USING SARIMAX TO FORECAST ELECTRICITY DEMAND AND CONSUMPTION IN UNIVERSITY BUILDINGS

by

Arash Shadkam

B.Sc., Civil Engineering, Sharif University of Technology, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Civil Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2020

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

USING SARIMAX TO FORECAST ELECTRICITY DEMAND AND CONSUMPTION IN UNIVERSITY BUILDINGS

---

submitted by Arash Shadkam in partial fulfillment of the requirements for

the degree of Master of Applied Science

in Civil Engineering

Examining Committee:

Dr. Sheryl Staub-French, Professor, Civil Engineering, University of British Columbia
Supervisor

Dr. Thomas M. Froese, Professor, Civil Engineering, University of Victoria
Supervisory Committee Member

Abstract

Forecasting electricity demand and consumption is critical to the optimal and cost-effective operation of buildings. Time-series forecasting methods identify and learn patterns with data sets and then use these patterns to predict future values. However, the traditional methods tend to fall short in working with seasonal data and external variables. As a result, many time-series forecasting methods are not applicable to electricity consumption data. This type of data is seasonal and highly affected by external factors such as outside air temperature or humidity. Seasonal AutoRegressive Integrated Moving Averages with eXogenous regressors (SARIMAX) is a class of time-series forecasting models that explicitly deals with seasonality in data and external variables.

This research used SARIMAX to predict daily average electricity consumption and daily peak demand in two university buildings. Electricity consumption data from 2017 to 2019 were split into training and test sets. The data from 2017 and 2018 were used as the training set, while values from 2019 were used as the test set. Daily average temperature and humidity were used as external variables. A grid search was conducted to find the best model for each building. Next, the residuals of the models were checked to see whether they satisfied the modelling assumptions. Afterwards, the models were used to predict values in 2019. The performance of the models was calculated using 2019 as test data.

The method was able to successfully use temperature and humidity as external variables and identify weekly patterns. The degree of forecast accuracy was different between the two buildings. The mean absolute percentage error (MAPE) of predicted values in 2019 was 4.1% in the first building and 12.8% in the second building. The models can be used to make informed decisions about the renovation or recommissioning activities in the building, detect abnormalities

in consumption trends, and quantify energy and cost-saving measures. They can also be used to identify and quantify the effects of sudden changes or disruptions in the system or the way occupants behave.

Lay Summary

Predicting electricity demand and consumption in buildings is critical to their optimal and cost-effective operation. This research applied a class of prediction models borrowed from the field of time-series analysis to forecast future electricity demand and consumption in university buildings. The models can be used to detect abnormalities in consumption trends, quantify disruptions in the system or behaviour of occupants, and help energy managers make informed decisions about recommissioning or renovation activities.

Preface

This dissertation is original, unpublished, independent work by the author, A. Shadkam, under the supervision of Dr. Thomas M. Froese from the University of Victoria and Dr. Sheryl Staub-French from the University of British Columbia.

This research did not require the approval of the UBC's Research Ethics Boards.

Table of Contents

List of Tables

List of Figures

List of Symbols

$actual_t$	actual value at time t
B^s	backshift operator such that $B^s y_t = y_{t-s}$
D	order of the seasonal differencing to make data stationary
d	order of differencing to make data stationary
e_t	error at time t
$forecast_t$	forecasted value at time t
n	number of observations in the sample
P	order of the seasonal AR term
p	order of the AR term
$\Theta(B)$	matrix polynomials in backshift operator B of order q
$\theta_1, \theta_2, \dots, \theta_q$	coefficients of the MA order
$\theta(B)$	polynomial in B of order q
Q	order of the seasonal MA term
q	order of the MA term
S	number of periods in a season
$\Phi(B)$	matrix polynomials in backshift operator B of order p
$\phi(B)(1-B)^d$	combined autoregressive operator
$\phi(B)$	polynomial in B of order P
y_t	value of the series at time t
Y_t	vector of values of series
z_t	purely random process, also known as white noise
z_t	error term with zero mean and variance
Z_t	vector of white noise

List of Abbreviations

ACF	AutoCorrelation Function
AIC	Akaike's Information Criterion
ANN	Artificial Neural Network
AR	Autoregressive
ARIMA	AutoRegressive Integrated Moving Average
ARIMAX	AutoRegressive Integrated Moving Average with eXogenous variables
ARMA	Autoregressive Moving Average
ARX	AutoRegressive with eXternal variables
CO_2	Carbon Dioxide
FNN	Fuzzy Neural Network
GFS	Generalized Fourier Series
HVAC	Heating, Ventilation, and Air Conditioning
KBES	Knowledge-Based Expert System
kWh	kilowatt-hour
LEED	Leadership in Energy and Environmental Design
MA	Moving Average
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
NNARX	Nonlinear Neural Network with External variables
PACF	Partial AutoCorrelation Function
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SARIMA	Seasonal AutoRegressive Moving Average
SARIMAX	Seasonal AutoRegressive Integrated Moving Averages with eXogenous regressors
TAR	Threshold AutoRegressive
TF	Transfer Function
UBC	University of British Columbia
VARMA	Vector AutoRegressive Moving Average
W	Watts

Acknowledgements

This thesis is the conclusion of my fulfilling and exciting journey at the University of British Columbia. Many individuals made it possible for me. I am not able to thank them enough.

First, I would like to thank Professor Thomas M. Froese. The interdisciplinary approach of this thesis was made possible because of his continuous support and encouragement to explore new and innovative ideas. His feedback and guidance gave my work its purpose and meaning. His openness to new ideas gave me the opportunity and courage to build expertise in areas I had limited knowledge in.

I would also like to thank Professor Staub-French for her tremendous support and feedback throughout my master's degree and research.

I am thankful to many other individuals who helped me during my master's degree. Arthur De Robert and Zach Danyluk from UBC energy and water services, helped me a lot in understanding UBC's energy systems and its digital infrastructure. Also, my colleague, Ariel Li, helped me navigate through the research process with her feedback.

I would like to sincerely thank my friends, Zahra, Ramtin, Panahi, GP, and Roozbeh, who have always supported me and brought joy into my life, especially throughout my master's degree.

Finally, I would like to give my most heartfelt thanks to my parents, Haleh and Nader. I would not have been able to be where I am and achieve what I have achieved without their unconditional love and support throughout my whole life, and especially during the challenging times of completing my master's degree. Their moral and financial support made all this possible for me. I would also like to thank my brother, Ashkan, for his support during this time.

To Haleh and Nader

Chapter 1: Introduction

1.1 Problem Background

Electricity load forecasting has become critical to the operation of power plants as a result of deregulation in the electricity generation industry and volatile energy prices [1]. Power plants use forecasts of varying time horizons and accuracy to ensure secure and optimal operation of the plant while planning for possible expansions of the facility to meet future demand [2]. Load forecasting is also important on a building level. According to the US Energy Information Administration, 40% of the total energy consumption comes from residential and commercial buildings [3]. Therefore, improving the energy efficiency of buildings is critical to lowering emissions. Analyzing the power consumption data in a building can lead to identifying energy and cost-saving opportunities such as demand response (DR) actions [4]. DR actions include basic tasks such as turning off electrical equipment to sophisticated computerized measures such as scheduling heating and cooling equipment in buildings [5]. As a result, forecasting energy demand at the building level has gained more interest in recent years [6]. Demand load forecasting enables the energy manager of a building to forecast future load shapes. Load shapes represent the load as a function of time and are used to identify values at each point in time [7]. It also enhances their ability to make informed decisions regarding the recommissioning or renovation activities of the building [1]. Moreover, load forecasting has the potential to be used to detect abnormalities, identify and quantify the impact of an intervention or change in a system, and quantify the potential associated energy or cost savings [8].

Accurately forecasting the short-term demand load of a building or power plant is a challenging task. Factors such as weather, season, day of the week, occupant behaviour, and social activities all influence the demand load [9]. In long term forecasting scenarios, economic factors

also play a role; for instance, growth in electricity demand in an era of a booming economy might be quite different during an economic downturn [9].

Load forecasting, on either the building or the power system level, is generally categorized based on time horizon into three groups: short term forecasting (from a few hours up to a week), medium-term (from a few weeks up to a year), and long term forecasting (from a year up to 20 years) [2]. Short term forecasts are mainly used for day to day operations of the power system or building. In contrast, the medium and long term forecasts are used for maintenance and major planning or renovations [2].

Demand forecasting techniques can alternatively be categorized into two major groups: a) classical methods and b) computational intelligence-based methods [2]. Classical time-series forecasting techniques model the demand as a function of historical data and predict future values [2]. These methods work under the assumption that data is linear and stationary [2]. Stationary data is defined as data for which the mean, variance and autocorrelation structures do not change as a function of time [10]. In other words, the statistical properties of a stationary series (e.g., mean and variance) do not depend on the time it was observed. However, electricity demand data is seasonal and non-stationary, which renders many classical techniques inadequate. Figure 1-1 shows the seasonality in the electricity demand from a group of buildings on the University of British Columbia's (UBC) campus.

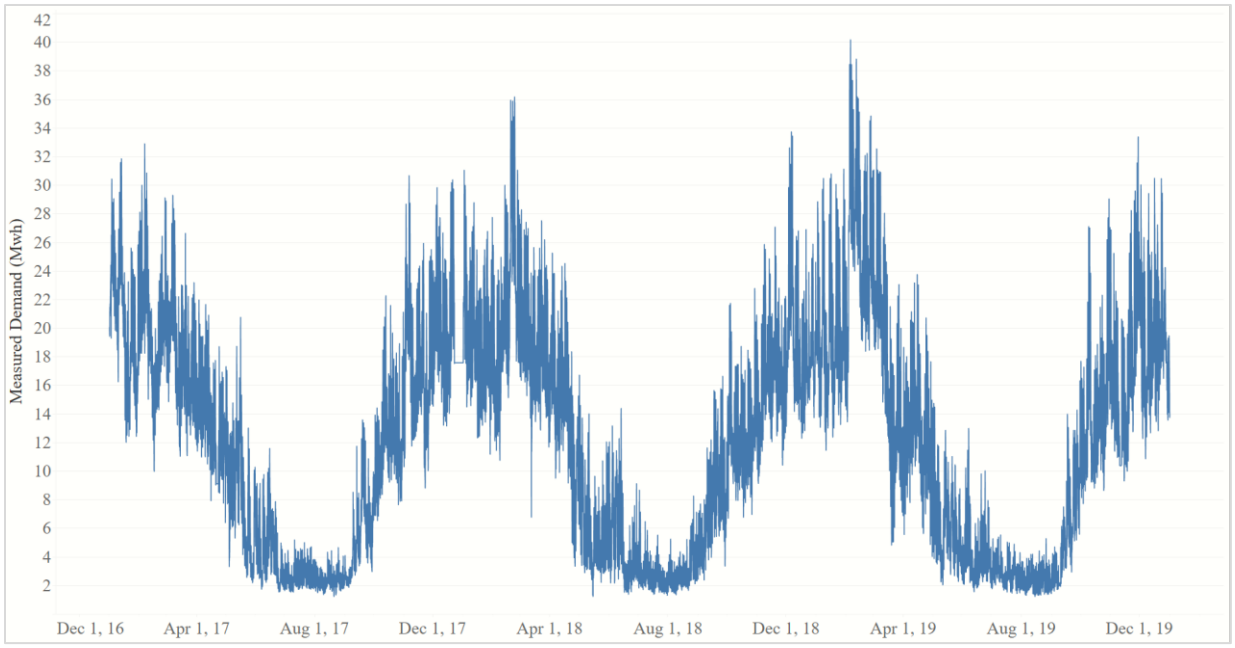

Figure 1-1 Seasonality of electricity load demand

An improvement to the classical methods is to combine the prediction models with external variables. In [11], Citroen et al. combined a classical method known as ARIMA with real GDP and demography as external variables to predict long term annual electricity demand in Morocco. In another study, Felice et al. predicted daily electricity demand in Italy using ARIMA with temperature as an external variable [12]. However, the mentioned methods do not specifically address the issue of seasonality in data.

The Seasonal AutoRegressive Integrated Moving Averages with eXogenous regressors (SARIMAX) method improves the traditional time-series methods by explicitly dealing with seasonality in the data and also accounting for external variables [13]. These features of SARIMAX make it an ideal class of models to be applied to electricity consumption and demand data in buildings. In [14], Papaioannou et al. applied the SARIMAX technique to predict the electricity demand in Greece using the day of the week, holidays, and temperature as external variables. In another research, Tas et al. used the SARIMAX method and showed that ambient

temperature and cloud cover highly influence the daily residential energy consumption in Turkey [15]. Both studies focused on using SARIMAX on energy consumption data on a national or regional scale.

This thesis focuses on applying the SARIMAX technique (SARIMA with external variables) to model and predict electricity consumption and peak demand on the building scale.

1.2 Research Objectives

The objective of this research is to predict the daily average electricity consumption and daily peak electricity demand in university buildings using the SARIMAX method and evaluate the performance of the models. This objective was broken down into the following sub-objectives:

1. Collect and preprocess the data.
2. Train forecasting models for two university buildings.
3. Evaluate the performance of the models.

1.3 Research Activities

The research activities are as follows.

1. Collect and preprocess data.
    a. Data collection
        - Collect electricity consumption data from two buildings on the UBC campus.
        - Collect weather data, i.e. daily mean outside air temperature and humidity.
    b. Data preprocessing
        - Split the data into training and test sets.
        - Identify and replace outliers.

2. Train SARIMAX models for two buildings using training data.
    a. Identify a suitable class of models.
    b. Estimate each model's coefficients and run diagnostic tests on the residuals.
    c. Forecast values.
3. Evaluate the performance of the models using test data.
    a. Calculate the performance metrics of each model.

1.4 Research Scope

This research focused on the application of the SARIMAX method to electricity consumption data and did not intend to compare different forecasting methods.

1.5 Research Methodology

This research adopted a Design Science method. This methodology was introduced by Richard Buckminster Fuller in 1963 [16].

There are six steps in the design science research methodology (DSRM): problem identification, objective definition for a solution, design and development, demonstration, evaluation, and communication [17]. This research follows the scheme mentioned below and depicted in Figure 1-2

1. Forecasting electricity demand and consumption using traditional time-series methods was identified as the problem.
2. Applying and evaluating the SARIMAX method was identified as the objective.
3. A framework was developed using existing frameworks.
4. The concept is demonstrated by implementing the forecasting models in this dissertation.

5. The models were evaluated using the model performance metrics, including mean absolute percentage error (MAPE), mean square error (MSE), and root mean square error (RMSE).
6. Communication of the results is done through this dissertation, other presentations and future publications.

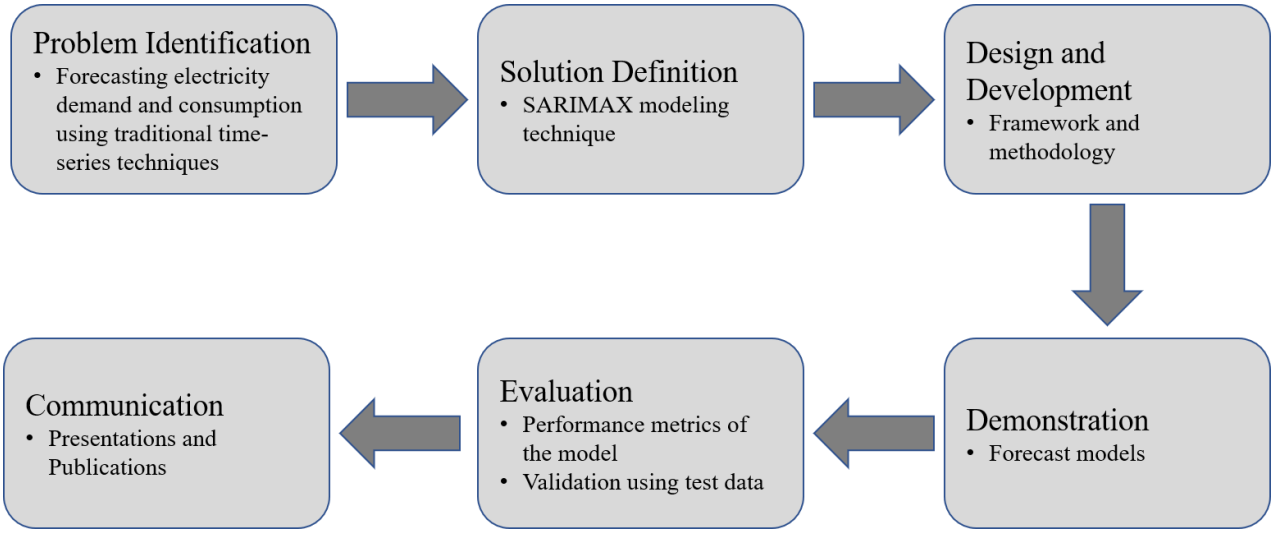

Figure 1-2 DSRM and research steps

1.6 Overview of the Dissertation – Reader's Guide

The rest of the thesis is structured as follows. Chapter 2 provides an overview of forecasting energy demand and consumption, the modelling techniques in this field, and the fundamental concepts in time-series modelling. Chapter 3 introduces the concepts in SARIMAX modelling. Chapter 4 focuses on the methodology of the thesis, collecting and preprocessing data, training SARIMAX models, and evaluating the performance of the models. Chapter 5 presents the results of the models. Chapter 6 provides a summary of the methodology and the results of the models, discusses the results, the contributions and limitations of the thesis, and discusses future work.

