6.41%**	**5.57%**
16	5.81%	7.99%	7.22%	7.47%	8.57%	9.25%	7.72%
17	6.52%	8.32%	7.34%	7.92%	7.58%	7.92%	7.60%
18	3.46%	5.27%	5.50%	5.99%	4.85%	5.81%	5.14%
19	2.99%	4.51%	5.29%	5.61%	4.90%	6.35%	4.94%
20	3.26%	4.79%	5.40%	5.71%	4.79%	5.77%	4.95%
21	6.29%	8.98%	8.84%	8.53%	11.13%	11.60%	9.23%
22	2.68%	4.68%	5.48%	5.38%	4.82%	4.75%	4.63%
23	2.52%	3.96%	5.82%	5.49%	4.56%	4.67%	4.50%
24	3.98%	5.44%	6.79%	6.02%	5.55%	5.82%	5.60%
25	4.32%	5.75%	6.94%	6.24%	5.82%	6.19%	5.88%
26	2.75%	4.40%	5.34%	5.24%	4.54%	4.49%	4.46%
27	4.23%	4.29%	6.49%	5.30%	6.58%	6.10%	5.50%
28	3.94%	3.98%	6.57%	5.44%	6.16%	5.65%	5.29%
29	4.35%	4.54%	5.76%	5.91%	5.87%	5.83%	5.38%
30	3.14%	5.22%	5.58%	6.13%	5.51%	5.32%	5.15%
Models 16-30	**4.02%**	**5.47%**	**6.29%**	**6.16%**	**6.08%**	**6.37%**	**5.73%**
31	2.17%	4.05%	7.27%	9.22%	8.72%	8.08%	6.59%
32	53.38%	66.51%	59.48%	58.75%	62.04%	63.38%	60.59%
33	7.98%	13.93%	21.16%	25.53%	24.64%	23.38%	19.44%
34	12.87%	20.27%	18.49%	19.15%	19.81%	18.96%	18.26%
35	11.51%	17.11%	14.90%	15.30%	15.42%	14.38%	14.77%
36	7.11%	4.80%	10.93%	13.26%	13.42%	12.12%	10.27%
37	1.42%	3.36%	6.28%	8.11%	7.60%	6.91%	5.61%
38	11.16%	15.73%	13.43%	13.99%	13.63%	12.73%	13.45%
39	12.06%	18.12%	15.86%	16.80%	16.49%	15.72%	15.84%
40	13.13%	20.69%	19.00%	20.15%	20.46%	19.71%	18.86%
41	1.45%	4.17%	7.71%	10.00%	9.55%	8.99%	6.98%

42	2.12%	3.42%	5.69%	7.13%	6.59%	5.96%	5.15%
43	2.03%	3.61%	6.49%	8.47%	8.02%	7.47%	6.02%
44	13.26%	17.29%	14.89%	15.05%	14.65%	13.41%	14.76%
45	1.59%	4.17%	7.61%	10.00%	9.46%	9.40%	7.04%
Models 31-45	**10.22%**	**14.48%**	**15.28%**	**16.73%**	**16.70%**	**16.04%**	**14.91%**
46	5.57%	5.16%	6.47%	7.36%	7.10%	6.12%	6.29%
47	50.20%	61.61%	60.67%	59.07%	61.45%	65.94%	59.83%
48	15.13%	18.44%	23.57%	26.44%	27.50%	25.54%	22.77%
49	17.71%	23.43%	21.57%	22.14%	23.02%	22.18%	21.68%
50	16.22%	19.93%	18.02%	18.08%	18.37%	17.33%	17.99%
51	7.21%	5.85%	10.42%	11.79%	12.53%	11.93%	9.95%
52	3.74%	4.34%	5.17%	6.47%	6.13%	5.73%	5.26%
53	16.02%	18.19%	15.67%	15.76%	15.70%	14.71%	16.01%
54	16.87%	20.32%	17.81%	18.08%	18.73%	17.30%	18.19%
55	17.38%	22.60%	20.33%	21.37%	21.99%	20.76%	20.74%
56	3.89%	4.56%	6.33%	8.22%	8.02%	7.87%	6.48%
57	5.56%	5.58%	6.22%	6.39%	6.41%	5.49%	5.94%
58	5.35%	4.96%	6.57%	7.34%	7.23%	6.16%	6.27%
59	17.58%	19.45%	16.64%	16.49%	16.31%	15.68%	17.02%
60	3.71%	4.50%	5.10%	5.98%	5.82%	5.29%	5.06%
Models 46-60	**13.48%**	**15.93%**	**16.04%**	**16.73%**	**17.09%**	**16.53%**	**15.97%**
61	4.66%	5.83%	6.12%	5.93%	5.26%	4.68%	5.41%
62	43.96%	81.70%	98.76%	84.71%	75.72%	72.80%	76.28%

Table B.5. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for New York over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	1.19%	1.63%	3.64%	4.13%	4.04%	3.51%	3.02%
2	1.56%	2.07%	3.61%	3.58%	3.93%	2.68%	2.91%
3	2.88%	2.97%	3.46%	3.46%	3.55%	3.62%	3.32%
4	6.09%	6.07%	7.50%	10.77%	11.63%	12.23%	9.05%
5	5.99%	6.09%	7.42%	10.83%	11.63%	11.83%	8.96%
6	2.48%	3.62%	3.80%	4.10%	5.20%	6.53%	4.29%
7	4.08%	3.61%	5.30%	5.54%	5.30%	5.35%	4.86%
8	6.58%	8.27%	9.45%	10.30%	12.84%	10.28%	9.62%
9	3.57%	3.92%	5.73%	6.84%	6.91%	6.11%	5.51%
10	3.26%	3.66%	6.00%	6.79%	8.78%	7.04%	5.92%
11	3.99%	3.64%	5.30%	5.49%	5.29%	5.23%	4.82%
12	1.96%	1.78%	4.04%	4.21%	4.64%	5.82%	3.74%
13	2.00%	1.65%	4.07%	4.30%	5.03%	6.05%	3.85%
14	2.63%	1.73%	3.96%	4.23%	4.25%	5.55%	3.73%
15	4.28%	4.38%	6.24%	6.11%	6.02%	6.31%	5.56%
Models 1-15	**3.50%**	**3.67%**	**5.30%**	**6.05%**	**6.60%**	**6.54%**	**5.28%**
16	2.51%	3.83%	3.91%	3.89%	5.08%	6.08%	4.22%
17	4.69%	6.09%	5.17%	5.12%	5.33%	6.06%	5.41%
18	1.19%	4.14%	4.08%	4.69%	4.45%	4.54%	3.85%

19	3.11%	4.43%	5.31%	4.81%	4.30%	3.86%	4.30%
20	3.29%	4.56%	5.31%	4.67%	4.29%	3.83%	4.33%
21	4.79%	7.83%	9.00%	8.32%	12.54%	14.53%	9.50%
22	1.55%	4.63%	4.72%	6.17%	6.56%	7.39%	5.17%
23	5.98%	6.32%	6.73%	8.08%	7.82%	7.22%	7.03%
24	2.12%	3.30%	4.88%	3.60%	4.00%	3.61%	3.59%
25	2.47%	3.57%	5.04%	3.73%	4.12%	3.65%	3.76%
26	1.75%	4.62%	4.68%	6.08%	6.57%	7.65%	5.22%
27	4.25%	5.59%	5.87%	5.89%	5.49%	6.58%	5.61%
28	2.55%	3.03%	5.18%	5.14%	4.52%	4.66%	4.18%
29	1.08%	2.03%	3.26%	3.61%	3.89%	3.17%	2.84%
30	2.27%	4.10%	4.12%	4.91%	5.24%	6.43%	4.51%
Models 16-30	**2.90%**	**4.54%**	**5.15%**	**5.25%**	**5.61%**	**5.95%**	**4.90%**
31	3.27%	4.72%	7.04%	8.86%	8.26%	7.63%	6.63%
32	56.72%	67.52%	61.79%	56.03%	58.18%	61.54%	60.30%
33	13.03%	18.72%	23.35%	26.73%	25.60%	24.33%	21.96%
34	8.52%	7.68%	7.27%	7.22%	7.44%	7.48%	7.60%
35	8.49%	7.35%	7.17%	7.28%	7.37%	7.38%	7.51%
36	5.52%	4.80%	8.50%	10.95%	11.69%	10.45%	8.65%
37	8.04%	9.00%	11.18%	12.68%	12.71%	12.47%	11.01%
38	8.54%	7.12%	7.07%	7.17%	7.28%	7.27%	7.41%
39	9.00%	7.37%	7.20%	7.21%	7.34%	7.40%	7.58%
40	8.97%	7.49%	7.06%	7.03%	7.30%	7.56%	7.57%
41	8.57%	9.95%	12.08%	13.69%	13.70%	13.34%	11.89%
42	3.21%	4.08%	6.34%	8.12%	7.69%	7.40%	6.14%
43	3.41%	4.57%	6.79%	8.73%	8.31%	7.90%	6.62%
44	9.32%	7.36%	7.19%	7.27%	7.18%	7.21%	7.59%
45	8.62%	9.90%	11.64%	13.27%	13.33%	13.54%	11.72%
Models 31-45	**10.88%**	**11.84%**	**12.78%**	**13.48%**	**13.56%**	**13.53%**	**12.68%**
46	2.52%	3.54%	5.44%	7.27%	6.81%	6.02%	5.27%
47	46.31%	45.79%	46.95%	42.64%	44.09%	47.55%	45.56%
48	17.76%	21.47%	23.89%	25.72%	26.44%	24.95%	23.37%
49	8.67%	9.34%	9.23%	8.86%	8.65%	8.42%	8.86%
50	9.06%	9.10%	9.02%	8.77%	8.53%	8.28%	8.79%
51	4.08%	4.92%	8.43%	10.99%	11.71%	10.80%	8.49%
52	5.46%	5.48%	7.12%	8.74%	8.81%	9.03%	7.44%
53	9.51%	8.86%	8.69%	8.79%	8.52%	8.17%	8.76%
54	9.24%	9.31%	9.11%	8.86%	8.72%	8.31%	8.92%
55	8.89%	9.58%	9.33%	8.94%	8.84%	8.46%	9.01%
56	6.08%	6.57%	8.36%	10.00%	10.05%	10.11%	8.53%
57	3.35%	4.34%	5.32%	6.37%	5.83%	4.92%	5.02%
58	2.84%	3.88%	5.34%	6.71%	6.23%	5.20%	5.03%
59	9.93%	9.01%	8.82%	8.82%	8.59%	8.37%	8.92%
60	8.57%	10.22%	11.44%	13.25%	13.41%	12.71%	11.60%
Models 46-60	**10.15%**	**10.76%**	**11.77%**	**12.31%**	**12.35%**	**12.09%**	**11.57%**
61	4.49%	5.29%	5.17%	5.04%	4.71%	4.37%	4.84%
62	44.57%	72.06%	92.19%	74.27%	70.83%	61.80%	69.29%