Chapter 2: Points of Departure

2.1 Overview

This chapter provides an overview of concepts regarding energy demand and consumption, modelling techniques used in this field and fundamental concepts in time-series modelling. Section 2.2 covers energy demand forecasting concepts, while section 2.3 presents fundamental concepts in time-series modelling and forecasting relevant to the scope of this thesis.

2.2 Energy Demand Forecasting

2.2.1 Energy Demand and Consumption

Energy is the capacity to do work while power is the rate at which work is done. If a person goes from point A to point B, the distance travelled would be analogous to energy while the speed at which they walked would be analogous to power. Power in the context of a building is usually measured in watts (W) or kilowatts, while energy is reported using watthour or kilowatt-hour (kWh). Energy demand is generally broken down into a) electricity demand and b) heating/cooling demand.

The electricity demand of a building is the amount of electricity needed to operate the electrical equipment in a building. Factors affecting the electricity demand load include the ventilation system, efficiency of the electrical equipment, and occupant behaviour.

In buildings, heating and cooling loads are the units of thermal energy needed to be added to or removed from a space by the heating, ventilation, and air conditioning system (HVAC) system to reach a defined level of comfort [18]. Factors such as location, orientation, time of the year, and indoor design conditions of a building influence the heating and cooling loads [18]. Both load data reside in the building management system and are commonly reported in kilowatts or megawatts.

2.2.2 Demand Load Forecasting Methods

Heating, ventilation, and air conditioning (HVAC) systems make up almost half of the energy consumption in buildings [19]. This fact, combined with the vast amount of energy data being generated in buildings, has resulted in an increased uptake of different data analytics and modelling techniques to predict energy demand load in buildings.

In [20], Zhao et al. grouped the energy demand forecasting methods into a) physics-based and b) data-driven methods. Another breakdown was given in [21] by Hahn et al., who grouped the modelling approaches into five categories, namely regression-based models, time-series models, neural networks, support vector machines, and hybrid approaches.

Moghram and Rahman in [22] applied five methods to predict peak daily loads of a US utility in summer and winter peak days. They showed that the transfer function (TF) method had the lowest percentage error for summer peak days, while for the wintertime, the knowledge-based expert system (KBES) worked well. Srinivasan et al. used fuzzy neural networks (FNN) in [23] to predict the day-ahead load in a power plant. FNN is a hybrid approach combining fuzzy knowledge-based models and neural networks [23]. Rahman and Bhatnagar utilized an expert system approach in [24] to predict six-hour and 24-hour ahead loads. The expert system approach used the knowledge and reasoning of human experts to build logical relationships between factors such as weather and load patterns. This approach revealed competitive results compared to other methods [24]. Amjady in [25] combined the autoregressive integrated moving average (ARIMA) modelling technique with knowledge of human experts to predict the daily peak load of a power plant in Iran. Amjady used temperature and day of the week as predictors. Amjady noted that the success of this method is highly dependent on the quality of the input provided by the expert human. Despite this possible risk, the results of this hybrid method outperformed the traditional

methods such as artificial neural networks (ANN) and ARIMA [25]. Rahman et al. in [26] used deep recurrent neural network (RNN) to predict electricity consumption in commercial and residential buildings in the medium to long term period. RNN method uses feedback loops to model temporal characteristics of time series. This method is prone to error when there is uncertainty regarding future values of temperature. However, the proposed RNN performs well in medium-term load forecasting of up to a year in commercial buildings in the US [26].

2.2.3 Demand Load Forecasting Using Time-series Methods

Time-series forecasting models and methods are commonly used in predicting load demand [1]. Easy interpretability and scalability of these methods and the chronological nature of demand load data are among the reasons that these techniques have been widely adopted in the industry. Kimbara et al. demonstrated the application of a time-series forecasting method via creating an online load profiling prediction system [27]. The time-series techniques come with limitations. The classical time series models can only be applied to stationary and linear data. However, in many real-world cases, data is non-stationary, nonlinear, and affected by external variables.

Electricity demand is seasonal and highly influenced by temporal and meteorological factors [28]. Temporal factors include the day of the week and holidays, while meteorological factors include temperature and humidity. The mentioned external factors and the seasonality in data are generally not captured in the traditional time-series methods. However, the SARIMAX method automatically works with seasonal data and external variables. As a result, the SARIMAX method is a suitable candidate for predicting electricity demand and consumption. Although the application of time-series forecasting methods is widely discussed in the literature, the application of the SARIMAX method to electricity demand and consumption is limited [29]. Papaioannou et al. [14] used the SARIMAX method with the day of the week and temperature as external variables

to predict the daily electricity demand in Greece on the national level. The SARIMAX method was noticeably successful in forecasting sudden increases in demand [14]. In another study, Felice et al. [12] used a non-seasonal time-series method to predict electricity demand at the national and regional level in Italy. It was demonstrated that using temperature as an external variable improved the prediction results [12].

An approach in time series modelling that works well with nonlinear data is the frequency-based modelling approach. This class of methods uses regression on sinusoids to create a model [30]. Reddy in [31] modified a generalized Fourier series (GFS) using outside air temperature to predict heating and cooling energy use in commercial buildings in the US. A key factor leading to the success of this method was using temperature frequency terms to deal with nonlinearity in the data caused by temperature.

Some techniques, such as ARIMA, work with non-stationary data [26]. Newsham and Birt [32] used occupancy data as an exogenous variable in an ARIMA model to improve the prediction of electricity use in an office building in Ontario, Canada. They used a collection of motion, wireless, carbon dioxide, and illuminance sensors to capture occupancy data. The results showed little improvement in the prediction results of the ARIMA model. One of the key reasons behind this outcome was noted to be the significant presence of process loads from the labs in the building, which are, by nature, not directly influenced by the number of occupants [32].

Lowry et al. [8] highlighted the importance of considering seasonality in forecasting by modelling water, gas, and electricity consumption of different types of buildings and comparing the results to traditional linear regression models. Lowry et al. [8] used monthly water and gas consumption data and hourly electricity consumption as inputs to the model.

Rios-Moreno et al. [8] used outside air temperature, relative humidity, air velocity, and global solar radiation flux as external variables to an autoregressive (AR) and an autoregressive moving average (ARMA) model. They successfully predicted the room temperature in a university classroom in Mexico. The results showed that the external variable older than 20 minutes did not improve the performance of the model [8].

Mustafaraj et al. [8] used a nonlinear neural network model with external variables (NNARX) and an autoregressive model with external variables (ARX) to predict room temperature and relative humidity in an open-plan office. The performance declined as the predictions went further into the future. However, using CO_2 concentration as an input to the model improved its performance. The NNARX model outperformed the ARX because of the nonlinear nature of room temperature and humidity [8].

Section 2.2 provided an overview of the time-series methods used in the field of energy forecasting. In the case of electricity consumption, the traditional methods are generally not suitable candidates for load forecasting. Electricity consumption is seasonal and highly influenced by many factors, including weather conditions. These characteristics of electricity consumption are generally not captured in the traditional time-series forecasting methods. The SARIMAX method was chosen in this thesis because of its ability to work with seasonal data and external variables. The thesis used the SARIMAX method to predict electricity demand and consumption in university buildings and evaluated the performance of the models.

2.3 Time-series Forecasting

2.3.1 Overview of Time-series Forecasting

A time series is a collection of observations recorded in a sequence through time [33]. Time series are generally categorized based on the frequency of measurement into continuous and discrete series [33]. This means that measurements are made either continuously through time or at discrete points in time. In many cases, a continuous time-series is transformed to a discrete series by either sampling from the series at regular intervals or aggregating the series over a period [33]. Electricity load demand data would be an example of a continuous time-series turned into a discrete series through sampling.

To be able to predict future values of a time series correctly, one must first understand the different types of forecasting methods. Forecasting methods are procedures that use present and past values to predict future values [33]. Based on this definition and whether additional variables are taken into account, forecasting methods are generally categorized into three groups [33]:

1) Judgmental forecasts: This category is based on subjective criteria such as judgment or intuition.
2) Univariate forecasts: This category uses only present and past values of the series to predict future values.
3) Multivariate forecasts: This category uses at least one additional variable to forecast future values. These additional variables are known as predictors or explanatory variables.

2.3.2 Time-series Models

Forecasting methods depend on underlying forecasting models. Forecasting models are generally categorized into two groups: a) univariate models, b) multivariate models [33]. Time

series models can also be divided into two general approaches, namely a) time-domain and b) frequency-domain [34]. The time-domain approach uses regression on past values, while the frequency-domain approach uses regression on sinusoids [30]. The rest of this chapter focuses on the former breakdown, which is used throughout the thesis.

2.3.2.1 Univariate Models

A univariate model describes a single variable based on its relationship with its past values and white noise [33]. The following sections discuss some of the widely adopted models.

2.3.2.1.1 Random Walk Model

A random walk model expresses the value of a time series at time t, y_t, to be the sum of the value of the variable at time t-1 and a purely random process, also known as white noise.

The mathematical representation of a random walk model is shown below [33]:

$$y_t = y_{t-1} + z_t \tag{2.1}$$

where:

- y_t denotes the value of the series at time t
- z_t denotes a purely random process, also known as white noise

The random walk model has applications in the finance industry, where the share price on a given day is equal to its value the day before plus or minus fluctuations in the value of the share price. Fluctuations, i.e. the difference in value from one day to another, of share prices have similar characteristics with a purely random process and are not predictable [33].

2.3.2.1.2 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) is a generalized form of an AutoRegressive Moving Average (ARMA) model and is one of the most widely used classes of time-series models. This class is widely associated with the Box-Jenkins method, named after

George Box and Gwilym Jenkins, who created the method [33]. ARIMA processes can be generally divided into two distinct processes, namely autoregressive (AR) processes, and moving average (MA) processes.

A purely autoregressive (AR) process deals with using only the past values of the time series to predict future values, and a purely moving average (MA) process uses the current and past values of a random process, also known as white noise, to predict future values. These two processes combined form the ARMA process. ARMA processes can only be applied to stationary data. However, in most real-world cases, the data is non-stationary. A method known as differencing is employed to deal with non-stationary data. This method replaces a value with its difference with previous values. The integrated (I) notation in ARIMA deals with this procedure.

Nonseasonal ARIMA models are written as ARIMA (*p, d, q*) where [33]:

- *p* is the order of the AR term.
- *d* is the order of differencing needed to make the data stationary.
- *q* is the order of the MA term.
- Here, the "order" is the number of previous values in the time series that are used in determining each term.

The mathematical representation of an ARIMA model and its building blocks are explained below.

An autoregressive process of order p, AR(*p*), is mathematically represented as a weighted linear sum of the past p values plus white noise [33]:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + z_t \qquad (2.2)$$

where:

- $\phi_1, \phi_2, \ldots, \phi_p$ denote the coefficients of the AR order.
- z_t denotes the error term with zero mean and variance.

Using the back-shift operator B , $By_t = y_{t-1}$, the AR(p) can be represented as [33]:

$$\phi(B)y_t = Z_t \tag{2.3}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2.4}$$

where:

- $\phi(B)$ is a polynomial in B of order P.

An MA(q) process is a weighted linear sum of the last q white noise error terms [33]:

$$y_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} + \cdots + \theta_q z_{t-q} \tag{2.5}$$

where:

- $\theta_1, \theta_2, \ldots, \theta_q$ denote the coefficients of the MA order.
- z_t denotes the white noise terms with zero mean and constant variance.

Using the back-shift operator B, the MA(q) can be written as [33]:

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q \tag{2.6}$$

where:

- $\theta(B)$ is a polynomial in B of order q.

An ARIMA (p, d, q) is a mixed autoregressive moving average model of p autoregressive terms and q moving average terms which are differenced d times [33]:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} + \cdots + \theta_q z_{t-q} \quad (2.7)$$

In other words:

Forecasted y_t = a constant + linear combination of lagged values + linear combination of current and past error values

Using the back-shift operator B, the ARIMA(*p, d, q*) can be written as [33]:

$$\phi(B)(1-B)^d y_t = \theta(B) z_t \quad (2.8)$$

where:

- $\phi(B)(1-B)^d$ is the combined autoregressive operator.

2.3.2.1.3 State Space Models

State-space models are a generalization of time-series regression models in which explanatory variables and parameters change over time [35]. The main feature of this class of time-series forecasting is that parameters can be linked together through time using dynamic linear regression, which in return can be used to create a correlation between observations [35].

State-space models are based on two components. The first component is called state variables and the other one state vectors. Each state vector is a linear combination of state variables [33]. In state-space models, different state variables such as trend or seasonality of the model can be modelled and studied separately. Kalman [36] first utilized the idea of analyzing time series with a state space approach and created what is known as the Kalman filter. Using the state-space modelling approach introduces flexibility in working with missing data and enables the user to fully decompose and scrutinize different components of the model [37]. One drawback of this method is the complicated process involved in creating the models [35].

An example of the state-space approach is to model a climatic signal such as global warming using two different datasets. Datasets can include global mean land-ocean temperature index and surface air temperature index, which theoretically are measuring the same underlying signal [37].

2.3.2.1.4 Non-linear Models

Although linear time-series models and methods work well in most cases, there are cases that the time series have characteristics that cannot be explained by a linear model. Examples of this type of time series can be found in economics. Economic series show different behaviours when going into recession versus coming out of one [33]. There are features such as time-changing variance and asymmetric cycle/structure that can help with identifying nonlinear time series [33]. However, defining a nonlinear time series model is not a straightforward task. Generally, a linear model is defined as a model that its values can be written as a linear sum of the present and past values of itself and a purely random process [33]. A model that does not fit this description is regarded as a nonlinear model. However, defining a nonlinear forecasting method is more straightforward. In a linear forecasting method, predictions at time t can be explained as a linear function of values up to time t [33].

Another class of models that works with nonlinear data is the threshold models [38]. The idea in this class of models is to fit local linear ARMA models based on a threshold criterion [37]. These models allow changes in the ARMA coefficients. In this class of models, each ARMA model is referred to as a regime. An example of a threshold autoregressive (TAR) approach in practice is modelling deaths caused by influenza [37]. This series is nonlinear, i.e. increases faster than it decreases, and have different characteristics based on the season. Thus, making it a good fit for varying coefficients for different regimes and hence the TAR approach.

2.3.2.2 Multivariate Models

Multivariate models are applied to multivariate datasets to explain the interrelationships between the time series [33]. For instance, outside air temperature and the building's occupancy rate can be used to build a multivariate model of the building's energy consumption. A more complicated example can be seen in economics, where an increase in prices leads to an increase in wages, which will lead to an increase in prices again [33]. This phenomenon that the outputs affect the inputs is present in closed-loop systems [33].

Data generated by a closed-loop system cannot be modelled using a single-equation approach. Modelling such interrelated variables uses a technique known as multiple time-series modelling [33]. Vector ARMA (VARMA) models are an example of a generalized class of the univariate autoregressive-moving average (ARMA) models. In VARMA models, each variable is a linear sum of the past and present values of all the variables and all the white noise terms. They are written as [33]:

$$\Phi(B)Y_t = \Theta(B)Z_t \tag{2.9}$$

where:

- Y_t and Z_t are vectors.
- $\Phi(B)$ and $\Theta(B)$ are matrix polynomials in backshift operator B of order p and q.

2.3.3 Forecasting Accuracy Measures

Accuracy of forecasting methods can be measured using in-sample errors and out-of-sample errors. In-sample mode refers to the training stage, while out-of-sample refers to the test stage. The in-sample error is a measure of how well the model fits the data. However, the out-of-sample error is the preferred option to measure and compare the strength of a forecasting method to predict future values [33].

MAPE is the most frequently used measure to evaluate the accuracy of a model [21]. Some measures, such as MSE and RMSE are dependent on the scale and unit of the data, while others such as MAPE are scale-independent, which makes them better choices for comparing models from different scales [39].

MSE calculates the mean of the errors squared. By taking the square of the errors, MSE disregards the direction of errors. It is calculated as [33]:

$$MSE = \frac{1}{n}\sum_{1}^{n} {e_t}^2 \tag{2.10}$$

where:

- n is the number of observations in the sample.
- e_t is the error at time t, such that $e_t = actual_t - forecast_t$.

RMSE takes the root of the MSE. Thus, it has the same unit of measurement as the data. It is calculated as [33]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n} {e_t}^2} \tag{2.11}$$

MAPE reports the average of the absolute errors as a percentage of the actual values. Thus, it is not dependent on the scale of the data. It is calculated as [33]:

$$MAPE = \frac{1}{n}\sum_{1}^{n}\left|\frac{e_t}{actual_t}\right| \times 100 \quad (2.12)$$

where:

- e_t is the error at time t, such that $e_t = actual_t - forecast_t$.
- $actual_t$ is the actual value at time t.

Section 2.3.3 covered the fundamental concepts in time-series modelling that are used throughout the thesis. It also introduced the forecast accuracy measures and the mathematical operations behind them, which are used in evaluating the performance of the models in the thesis.

Chapter 3: Seasonal Autoregressive Integrated Moving Averages with eXogenous Regressors Modelling

3.1 Overview

This chapter introduces the concepts in SARIMAX modelling that were used in this thesis. Section 3.2 introduces some of the concepts and assumptions of SARIMAX modelling. Section 3.3 covers the general form of a SARIMAX model. Section 3.4 covers a highly adopted forecasting methodology in time-series forecasting.

3.2 Introduction to SARIMAX

Autoregressive and moving average models work with stationary and linear data. However, in many cases, the data is non-stationary. Autoregressive Integrated Moving Average (ARIMA) models are used to deal with non-stationary data. An ARIMA model consists of three parts, namely autoregressive (AR) terms, moving average (MA) terms and differencing operations (I). The differencing operation is used to create a stationary series for modelling [33]. In this operation, a value is replaced with the difference of the value and its previous value [33].

A generalized form of the ARIMA model known as the Seasonal ARIMA (SARIMA) is used to handle seasonality in data. This class of ARIMA models deals explicitly with seasonality in data by using seasonal AR, MA, and differencing terms in the model.

External variables can also be added to the model through an exogenous regressor term. Seasonal ARIMA with exogenous regressors (SARIMAX) enables the user to add the effects of external variables to the model. Exogenous variables are defined as variables that influence a model but are not influenced by it. The weather is considered an exogenous variable in the context of an energy consumption model of a building.

3.3 General Form of SARIMAX

A SARIMAX model is written as SARIMAX (p, d, q) $(P, D, Q)_s$ where:

- p is the order of the AR term.
- d is the order of differencing needed to make the data stationary.
- q is the order of the MA term.
- P is the order of the seasonal AR term.
- D is the order of the seasonal differencing needed to make data stationary.
- Q is the order of the seasonal MA term.
- S is the number of periods in a season.

A SARIMAX (p, d, q) $(P, D, Q)_s$ is mathematically represented as [13]:

$$\begin{aligned} & y_t \\ & = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} \\ & + \frac{\left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right)\left(1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}\right)}{\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right)\left(1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}\right)} z_t \end{aligned} \tag{3.1}$$

where:

- y_t denotes the value of the series at time t.
- $X_{1,t}, X_{2,t}, \dots, X_{k,t}$ denote observations of the exogenous variables.
- $\beta_0, \beta_1, \dots, \beta_k$ denote the parameters of the regression part.
- $\phi_1, \phi_2, \dots, \phi_p$ denote the weight of the nonseasonal autoregressive terms.
- $\Phi_1, \Phi_2, \dots, \Phi_P$ denote the weight of the seasonal autoregressive terms.
- $\theta_1, \theta_2, \dots, \theta_q$ denote the weight of the nonseasonal moving average terms.
- $\Theta_1, \Theta_2, \dots, \Theta_Q$ denote the weight of the seasonal moving average terms.
- B^s denotes the backshift operator such that $B^s y_t = y_{t-s}$.
- z_t denotes the white noise terms.

For example, a SARIMAX (*1, 1, 1*) (*1, 1, 1*)$_{12}$ with one exogenous variable and weights in Table 3-1 is represented as:

$$y_t = 0.4(Exogenous\ variable) + \frac{(1 + 5.3B)(1 - 0.33B^7)}{(1 - 0.02B)(1 - 2.1B^7)} z_t \quad (3.1)$$

Table 3-1 Example coefficients

Term	Coefficient
Autoregressive (1)	0.02
Moving Average (1)	-5.3
Seasonal Autoregressive (1)	2.1
Seasonal Moving Average (1)	0.33
Exogenous Variable	0.4

3.4 Forecasting Methodology

The Box-Jenkins methodology is one of the most adopted forecasting methods using ARIMA models and is applicable to several domains. According to Box and Jenkins [40], the modelling process is broken down into three iterative steps, namely identification, estimation, and diagnostic checking Figure 3-1 is a schematic view of the process taken from Makridakis [41].

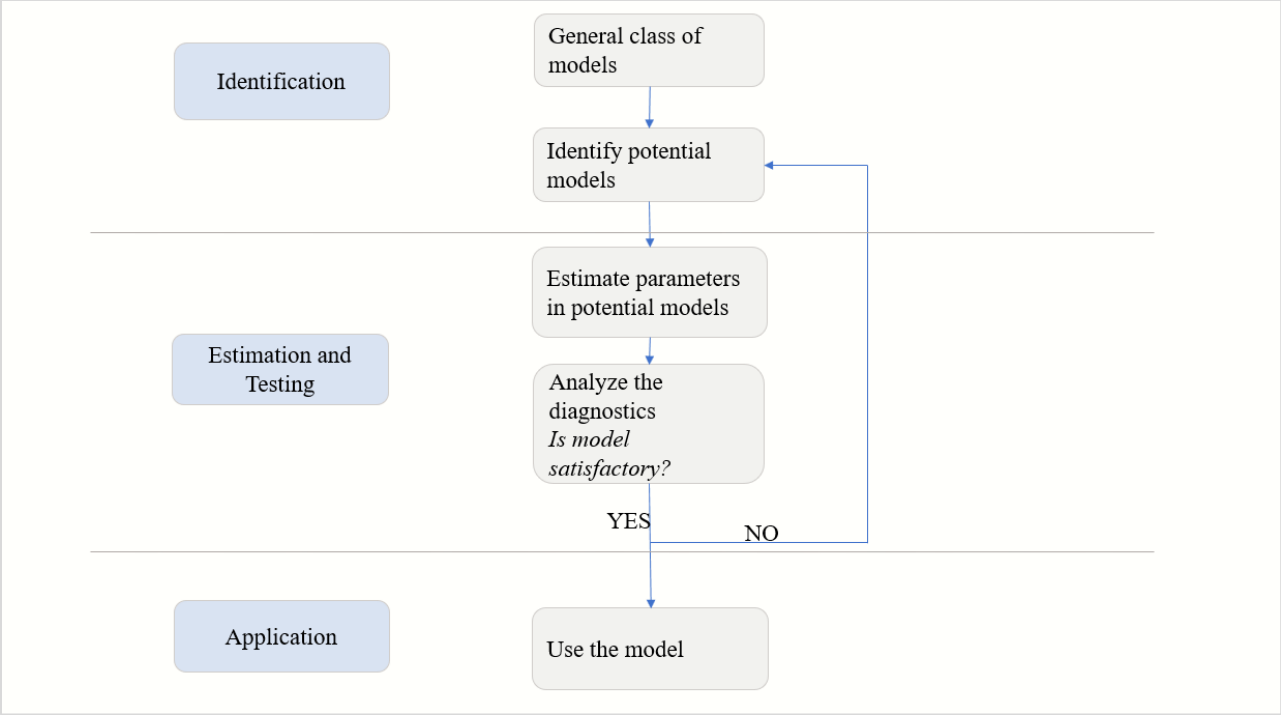

Figure 3-1 Schematic view of the Box-Jenkins methodology

More details about the methodology are provided in the following sections.

3.4.1 Identification

This step uses the data and knowledge regarding how it was generated to identify a subclass of models that best suit the data. This step can be broken down into the following two steps:

a. Data preparation

Differencing operations are applied to make the series stationary.

b. Model selection

Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are used to identify potential models. These functions explain the correlation of a value in the series with its lagged values.

3.4.2 Estimation and Diagnostics

This step uses the data to train the model, estimate the coefficients and check the residuals to see if they adhere to the assumptions. This step can be divided into the following two steps:

a. Estimation

Estimates are made of the parameters, and the best model is chosen based on a criterion. The outcome of the model selection step informs this step.

Akaike's Information Criterion (AIC) is the most commonly used model selection criterion [33]. AIC essentially measures the goodness of fit of a model (how well it fits the data) while penalizing complexity. Therefore, AIC reduces the risk of both overfitting and underfitting. A model that fits the data well and uses many predictors will have a larger AIC compared to a model that has the same goodness of fit but uses fewer predictors. Therefore, when comparing models, the one with the least AIC is chosen as the winner. It should be emphasized that the AIC of a model is a relative measure and is meaningful when compared to other models. AIC is calculated as [33]:

$$AIC = -2\ln(maximum\ likelihood) + 2p \tag{3.3}$$

where:

- p denotes the number of independent parameters estimated.

AIC can also be used to compare non-nested models, i.e. comparing an ARIMA model with a neural network model.

b. Diagnostic checking

The significance test is used for parameters to identify unnecessary parameters. The goodness-of-fit metrics are used to enable the user to compare models with each other. Lastly, statistical tests are used to check if the residuals are aligned with the assumptions of the modelling technique.

The tests on whether the residuals are white noise reveal valuable insights regarding the model. If the residuals are correlated, a more complex model is needed to capture all the information in the data. If residuals do not have a mean of zero, the forecast is biased.

3.4.3 Forecasting

In this step, test data is used to generate forecasts from the model, and the performance of the model is evaluated using forecast accuracy measures such as MSE, RMSE, MAPE, etc. A more thorough discussion of in-sample and out-of-sample forecast accuracy measures was provided in section 2.3.3.

Section 3 covered the concepts in the SARIMAX modelling technique and the Box-Jenkins methodology that are used throughout the thesis. The SARIMAX technique was chosen because of its ability to handle the characteristics of electricity consumption trends, i.e. seasonality and being highly influenced by external factors. The Box-Jenkins methodology was chosen as the basis of the modelling methodology because of its wide compatibility with the SARIMAX technique and generalizability to other domains. The methodology was modified to fit the requirements of this research better. Details about the modified methodology are discussed in chapter 4.

Chapter 4: Methodology

4.1 Introduction

This chapter focuses on the adopted methodology of the thesis. Section 4.1 introduces the Box-Jenkins methodology used in time-series modelling. Section 4.2 focuses on collecting and preprocessing data. Section 4.3 focuses on training SARIMAX models. Section 4.4 focuses on evaluating the performance of the models.

The Box-Jenkins methodology breaks down the modelling process into three iterative steps, namely, identification, estimation, and diagnostic checking [40]. Details about the Box-Jenkins methodology were discussed in section 3.4.

Figure 4-1 illustrates the methodology used in this thesis. This methodology was created by adding data collection and preprocessing, and performance evaluation steps to the Box-Jenkins methodology.

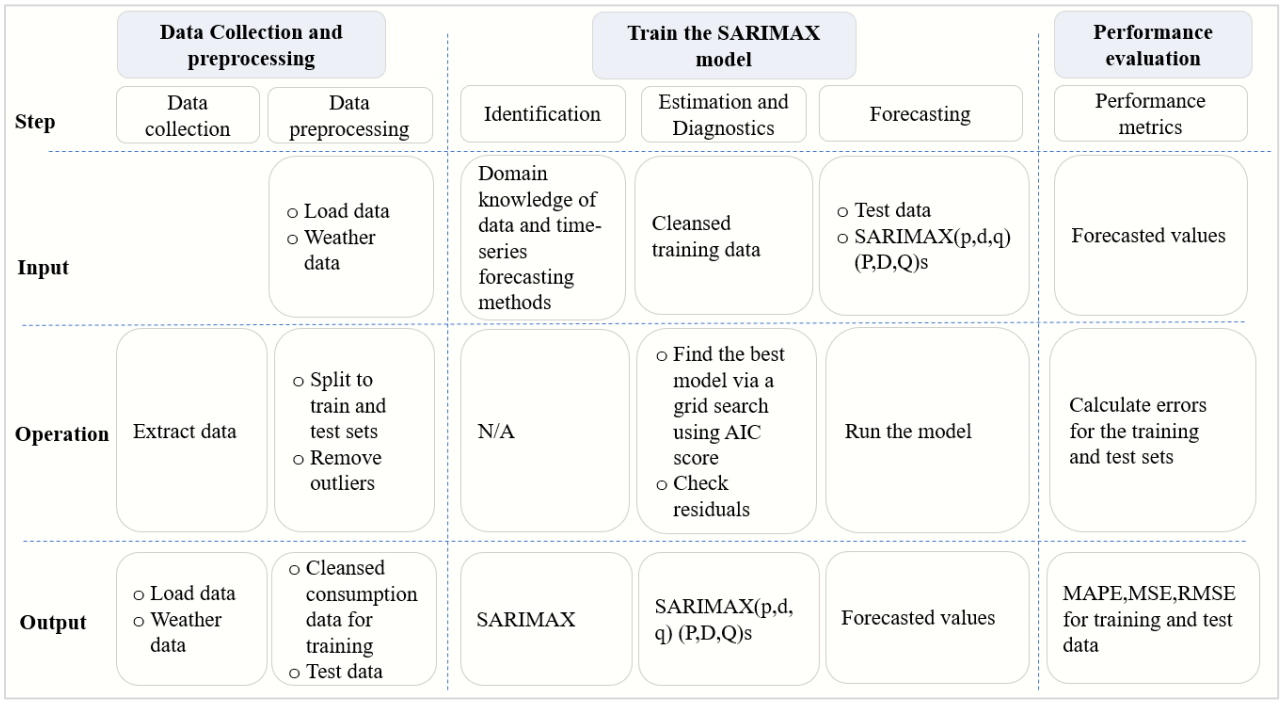

Figure 4-1 Modified Box-Jenkins methodology used in the thesis

The following sections describe each step in detail.

4.2 Data Collection and Preprocessing

4.2.1 Data Collection

Energy data and weather data were collected for modelling. Both datasets were collected from the university's energy data management software called SkySpark in hourly values, and these were resampled to obtain daily average or maximum values. Data covered the time period from the beginning of 2017 to the end of 2019. Energy data included electricity consumption of two buildings on campus. Weather data included hourly temperature and humidity from a weather station on the university campus. For each category, there were 26280 data points collected in total, 8760 for each year.

4.2.2 Data Preprocessing

First, the data was split into training and testing sets. The 70/30 approach mentioned in [42] was adopted to split the data. The first 70% of the data was used as the training set, while the last 30% of it was used for testing the models. Specifically, the 2017 and 2018 data were used for training, while the 2019 data were reserved for testing. Afterwards, the outliers were identified and replaced in the training set before moving to the modelling stage.

An outlier is a data point that is substantially different from other observations [43]. A widely used method for outlier detection and replacement is the Hampel filter. The Hampel filter identifies data points that differ from the median of the values in the window by more than three standard deviations and replaces them with median value [44]. In this thesis, the Hampel filter was applied to the time series to deal with the outliers in the same manner. One outlier was detected in total.

4.3 Training a SARIMAX Model

All of the modelling steps were carried out in a Jupyter notebook [45] using the Python language. The modelling was done using the SARIMAX class in the statsmodels [44] module.

4.3.1 Identification

From visual inspection and prior knowledge of the nature of each time series in the experiment, the author identified the data to be seasonal and non-stationary. There were generally weekly and yearly seasonal patterns in the data. Therefore, SARIMAX was identified as a suitable modelling technique that automatically handles seasonal and non-stationary data. The weekly patterns were categorized as seasonality, and the yearly patterns were dealt with using terms for trend.