Table B.6. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Chicago over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	3.97%	6.51%	10.38%	12.21%	13.20%	14.55%	10.14%
2	3.71%	3.66%	4.78%	4.34%	4.09%	3.13%	3.95%
3	5.61%	6.49%	6.62%	7.07%	9.54%	13.52%	8.14%
4	5.90%	8.02%	11.38%	14.88%	16.05%	18.31%	12.42%
5	6.71%	9.07%	11.39%	15.99%	14.36%	15.19%	12.12%
6	4.29%	3.52%	4.04%	3.73%	4.51%	5.04%	4.19%
7	5.12%	7.62%	10.74%	12.07%	13.80%	14.54%	10.65%
8	6.43%	10.42%	14.88%	17.05%	21.21%	23.56%	15.59%
9	5.48%	8.39%	10.88%	10.09%	10.03%	9.71%	9.10%
10	5.33%	8.77%	13.28%	16.46%	21.64%	23.76%	14.87%
11	5.02%	7.49%	10.59%	11.71%	13.17%	13.57%	10.26%
12	3.79%	4.54%	6.48%	6.72%	5.63%	7.41%	5.76%
13	3.60%	3.90%	6.08%	6.04%	5.83%	6.76%	5.37%
14	3.32%	3.90%	5.99%	5.35%	4.81%	6.47%	4.97%
15	4.49%	6.07%	8.44%	8.45%	7.67%	9.08%	7.37%
Models 1-15	**4.85%**	**6.56%**	**9.06%**	**10.14%**	**11.04%**	**12.31%**	**8.99%**
16	4.84%	5.24%	5.10%	4.85%	5.11%	4.49%	4.94%
17	7.03%	7.03%	5.42%	6.04%	7.02%	7.63%	6.69%
18	3.20%	3.04%	2.96%	4.25%	4.02%	3.84%	3.55%
19	3.03%	3.48%	4.21%	4.80%	4.98%	4.81%	4.22%
20	2.84%	3.38%	4.43%	4.96%	5.03%	4.80%	4.24%
21	7.78%	9.39%	8.64%	10.62%	12.46%	13.66%	10.42%
22	3.03%	3.84%	3.43%	5.30%	5.25%	5.99%	4.47%
23	2.59%	4.13%	6.18%	7.41%	7.50%	6.62%	5.74%
24	3.88%	3.87%	4.98%	4.18%	4.13%	4.46%	4.25%
25	4.94%	4.64%	4.80%	4.32%	4.09%	4.82%	4.60%
26	2.75%	3.69%	3.56%	5.34%	5.35%	5.36%	4.34%
27	5.87%	6.02%	5.66%	6.30%	5.62%	6.34%	5.97%
28	5.55%	5.39%	5.39%	5.67%	5.26%	5.98%	5.54%
29	4.12%	3.81%	4.76%	4.34%	4.70%	4.83%	4.43%
30	1.37%	1.94%	2.61%	3.63%	3.86%	3.69%	2.85%
Models 16-30	**4.19%**	**4.59%**	**4.81%**	**5.47%**	**5.63%**	**5.82%**	**5.08%**
31	2.94%	5.35%	8.79%	10.59%	10.49%	9.74%	7.98%
32	52.88%	63.71%	60.24%	62.18%	64.75%	66.78%	61.76%
33	10.65%	16.65%	21.74%	25.47%	23.82%	21.81%	20.02%
34	7.16%	7.87%	9.21%	9.26%	9.10%	9.67%	8.71%
35	7.80%	8.58%	10.78%	10.71%	10.17%	10.97%	9.83%
36	3.17%	5.79%	9.39%	11.97%	12.41%	11.05%	8.96%
37	6.04%	9.67%	13.11%	15.22%	15.50%	14.68%	12.37%
38	7.63%	8.42%	10.42%	9.96%	9.84%	10.47%	9.46%
39	9.56%	9.64%	10.71%	9.95%	8.84%	9.00%	9.62%
40	8.13%	8.47%	9.40%	8.86%	8.51%	8.70%	8.68%

41	6.43%	10.59%	14.59%	16.81%	16.90%	16.06%	13.56%
42	3.00%	5.54%	8.37%	10.66%	10.48%	10.13%	8.03%
43	2.97%	5.53%	8.76%	11.06%	10.74%	10.31%	8.23%
44	9.25%	9.18%	10.30%	9.72%	8.73%	8.98%	9.36%
45	6.19%	9.90%	13.00%	15.46%	15.60%	15.54%	12.62%
Models 31-45	**9.59%**	**12.33%**	**14.59%**	**15.86%**	**15.72%**	**15.59%**	**13.95%**
46	2.65%	3.62%	6.15%	7.92%	8.12%	7.62%	6.01%
47	61.77%	73.51%	71.39%	73.00%	74.06%	75.50%	71.54%
48	15.44%	20.20%	21.53%	24.52%	24.97%	23.90%	21.76%
49	10.07%	9.54%	9.69%	9.37%	10.07%	10.89%	9.94%
50	10.45%	9.99%	10.59%	10.39%	11.02%	12.01%	10.74%
51	4.40%	6.82%	9.64%	12.15%	12.77%	11.41%	9.53%
52	5.10%	7.14%	9.55%	11.93%	12.42%	11.93%	9.68%
53	10.89%	10.22%	11.08%	11.06%	11.47%	12.11%	11.14%
54	14.78%	13.58%	12.65%	11.77%	10.82%	10.45%	12.34%
55	14.17%	13.48%	12.13%	11.19%	10.44%	10.16%	11.93%
56	5.90%	8.45%	11.66%	14.33%	14.74%	13.95%	11.51%
57	2.60%	3.44%	5.52%	7.00%	7.19%	6.45%	5.37%
58	2.74%	3.63%	6.32%	7.86%	8.11%	7.20%	5.98%
59	15.43%	13.89%	13.37%	12.29%	11.34%	11.06%	12.90%
60	6.14%	9.07%	12.62%	14.66%	15.95%	14.42%	12.14%
Models 46-60	**12.17%**	**13.77%**	**14.93%**	**15.96%**	**16.23%**	**15.94%**	**14.83%**
61	4.50%	4.33%	5.31%	5.28%	5.26%	5.22%	4.98%
62	144.88%	135.60%	110.41%	112.72%	122.51%	115.74%	123.64%

Table B.7. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Philadelphia over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	3.15%	5.17%	7.45%	7.32%	6.89%	6.42%	6.07%
2	5.05%	6.00%	8.01%	10.21%	10.26%	9.10%	8.11%
3	3.61%	5.40%	6.73%	5.94%	5.89%	5.68%	5.54%
4	7.25%	7.37%	9.47%	10.74%	11.37%	11.27%	9.58%
5	7.07%	8.38%	10.16%	13.85%	11.02%	10.08%	10.09%
6	2.91%	5.11%	6.67%	5.73%	5.80%	6.24%	5.41%
7	5.25%	6.65%	8.58%	8.96%	8.29%	7.54%	7.54%
8	6.58%	9.79%	10.90%	15.14%	10.32%	9.76%	10.41%
9	5.69%	9.23%	8.95%	11.62%	6.99%	7.35%	8.30%
10	5.76%	9.17%	8.92%	11.77%	7.07%	7.57%	8.38%
11	5.02%	6.57%	8.56%	9.03%	8.32%	7.56%	7.51%
12	3.92%	4.93%	7.47%	7.06%	7.25%	7.51%	6.36%
13	3.94%	4.94%	7.54%	7.31%	7.45%	7.68%	6.48%
14	3.49%	4.70%	7.22%	6.86%	6.85%	7.35%	6.08%
15	6.17%	6.31%	8.85%	9.72%	9.56%	9.01%	8.27%
Models 1-15	**4.99%**	**6.65%**	**8.37%**	**9.42%**	**8.22%**	**8.01%**	**7.61%**
16	2.87%	6.03%	6.70%	6.24%	7.03%	7.74%	6.10%
17	2.82%	5.73%	6.68%	6.10%	5.57%	6.01%	5.48%

18	3.93%	6.75%	7.50%	6.98%	5.68%	5.55%	6.07%
19	4.63%	7.38%	8.26%	7.27%	6.63%	6.20%	6.73%
20	4.66%	7.41%	8.28%	7.30%	6.63%	6.29%	6.76%
21	2.41%	6.28%	7.27%	6.41%	7.32%	7.59%	6.21%
22	4.39%	6.62%	7.42%	6.68%	5.84%	5.31%	6.04%
23	6.44%	8.96%	10.15%	12.11%	8.84%	7.07%	8.93%
24	2.75%	5.17%	7.10%	6.47%	6.75%	5.57%	5.63%
25	2.58%	5.12%	7.08%	6.41%	6.69%	5.52%	5.57%
26	4.68%	6.81%	7.62%	6.88%	6.04%	5.36%	6.23%
27	2.70%	5.69%	6.83%	6.41%	6.32%	7.05%	5.83%
28	2.70%	5.57%	6.83%	6.25%	6.12%	6.72%	5.70%
29	3.61%	7.29%	9.91%	14.90%	10.34%	9.15%	9.20%
30	4.30%	7.05%	7.24%	6.82%	6.22%	6.21%	6.31%
Models 16-30	**3.70%**	**6.53%**	**7.66%**	**7.55%**	**6.80%**	**6.49%**	**6.45%**
31	5.48%	7.89%	9.85%	11.62%	10.81%	9.36%	9.17%
32	52.02%	59.03%	53.58%	50.19%	51.96%	53.70%	53.42%
33	10.40%	17.53%	22.57%	26.49%	25.10%	22.54%	20.77%
34	6.36%	10.01%	11.68%	11.69%	11.58%	10.89%	10.37%
35	5.61%	9.33%	10.98%	10.61%	10.04%	9.49%	9.34%
36	4.27%	6.73%	11.23%	14.19%	14.77%	13.03%	10.70%
37	9.67%	11.83%	13.42%	14.99%	14.07%	12.63%	12.77%
38	4.97%	9.42%	10.97%	11.07%	10.47%	9.95%	9.48%
39	5.03%	9.84%	11.50%	11.49%	10.96%	10.50%	9.89%
40	5.63%	10.85%	12.25%	12.98%	12.91%	12.68%	11.22%
41	8.89%	11.90%	14.01%	16.07%	15.13%	13.41%	13.23%
42	6.22%	8.11%	9.57%	11.19%	10.76%	9.50%	9.22%
43	5.14%	7.83%	9.86%	12.22%	11.60%	10.34%	9.50%
44	5.25%	9.75%	11.66%	11.52%	11.03%	10.35%	9.93%
45	9.39%	11.85%	13.19%	15.22%	14.21%	13.06%	12.82%
Models 31-45	**9.62%**	**13.46%**	**15.09%**	**16.10%**	**15.69%**	**14.76%**	**14.12%**
46	2.31%	4.12%	6.10%	8.05%	7.93%	7.03%	5.92%
47	44.50%	48.23%	48.81%	45.21%	47.71%	51.57%	47.67%
48	15.34%	19.06%	22.63%	25.49%	25.53%	22.99%	21.84%
49	8.56%	12.69%	13.74%	13.84%	14.22%	13.98%	12.84%
50	9.02%	11.42%	12.40%	11.68%	11.51%	11.12%	11.19%
51	1.90%	4.83%	9.86%	13.02%	13.91%	12.71%	9.37%
52	7.07%	8.81%	10.64%	12.18%	11.79%	10.94%	10.24%
53	9.49%	11.67%	12.14%	11.60%	11.59%	11.10%	11.27%
54	10.38%	12.91%	13.71%	12.84%	12.73%	12.52%	12.52%
55	10.04%	14.28%	14.88%	14.81%	15.52%	15.59%	14.19%
56	7.19%	9.55%	11.80%	13.90%	13.63%	12.78%	11.47%
57	2.20%	3.72%	5.93%	7.43%	7.45%	6.26%	5.50%
58	2.32%	4.44%	7.03%	9.01%	9.03%	7.48%	6.55%
59	10.45%	13.37%	13.46%	12.99%	12.89%	12.98%	12.69%
60	7.03%	9.67%	11.53%	13.67%	13.52%	11.98%	11.23%
Models 46-60	**9.85%**	**12.59%**	**14.31%**	**15.05%**	**15.26%**	**14.73%**	**13.63%**
61	4.60%	7.21%	7.69%	7.31%	6.60%	5.76%	6.53%
62	63.80%	105.07%	116.98%	97.40%	89.13%	79.84%	92.04%