4.3.2 Estimation and Diagnostics

This step has two parts, namely a) estimation and b) diagnostics. First, the values of the model's coefficients are estimated, and second, the residual values of the model are checked against the assumptions of SARIMAX modelling.

a. Estimation

A grid search was carried out using AIC as the model selection criterion to find the best model. Grid search is terminology in machine learning that refers to an automated process of training and evaluating a model [46]. In the case of SARIMAX, models were automatically created with different combinations of seasonal (P, D, Q) and non-seasonal (p, d, q) terms selected from a range of values, e.g. zero and one, and their respective AIC values were automatically calculated. A few examples of the combinations are: SARIMAX (0,0,0) $(0,0,0)_{7,}$ SARIMAX (0,0,0) $(0,0,1)_{7,}$ SARIMAX (0,0,0) $(0,1,0)_{7,}$ and SARIMAX (1,0,1) $(1,0,0)_{7.}$ All combinations of parameters and

the respective AIC values for each building are given in sections 5.3.3.2 and 5.2.3.2. Details of the AIC method were discussed in section 3.4.2.

A rule of thumb created by Burnham et al. in [47] suggests that models with $\Delta_i \leq 2$ ($\Delta_i = AIC_i - AIC_{minimum}$) have a strong possibility to be the best model, while $4 \leq \Delta_i \leq 7$ have a much lower possibility, and $\Delta_i > 10$ have no possibility. This rule was applied to select the best model in this thesis.

The process is as follows [47]:

- AIC values of all the possible parameter combinations are calculated.
- $AIC_{minimum}$ is subtracted from all values.
- Parameters of AIC values with $\Delta_i \leq 2$ are used to fit models.
- The model with the least dynamic MAPE is chosen.

b. Diagnostics

First, the estimated coefficients from the previous step and their significance (p-values) are analyzed. Coefficients are dropped from the model if their value is too low or are not significant, i.e. p-value > 0.005.

Second, four statistical tests are carried out on the residuals, and the results are visualized. These tests are as follow [48]:

1. Standardized residuals
2. Histogram plus estimated density
3. Normal Q-Q
4. Correlogram

Figure 4-2 shows the respective graphs for each test using sample data from statsmodels [49].

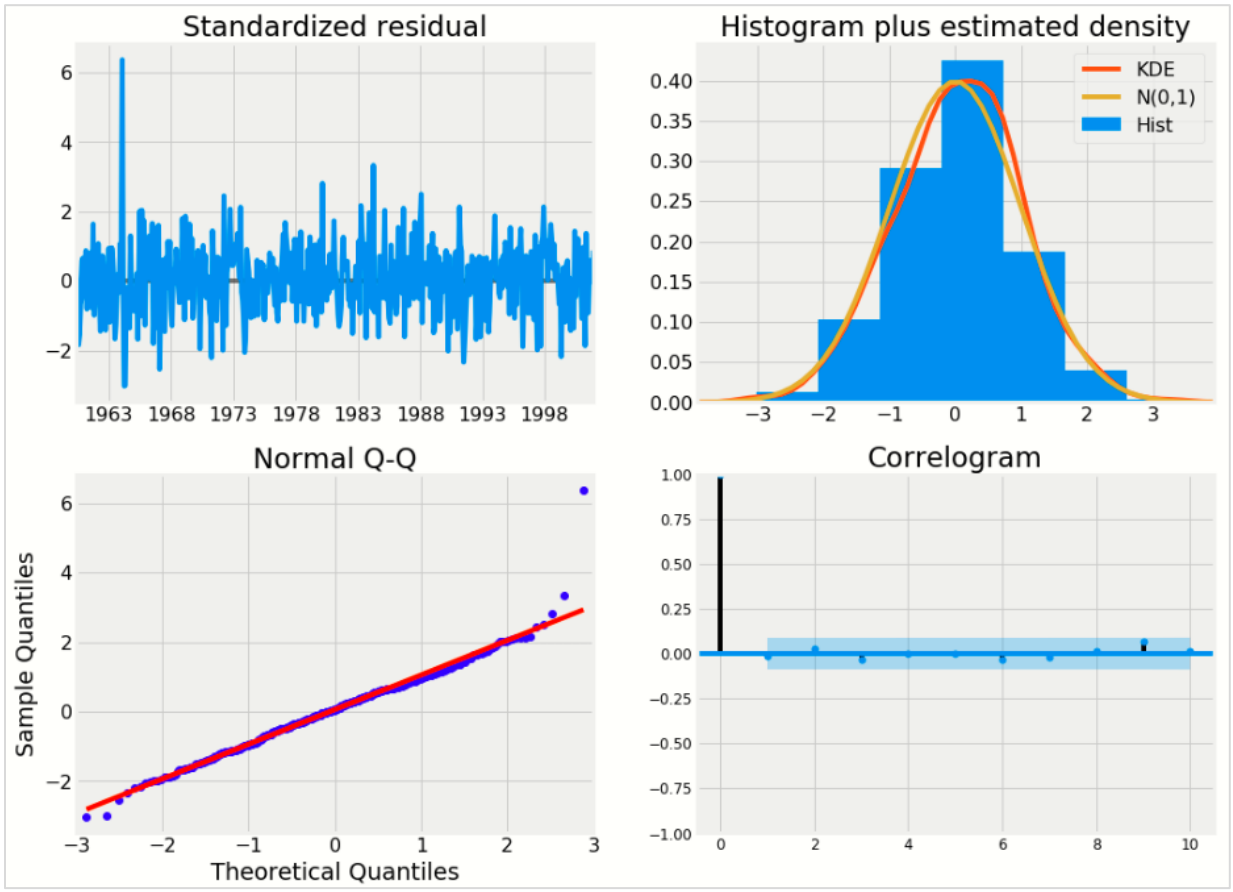

Figure 4-2 Diagnostic tests on the residuals of a sample model from statsmodels [49]

The standardized residual and residual correlogram graphs show whether the residuals are white noise. If that is not the case, it means the model cannot explain the series well, and other factors might need to be considered.

The histogram shows whether the residuals are normally distributed. The N(0,1) line is a normal distribution with a mean equal to zero and a standard deviation of one.

The Q-Q plot shows whether the residuals are normally distributed. Blue dots are the ordered distribution of residuals, and the red dots are samples taken from a normal distribution

with N(0,1). If the blue dots closely follow the red dots, it is a strong sign that the residuals are normally distributed.

4.3.3 Forecasting

The daily electricity consumption/power for 2019 was forecasted in two individual buildings with daily average temperature and humidity as exogenous variables. The predicted values were compared to the testing set values.

4.4 Performance Evaluation

In-sample and out-of-sample errors were considered to evaluate the performance of the models. The in-sample error refers to the error that comes from the training set, while out-of-sample error refers to the testing set.

Errors are calculated for two types of forecasts: one-step ahead and dynamic forecasts. One-step-ahead forecasts use the full history of the data up to the point of prediction. Dynamic forecasts use data only up to a certain point in time, and then predict values for future times beyond that point. In the one-step-ahead forecast, the number of input values increases as the prediction goes further into the future. In contrast, the number of input values remains the same for dynamic forecasts.

4.4.1 Performance Metrics

MAPE, MSE, and RMSE were used as performance evaluation metrics. The selected metrics help gauge the performance of each model using both absolute and percentage values. MSE and RMSE were used to gauge the performance of models individually using the absolute values, while MAPE was used to gauge the errors as a percentage of absolute values. MAPE also enabled comparison of a model to other models (MAPE). The respective calculations were discussed in more detail in sections 2.3.3 and 3.4.3.

Chapter 5: Results and Discussion

5.1 Introduction

This chapter presents the results of the different steps of the applied methodology and the models. Two buildings with different compositions were modelled in order to investigate the level of success of the modelling approach for different types of buildings. The first building, Pharmaceutical Sciences Building, has labs, classes, and office spaces. In contrast, the second building, Robert H. Lee Alumni Centre, has a large venue for events, a café, classrooms and social spaces. Section 5.2 presents the results for the Pharmaceutical Sciences Building. Section 5.3 presents the results for the Robert H. Lee Alumni Centre. Section 5.5 provides a summary of the chapter.

5.2 Pharmaceutical Sciences Building

5.2.1 Introduction

The Pharmaceutical Sciences Building was opened in 2012 and is Leadership in Energy and Environmental Design (LEED) [50] gold certified [51]. The building is a mix of labs, classes, and office spaces. It also houses the main data centre on campus. In this building, electricity consumption is mainly from plug loads (appliances and equipment) and the HVAC system. The building also uses a heat recovery system that captures and repurposes the heat generated in the data centre. The daily average electricity consumption of the building was modelled using consumption and weather data from 2017 and 2018. The model was used to forecast daily average electricity consumption in 2019. The predicted values were compared to the actual daily average electricity consumption in 2019. The modelling steps and the results are presented in the following sections.

The modelling steps and the results are presented in the following sections.

5.2.2 Data Collection and Preprocessing

5.2.2.1 Data Collection

Hourly electricity consumption data were collected from SkySpark software from 2017 to 2019. Data were reported in kilowatt-hour and resampled to average daily values. Weather data were collected from 2017 to 2019. Weather data included hourly temperature and humidity and were resampled to average daily values. The weather station that the data was recorded from was on campus, as noted in 4.3.1. The data from 2017 to the end of 2018 were used as the training set, while the 2019 data were used as the testing set.

Figure 5-1 shows the average daily electricity consumption of the Pharmaceutical Sciences Building in 2017 and 2018.

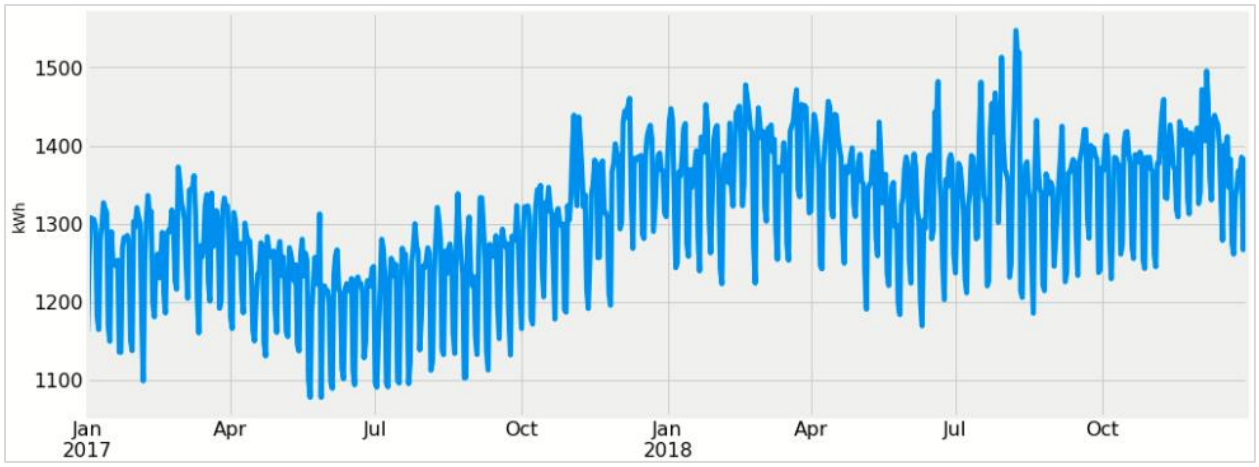

Figure 5-1 Average daily electricity consumption of the pharmaceutical building

Figure 5-2 shows the temperature data in Celsius for the same period.

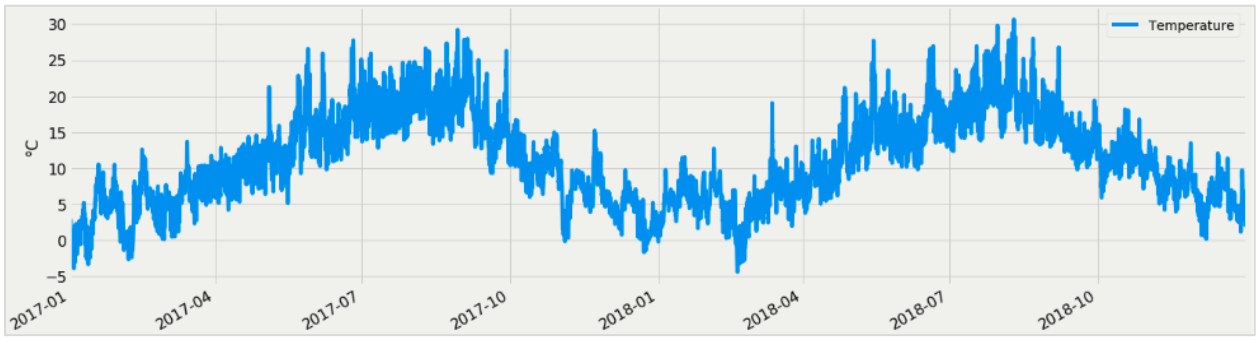

Figure 5-2 Average daily temperature for 2017 and 2018

Every time series can be broken down into three components, namely trend, seasonal patterns, and residuals. In an additive decomposition, the value of time-series at time t is the sum of the values of the three components at time t. Trend is a gradual increase or decrease in values,

a seasonal pattern is a repeating short-term cycle which is affected by seasonal factors such as day of the week, residual is what is left of the series after trend and seasonality are taken out.

Figure 5-3 shows the electricity consumption data decomposed into the trend, seasonal (weekly period), and residual values.

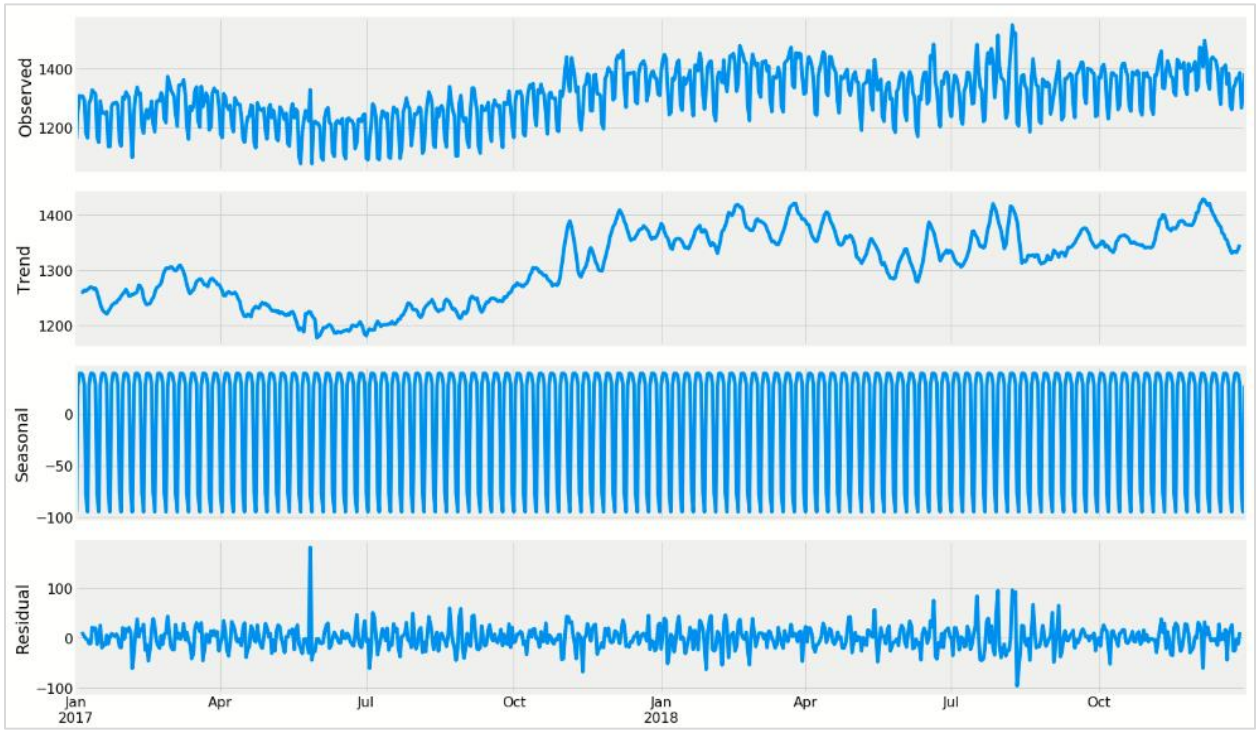

Figure 5-3 Decomposed view of the daily average electricity consumption

The trend appears to show an annual pattern of lower daily average electricity consumption in the summer and higher values in the winter, along with a gradual increase from the beginning of 2017 to the end of 2018. This annual pattern was treated as a trend component, while the weekly pattern was treated as a seasonal component. To better understand the weekly seasonal graph above, Figure 5-4 shows a detailed view of the seasonal segment for the first month of 2017, starting from Monday, January 2nd, 2017. The weekly pattern in the data is easily noted in this figure.