Table B.8. Forecasting accuracy (MAPE) of pre-COVID trained models 1-62 for Kansas City over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	3.25%	7.61%	12.84%	15.93%	15.62%	17.36%	12.10%
2	3.82%	7.04%	10.08%	10.93%	11.00%	9.42%	8.71%
3	4.97%	7.89%	10.88%	13.16%	13.27%	13.13%	10.55%
4	5.83%	10.14%	14.49%	18.73%	19.29%	19.03%	14.59%
5	5.84%	10.40%	14.77%	19.14%	19.83%	19.70%	14.95%
6	3.94%	6.97%	8.88%	7.01%	9.46%	11.78%	8.01%
7	4.65%	9.31%	13.81%	15.44%	16.23%	17.01%	12.74%
8	5.16%	9.92%	14.45%	16.06%	14.20%	15.82%	12.60%
9	3.68%	8.62%	13.22%	15.43%	15.91%	13.29%	11.69%
10	3.55%	7.08%	10.50%	12.09%	11.50%	11.06%	9.30%
11	4.63%	9.42%	14.03%	16.20%	16.69%	18.00%	13.16%
12	3.50%	6.63%	10.10%	11.23%	10.91%	11.05%	8.90%
13	3.22%	6.50%	10.20%	11.53%	11.27%	11.62%	9.06%
14	3.51%	7.25%	10.14%	10.33%	10.22%	10.36%	8.64%
15	4.32%	7.54%	10.91%	11.61%	11.69%	12.33%	9.73%
Models 1-15	**4.26%**	**8.15%**	**11.95%**	**13.65%**	**13.81%**	**14.06%**	**10.98%**
16	3.86%	7.50%	10.64%	11.73%	11.31%	10.74%	9.30%
17	6.61%	9.12%	10.11%	9.54%	9.82%	8.13%	8.89%
18	5.97%	8.40%	9.13%	9.61%	9.39%	8.51%	8.50%
19	6.70%	9.32%	11.33%	12.36%	12.27%	11.23%	10.54%
20	6.41%	9.28%	11.37%	12.58%	12.52%	11.55%	10.62%
21	6.47%	9.84%	10.80%	8.66%	9.96%	9.52%	9.21%
22	4.10%	8.95%	12.17%	14.65%	12.96%	13.82%	11.11%
23	5.60%	8.77%	11.98%	15.53%	15.97%	15.02%	12.15%
24	4.10%	7.55%	10.60%	11.36%	11.61%	10.22%	9.24%
25	4.92%	8.14%	10.44%	11.51%	10.61%	9.40%	9.17%
26	4.24%	9.82%	13.74%	16.82%	15.21%	16.92%	12.79%
27	2.92%	8.25%	15.72%	18.52%	14.98%	15.25%	12.61%
28	2.74%	8.34%	15.78%	18.87%	15.28%	15.67%	12.78%
29	3.47%	7.04%	10.60%	10.45%	10.88%	8.28%	8.45%
30	3.95%	9.73%	13.52%	17.89%	17.06%	18.62%	13.46%
Models 16-30	**4.80%**	**8.67%**	**11.86%**	**13.34%**	**12.66%**	**12.19%**	**10.59%**
31	5.25%	8.49%	11.19%	13.63%	13.25%	11.73%	10.59%
32	40.71%	60.17%	52.77%	50.61%	54.81%	57.23%	52.72%
33	13.40%	21.56%	22.80%	25.44%	23.94%	21.41%	21.42%
34	7.55%	18.11%	20.88%	21.08%	18.63%	18.03%	17.38%
35	7.80%	18.45%	22.01%	22.88%	20.63%	20.44%	18.70%
36	4.27%	8.26%	9.41%	10.86%	12.36%	12.18%	9.55%
37	11.63%	17.00%	19.89%	22.32%	21.14%	19.13%	18.52%
38	8.18%	18.61%	22.10%	22.14%	20.22%	19.66%	18.49%
39	6.30%	16.16%	18.12%	16.28%	15.04%	13.30%	14.20%
40	6.06%	14.90%	17.30%	15.54%	14.16%	12.21%	13.36%

41	11.98%	17.05%	20.53%	23.09%	21.89%	20.10%	19.11%
42	6.07%	10.30%	12.82%	15.53%	14.80%	13.48%	12.17%
43	5.09%	8.46%	10.97%	13.98%	13.65%	12.35%	10.75%
44	5.77%	15.57%	17.99%	16.60%	15.24%	13.86%	14.17%
45	11.90%	17.38%	19.95%	22.83%	21.54%	20.98%	19.10%
Models 31-45	**10.13%**	**18.03%**	**19.91%**	**20.85%**	**20.09%**	**19.07%**	**18.01%**
46	2.64%	5.07%	8.56%	10.92%	11.34%	10.52%	8.18%
47	42.79%	66.89%	62.00%	60.09%	58.70%	61.12%	58.60%
48	14.98%	24.02%	22.66%	23.93%	24.76%	23.22%	22.26%
49	13.97%	17.86%	20.56%	19.89%	20.73%	20.43%	18.91%
50	12.52%	17.84%	21.16%	21.11%	21.78%	21.87%	19.38%
51	5.42%	4.98%	7.48%	9.67%	11.61%	11.88%	8.51%
52	9.86%	15.00%	16.37%	18.49%	17.90%	17.36%	15.83%
53	12.51%	17.60%	20.59%	20.61%	21.59%	21.64%	19.09%
54	12.24%	19.78%	20.18%	17.64%	17.26%	15.81%	17.15%
55	13.03%	19.68%	20.07%	16.71%	16.69%	15.58%	16.96%
56	9.70%	14.24%	16.26%	19.27%	19.20%	18.04%	16.12%
57	2.70%	5.14%	8.54%	10.93%	11.58%	9.84%	8.12%
58	2.67%	4.91%	8.39%	10.90%	11.82%	10.08%	8.13%
59	12.18%	20.49%	20.63%	18.01%	17.68%	16.34%	17.56%
60	9.84%	14.27%	16.79%	19.14%	20.08%	18.10%	16.37%
Models 46-60	**11.80%**	**17.85%**	**19.35%**	**19.82%**	**20.18%**	**19.46%**	**18.08%**
61	5.27%	6.83%	9.55%	9.90%	10.46%	10.23%	8.71%
62	70.70%	122.37%	157.27%	143.49%	146.32%	138.44%	129.77%

Table B.9. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Los Angeles over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	1.74%	1.28%	1.70%	3.82%	3.16%	4.46%	2.69%
2	3.07%	3.12%	3.05%	3.21%	2.89%	3.65%	3.17%
3	2.68%	1.94%	2.50%	2.39%	2.98%	3.64%	2.69%
4	2.82%	3.12%	2.85%	3.39%	3.99%	3.40%	3.26%
5	2.87%	3.18%	3.02%	3.19%	4.07%	3.53%	3.31%
6	2.16%	2.10%	3.26%	3.73%	2.43%	3.24%	2.82%
7	3.04%	3.00%	2.60%	2.92%	2.70%	2.73%	2.83%
8	2.99%	3.14%	3.19%	3.48%	3.61%	3.04%	3.24%
9	3.65%	3.57%	3.31%	3.62%	4.34%	4.09%	3.77%
10	3.62%	3.73%	2.38%	3.12%	4.43%	3.73%	3.50%
11	2.75%	2.63%	1.36%	1.97%	2.37%	2.73%	2.30%
12	2.52%	1.99%	2.72%	2.49%	2.62%	3.10%	2.57%
13	2.39%	1.79%	3.18%	3.82%	2.39%	2.88%	2.74%
14	2.57%	2.03%	4.20%	3.93%	2.76%	2.63%	3.02%
15	2.07%	1.64%	1.55%	2.65%	2.41%	3.62%	2.32%
Models 1-15	**2.73%**	**2.55%**	**2.72%**	**3.18%**	**3.14%**	**3.37%**	**2.95%**
16	2.22%	1.25%	1.25%	2.24%	2.04%	2.61%	1.93%
17	3.00%	2.31%	2.71%	2.12%	2.53%	3.61%	2.71%

18	3.51%	3.13%	3.13%	3.17%	3.42%	3.79%	3.36%
19	3.56%	2.94%	3.14%	3.19%	3.47%	4.19%	3.42%
20	3.28%	2.75%	3.42%	3.63%	2.91%	3.66%	3.28%
21	2.46%	1.87%	3.27%	3.91%	4.41%	4.64%	3.43%
22	3.42%	2.93%	3.31%	3.59%	3.50%	4.01%	3.46%
23	2.83%	2.35%	2.11%	2.48%	2.83%	3.31%	2.65%
24	2.45%	2.22%	3.02%	2.32%	3.74%	3.78%	2.92%
25	2.72%	2.22%	2.33%	2.29%	3.58%	3.33%	2.74%
26	3.24%	2.50%	1.48%	2.27%	3.02%	3.59%	2.68%
27	3.01%	2.97%	3.77%	3.48%	4.28%	4.75%	3.71%
28	2.76%	2.64%	3.14%	3.11%	3.67%	4.09%	3.24%
29	3.38%	2.97%	2.47%	3.45%	3.40%	3.69%	3.23%
30	2.70%	2.03%	2.46%	3.02%	2.58%	3.62%	2.73%
Models 16-30	**2.97%**	**2.47%**	**2.73%**	**2.95%**	**3.29%**	**3.78%**	**3.03%**
31	6.65%	6.33%	5.66%	6.08%	6.84%	6.28%	6.31%
32	8.96%	8.52%	8.91%	10.23%	13.14%	11.68%	10.24%
33	6.76%	6.19%	7.51%	9.85%	7.58%	6.93%	7.47%
34	9.99%	10.43%	5.41%	6.53%	13.80%	12.88%	9.84%
35	8.12%	8.24%	7.47%	9.55%	12.37%	11.47%	9.54%
36	7.07%	5.76%	6.93%	6.71%	6.48%	6.68%	6.60%
37	6.90%	6.72%	5.29%	5.82%	6.67%	6.40%	6.30%
38	8.37%	8.73%	5.19%	5.58%	12.03%	11.62%	8.59%
39	7.62%	7.36%	9.60%	10.63%	11.21%	10.79%	9.54%
40	9.46%	9.58%	5.49%	6.02%	12.39%	12.17%	9.18%
41	6.62%	6.22%	10.25%	11.10%	6.77%	6.47%	7.90%
42	6.83%	6.50%	5.96%	6.09%	6.79%	6.55%	6.46%
43	6.49%	5.95%	9.03%	11.05%	6.73%	6.45%	7.62%
44	7.89%	7.89%	7.55%	7.41%	11.24%	10.73%	8.79%
45	6.29%	6.25%	6.19%	6.29%	6.77%	6.58%	6.39%
Models 31-45	**7.60%**	**7.38%**	**7.10%**	**7.93%**	**9.39%**	**8.91%**	**8.05%**
46	6.00%	5.23%	6.26%	6.10%	7.03%	6.94%	6.26%
47	9.47%	9.15%	5.86%	5.76%	12.92%	12.08%	9.21%
48	9.51%	8.37%	7.54%	9.39%	8.80%	9.06%	8.78%
49	8.89%	9.66%	6.86%	8.71%	11.92%	11.70%	9.62%
50	7.13%	7.75%	7.34%	9.33%	10.59%	10.24%	8.73%
51	8.18%	6.22%	6.42%	6.57%	7.75%	7.56%	7.12%
52	5.31%	4.86%	9.69%	10.61%	5.53%	6.22%	7.03%
53	7.21%	7.23%	8.59%	8.65%	10.85%	10.31%	8.81%
54	7.26%	6.86%	10.06%	10.94%	10.70%	9.71%	9.26%
55	8.52%	8.07%	5.85%	5.63%	11.52%	10.54%	8.35%
56	4.61%	4.32%	6.22%	6.24%	5.75%	6.15%	5.55%
57	6.17%	5.14%	8.89%	11.07%	6.29%	6.26%	7.30%
58	5.97%	4.98%	6.67%	7.15%	6.77%	6.51%	6.34%
59	6.52%	6.34%	7.85%	9.65%	10.35%	9.62%	8.39%
60	5.67%	5.14%	11.14%	12.55%	7.33%	7.20%	8.17%
Models 46-60	**7.09%**	**6.62%**	**7.68%**	**8.56%**	**8.94%**	**8.67%**	**7.93%**