Figure 5-4 Detailed view of daily average electricity consumption in the first month of 2017, starting from Monday, January 2nd

The residual shows random fluctuations around zero.

5.2.2.2 Data Preprocessing

A Hampel filter was applied to identify and replace outliers. One data point was identified to be away from the median by more than three standard deviations (greyed-out band) and was replaced by the median value of the series. Figure 5-5 illustrates the band and the median value.

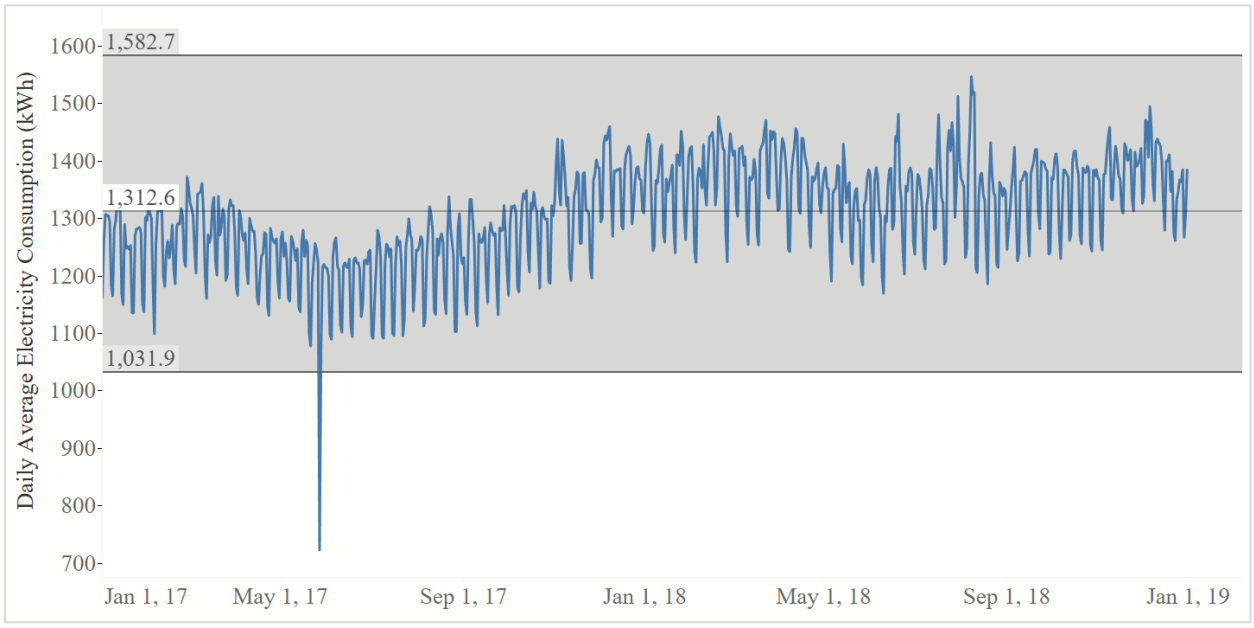

Figure 5-5 Daily average electricity consumption with three standard deviations from the median highlighted in grey

Table 5-1 shows the data point of the outlier, its value, and the value it was replaced with.

Table 5-1 Value of the outlier and its replaced value

Date	Value	Replaced value
05/27/2017	722.8	1312.6

5.2.3 Training the SARIMAX model

5.2.3.1 Identification

From domain knowledge, it was known that the electricity consumption data was seasonal and non-stationary. This prior knowledge resulted in identifying SARIMAX as a suitable modelling technique. SARIMAX automatically handles seasonal and non-stationary data.

5.2.3.2 Estimation and Diagnostics

a. Estimation

To find the best SARIMAX (p, d, q) $(P, D, Q)_s$ model, a grid search was conducted on 60 different combinations of seasonal (P, D, Q) and nonseasonal (p, d, q) parameters. These parameters took either the value of zero or one. The seasonal parameter(s) was set as 7 to be able to extract weekly patterns.

Table 5-2 shows all the candidate models and their AIC values.

Table 5-2 AIC values of the candidate models for the Pharmaceutical Building

SARIMAX (p, d, q) $(P, D, Q)_s$	AIC Values
SARIMAX(0, 0, 0)x(0, 0, 1, 7)	AIC:9276.718798034955
SARIMAX(0, 0, 0)x(0, 1, 1, 7)	AIC:7262.314631598504
SARIMAX(0, 0, 0)x(1, 0, 0, 7)	AIC:7661.866417811521
SARIMAX(0, 0, 0)x(1, 0, 1, 7)	AIC:8551.60381861628
SARIMAX(0, 0, 0)x(1, 1, 0, 7)	AIC:7316.779514003685
SARIMAX(0, 0, 0)x(1, 1, 1, 7)	AIC:7255.682574383288
SARIMAX(0, 0, 1)x(0, 0, 0, 7)	AIC:9078.808068973403

SARIMAX (p, d, q) $(P, D, Q)_s$	AIC Values
SARIMAX(0, 0, 1)x(0, 0, 1, 7)	AIC:8904.961470783157
SARIMAX(0, 0, 1)x(0, 1, 0, 7)	AIC:7406.445941296425
SARIMAX(0, 0, 1)x(0, 1, 1, 7)	AIC:7000.079099855881
SARIMAX(0, 0, 1)x(1, 0, 0, 7)	AIC:7548.335255191223
SARIMAX(0, 0, 1)x(1, 0, 1, 7)	AIC:7922.370376039563
SARIMAX(0, 0, 1)x(1, 1, 0, 7)	AIC:7120.723543500815
SARIMAX(0, 0, 1)x(1, 1, 1, 7)	AIC:7004.148782584813
SARIMAX(0, 1, 0)x(0, 0, 1, 7)	AIC:7826.0188833889015
SARIMAX(0, 1, 0)x(0, 1, 1, 7)	AIC:6953.5341654387685
SARIMAX(0, 1, 0)x(1, 0, 0, 7)	AIC:7436.962293176768
SARIMAX(0, 1, 0)x(1, 0, 1, 7)	AIC:7245.032001712807
SARIMAX(0, 1, 0)x(1, 1, 0, 7)	AIC:7217.105637768912
SARIMAX(0, 1, 0)x(1, 1, 1, 7)	AIC:6943.701976384851
SARIMAX(0, 1, 1)x(0, 0, 0, 7)	AIC:8216.18748980006
SARIMAX(0, 1, 1)x(0, 0, 1, 7)	AIC:7819.64819622111
SARIMAX(0, 1, 1)x(0, 1, 0, 7)	AIC:7460.532925484178
SARIMAX(0, 1, 1)x(0, 1, 1, 7)	AIC:7040.169965981815
SARIMAX(0, 1, 1)x(1, 0, 0, 7)	AIC:7411.792938577747
SARIMAX(0, 1, 1)x(1, 0, 1, 7)	AIC:7320.324659190166
SARIMAX(0, 1, 1)x(1, 1, 0, 7)	AIC:7154.452119376351
SARIMAX(0, 1, 1)x(1, 1, 1, 7)	AIC:7040.892796217946
SARIMAX(1, 0, 0)x(0, 0, 0, 7)	AIC:8257.401632190315
SARIMAX(1, 0, 0)x(0, 0, 1, 7)	AIC:7840.019732373621
SARIMAX(1, 0, 0)x(0, 1, 0, 7)	AIC:7348.361863511855
SARIMAX(1, 0, 0)x(0, 1, 1, 7)	AIC:6872.19548542968
SARIMAX(1, 0, 0)x(1, 0, 0, 7)	AIC:7505.5692603443795
SARIMAX(1, 0, 0)x(1, 0, 1, 7)	AIC:8409.361284865627
SARIMAX(1, 0, 0)x(1, 1, 0, 7)	AIC:7042.473335226991
SARIMAX(1, 0, 0)x(1, 1, 1, 7)	AIC:6965.03655409581
SARIMAX(1, 0, 1)x(0, 0, 0, 7)	AIC:8198.07994888641
SARIMAX(1, 0, 1)x(0, 0, 1, 7)	AIC:7953.419839725032
SARIMAX(1, 0, 1)x(0, 1, 0, 7)	AIC:7336.633648931907
SARIMAX(1, 0, 1)x(0, 1, 1, 7)	AIC:6909.295579315535
SARIMAX(1, 0, 1)x(1, 0, 0, 7)	AIC:7818.668912373695
SARIMAX(1, 0, 1)x(1, 0, 1, 7)	AIC:8347.326864031435

SARIMAX (p, d, q) (P, D, Q)$_s$	AIC Values
SARIMAX(1, 0, 1)x(1, 1, 0, 7)	AIC:7043.981538708403
SARIMAX(1, 0, 1)x(1, 1, 1, 7)	AIC:6939.843432713456
SARIMAX(1, 1, 0)x(0, 0, 0, 7)	AIC:8228.451578054255
SARIMAX(1, 1, 0)x(0, 0, 1, 7)	AIC:7827.73094724716
SARIMAX(1, 1, 0)x(0, 1, 0, 7)	AIC:7479.184697704397
SARIMAX(1, 1, 0)x(0, 1, 1, 7)	AIC:7060.369754016028
SARIMAX(1, 1, 0)x(1, 0, 0, 7)	AIC:7407.591972774075
SARIMAX(1, 1, 0)x(1, 0, 1, 7)	AIC:7194.359068134883
SARIMAX(1, 1, 0)x(1, 1, 0, 7)	AIC:7166.095887119316
SARIMAX(1, 1, 0)x(1, 1, 1, 7)	AIC:7000.994808072945
SARIMAX(1, 1, 1)x(0, 0, 0, 7)	AIC:8043.5061215054375
SARIMAX(1, 1, 1)x(0, 0, 1, 7)	AIC:7747.674124806013
SARIMAX(1, 1, 1)x(0, 1, 0, 7)	AIC:7345.317486390242
SARIMAX(1, 1, 1)x(0, 1, 1, 7)	AIC:7003.518144720236
SARIMAX(1, 1, 1)x(1, 0, 0, 7)	AIC:7633.4063843495915
SARIMAX(1, 1, 1)x(1, 0, 1, 7)	AIC:7549.2415673597225
SARIMAX(1, 1, 1)x(1, 1, 0, 7)	AIC:7117.994277679479
SARIMAX(1, 1, 1)x(1, 1, 1, 7)	AIC:7016.499279148635

Table 5-3 shows the least AIC value calculated. No values were within two units from the minimum value, i.e. $\Delta_i \leq 2$ ($\Delta_i = AIC_i - AIC_{minimum}$).

Table 5-3 Lowest AIC value and its corresponding seasonal and non-seasonal parameters

Rank	SARIMAX (p, d, q) (P, D, Q)$_s$	AIC
1	SARIMAX (1,0,0) (0,1,1) $_7$	6872.19548542968

The interpretation of SARIMAX (1,0,0) $(0,1,1)_7$ is as follows.

- $p = 1$: Model uses the consumption value of the day before.
- $d = 0$: Model does not need differencing to make the data stationary.
- $q = 0$: Model does not use lagged forecast errors.
- $P = 0$: Model does not use the consumption values of previous seasons, i.e. previous weeks.
- $D = 1$: Model needs to seasonally (one week) difference the data once to make it stationary.
- $Q = 1$: Model uses one seasonally lagged (one week) forecast error.
- $s = 7$: Model uses weekly seasonality.

The estimated coefficients of SARIMAX (1,0,0) $(0,1,1)_7$ are given in Table 5-4.

Table 5-4 Estimated coefficients of SARIMAX (1,0,0) $(0,1,1)_7$ and corresponding standard errors and p values

Term	Coefficient	Standard error	P-value
Drift	0.0008	0.000	0.020
Average daily temperature	-3.7626	0.510	0.000
Average daily humidity	-1.0348	0.150	0.000
AR (1)	0.7045	0.021	0.000
SMA (1)	-0.8953	0.020	0.000
σ^2	851.3310	30.767	0.000

A brief explanation of the categories in Table 5-4 is as follows.

- The coefficient column shows the importance (weight) of each term. The negative sign of a coefficient indicates a reverse relationship between the term and the

current value. The higher the absolute value of the coefficient is, the bigger impact that variable has on electricity consumption. For (σ^2), the coefficient value represents the variance of the error term.

- The standard error shows the standard deviation of the coefficient. The smaller the standard error, the more precise the estimate is. A more precise coefficient results in a more accurate electricity consumption model.
- The p-value shows the significance of each term's weight. A significance of 0.05 or lower shows strong evidence against the null hypothesis [52]. In this case, the null hypothesis is that the term's weight is equal to zero. If the p-value is less than 0.05, then the null hypothesis is rejected, and the coefficient's weight is deemed significant. For instance, if the p-value of the average daily temperature is more than 0.05, then the average daily temperature is deemed insignificant to the electricity consumption model and can be dropped.

The terms used in the model are explained next.

- The drift characterizes the slope of the trend used in the model. Despite its small coefficient it is significant (p-value < 0.05).
- The average daily temperature has the highest weight. The value of -3.7626 as the coefficient for the average daily temperature can be interpreted as that one unit, in this case, centigrade, increase in the average daily temperature corresponds to a 3.7626 unit, in this case kwh, decrease in the average daily electricity consumption. That is if none of the other terms in the model change.

- Regarding the daily average humidity, the -1.0348 value of the coefficient means that a one unit, in this case percent, increase in daily average humidity results in a 1.0348 unit, in this case kwh, decrease in average daily electricity consumption.
- The estimated value of 0.7045 for the autoregressive term (AR (1)) shows that if the previous value, e.g. yesterdays' electricity consumption, increases by one, the current value, e.g. today's electricity consumption, increases by 0.7045.
- The estimated value of -0.8953 for the seasonal moving average term (SMA (1)) shows that if the error in the previous season, previous week, increases by one, the current value, e.g. today's electricity consumption, decreases by 0.8953.
- σ^2 is the variance of the error term, and its value is dependent on the scale of the data.

All the coefficients are significant (p-value < 0.05) and estimated with reasonable precision, i.e. low standard errors. Therefore, all terms are kept in the model.

a. Diagnostics

The results of the diagnostic tests on SARIMAX (1,0,0) $(0,1,1)_7$ are shown in Figure 5-6. The standardized residual and correlogram show that the residuals are white noise. The KDE line on the histogram and the Q-Q plot show that the residuals roughly follow a normal distribution. The results of the diagnostic tests satisfy the assumptions of SARIMAX modelling.