Table B.10. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Houston over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	2.60%	3.01%	3.65%	3.62%	3.77%	4.87%	3.59%
2	3.20%	3.28%	3.56%	5.02%	5.01%	6.89%	4.49%
3	2.55%	3.07%	4.38%	3.49%	4.73%	9.64%	4.64%
4	2.54%	2.68%	3.44%	3.61%	3.69%	4.98%	3.49%
5	2.54%	2.80%	3.46%	4.00%	4.07%	5.10%	3.66%
6	2.81%	3.65%	3.65%	5.18%	7.55%	5.69%	4.76%
7	2.32%	3.91%	3.51%	4.30%	4.84%	5.30%	4.03%
8	2.74%	3.09%	3.78%	3.54%	3.77%	4.98%	3.65%
9	2.74%	3.22%	4.02%	3.91%	4.25%	5.26%	3.90%
10	2.69%	3.11%	3.96%	3.93%	4.41%	5.33%	3.90%
11	2.43%	3.90%	3.50%	4.27%	4.82%	5.24%	4.03%
12	2.30%	3.67%	3.98%	5.31%	7.61%	6.47%	4.89%
13	2.33%	3.59%	3.96%	5.23%	7.63%	6.40%	4.86%
14	2.64%	2.75%	3.79%	3.71%	3.71%	4.96%	3.59%
15	2.04%	4.19%	4.74%	7.22%	6.40%	7.46%	5.34%
Models 1-15	**2.56%**	**3.33%**	**3.82%**	**4.42%**	**5.08%**	**5.90%**	**4.19%**
16	2.14%	3.20%	3.75%	3.74%	3.46%	4.69%	3.50%
17	2.35%	3.15%	4.63%	4.60%	5.36%	7.00%	4.52%
18	2.51%	4.00%	4.46%	4.95%	4.50%	4.94%	4.23%
19	3.09%	3.16%	4.73%	4.26%	3.83%	4.50%	3.93%
20	2.85%	3.11%	4.56%	4.17%	3.78%	4.46%	3.82%
21	3.49%	3.33%	4.34%	5.48%	7.91%	13.86%	6.40%
22	3.10%	3.47%	4.76%	5.26%	8.01%	12.79%	6.23%
23	4.03%	3.89%	4.67%	4.62%	4.13%	5.02%	4.39%
24	3.82%	3.42%	3.66%	5.26%	5.03%	5.40%	4.43%
25	2.70%	2.69%	3.89%	4.42%	5.15%	4.86%	3.95%
26	2.71%	3.29%	4.72%	5.16%	7.45%	11.93%	5.88%
27	2.56%	3.20%	4.58%	3.83%	3.63%	4.44%	3.70%
28	3.20%	3.67%	3.86%	3.91%	4.54%	4.52%	3.95%
29	3.23%	3.83%	5.41%	4.29%	6.01%	5.29%	4.68%
30	2.80%	3.81%	4.67%	3.75%	3.68%	5.43%	4.02%
Models 16-30	**2.97%**	**3.41%**	**4.45%**	**4.51%**	**5.10%**	**6.61%**	**4.51%**
31	7.51%	9.02%	8.15%	9.98%	11.23%	12.43%	9.72%
32	15.08%	16.28%	14.16%	15.84%	17.47%	19.28%	16.35%
33	12.44%	10.10%	12.08%	11.94%	11.90%	12.52%	11.83%
34	7.33%	8.90%	9.78%	13.71%	14.81%	15.72%	11.71%
35	5.14%	7.64%	9.29%	13.16%	14.33%	15.34%	10.82%
36	10.71%	10.90%	10.07%	10.40%	11.29%	12.76%	11.02%
37	5.62%	7.90%	9.28%	11.83%	13.12%	13.29%	10.18%
38	6.15%	8.01%	9.17%	13.29%	14.82%	15.47%	11.15%
39	7.47%	9.44%	9.15%	12.51%	14.06%	14.97%	11.27%
40	8.87%	10.35%	9.51%	12.81%	14.45%	15.42%	11.90%
41	5.07%	8.41%	9.99%	12.90%	13.83%	13.86%	10.68%

42	8.06%	9.35%	8.33%	10.07%	11.19%	12.41%	9.90%
43	7.26%	9.08%	8.22%	10.13%	11.40%	12.45%	9.76%
44	6.95%	9.04%	9.22%	12.69%	13.98%	14.81%	11.12%
45	5.59%	9.15%	10.77%	13.76%	14.50%	14.16%	11.32%
Models 31-45	**7.95%**	**9.57%**	**9.81%**	**12.33%**	**13.49%**	**14.33%**	**11.25%**
46	7.91%	9.36%	8.38%	9.73%	10.80%	11.97%	9.69%
47	17.66%	18.39%	16.16%	17.76%	19.92%	22.02%	18.65%
48	15.70%	14.27%	15.90%	14.86%	14.37%	14.96%	15.01%
49	13.24%	19.60%	20.88%	23.50%	23.80%	24.13%	20.85%
50	9.84%	14.83%	17.06%	19.87%	19.86%	19.65%	16.85%
51	11.48%	11.13%	10.30%	11.13%	12.19%	13.84%	11.68%
52	12.94%	17.06%	16.87%	18.87%	19.58%	19.57%	17.48%
53	7.13%	12.94%	15.08%	18.96%	19.22%	19.53%	15.48%
54	6.52%	11.75%	11.14%	13.54%	14.92%	16.31%	12.36%
55	8.27%	14.34%	12.54%	14.76%	15.96%	17.33%	13.87%
56	10.99%	14.32%	14.81%	17.42%	17.54%	17.33%	15.40%
57	8.32%	9.56%	8.45%	9.57%	10.56%	11.82%	9.71%
58	7.93%	9.35%	8.34%	9.61%	10.67%	11.88%	9.63%
59	4.88%	11.15%	11.77%	14.64%	16.03%	16.43%	12.48%
60	7.16%	11.51%	13.21%	16.54%	16.87%	16.17%	13.58%
Models 46-60	**10.00%**	**13.30%**	**13.39%**	**15.38%**	**16.15%**	**16.86%**	**14.18%**

Table B.11. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Boston over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	3.90%	4.12%	6.86%	6.24%	6.36%	5.35%	5.47%
2	4.99%	5.96%	8.11%	9.53%	9.22%	9.61%	7.90%
3	5.72%	6.20%	7.32%	9.68%	10.44%	17.41%	9.46%
4	5.02%	6.26%	8.03%	9.62%	9.32%	9.76%	8.00%
5	5.02%	6.33%	8.13%	9.74%	9.37%	9.82%	8.07%
6	4.88%	5.43%	8.21%	6.78%	7.94%	5.53%	6.46%
7	5.71%	5.80%	8.04%	11.85%	8.31%	8.22%	7.99%
8	4.59%	5.98%	8.74%	8.51%	7.89%	8.14%	7.31%
9	4.53%	6.26%	9.28%	9.05%	8.33%	8.27%	7.62%
10	4.49%	6.13%	9.08%	8.86%	8.18%	8.12%	7.48%
11	5.71%	5.84%	8.11%	11.86%	8.33%	8.22%	8.01%
12	3.44%	4.61%	8.85%	6.21%	6.71%	4.26%	5.68%
13	3.46%	4.63%	8.87%	6.19%	6.79%	4.24%	5.69%
14	4.03%	4.78%	6.45%	5.94%	5.78%	4.43%	5.23%
15	3.69%	3.20%	4.76%	4.97%	6.41%	7.38%	5.07%
Models 1-15	**4.61%**	**5.44%**	**7.92%**	**8.34%**	**7.96%**	**7.92%**	**7.03%**
16	3.58%	4.25%	6.16%	4.94%	5.06%	4.77%	4.79%
17	3.77%	4.65%	7.42%	6.71%	5.89%	6.51%	5.83%
18	2.73%	4.78%	6.63%	6.82%	6.12%	5.72%	5.47%
19	3.85%	4.93%	7.40%	6.26%	5.59%	5.63%	5.61%
20	3.84%	4.86%	7.12%	6.16%	5.58%	5.34%	5.48%

21	3.32%	3.39%	8.09%	5.52%	4.18%	6.90%	5.23%
22	3.39%	4.12%	7.89%	7.96%	5.03%	8.79%	6.20%
23	3.16%	4.66%	6.22%	7.28%	6.30%	4.93%	5.43%
24	3.92%	4.43%	6.54%	7.65%	6.82%	5.21%	5.76%
25	3.89%	4.47%	6.59%	7.79%	6.96%	5.33%	5.84%
26	3.36%	4.06%	7.66%	7.75%	4.84%	8.56%	6.04%
27	3.88%	4.82%	7.92%	6.48%	6.61%	4.94%	5.77%
28	3.84%	4.70%	7.75%	6.31%	6.43%	4.81%	5.64%
29	3.47%	4.79%	8.01%	6.74%	6.16%	5.18%	5.73%
30	2.76%	6.37%	8.47%	9.00%	7.99%	8.05%	7.11%
Models 16-30	**3.52%**	**4.62%**	**7.32%**	**6.89%**	**5.97%**	**6.04%**	**5.73%**
31	1.32%	2.91%	4.81%	6.06%	5.89%	5.69%	4.45%
32	10.60%	11.27%	16.17%	16.32%	16.83%	16.26%	14.58%
33	5.35%	5.15%	7.18%	8.88%	9.07%	8.52%	7.36%
34	9.54%	10.91%	15.55%	18.44%	17.25%	16.58%	14.71%
35	7.81%	9.57%	14.04%	16.44%	15.80%	15.56%	13.20%
36	4.28%	4.77%	6.49%	7.69%	7.88%	7.73%	6.47%
37	2.16%	3.37%	5.33%	6.76%	6.58%	6.52%	5.12%
38	7.74%	9.59%	14.13%	16.01%	15.31%	15.74%	13.08%
39	8.53%	9.00%	13.40%	14.29%	13.45%	13.45%	12.02%
40	9.58%	9.73%	14.27%	15.40%	14.10%	13.95%	12.84%
41	1.31%	2.88%	4.91%	6.17%	5.89%	5.83%	4.50%
42	2.15%	3.39%	5.38%	6.74%	6.67%	6.59%	5.16%
43	1.30%	2.99%	5.03%	6.22%	6.01%	5.95%	4.58%
44	7.28%	8.33%	12.71%	13.73%	13.28%	13.28%	11.44%
45	1.08%	2.90%	5.08%	6.46%	6.10%	6.01%	4.60%
Models 31-45	**5.34%**	**6.45%**	**9.63%**	**11.04%**	**10.67%**	**10.51%**	**8.94%**
46	1.59%	3.87%	6.17%	7.39%	7.09%	6.84%	5.49%
47	12.47%	11.62%	16.78%	16.44%	17.20%	16.69%	15.20%
48	6.21%	9.52%	11.57%	12.45%	12.36%	11.07%	10.53%
49	10.06%	11.58%	16.18%	18.01%	16.79%	16.08%	14.78%
50	9.42%	11.05%	15.31%	16.68%	15.62%	15.66%	13.96%
51	3.81%	6.92%	9.70%	10.68%	10.46%	9.66%	8.54%
52	2.47%	4.59%	6.66%	7.78%	7.56%	7.43%	6.08%
53	10.69%	11.85%	16.14%	16.91%	16.11%	16.39%	14.68%
54	9.07%	8.62%	12.80%	13.57%	13.33%	12.96%	11.73%
55	10.18%	9.31%	13.62%	14.56%	13.44%	13.53%	12.44%
56	2.20%	4.44%	6.35%	7.22%	6.93%	6.83%	5.66%
57	2.28%	4.07%	6.51%	7.57%	7.26%	7.10%	5.80%
58	1.64%	3.76%	6.00%	7.02%	6.81%	6.64%	5.31%
59	8.85%	8.13%	12.49%	13.23%	13.08%	13.26%	11.51%
60	2.39%	4.15%	6.03%	7.03%	7.01%	6.73%	5.56%
Models 46-60	**6.22%**	**7.57%**	**10.82%**	**11.77%**	**11.40%**	**11.12%**	**9.82%**