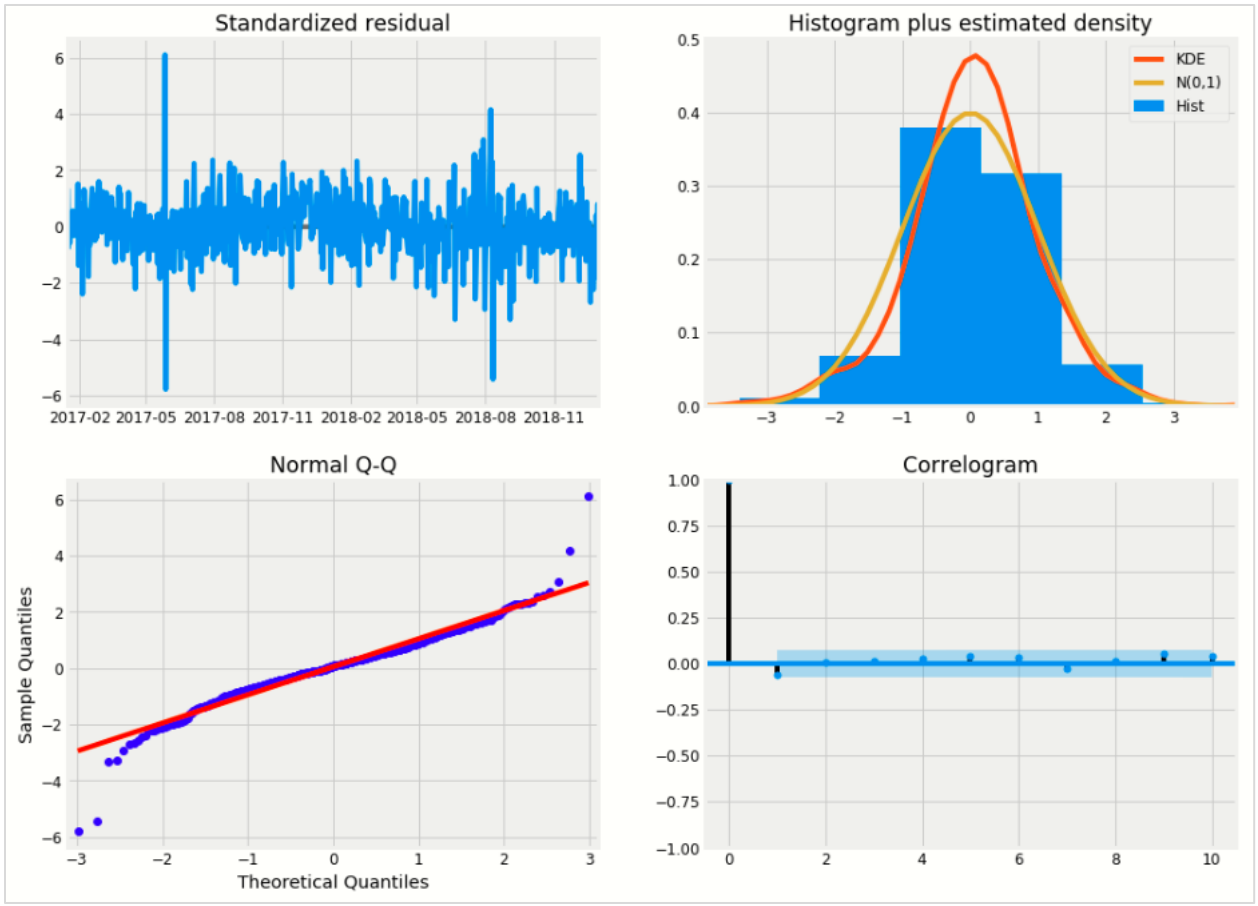

Figure 5-6 Diagnostic tests on the residuals of the model

5.2.3.3 Forecasting

The average daily electricity consumption of 2019 was forecasted using the model from the previous step. Data for 2019 were used as the testing set and were not included in the modelling steps.

Figure 5-7 shows forecasted values in 2019 overlaid with the actual values. The blue line shows actual values, and the orange line shows forecasted values in kWh.

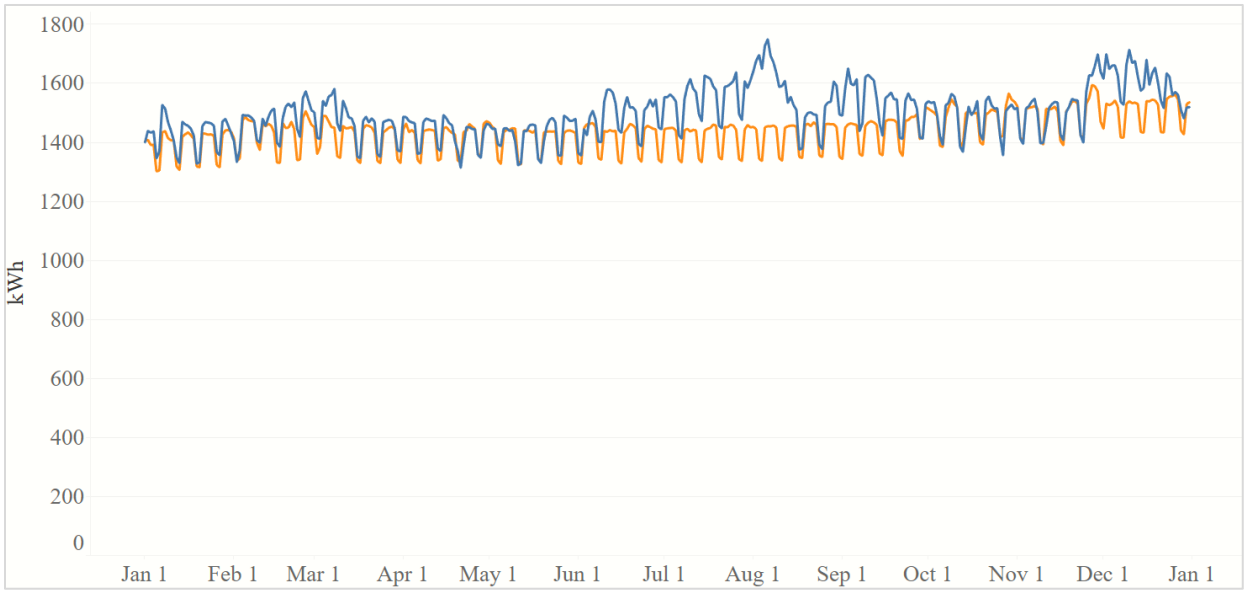

Figure 5-7 Forecasted and actual values of daily average electricity consumption in 2019

The model was able to capture the weekly trend successfully. However, it was not able to capture the spikes in August and December. This area is discussed in more detail in section 5.4.

The MAPE was 4.1%. Accuracy measures are discussed in more detail in the next section, while discussions regarding the interpretation of the results are covered in section 5.4.

5.2.4 Performance Evaluation

5.2.4.1 Performance Metrics

The performance of the model was evaluated using the testing set. The electricity consumption data for 2019 were designated as the testing set. The results are shown in Table 5-4. Figure 5-8 and Figure 5-9 show the one-step ahead and dynamic forecast errors of the model.

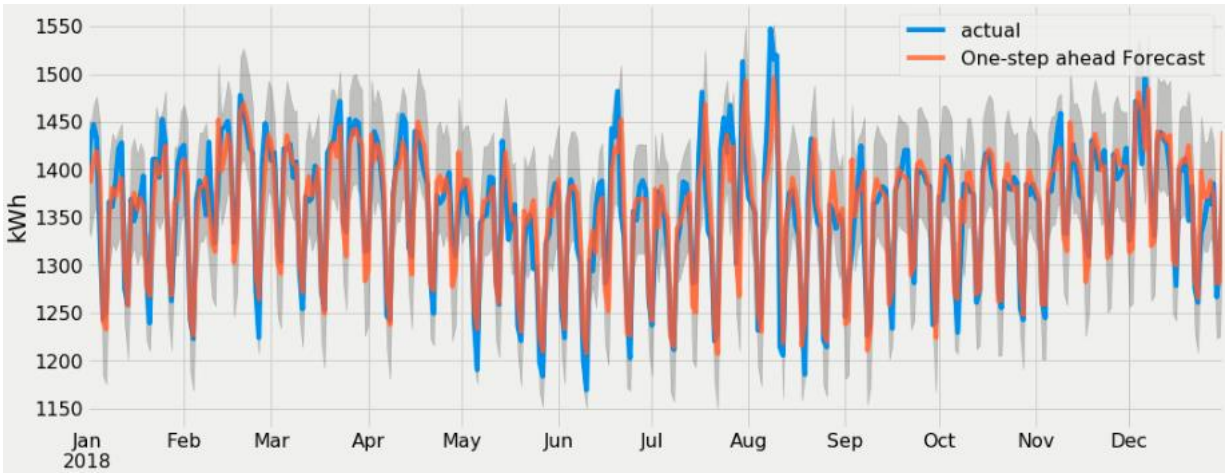

Figure 5-8 One-step-ahead error of the model with confidence bands

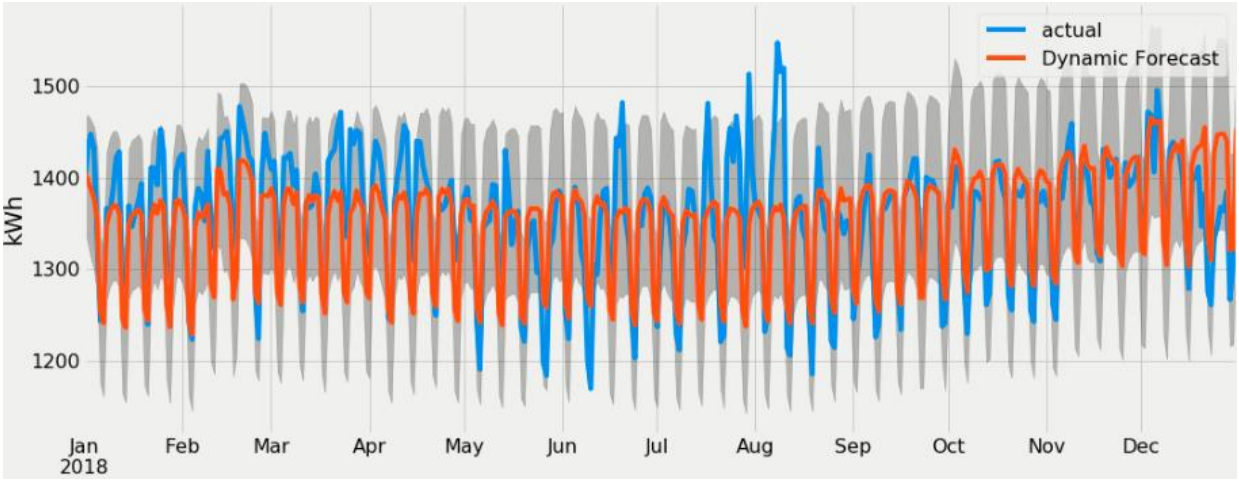

Figure 5-9 Dynamic error of the model with confidence bands

Table 5-5 Results of the performance evaluation of the model

	Training set: One-step ahead	Training set: Dynamic	Testing set
MSE	919.4	1859.3	7428.1
RMSE	30.3	43.11	86.2
MAPE	1.6%	2.4%	4.1%

Table 5-5 shows that, in general, the error is lower in the training set than in the testing set, as is to be expected. If the error in the testing set were lower than the error in the training set, it would have been concluded that the model was overfitted. Overfit models follow a set of data too closely and thus are not reliable for future predictions with additional data.

The MSE calculates the mean of the errors squared and disregards the direction of errors. As expected, this error is higher for the training set using the dynamic method than for the training set using the one-step-ahead method, and it is highest for the testing set. It shows that the average of prediction errors squared in 2019 was 7428.1. The RMSE takes the root of the MSE. Thus, it has the same unit of measurement as the data, kWh in this case. It shows that the predictions in 2019 were on average off by 86.2 kWh compared to the actuals. The MAPE reports the average of the absolute errors as a percentage of the actual values. This value was 4.1% using the 2019 test data. Because of using the percentage, MAPE is not dependent on the scale of the data. The MAPE also increases from the training set to the testing set, as it was expected. This model has a satisfactory performance. The performance of the model is discussed in more detail in section 5.4.

5.3 Robert H. Lee Alumni Centre

5.3.1 Introduction

The Robert H. Lee Alumni Centre was opened in the spring of 2015 on UBC campus with LEED gold certification. It has a large venue for events, a café, meeting rooms, classrooms, and social spaces [53]. Electricity consumption is mainly used for plug loads (appliances and equipment) and the HVAC system in the Alumni Centre. The peak electricity load of the building was modelled using hourly demand and weather data from 2017 and 2018. The model was used to forecast the building's daily peak electricity load in 2019. The predicted values were compared to the actual daily peak loads in 2019. The modelling steps and the results are presented in the following sections.

5.3.2 Data Collection and Preprocessing

5.3.2.1 Data Collection

Hourly electricity consumption data were collected from SkySpark software from 2017 to 2019. Data were reported in kilowatt-hour and resampled to maximum daily values. The maximum daily values also represent daily peak loads of the building. Since the consumption data collected were in hourly timestamped values, they had the same numerical values as the hourly demand load of the building. This was validated by comparing both the consumption and demand data.

Weather data were also collected from 2017 to 2019. Weather data included hourly temperature and humidity and were resampled to average daily values. The weather station that the data were recorded from was on campus, as noted in 4.3.1. The data from 2017 and 2018 was used as the training set, while the 2019 data were used as the testing set.

Figure 5-10 shows the daily peak loads of the Alumni centre building from 2017 to the end of 2018.

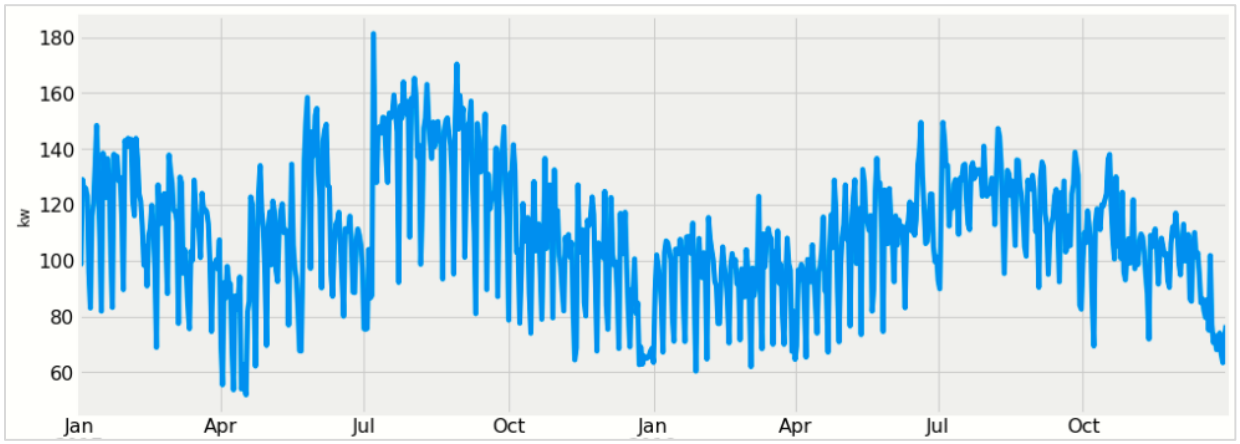

Figure 5-10 Peak daily electricity demand of the Alumni Center building

Figure 5-11 shows the temperature data in Celsius for the same period.

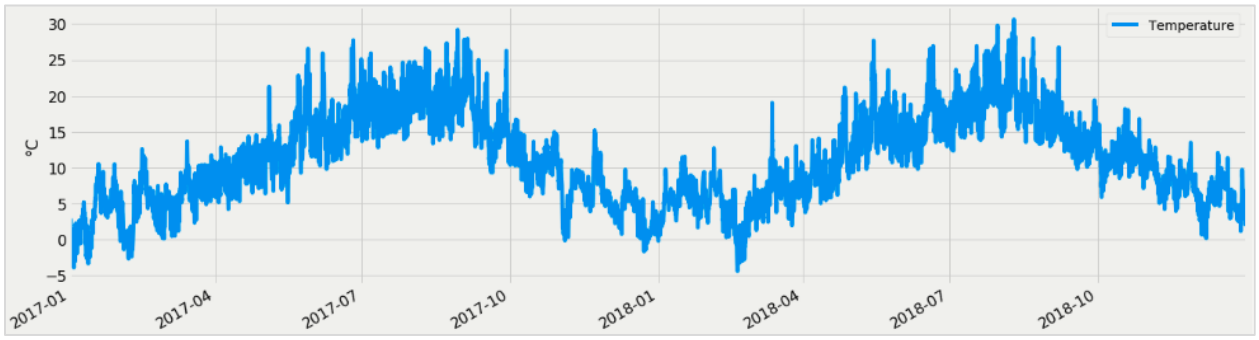

Figure 5-11 Average daily temperature of 2017 and 2018

As discussed in section 5.2.2.1, a time series can be broken down into three components, namely trend, seasonal patterns, and residuals. Figure 5-12 shows the decomposed view of the daily peak electricity demand of the Alumni Centre.

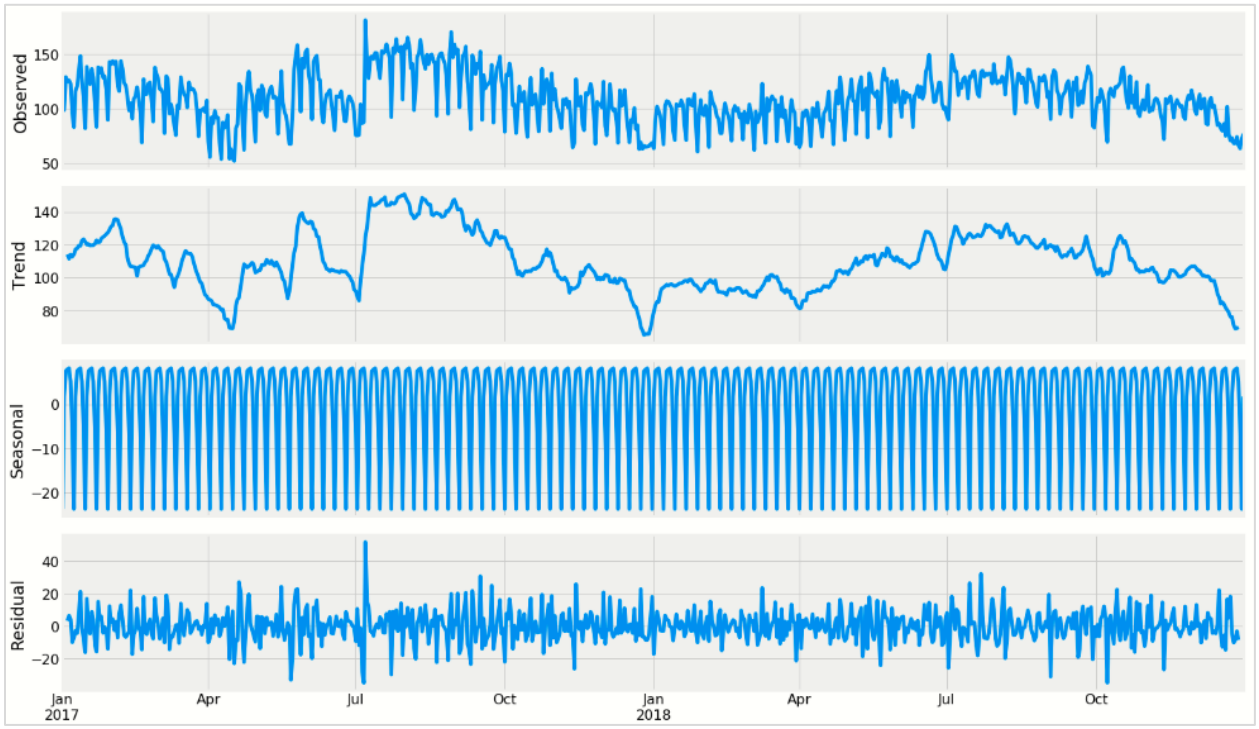

Figure 5-12 Decomposed view of the daily peak electricity demand of the Alumni Center

The graph for trend shows a lack of a dominant trend in the data. The trend graph shows irregular increases and decreases in value with no gradual upward or downward trend. To better understand the seasonal graph above, Figure 5-13 shows a detailed view of the seasonal segment for the first month of 2017, starting from Monday, January 2nd, 2017. The weekly pattern in the data is easily noted in this figure.

Figure 5-13 Detailed view of the daily peak electricity demand of the Alumni Center in the first month of 2017, starting from Monday, January 2nd

The residual shows random fluctuations around zero.

5.3.2.2 Data Preprocessing

A Hampel filter was applied to identify and replace outliers. No outlier was detected. Figure 5-14 shows the daily peak loads of the Alumni Center with the blue line and three standard deviations from the median value with the grey band.

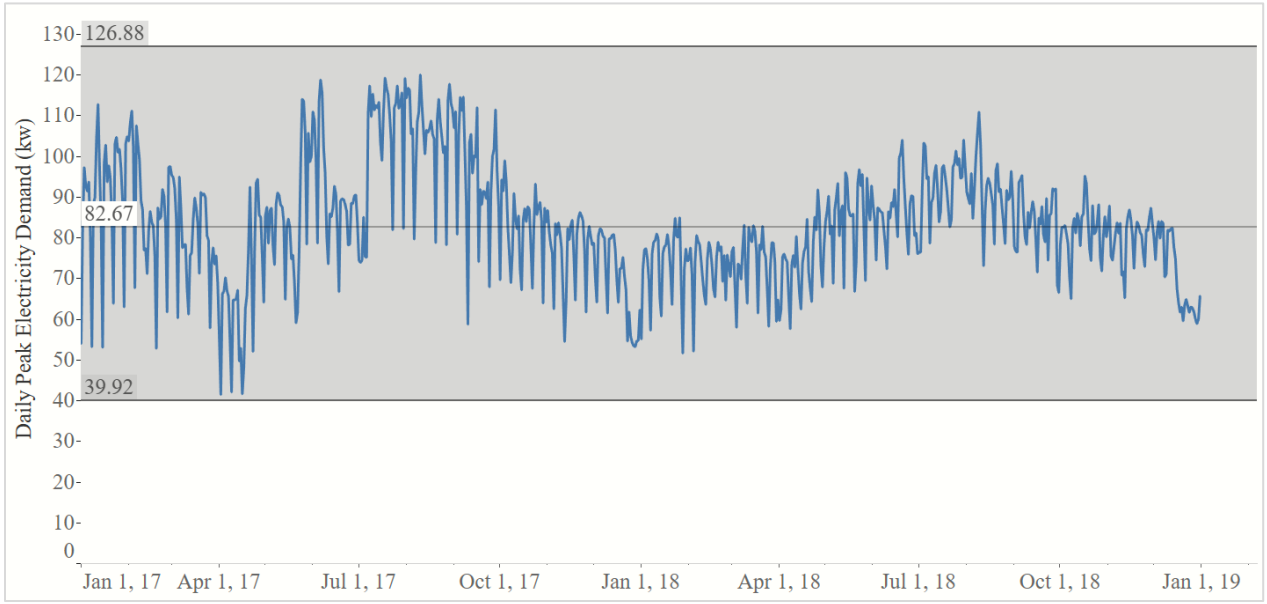

Figure 5-14 Daily peak electricity demand of the Alumni Center with three standard deviations from the median highlighted in grey

5.3.3 Training the SARIMAX model

5.3.3.1 Identification

From domain knowledge, it was known that the peak demand data was seasonal and non-stationary. This prior knowledge resulted in identifying SARIMAX as a suitable modelling technique. SARIMAX automatically handles seasonal and non-stationary data.

5.3.3.2 Estimation and Diagnostics

b. Estimation

To find the best SARIMAX (p, d, q) $(P, D, Q)_s$ model, a grid search was conducted on 60 different combinations of seasonal (P, D, Q) and nonseasonal (p, d, q) parameters. These parameters took either the value of zero or one. The s parameter was set as 7 to be able to extract weekly patterns.

Table 5-6 shows all the candidate models and their AIC values.

Table 5-6 AIC values of the candidate models for the Alumni Centre

SARIMAX (p, d, q) $(P, D, Q)_s$	AIC Values
SARIMAX(0, 0, 0)x(0, 0, 1, 7)	AIC:6412.285842719054
SARIMAX(0, 0, 0)x(0, 1, 1, 7)	AIC:5942.4864505774985
SARIMAX(0, 0, 0)x(1, 0, 0, 7)	AIC:6342.665769889736
SARIMAX(0, 0, 0)x(1, 0, 1, 7)	AIC:5996.188064534124
SARIMAX(0, 0, 0)x(1, 1, 0, 7)	AIC:6039.248679847084
SARIMAX(0, 0, 0)x(1, 1, 1, 7)	AIC:5929.19037374907
SARIMAX(0, 0, 1)x(0, 0, 0, 7)	AIC:6413.911503244104
SARIMAX(0, 0, 1)x(0, 0, 1, 7)	AIC:6209.146316516963
SARIMAX(0, 0, 1)x(0, 1, 0, 7)	AIC:6074.053174931856
SARIMAX(0, 0, 1)x(0, 1, 1, 7)	AIC:5777.7232064760865
SARIMAX(0, 0, 1)x(1, 0, 0, 7)	AIC:6153.237625642477
SARIMAX(0, 0, 1)x(1, 0, 1, 7)	AIC:5830.5269936089335
SARIMAX(0, 0, 1)x(1, 1, 0, 7)	AIC:5925.487601243984
SARIMAX(0, 0, 1)x(1, 1, 1, 7)	AIC:5772.917641082843
SARIMAX(0, 1, 0)x(0, 0, 1, 7)	AIC:6169.591505933586

SARIMAX (p, d, q) $(P, D, Q)_s$	AIC Values
SARIMAX(0, 1, 0)x(0, 1, 1, 7)	AIC:5835.521526581343
SARIMAX(0, 1, 0)x(1, 0, 0, 7)	AIC:6099.950876237204
SARIMAX(0, 1, 0)x(1, 0, 1, 7)	AIC:5891.479166058855
SARIMAX(0, 1, 0)x(1, 1, 0, 7)	AIC:6084.529095251786
SARIMAX(0, 1, 0)x(1, 1, 1, 7)	AIC:5836.44698472607
SARIMAX(0, 1, 1)x(0, 0, 0, 7)	AIC:6197.858681092486
SARIMAX(0, 1, 1)x(0, 0, 1, 7)	AIC:6004.055277015097
SARIMAX(0, 1, 1)x(0, 1, 0, 7)	AIC:6098.155407429591
SARIMAX(0, 1, 1)x(0, 1, 1, 7)	AIC:5613.298457274782
SARIMAX(0, 1, 1)x(1, 0, 0, 7)	AIC:5929.76554072578
SARIMAX(0, 1, 1)x(1, 0, 1, 7)	AIC:5657.139363190072
SARIMAX(0, 1, 1)x(1, 1, 0, 7)	AIC:5889.086558347523
SARIMAX(0, 1, 1)x(1, 1, 1, 7)	AIC:5614.914930767051
SARIMAX(1, 0, 0)x(0, 0, 0, 7)	AIC:6336.549340968721
SARIMAX(1, 0, 0)x(0, 0, 1, 7)	AIC:6142.362602315021
SARIMAX(1, 0, 0)x(0, 1, 0, 7)	AIC:6052.181242751619
SARIMAX(1, 0, 0)x(0, 1, 1, 7)	AIC:5693.789041440536
SARIMAX(1, 0, 0)x(1, 0, 0, 7)	AIC:6064.098733560891
SARIMAX(1, 0, 0)x(1, 0, 1, 7)	AIC:5750.011868844086
SARIMAX(1, 0, 0)x(1, 1, 0, 7)	AIC:5864.9159477330395
SARIMAX(1, 0, 0)x(1, 1, 1, 7)	AIC:5688.216438424369
SARIMAX(1, 0, 1)x(0, 0, 0, 7)	AIC:6321.575124299099
SARIMAX(1, 0, 1)x(0, 0, 1, 7)	AIC:6131.504821235914
SARIMAX(1, 0, 1)x(0, 1, 0, 7)	AIC:6025.157729956926
SARIMAX(1, 0, 1)x(0, 1, 1, 7)	AIC:5603.348455609514
SARIMAX(1, 0, 1)x(1, 0, 0, 7)	AIC:5929.551581941353
SARIMAX(1, 0, 1)x(1, 0, 1, 7)	AIC:5656.076344815927
SARIMAX(1, 0, 1)x(1, 1, 0, 7)	AIC:5832.609449109848
SARIMAX(1, 0, 1)x(1, 1, 1, 7)	AIC:5605.25845881479
SARIMAX(1, 1, 0)x(0, 0, 0, 7)	AIC:6317.798122090186
SARIMAX(1, 1, 0)x(0, 0, 1, 7)	AIC:6102.120102715783
SARIMAX(1, 1, 0)x(0, 1, 0, 7)	AIC:6171.983364857979
SARIMAX(1, 1, 0)x(0, 1, 1, 7)	AIC:5699.869963776844
SARIMAX(1, 1, 0)x(1, 0, 0, 7)	AIC:5991.975741137154
SARIMAX(1, 1, 0)x(1, 0, 1, 7)	AIC:5744.354806784388

SARIMAX (p, d, q) $(P, D, Q)_s$	AIC Values
SARIMAX(1, 1, 0)x(1, 1, 0, 7)	AIC:5943.232112169653
SARIMAX(1, 1, 0)x(1, 1, 1, 7)	AIC:5701.833306402739
SARIMAX(1, 1, 1)x(0, 0, 0, 7)	AIC:6133.762946777368
SARIMAX(1, 1, 1)x(0, 0, 1, 7)	AIC:5938.804783846879
SARIMAX(1, 1, 1)x(0, 1, 0, 7)	AIC:6040.265246956465
SARIMAX(1, 1, 1)x(0, 1, 1, 7)	AIC:5612.493661825187
SARIMAX(1, 1, 1)x(1, 0, 0, 7)	AIC:5870.4088259935725
SARIMAX(1, 1, 1)x(1, 0, 1, 7)	AIC:5657.60081406948
SARIMAX(1, 1, 1)x(1, 1, 0, 7)	AIC:5860.121315018765
SARIMAX(1, 1, 1)x(1, 1, 1, 7)	AIC:5614.077220561632

Table 5-7 shows the lowest two AIC values calculated. One value was within two units from the minimum value, i.e. $\Delta_i \leq 2$ ($\Delta_i = AIC_i - AIC_{minimum}$).

Table 5-7 Lowest AIC values and corresponding seasonal and non-seasonal parameters

Rank	SARIMAX (p, d, q) $(P, D, Q)_s$	AIC
1	SARIMAX (1,0,1) $(0,1,1)_7$	5603.348455609514
2	SARIMAX (1,0,1) $(1,1,1)_7$	5605.25845881479

Both models were quite similar in structure and prediction accuracy. In the end, SARIMAX (1,0,1) $(0,1,1)_7$ was chosen as the best model since it produced slightly better results.

The interpretation of SARIMAX (1,0,1) $(0,1,1)_7$ is as follows.

- $p = 1$: Model uses the peak demand value of the day before.
- $d = 0$: Model does not need differencing to make the data stationary.
- $q = 1$: Model uses one lagged forecast error.
- $P = 0$: Model does not use peak demand values of previous seasons, i.e. previous weeks.

- $D = 1$: Model needs to seasonally (one week) difference the data once to make it stationary.
- $Q = 1$: Model uses one seasonally lagged (one week) forecast error.
- $s = 7$: Model uses weekly seasonality.

The estimated coefficients of SARIMAX (1,0,1) $(0,1,1)_7$ are given in Table 5-8.

Table 5-8 Estimated coefficients of SARIMAX (1,0,1) $(0,1,1)_7$ and corresponding standard errors and p values

Term	Coefficient	Standard error	P-value
Average daily temperature	0.9433	0.232	0.000
Average daily humidity	-0.2143	0.055	0.000
AR (1)	0.9250	0.016	0.000
MA (1)	-0.5591	0.031	0.000
SMA (1)	-0.9203	0.019	0.000
σ^2	145.5655	4.706	0.000

A brief explanation of the categories in Table 5-4 is as follows.

- The coefficient column shows the importance (weight) of each term. The negative sign of a coefficient indicates a reverse relationship between the term and current value. The higher the absolute value of the coefficient is, the bigger impact that

variable has on peak demand For (σ^2), the coefficient value represents the variance of the error term.

- The standard error shows the standard deviation of the coefficient. The smaller the standard error, the more precise the estimate is. A more precise coefficient results in a more accurate peak demand model.
- The p-value shows the significance of each term's weight. A significance of 0.05 or lower shows strong evidence against the null hypothesis [52]. In this case, the null hypothesis is that the term's weight is equal to zero. If the p-value is less than 0.05, then the null hypothesis is rejected, and the coefficient's weight is deemed significant. For instance, if the p-value of the average daily humidity is more than 0.05, then the average daily humidity is deemed insignificant to the peak demand model and can be dropped.

The terms used in the model are explained next.

- The average daily temperature has the highest weight. The estimated value of 0.9433 as the coefficient for the average daily temperature can be interpreted as that one unit, in this case centigrade, increase in the average daily temperature corresponds to a 0.9433 unit, in this case kwh, increase in the daily peak demand. That is if none of the other terms in the model change.
- Regarding the daily average humidity, the -0.2143 value of the coefficient means that a one unit, in this case percent, increase in daily average humidity results in a 0.2143unit, in this case kwh, decrease in daily peak demand.

- The estimated value of 0.9250 for the autoregressive term (AR (1)) shows that if the previous value, e.g. yesterday's peak demand, increases by one, the current value, e.g. today's peak demand, increases by 0.9250.
- The estimated value of -0.5591 for the moving average term (MA (1)) shows that if the error in the previous timestamp increases by one, the current value, e.g. today's peak demand, decreases by 0.5591.
- The estimated value of -0.9203 for the seasonal moving average term (SMA (1)) shows that if the error in the previous season, e.g. previous week, increases by one, the current value, e.g. today's peak demand, decreases by 0.9203.
- σ^2 is the variance of the error term, and its value is dependent on the scale of the data.