Table B.12. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for New York over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	1.45%	2.02%	4.07%	3.71%	4.00%	3.53%	3.13%
2	4.45%	6.22%	8.36%	9.41%	10.10%	10.16%	8.12%
3	4.57%	5.84%	7.71%	8.85%	9.02%	10.27%	7.71%
4	4.73%	6.08%	8.45%	9.13%	9.89%	10.05%	8.05%
5	3.74%	5.05%	7.60%	8.12%	8.42%	8.97%	6.98%
6	4.18%	5.59%	8.12%	7.35%	8.16%	7.94%	6.89%
7	4.99%	6.14%	8.15%	7.67%	7.76%	8.04%	7.13%
8	2.13%	2.90%	4.84%	4.38%	4.52%	4.76%	3.92%
9	4.66%	5.81%	7.90%	7.35%	8.82%	8.23%	7.13%
10	5.16%	6.43%	9.08%	9.04%	9.62%	9.96%	8.22%
11	4.47%	5.56%	7.47%	6.91%	6.99%	7.18%	6.43%
12	4.31%	5.00%	7.71%	7.28%	7.34%	7.89%	6.59%
13	3.88%	4.51%	7.13%	6.62%	6.67%	7.15%	5.99%
14	1.63%	2.12%	3.26%	3.52%	3.56%	3.07%	2.86%
15	2.44%	2.87%	4.71%	4.66%	4.59%	4.54%	3.97%
Models 1-15	**3.79%**	**4.81%**	**6.97%**	**6.93%**	**7.30%**	**7.45%**	**6.21%**
16	1.51%	2.35%	4.14%	3.94%	4.84%	5.32%	3.68%
17	2.77%	4.43%	6.85%	7.32%	7.19%	8.27%	6.14%
18	3.85%	5.44%	7.86%	8.69%	8.05%	9.17%	7.18%
19	2.92%	4.68%	6.70%	5.90%	5.19%	6.12%	5.25%
20	3.08%	4.53%	5.88%	5.00%	4.26%	4.89%	4.60%
21	4.21%	5.15%	10.34%	8.61%	6.69%	14.10%	8.18%
22	4.02%	5.69%	8.48%	8.33%	7.92%	7.88%	7.05%
23	5.04%	5.79%	5.45%	5.98%	5.73%	6.59%	5.76%
24	1.71%	3.26%	5.74%	6.17%	6.57%	6.66%	5.02%
25	2.29%	3.75%	6.40%	7.09%	7.51%	7.58%	5.77%
26	3.12%	4.91%	6.82%	6.40%	5.81%	5.48%	5.42%
27	4.11%	5.16%	8.47%	8.80%	9.00%	8.85%	7.40%
28	3.41%	4.37%	7.49%	7.81%	7.95%	7.73%	6.46%
29	1.71%	2.48%	3.83%	3.38%	3.05%	2.76%	2.87%
30	4.02%	4.00%	4.55%	3.88%	3.78%	4.11%	4.06%
Models 16-30	**3.18%**	**4.40%**	**6.60%**	**6.49%**	**6.24%**	**7.03%**	**5.66%**
31	2.19%	3.76%	5.14%	5.90%	6.00%	5.42%	4.73%
32	7.32%	8.59%	14.00%	16.09%	18.09%	17.15%	13.54%
33	6.92%	8.35%	9.43%	10.98%	10.77%	9.49%	9.33%
34	6.74%	7.54%	14.25%	15.93%	16.85%	17.04%	13.06%
35	6.96%	7.84%	13.61%	14.78%	15.84%	16.08%	12.52%
36	5.01%	6.81%	7.56%	8.62%	8.69%	8.01%	7.45%
37	2.65%	3.74%	5.20%	6.31%	5.87%	5.44%	4.87%
38	8.24%	8.84%	13.72%	14.51%	15.57%	16.00%	12.82%
39	5.13%	6.28%	11.57%	12.72%	14.13%	14.18%	10.67%
40	5.28%	6.29%	12.22%	13.90%	15.20%	15.17%	11.34%
41	2.49%	3.66%	5.00%	5.83%	5.47%	5.13%	4.60%

42	2.47%	3.73%	5.50%	6.48%	6.24%	6.03%	5.07%
43	2.31%	3.63%	5.30%	5.95%	5.83%	5.69%	4.78%
44	5.12%	6.61%	10.96%	11.80%	13.35%	13.42%	10.21%
45	2.64%	3.77%	5.12%	5.90%	5.84%	5.62%	4.81%
Models 31-45	**4.76%**	**5.96%**	**9.24%**	**10.38%**	**10.92%**	**10.66%**	**8.65%**
46	2.43%	3.92%	5.51%	6.51%	6.83%	6.37%	5.26%
47	16.85%	18.29%	25.94%	27.06%	29.22%	29.19%	24.42%
48	11.92%	13.66%	14.69%	15.25%	14.94%	14.16%	14.10%
49	8.35%	10.58%	17.27%	18.57%	19.61%	19.90%	15.71%
50	9.67%	11.17%	16.67%	17.38%	18.54%	18.92%	15.39%
51	7.88%	9.71%	11.37%	12.57%	12.29%	11.22%	10.84%
52	3.30%	5.03%	6.62%	7.55%	7.17%	6.86%	6.09%
53	11.48%	12.33%	17.58%	17.71%	18.70%	19.26%	16.18%
54	6.80%	7.39%	13.12%	14.46%	15.77%	15.73%	12.21%
55	7.11%	7.40%	13.84%	15.65%	16.85%	16.77%	12.94%
56	3.52%	5.24%	6.54%	7.14%	7.08%	6.42%	5.99%
57	2.17%	3.61%	5.75%	7.17%	7.22%	6.64%	5.43%
58	2.22%	3.62%	5.55%	6.62%	6.73%	6.22%	5.16%
59	7.46%	8.13%	13.14%	13.88%	15.23%	15.56%	12.23%
60	3.85%	5.39%	6.13%	6.17%	6.15%	5.78%	5.58%
Models 46-60	**7.00%**	**8.36%**	**11.98%**	**12.91%**	**13.49%**	**13.27%**	**11.17%**

Table B.13. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Chicago over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-days	Average
1	5.91%	4.90%	3.96%	4.64%	5.10%	5.97%	5.08%
2	5.86%	5.05%	4.38%	4.35%	4.84%	5.47%	4.99%
3	5.73%	5.00%	4.39%	4.35%	5.78%	7.27%	5.42%
4	4.95%	4.71%	4.92%	4.86%	4.98%	6.19%	5.10%
5	4.84%	4.69%	4.93%	4.86%	4.91%	6.08%	5.05%
6	5.11%	4.38%	4.18%	4.48%	5.48%	6.58%	5.04%
7	5.39%	4.39%	4.19%	4.41%	5.44%	6.26%	5.02%
8	4.02%	3.56%	3.80%	5.28%	6.06%	6.08%	4.80%
9	5.12%	3.93%	3.66%	4.64%	5.49%	5.32%	4.69%
10	4.92%	3.33%	3.99%	4.86%	5.88%	6.72%	4.95%
11	5.24%	4.25%	4.06%	4.28%	5.33%	6.16%	4.89%
12	4.64%	3.73%	3.86%	4.20%	4.85%	6.13%	4.57%
13	4.45%	3.59%	3.77%	4.12%	4.79%	6.05%	4.46%
14	3.14%	3.82%	5.02%	4.73%	4.09%	14.63%	5.90%
15	5.52%	4.96%	3.97%	4.38%	4.88%	6.02%	4.96%
Models 1-15	**4.99%**	**4.29%**	**4.21%**	**4.56%**	**5.19%**	**6.73%**	**4.99%**
16	5.62%	6.12%	5.11%	4.60%	5.04%	6.72%	5.53%
17	9.38%	8.58%	6.70%	6.67%	6.36%	5.62%	7.22%
18	3.83%	5.81%	6.78%	7.55%	9.41%	11.69%	7.51%
19	3.66%	4.33%	4.91%	5.08%	6.45%	8.25%	5.45%
20	3.94%	4.47%	5.52%	5.59%	6.73%	8.84%	5.85%

21	7.60%	6.88%	5.70%	5.03%	6.83%	8.55%	6.76%
22	3.58%	5.40%	7.13%	7.41%	9.82%	11.79%	7.52%
23	8.26%	6.93%	6.71%	7.61%	7.40%	9.40%	7.72%
24	5.60%	5.13%	5.39%	5.09%	5.92%	7.63%	5.79%
25	6.34%	5.76%	5.73%	5.35%	6.04%	7.71%	6.16%
26	3.81%	5.32%	7.07%	7.50%	9.91%	11.87%	7.58%
27	5.75%	3.87%	5.01%	5.80%	7.32%	6.79%	5.76%
28	5.02%	3.24%	4.60%	5.55%	7.12%	6.59%	5.35%
29	5.61%	3.73%	4.17%	4.65%	5.10%	5.55%	4.80%
30	5.07%	5.66%	6.84%	7.19%	8.51%	11.36%	7.44%
Models 16-30	**5.54%**	**5.42%**	**5.82%**	**6.04%**	**7.20%**	**8.56%**	**6.43%**
31	11.74%	9.91%	7.91%	8.35%	7.62%	7.00%	8.76%
32	35.65%	34.96%	27.31%	24.16%	20.77%	18.66%	26.92%
33	14.62%	14.02%	11.35%	11.52%	10.06%	8.91%	11.75%
34	18.64%	13.30%	12.52%	12.42%	13.79%	14.34%	14.17%
35	10.46%	8.05%	8.30%	8.76%	10.20%	10.67%	9.41%
36	15.90%	15.07%	12.59%	12.91%	11.31%	9.72%	12.92%
37	15.45%	12.03%	9.04%	9.07%	8.43%	7.70%	10.29%
38	9.33%	7.56%	8.08%	7.98%	9.45%	10.18%	8.76%
39	13.99%	11.12%	9.75%	8.89%	9.63%	9.83%	10.54%
40	21.96%	17.07%	13.94%	12.36%	12.45%	12.35%	15.02%
41	12.11%	9.98%	7.71%	7.95%	7.57%	6.97%	8.72%
42	15.33%	12.16%	9.89%	9.84%	8.55%	7.92%	10.61%
43	12.21%	10.20%	8.56%	8.76%	7.78%	7.21%	9.12%
44	10.58%	8.40%	8.43%	7.87%	8.38%	8.78%	8.74%
45	12.58%	10.15%	8.12%	8.32%	7.48%	7.02%	8.95%
Models 31-45	**15.37%**	**12.93%**	**10.90%**	**10.61%**	**10.23%**	**9.82%**	**11.64%**
46	13.76%	12.13%	10.16%	9.97%	8.79%	7.87%	10.45%
47	43.87%	41.51%	34.78%	29.40%	26.40%	24.85%	33.47%
48	18.23%	16.58%	15.08%	15.24%	14.38%	12.13%	15.27%
49	16.67%	13.31%	14.60%	14.58%	16.82%	18.26%	15.71%
50	11.28%	9.41%	11.46%	11.78%	13.81%	15.02%	12.13%
51	22.93%	21.97%	18.75%	17.01%	15.11%	14.05%	18.31%
52	7.24%	6.74%	6.54%	7.72%	7.86%	7.71%	7.30%
53	10.67%	7.64%	9.98%	10.22%	12.16%	12.92%	10.60%
54	16.42%	12.02%	11.01%	10.62%	11.40%	11.11%	12.10%
55	23.36%	17.23%	14.36%	13.17%	13.60%	13.08%	15.80%
56	5.26%	5.30%	5.92%	7.14%	7.50%	7.52%	6.44%
57	16.86%	14.21%	11.19%	10.64%	9.68%	8.50%	11.85%
58	13.83%	12.19%	9.82%	9.56%	8.84%	7.86%	10.35%
59	13.27%	10.20%	10.36%	9.86%	10.87%	10.88%	10.91%
60	4.57%	4.28%	5.35%	6.64%	7.20%	7.39%	5.90%
Models 46-60	**15.88%**	**13.65%**	**12.62%**	**12.24%**	**12.29%**	**11.95%**	**13.10%**

Table B.14. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Philadelphia over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	4.30%	5.70%	5.12%	5.73%	11.62%	4.26%	6.12%
2	4.14%	5.82%	6.52%	6.42%	6.22%	7.79%	6.15%
3	4.12%	3.65%	5.57%	4.57%	5.50%	8.58%	5.33%
4	3.73%	5.04%	6.10%	6.04%	5.87%	7.67%	5.74%
5	3.55%	4.99%	6.01%	6.12%	5.89%	7.58%	5.69%
6	4.84%	5.62%	5.41%	5.56%	5.34%	6.98%	5.63%
7	4.37%	4.50%	5.17%	5.02%	5.71%	7.39%	5.36%
8	3.06%	3.98%	3.98%	5.28%	4.97%	5.63%	4.48%
9	4.13%	5.08%	5.02%	6.31%	5.56%	5.88%	5.33%
10	4.93%	4.60%	4.61%	5.49%	5.42%	6.29%	5.22%
11	4.15%	4.28%	5.05%	4.93%	5.67%	7.42%	5.25%
12	3.68%	3.61%	4.71%	4.79%	5.22%	7.82%	4.97%
13	3.46%	3.37%	4.60%	4.71%	5.29%	7.90%	4.89%
14	6.99%	4.22%	7.69%	3.86%	6.14%	18.94%	7.97%
15	3.38%	3.18%	4.39%	4.20%	4.81%	8.03%	4.66%
Models 1-15	**4.19%**	**4.51%**	**5.33%**	**5.27%**	**5.95%**	**7.88%**	**5.52%**
16	3.59%	5.06%	5.59%	4.78%	5.32%	7.64%	5.33%
17	6.09%	7.63%	6.28%	6.32%	5.94%	6.53%	6.46%
18	4.51%	3.79%	5.17%	5.54%	6.63%	6.98%	5.44%
19	5.37%	4.07%	5.61%	5.48%	6.64%	8.88%	6.01%
20	5.72%	4.38%	6.00%	5.73%	6.66%	9.69%	6.36%
21	3.51%	4.35%	5.27%	7.38%	7.45%	16.69%	7.44%
22	5.73%	6.30%	9.53%	12.48%	12.22%	21.45%	11.29%
23	6.74%	5.48%	6.22%	6.01%	6.71%	10.23%	6.90%
24	3.41%	3.95%	5.71%	5.26%	6.26%	9.27%	5.64%
25	3.61%	4.12%	5.82%	5.38%	6.29%	9.26%	5.75%
26	5.89%	6.27%	9.37%	11.94%	11.79%	21.14%	11.07%
27	4.45%	4.57%	4.72%	5.61%	6.21%	7.18%	5.46%
28	4.19%	4.21%	4.64%	5.50%	6.23%	7.28%	5.34%
29	4.10%	3.70%	4.47%	5.25%	5.76%	7.19%	5.08%
30	5.59%	6.06%	8.21%	6.60%	9.16%	12.72%	8.06%
Models 16-30	**4.83%**	**4.93%**	**6.17%**	**6.62%**	**7.28%**	**10.81%**	**6.77%**
31	5.60%	6.71%	6.97%	7.14%	7.64%	9.58%	7.27%
32	10.47%	12.46%	12.58%	12.11%	13.36%	15.90%	12.81%
33	7.33%	9.55%	9.39%	9.73%	9.76%	11.21%	9.50%
34	8.56%	6.82%	10.26%	10.73%	13.31%	17.09%	11.13%
35	7.81%	5.87%	8.28%	8.37%	11.02%	14.74%	9.35%
36	9.56%	12.27%	11.38%	11.91%	11.45%	12.24%	11.47%
37	7.81%	8.33%	8.35%	8.19%	8.61%	10.43%	8.62%
38	8.07%	6.19%	7.96%	7.49%	10.28%	14.19%	9.03%
39	6.12%	5.53%	7.38%	7.18%	9.64%	13.13%	8.16%
40	6.42%	6.27%	8.47%	8.72%	10.74%	14.17%	9.13%
41	5.48%	6.67%	6.74%	6.80%	7.53%	9.55%	7.13%

42	7.25%	7.65%	8.17%	8.15%	8.64%	10.06%	8.32%
43	5.30%	6.72%	6.42%	6.97%	7.58%	9.40%	7.06%
44	6.95%	5.36%	7.01%	6.69%	9.06%	12.60%	7.94%
45	5.94%	6.92%	7.24%	7.39%	7.54%	9.69%	7.45%
Models 31-45	**7.24%**	**7.55%**	**8.44%**	**8.50%**	**9.74%**	**12.26%**	**8.96%**
46	6.57%	7.76%	8.02%	8.10%	8.19%	10.15%	8.13%
47	12.84%	14.14%	16.31%	16.40%	17.83%	22.67%	16.70%
48	9.75%	12.12%	13.33%	13.27%	13.52%	14.71%	12.79%
49	15.18%	10.30%	14.50%	14.31%	17.10%	21.80%	15.53%
50	14.27%	9.16%	12.59%	12.29%	15.12%	19.68%	13.85%
51	11.25%	14.34%	12.71%	12.66%	12.03%	12.47%	12.58%
52	8.38%	9.48%	9.62%	9.86%	10.13%	11.52%	9.83%
53	14.12%	9.04%	11.44%	11.02%	13.76%	17.88%	12.88%
54	10.86%	7.06%	9.75%	10.05%	12.52%	16.19%	11.07%
55	11.18%	7.70%	10.83%	11.37%	13.69%	17.29%	12.01%
56	6.93%	8.07%	8.23%	8.51%	9.13%	10.80%	8.61%
57	7.32%	8.66%	8.51%	8.33%	8.82%	10.59%	8.70%
58	6.10%	7.73%	7.62%	7.57%	8.22%	10.11%	7.89%
59	10.94%	7.47%	9.61%	9.31%	11.98%	15.97%	10.88%
60	7.02%	8.50%	8.81%	8.77%	9.49%	11.41%	9.00%
Models 46-60	**10.18%**	**9.44%**	**10.79%**	**10.79%**	**12.10%**	**14.88%**	**11.36%**

Table B.15. Forecasting accuracy (MAPE) of post-COVID trained models 1-60 for Kansas City over a 7- to 42-day time window.

Models	7-day	14-day	21-day	28-day	35-day	42-day	Average
1	5.13%	6.76%	7.75%	10.27%	6.76%	8.43%	7.52%
2	4.59%	6.47%	7.59%	8.22%	7.09%	7.98%	6.99%
3	3.74%	6.57%	8.04%	8.50%	12.63%	6.98%	7.75%
4	4.04%	7.75%	10.71%	11.41%	12.01%	12.44%	9.73%
5	4.12%	7.34%	10.01%	11.61%	12.20%	12.18%	9.58%
6	2.33%	5.33%	4.97%	8.05%	6.98%	8.20%	5.98%
7	5.07%	7.86%	9.56%	13.64%	13.38%	15.69%	10.87%
8	4.50%	6.17%	7.98%	10.71%	11.09%	13.88%	9.06%
9	3.95%	5.34%	6.80%	9.22%	9.30%	11.39%	7.67%
10	3.85%	5.34%	6.95%	9.51%	9.89%	12.21%	7.96%
11	5.66%	7.87%	9.78%	12.81%	12.28%	14.00%	10.40%
12	4.26%	8.57%	6.57%	12.82%	11.30%	12.27%	9.30%
13	4.90%	6.96%	8.43%	10.55%	10.31%	11.18%	8.72%
14	4.79%	8.50%	10.21%	12.67%	12.40%	16.09%	10.78%
15	5.09%	9.25%	9.77%	13.33%	14.32%	17.54%	11.55%
Models 1-15	**4.40%**	**7.07%**	**8.34%**	**10.89%**	**10.80%**	**12.03%**	**8.92%**
16	4.84%	6.34%	6.60%	7.55%	7.38%	7.46%	6.69%
17	5.72%	7.21%	7.41%	8.75%	8.33%	8.04%	7.58%
18	7.39%	9.77%	9.84%	12.44%	12.79%	11.28%	10.58%
19	6.18%	9.24%	9.29%	11.51%	10.96%	11.38%	9.76%
20	6.67%	9.48%	9.54%	11.56%	11.55%	11.92%	10.12%

21	2.94%	5.21%	5.10%	6.95%	8.83%	7.43%	6.08%
22	2.44%	4.65%	5.85%	7.29%	7.93%	8.38%	6.09%
23	3.84%	7.22%	9.34%	9.51%	14.46%	8.19%	8.76%
24	4.29%	7.25%	9.60%	9.55%	9.07%	9.64%	8.23%
25	4.19%	7.17%	9.41%	9.76%	9.00%	9.89%	8.24%
26	4.30%	7.69%	7.77%	9.53%	8.98%	8.56%	7.81%
27	4.01%	4.43%	6.18%	6.78%	7.23%	6.76%	5.90%
28	4.09%	4.51%	6.12%	6.75%	7.34%	6.75%	5.93%
29	5.11%	6.00%	7.17%	9.29%	8.56%	10.65%	7.80%
30	3.66%	5.15%	7.19%	7.39%	7.49%	6.66%	6.26%
Models 16-30	**4.64%**	**6.75%**	**7.76%**	**8.97%**	**9.33%**	**8.87%**	**7.72%**
31	2.52%	3.81%	4.09%	6.92%	6.96%	8.79%	5.52%
32	15.59%	18.36%	15.09%	17.45%	17.27%	18.53%	17.05%
33	6.02%	5.28%	5.40%	7.78%	10.48%	11.80%	7.79%
34	8.92%	8.11%	9.89%	14.24%	16.30%	14.71%	12.03%
35	7.31%	7.24%	9.20%	13.88%	16.05%	14.61%	11.38%
36	3.91%	5.26%	4.41%	6.42%	8.10%	9.94%	6.34%
37	3.93%	4.28%	5.24%	8.44%	9.31%	10.59%	6.96%
38	5.92%	6.46%	8.60%	13.70%	16.69%	14.35%	10.95%
39	4.55%	5.59%	6.66%	11.09%	13.31%	12.42%	8.94%
40	5.94%	6.45%	7.10%	11.90%	13.64%	12.94%	9.66%
41	3.88%	4.46%	5.70%	8.89%	8.97%	9.99%	6.98%
42	2.73%	3.96%	4.19%	7.04%	7.29%	8.88%	5.68%
43	2.47%	3.92%	4.46%	7.13%	7.11%	8.70%	5.63%
44	5.23%	5.90%	7.36%	11.82%	13.52%	13.51%	9.56%
45	3.98%	4.49%	5.73%	8.93%	9.11%	10.12%	7.06%
Models 31-45	**5.53%**	**6.24%**	**6.87%**	**10.37%**	**11.61%**	**11.99%**	**8.77%**
46	3.49%	4.66%	4.16%	5.96%	6.01%	7.89%	5.36%
47	23.39%	28.51%	25.62%	26.76%	27.11%	27.71%	26.52%
48	10.90%	10.44%	9.96%	11.41%	12.89%	13.30%	11.49%
49	19.39%	19.51%	20.87%	23.53%	22.78%	22.42%	21.42%
50	16.43%	16.86%	19.55%	23.45%	22.88%	22.47%	20.27%
51	6.21%	5.89%	5.46%	6.70%	7.47%	8.86%	6.76%
52	10.42%	12.03%	12.55%	14.85%	17.52%	19.16%	14.42%
53	13.60%	12.93%	15.33%	20.17%	20.71%	19.39%	17.02%
54	9.20%	9.69%	9.68%	13.21%	14.35%	14.12%	11.71%
55	10.66%	11.09%	10.61%	13.83%	14.58%	14.62%	12.57%
56	9.73%	11.73%	12.48%	14.84%	16.34%	18.11%	13.87%
57	3.51%	4.50%	3.93%	5.95%	6.43%	8.35%	5.45%
58	3.61%	4.69%	4.16%	6.10%	6.16%	7.99%	5.45%
59	8.16%	9.29%	9.40%	13.30%	14.91%	14.19%	11.54%
60	7.98%	10.35%	11.14%	14.19%	14.68%	15.77%	12.35%
Models 46-60	**10.45%**	**11.48%**	**11.66%**	**14.28%**	**14.99%**	**15.62%**	**13.08%**

Table B.16. Daily difference between actual and estimated usage (kWh) for the best fitting pre-COVID null model over a 7- to 42-day time window

City	Model	7-day	14-day	21-day	28-day	35-day	42-day
Los Angeles	61	–1274 (–1.9%)	–2489 (–3.7%)	–4089 (–6.1%)	–3507 (–5.3%)	–4967 (–7.4%)	–4274 (–6.4%)
Houston	61	9681 (3.6%)	10503 (3.8%)	2683 (1.0%)	5526 (2.1%)	–4377 (–1.6%)	–219 (–0.1%)
Boston	61	88 (0.2%)	–2054 (–3.5%)	–1354 (–2.3%)	–1927 (–3.3%)	–4071 (–6.8%)	–3108 (–5.3%)
New York	61	–2116 (–1.7%)	–5380(–4.4%)	–7196 (–5.9%)	–6401 (–5.4%)	–8859 (–7.4%)	–8698 (–7.3%)
Chicago	61	–5800 (–2.4%)	–10165 (–4.1%)	–15275 (–6.2%)	–12672 (–5.4%)	–18109 (–7.5%)	–19773 (–8.2%)
Philadelphia	61	–198 (–0.2%)	–1062 (–1.1%)	–779 (–0.9%)	280 (0.3%)	–1429 (–1.6%)	–1381 (–1.5%)
Kansas City	61	3725 (10.0%)	1829 (4.9%)	780 (2.1%)	1670 (4.7%)	813 (2.3%)	603 (1.7%)

C. Supplementary Figures and Tables for Chapter 4

Table C.1. The ACORN label description.

Acorn Category	Acorn Group	Acorn Type
1. Affluent Achievers	A. Lavish Lifestyles	1. Exclusive enclaves
		2. Metropolitan money
		3. Large house luxury
	B. Executive Wealth	4. Asset rich families
		5. Wealthy countryside commuters
		6. Financially comfortable families
		7. Affluent professionals
		8. Prosperous suburban families
		9. Well-off edge of towners
	C. Mature Money	10. Better-off villagers
		11. Settled suburbia, older people
		12. Retired and empty nesters
		13. Upmarket downsizers
2. Rising Prosperity	D. City Sophisticates	14. Townhouse cosmopolitans
		15. Younger professionals in smaller flats
		16. Metropolitan professionals
		17. Socializing young renters
	E. Career Climbers	18. Career driven young families
		19. First time buyers in small, modern homes
		20. Mixed metropolitan areas
3. Comfortable Communities	F. Countryside Communities	21. Farms and cottages
		22. Older couples and families in rural areas
		23. Owner occupiers in small towns and villages
	G. Successful Suburbs	24. Comfortably-off families in modern housing
		25. Larger family homes, multi-ethnic areas
		26. Semi-professional families, owner occupied neighborhoods
	H. Steady Neighbourhoods	27. Suburban semis, conventional attitudes
		28. Owner occupied terraces, average income
		29. Established suburbs, older families
	I. Comfortable Seniors	30. Older people, neat and tidy neighborhoods
		31. Elderly singles in purpose-built accommodation
	J. Starting Out	32. Educated families in terraces, young children
		33. Smaller houses and starter homes
4. Financially Stretched	K. Student Life	34. Student flats and halls of residence
		35. Term-times terraces
		36. Educated young people in flats and tenements

	L. Modest Means	37. Low costs flats in suburban areas
		38. Semi-skilled workers in traditional neighborhoods
		39. Fading owner occupied terraces
		40. High occupancy terraces, culturally diverse family areas
	M. Striving Families	41. Labouring semi-rural estates
		42. Struggling young families in post-war terraces
		43. Families in right-to-bury estates
		44. Post-war estates, limited means
	N. Poorer Families	45. Pensioners in social housing, semis and terraces
		46. Elderly people in social rented flats
		47. Low income older people in smaller semis
		48. Pensioners and singles in social rented flats
5. Urban Adversity	O. Young Hardship	49. Young families in low cost private flats
		50. Struggling younger people in mixed tenure
		51. Young people in small, low cost terraces
	P. Struggling Estates	52. Poorer families, many children, terraced housing
		53. Low income terraces
		54. Multi-ethnic, purpose-built estates
		55. Deprived and ethnically diverse in flats
		56. Low income large families in social rented semis
	Q. Difficult Circumstances	57. Social rented flats, families and single parents
		58. Singles and young families, some receiving benefits
		59. Deprived areas and high-rise flats
6. Not Private Households	R. Not Private Households	60. Active communal population
		61. Inactive communal population
		62. Business areas without resident population

Table C.2. Performance of variable reduction technique (VR) rules SV1-SV3 on benchmark datasets.

		Without VR		Rule MV1			Rule MV2			Rule MV3		
Instance	Configuration	Obj	Time(s)	Obj	Time(s)	Gap(%)	Obj	Time(s)	Gap(%)	Obj	Time(s)	Gap(%)
Soybean	1	113.84	0.23	117.94	0.08	3.60	115.49	0.07	1.45	115.49	0.11	1.45
	2	115.69	0.14	117.49	0.04	1.56	116.57	0.07	0.76	116.57	0.10	0.76
Protein	1	1269.33	0.70	1296.32	0.17	2.13	1279.68	0.19	0.82	1270.23	0.32	0.07
	2	1262.61	0.54	1304.07	0.12	3.28	1273.63	0.14	0.87	1271.12	0.28	0.67
Iris	1	98.28	2.54	106.49	0.09	8.35	103.69	0.22	5.50	102.45	0.24	4.24
	2	99.56	2.37	107.82	0.09	8.30	104.80	0.11	5.26	102.63	0.19	3.08
Wine	1	17,493.07	3.62	17,869.32	0.09	2.15	17,618.14	0.17	0.71	17,547.36	0.27	**0.31**
	2	19,077.46	5.66	19,817.38	0.08	3.88	19,204.96	0.13	0.67	19,128.19	0.27	**0.27**
Ionosphere	1	817.34	17.46	838.02	0.13	2.53	848.92	0.21	3.86	827.69	0.35	1.27
	2	838.53	15.96	856.76	0.08	2.17	866.84	0.21	3.38	848.17	0.37	1.15
Control	1	26,934.41	36.76	27,354.17	0.46	1.56	27,206.07	0.93	1.01	27,242.23	2.18	1.14
	2	26,939.45	33.59	27,358.43	0.48	1.56	27,196.99	0.93	0.96	27,240.56	2.07	1.12
Balance	1	1466.41	225.97	1498.45	0.41	2.18	1572.22	0.79	7.22	1499.01	1.24	2.22
	2	1471.80	206.73	1503.44	0.31	2.15	1586.60	0.74	7.80	1502.00	1.14	2.05
Yeast	1	252.22	14,835.61	261.10	2.23	3.52	261.08	10.18	3.51	257.89	46.81	2.25
	2	260.32	2281.82	268.79	2.10	3.25	268.91	10.85	3.30	265.85	36.14	2.12

Table C.3. Performance of variable reduction technique (VR) rules MV1-MV3 on benchmark datasets.

		Without VR		Rule MV1			Rule MV2			Rule MV3		
Instance	Configuration	Obj	Time(s)	Obj	Time(s)	Gap(%)	Obj	Time(s)	Gap(%)	Obj	Time(s)	Gap(%)
Soybean	1	113.84	0.23	113.84	0.10	**0.00**	113.84	0.06	0.00	113.84	0.06	0.00
	2	115.69	0.14	115.69	0.06	**0.00**	115.69	0.05	0.00	115.69	0.06	0.00
Protein	1	1,269.33	0.70	1269.33	0.17	**0.00**	1270.23	0.12	0.07	1269.33	0.12	0.00
	2	1,262.61	0.54	1262.61	0.11	**0.00**	1269.68	0.11	0.56	1269.02	0.12	0.51
Iris	1	98.28	2.54	98.28	0.10	**0.00**	98.28	0.07	0.00	98.28	0.18	0.00
	2	99.56	2.37	99.56	0.10	**0.00**	99.56	0.07	0.00	99.56	0.10	0.00
Wine	1	17,493.07	3.62	17,704.69	0.15	1.21	17,823.49	0.11	1.89	17,772.26	0.12	1.60
	2	19,077.46	5.66	19,237.92	0.15	0.84	19,237.92	0.19	0.84	19,423.55	0.20	1.81
Ionosphere	1	817.34	17.46	817.34	0.22	**0.00**	817.34	0.11	0.00	817.34	0.18	0.00
	2	838.53	15.96	838.53	0.22	**0.00**	838.53	0.10	0.00	838.53	0.23	0.00
Control	1	26,934.41	36.76	27,000.07	0.44	0.24	26,934.41	0.64	0.00	26,934.41	1.01	**0.00**
	2	26,939.45	33.59	27,021.48	0.46	0.30	26,939.45	0.60	0.00	26,939.45	1.03	**0.00**
Balance	1	1466.41	225.97	1474.83	1.19	**0.57**	1476.37	0.52	0.68	1474.83	0.73	0.57
	2	1471.80	206.73	1482.55	0.96	**0.73**	1482.55	0.56	0.73	1487.69	0.89	1.08
Yeast	1	252.22	14,835.61	254.13	10.06	**0.76**	254.61	6.63	0.95	254.47	11.79	0.89
	2	260.32	2,281.82	262.42	8.32	**0.81**	263.87	6.46	1.36	262.91	10.74	0.99

Table C.4. Lagrangian relaxation performance at each iteration for the large UK Power Networks dataset (UKPN-L).

Itr	LB	UB	Best LB	Best UB	Gap (%)	Times(s)
1	-8,834,209.2	427,686.8	-8,834,209.2	427,686.8	2165.58	11.3
2	-130,841.6	1,063,037.4	-130,841.6	427,686.8	130.59	12
3	-266,636.2	412,347.5	-130,841.6	412,347.5	131.73	9.8
4	137,752.3	591,621.4	137,752.3	412,347.5	66.59	11.7
5	-1,145,948.5	412,347.5	137,752.3	412,347.5	66.59	10.7
6	77,900.8	1,008,482.2	137,752.3	412,347.5	66.59	10
7	-3,633,458.4	427,686.8	137,752.3	412,347.5	66.59	10.6
8	6,643.2	887,327.0	137,752.3	412,347.5	66.59	12.6
9	170,128.2	395,062.0	170,128.2	395,062.0	56.94	9.7
10	142,098.8	409,624.0	170,128.2	395,062.0	56.94	10.9
11	223,160.4	466,977.4	223,160.4	395,062.0	43.51	11.1
12	-461,465.1	412,347.5	223,160.4	395,062.0	43.51	10.1
13	190,888.5	387,790.3	223,160.4	387,790.3	42.45	11.5
14	298,257.5	401,520.7	298,257.5	387,790.3	23.09	11
15	299,402.4	385,215.0	299,402.4	385,215.0	22.28	11.5
16	306,973.7	390,555.6	306,973.7	385,215.0	20.31	10.7
17	317,459.1	406,711.0	317,459.1	385,215.0	17.59	10.5
18	295,391.6	427,686.8	317,459.1	385,215.0	17.59	10.7
19	330,248.0	439,143.9	330,248.0	385,215.0	14.27	11.8
20	297,631.6	405,332.1	330,248.0	385,215.0	14.27	11.5
21	336,769.5	383,633.4	336,769.5	383,633.4	12.22	10.9
22	348,000.0	409,256.8	348,000.0	383,633.4	9.29	10.1
23	346,615.4	396,855.5	348,000.0	383,633.4	9.29	11.5
24	353,502.6	408,537.5	353,502.6	383,633.4	7.85	10.9
25	347,759.9	393,181.1	353,502.6	383,633.4	7.85	11.3
26	355,453.9	375,237.4	355,453.9	375,237.4	5.27	10.6
27	357,898.8	381,921.0	357,898.8	375,237.4	4.62	9.7
28	357,449.3	381,193.0	357,898.8	375,237.4	4.62	10.5
29	358,943.7	368,066.8	358,943.7	368,066.8	2.48	10.6
30	359,425.0	384,215.6	359,425.0	368,066.8	2.35	11.5
31	359,809.4	368,978.9	359,809.4	368,066.8	2.24	10.7
32	359,658.1	380,069.8	359,809.4	368,066.8	2.24	10.3
33	360,413.2	367,855.7	360,413.2	367,855.7	2.02	11.4
34	360,499.8	366,984.0	360,499.8	366,984.0	1.77	11
35	360,480.2	386,179.2	360,499.8	366,984.0	1.77	11.3
36	360,754.2	367,982.2	360,754.2	366,984.0	1.70	10.8
37	360,520.7	378,331.1	360,754.2	366,984.0	1.70	10.1
38	360,842.2	365,659.7	360,842.2	365,659.7	1.32	10.5
39	360,924.2	366,651.9	360,924.2	365,659.7	1.30	11.2
40	360,908.2	366,888.3	360,924.2	365,659.7	1.30	10.9
41	360,965.1	368,555.2	360,965.1	365,659.7	1.28	10.7
42	360,937.8	366,450.8	360,965.1	365,659.7	1.28	9.8
43	361,077.5	364,569.5	361,077.5	364,569.5	0.96	10.6
44	361,080.8	369,862.8	361,080.8	364,569.5	0.96	10.2
45	361,122.7	370,139.0	361,122.7	364,569.5	0.95	11
46	361,108.9	367,798.2	361,122.7	364,569.5	0.95	10.4

47	361,138.8	366,760.3	361,138.8	364,569.5	0.94	9.7
48	361,126.1	367,798.2	361,138.8	364,569.5	0.94	10.2
49	361,169.0	364,918.7	361,169.0	364,569.5	0.93	10.5
50	361,171.2	364,464.4	361,171.2	364,464.4	0.90	10.7
51	361,193.2	366,186.2	361,193.2	364,464.4	0.90	10.3
52	361,198.5	366,760.3	361,198.5	364,464.4	0.90	9.7
53	361,205.4	367,975.9	361,205.4	364,464.4	0.89	10.7
54	361,203.1	370,139.0	361,205.4	364,464.4	0.89	10.1
55	361,216.1	363,565.9	361,216.1	363,565.9	0.65	10.7
56	361,218.1	366,360.2	361,218.1	363,565.9	0.65	10.2
57	361,233.3	366,819.0	361,233.3	363,565.9	0.64	9.8
58	361,228.5	369,079.2	361,233.3	363,565.9	0.64	10.1
59	361,246.7	366,106.7	361,246.7	363,565.9	0.64	10.6
60	361,248.6	364,941.8	361,248.6	363,565.9	0.64	10.6
61	361,250.8	367,508.7	361,250.8	363,565.9	0.64	10.3
62	361,255.2	366,885.8	361,255.2	363,565.9	0.64	9.6
63	361,253.6	366,824.3	361,255.2	363,565.9	0.64	10.7
64	361,255.6	366,280.9	361,255.6	363,565.9	0.64	10.1
65	361,261.1	364,777.1	361,261.1	363,565.9	0.63	10.9
66	361,257.9	364,423.7	361,261.1	363,565.9	0.63	10.8
67	361,267.8	364,366.0	361,267.8	363,565.9	0.63	10.3
68	361,268.7	366,454.3	361,268.7	363,565.9	0.63	10.9
69	361,270.3	366,105.2	361,270.3	363,565.9	0.63	11.6
70	361,272.2	367,975.9	361,272.2	363,565.9	0.63	11.2
71	361,271.6	370,139.0	361,272.2	363,565.9	0.63	10.8
72	361,273.4	365,098.2	361,273.4	363,565.9	0.63	10
73	361,273.7	369,079.2	361,273.7	363,565.9	0.63	11
74	361,274.9	367,975.9	361,274.9	363,565.9	0.63	10.3
75	361,274.8	366,234.4	361,274.9	363,565.9	0.63	11.2
76	361,279.2	366,760.3	361,279.2	363,565.9	0.63	10.7
77	361,278.3	367,530.0	361,279.2	363,565.9	0.63	11
78	361,280.0	364,107.0	361,280.0	363,565.9	0.63	10.8
79	361,279.3	368,263.9	361,280.0	363,565.9	0.63	10.4
80	361,280.5	366,275.3	361,280.5	363,565.9	0.63	11.2
81	361,281.6	367,061.4	361,281.6	363,565.9	0.63	9.9
82	361,281.9	363,462.6	361,281.9	363,462.6	0.60	10.9
83	361,282.1	365,331.9	361,282.1	363,462.6	0.60	10.3
84	361,283.0	363,462.6	361,283.0	363,462.6	0.60	11
85	361,283.0	365,235.5	361,283.0	363,462.6	0.60	9.9
86	361,283.9	364,014.9	361,283.9	363,462.6	0.60	10.3
87	361,283.4	367,975.9	361,283.9	363,462.6	0.60	11.4
88	361,284.2	366,792.4	361,284.2	363,462.6	0.60	9.9
89	361,284.5	368,168.1	361,284.5	363,462.6	0.60	11
90	361,285.2	366,237.8	361,285.2	363,462.6	0.60	10.4
91	361,285.1	364,107.0	361,285.2	363,462.6	0.60	10.8
92	361,285.4	364,868.2	361,285.4	363,462.6	0.60	11.7
93	361,285.6	363,781.4	361,285.6	363,462.6	0.60	11.4
94	361,285.7	370,139.0	361,285.7	363,462.6	0.60	10.3
95	361,285.7	366,087.7	361,285.7	363,462.6	0.60	11.1
96	361,286.2	366,238.8	361,286.2	363,462.6	0.60	11.8

97	361,286.2	370,139.0	361,286.2	363,462.6	0.60	10.2
98	361,286.3	367,798.2	361,286.3	363,462.6	0.60	12.2
99	361,286.3	370,139.0	361,286.3	363,462.6	0.60	10
100	361,286.4	366,467.2	361,286.4	363,462.6	0.60	11.2
101	361,286.4	368,524.5	361,286.4	363,462.6	0.60	10.4
102	361,286.5	363,794.3	361,286.5	363,462.6	0.60	10.8
103	361,286.5	365,249.4	361,286.5	363,462.6	0.60	11
104	361,286.7	366,205.7	361,286.7	363,462.6	0.60	9.9
105	361,286.7	366,824.3	361,286.7	363,462.6	0.60	12.6
106	361,286.7	367,814.6	361,286.7	363,462.6	0.60	10.2

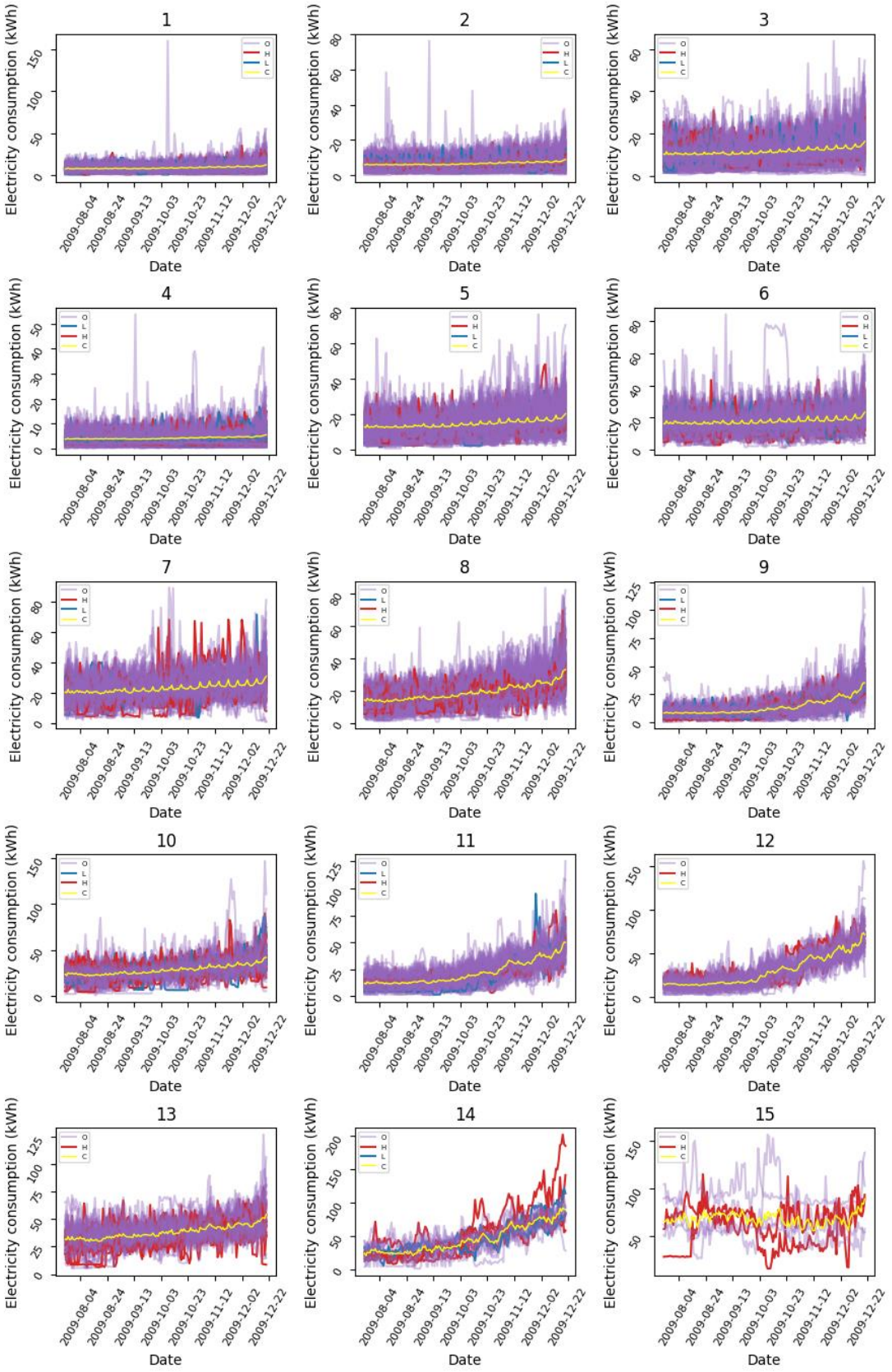


Figure C.1. Daily electricity consumption for k-means clusters 1-15 (24 July 2009 to 20 December 2009). In each subplot, labels H, L, and O refer to high-, low-, and other-income households, respectively. Label C refers to the cluster average.

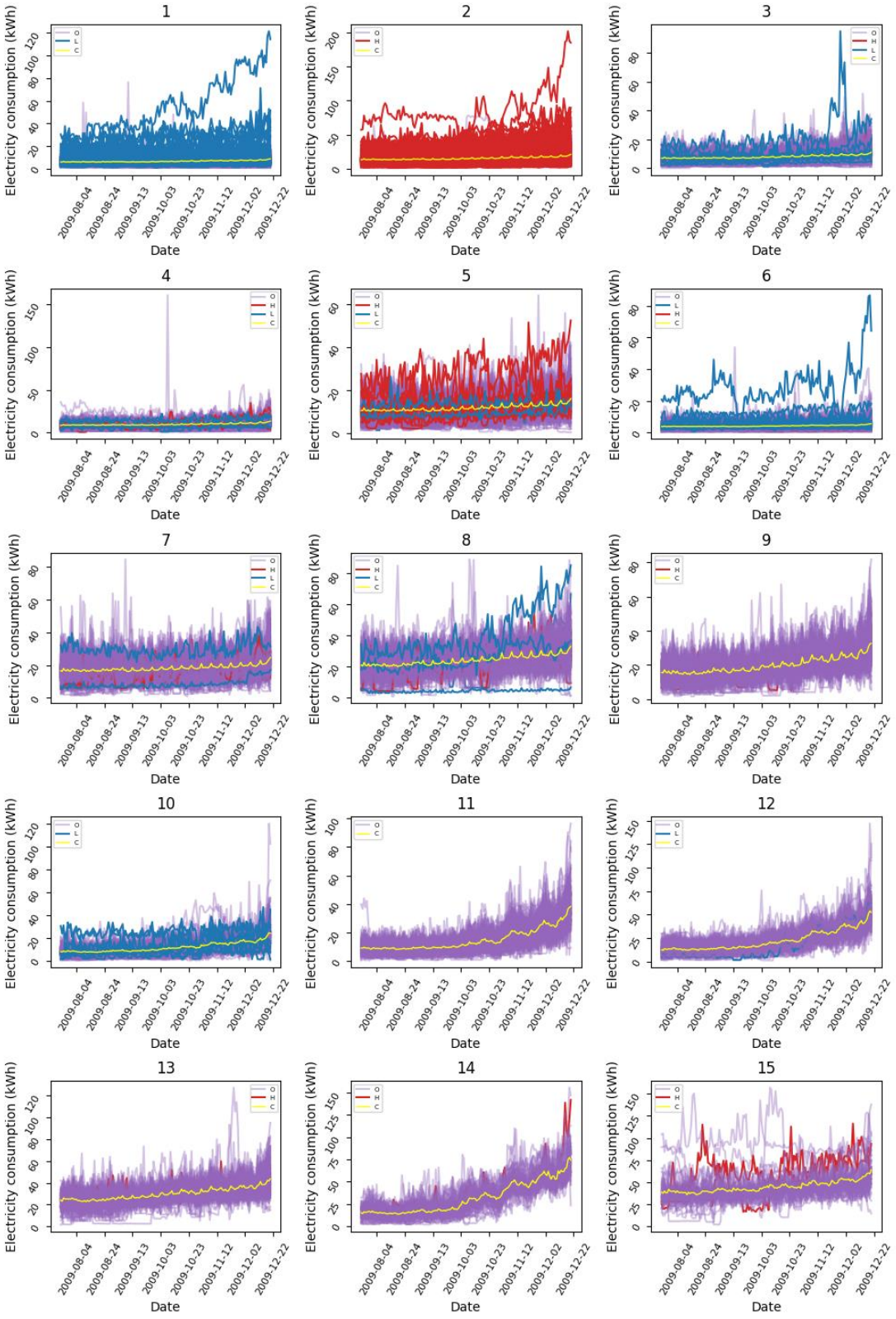


Figure C.2. Daily electricity consumption for constrained p-median clusters 1-15 (24 July 2009 to 20 December 2009). In each subplot, labels H, L, and O refer to high-, low-, and other-income households, respectively. Label C refers to the cluster average.