All the coefficients are significant (p-value < 0.05) and estimated with reasonable precision, i.e. low standard errors. Therefore, all terms are kept in the model.

b. Diagnostics

The results of the diagnostic tests on SARIMAX (1,0,1) (0,1,1) $_7$ are shown in Figure 5-15.The standardized residual and correlogram show that the residuals are white noise. The KDE line on the histogram and the Q-Q plot show the residuals roughly follow a normal distribution. Therefore, the results of the diagnostic tests satisfy the assumptions of SARIMAX modelling.

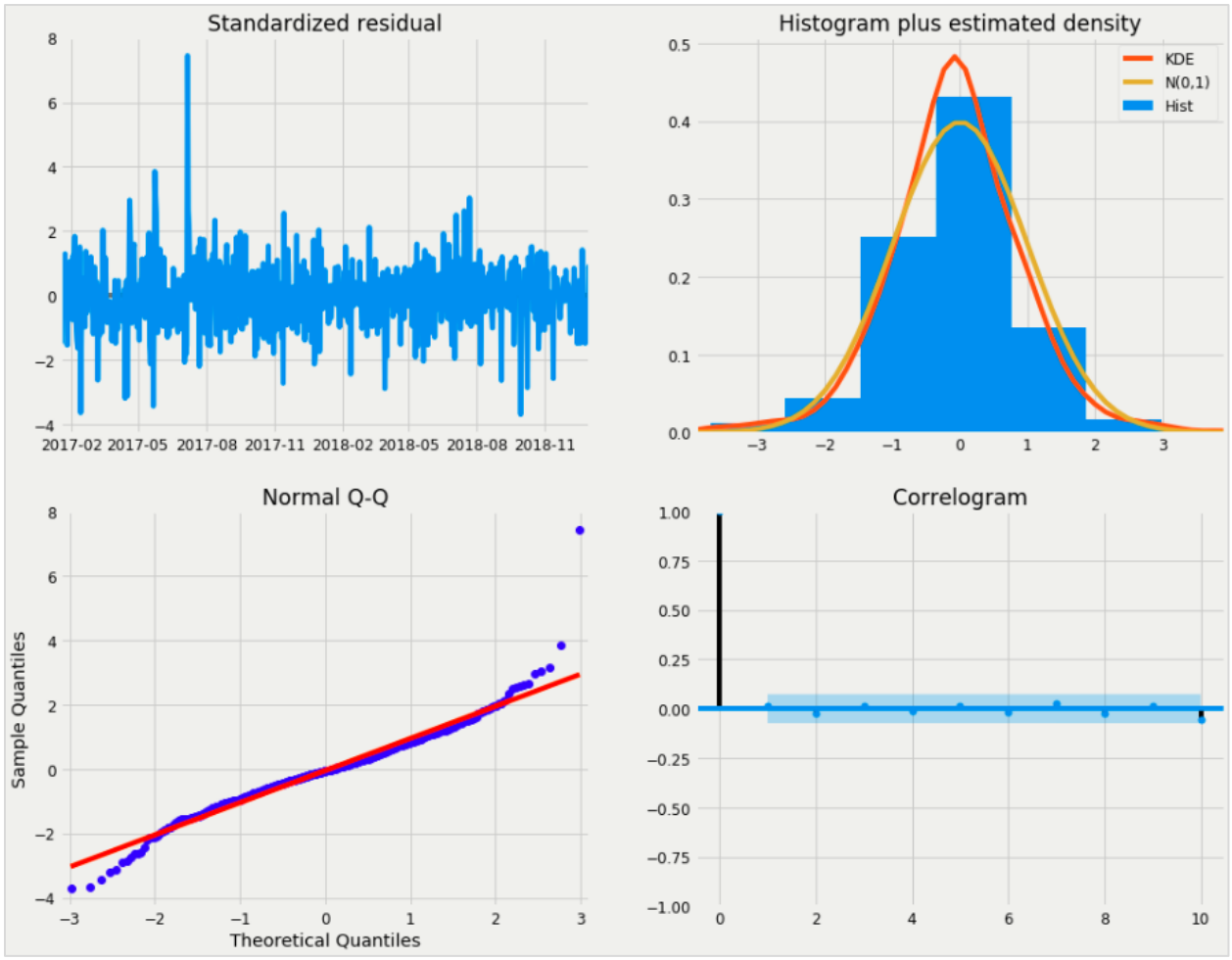

Figure 5-15 Diagnostic tests on the residuals of the model for the Alumni Center

5.3.3.3 Forecasting

Daily peak demand of the Alumni Center in 2019 was forecasted using the model from the previous step. Data for 2019 were used as the testing set and were not included in the modelling steps.

Figure 5-16 shows the forecasted values in 2019 overlaid with the actual values. The blue line shows actual values, and the orange line shows forecasted values.

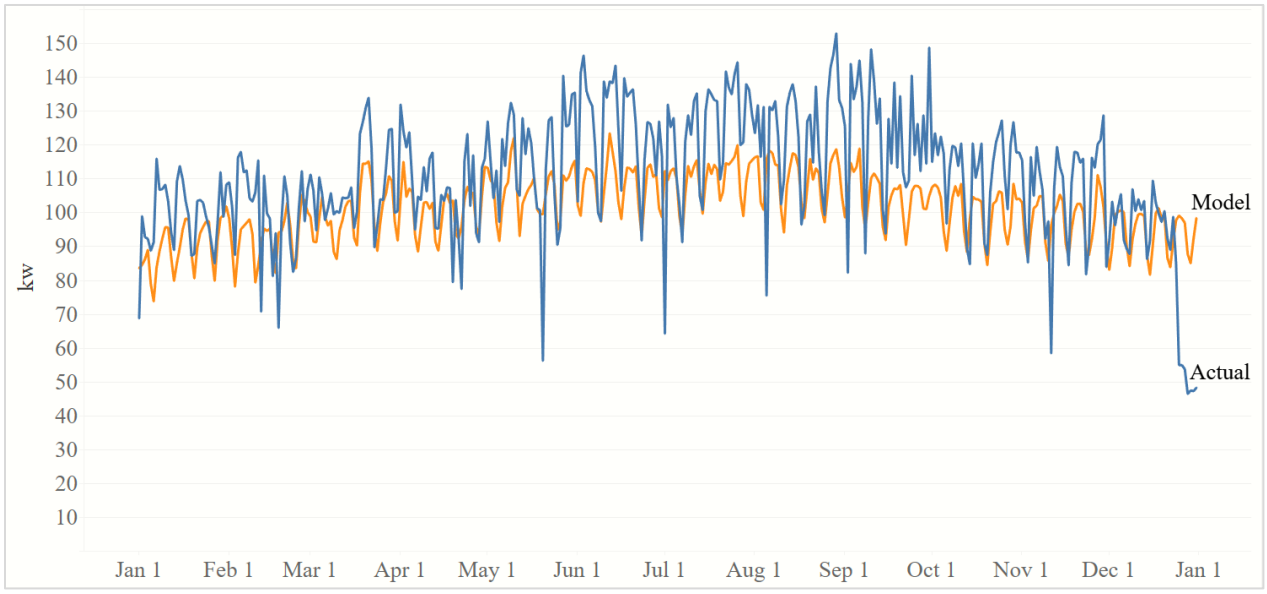

Figure 5-16 Forecasted and actual values of the daily peak electricity demand of the Alumni Centre in 2019

The model was able to capture the weekly trend successfully. However, it was not able to predict the magnitude of the spikes or dips. This area is discussed in more detail in section 5.4.

The MAPE was 12.8%. Accuracy measures are discussed in more detail in the next section. Interpretation and discussion of the results are covered in section 5.4.

5.3.4 Performance Evaluation

5.3.4.1 Performance Metrics

The performance of the model was evaluated using the testing set. The daily peak electricity demand data for 2019 were designated as the testing set. The results are shown in Table 5-9. Figure 5-17 and Figure 5-18 show the graphs of one-step ahead and dynamic forecast errors of the model.

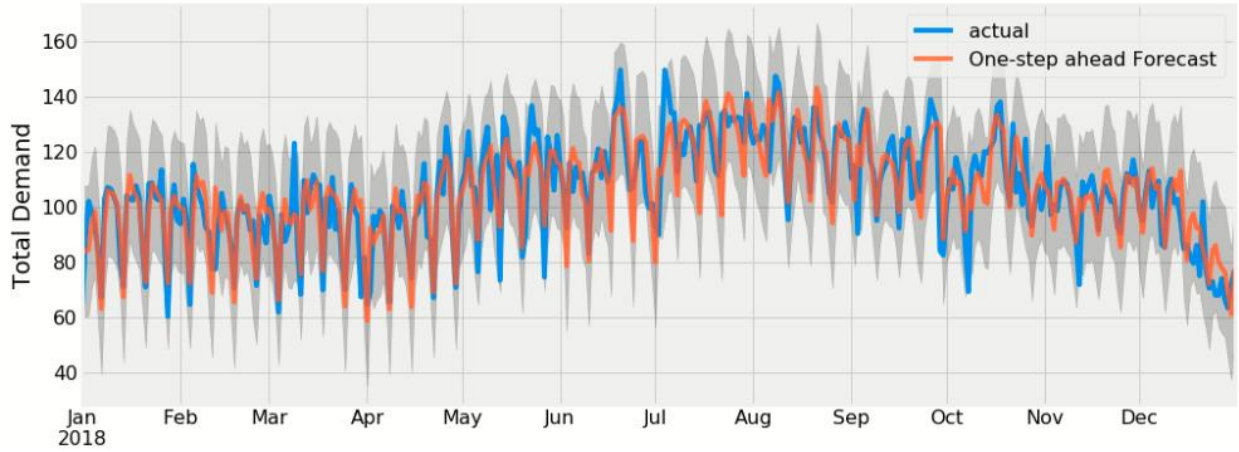

Figure 5-17 One-step ahead error of the Alumni Center model with confidence bands

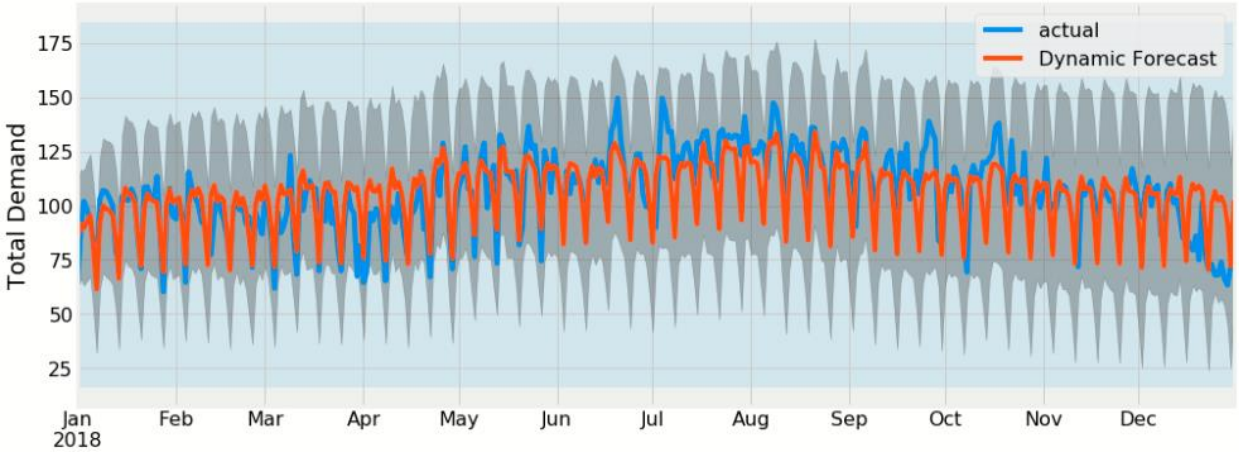

Figure 5-18 Dynamic error of the Alumni Center model with confidence bands

Table 5-9 Results of the performance evaluation of the alumni center model

	Training set: One-step ahead	Training set: Dynamic	Testing set
MSE	97.1	227.0	288.0
RMSE	9.8	15.1	17.0
MAPE	7.7%	12.2%	12.8%

Table 5-9 shows that in general, the error is lower in the training set, as it was expected. The model produced a 12.8% MAPE using the testing set. The MAPE reflects the average percentage that the predictions were off in 2019 compared to the actuals. It is not dependent on the scale of the data. The MSE shows that the average of the errors squared in 2019 was equal to 288.

The RMSE shows that the predictions in 2019 were, on average, 17 kWh off compared to the actuals in 2019. The performance of the model is discussed in more detail in the next section.

5.4 Discussion

Several factors affect electricity consumption in a building. The factors that generally have the strongest influence on electricity consumption include the end-use of electricity, weather conditions, number of occupants, and day of the week.

Buildings used as case studies in this work have similarities and differences. They are both LEED gold certified, and they both use electricity for plug loads and the HVAC system. However, they are different in their composition. The Pharmaceutical Sciences Building has a data centre, classrooms, labs, and office spaces, while the Alumni Centre is mostly compromised of a large venue for events, social spaces, and a café.

The differences in the composition of the buildings result in different consumption patterns and variables affecting the consumption. The electricity consumption in the Pharmaceutical Building followed a strict weekly pattern because of having classrooms, labs, and office spaces. The electricity consumption in the Alumni Centre followed a weekly pattern. However, it also fluctuated depending on the time of the year and the number of events in that period.

Regarding the end-use of electricity, the Alumni Centre used electricity for heating and cooling (HVAC system). The HVAC system started working when the temperature hit certain thresholds. Thus, outside air temperature and electricity consumption were not strictly proportional to each other. However, they shared a direct relationship, i.e. the consumption increased as the temperature increased. This direct relationship was captured in the positive coefficient of the average daily temperature term in the model. Similarly, the Pharmaceutical Building used electricity for its HVAC system. However, the outside air temperature and electricity consumption

had an inverse relationship in this building, i.e. electricity consumption decreased as the temperature increased. A reason behind this inverse relationship is speculated to be the fact that the building uses electricity for both cooling and heating, and captures and repurposes the heat generated in the data centre. This inverse relationship was reflected in the negative coefficient for the average daily air temperature in the model.

Another factor that typically has a strong influence on electricity consumption is the number of occupants in the building. A drop in the occupancy numbers over the weekend, which is generally associated with a decline in the level of activity, causes a significant drop in electricity consumption. The Pharmaceutical Building houses a data centre and several labs, all of which introduce substantial process loads. These types of process loads are not directly influenced by the number of occupants. For instance, the equipment in the lab runs at the same level of consumption, whether there are ten people in the lab or twenty people. Even though the direct correlation between occupancy numbers and electricity consumption was not reflected in the model, the presence of stable and predictable process loads resulted in a low MAPE for the Pharmaceutical building. In contrast, the Alumni Centre houses a large venue for events and social spaces, which by its nature introduces a high dependence on the number of occupants. This fact, coupled with the yearly patterns such as events in certain periods of the year, resulted in a higher MAPE for the Alumni Centre compared to the Pharmaceutical Building.

A final major influence is the day of the week. As the main sources of electricity consumption were not present over weekends in the two buildings, the levels of consumption dropped in both buildings. The Alumni Centre was mostly vacant over weekends resulting in a significant drop in plug loads and cooling load. The level of activity and electricity consumption

also dropped significantly in the Pharmaceutical Building over weekends as it was made up of classes, labs, and office spaces. Both models successfully predicted weekly patterns.

Traditional time-series forecasting methods such as ARIMA or ARMA work with the assumption that data are stationary and non-seasonal, and they do not use external variables in the model. In contrast, SARIMAX automatically deals with seasonality in data and works with external variables. However, SARIMAX falls short when there is multiple seasonality present in the data, e.g. weekly and yearly, or when the length of the season is large, e.g. s = 365. Trigonometric seasonality, Box-Cox transformation, ARMA errors, Trend and Seasonal components (TBATS) is a method that can handle multiple seasonality and long seasons. However, TBATS models generally suffer from slow computation time because of having to evaluate many models with many different attributes. Another technique to deal with multiple seasonality is to use Fourier terms. This method adds a layer of complexity to the model but has the potential to improve the results. Techniques from the field of machine learning, such as Neural Networks, can potentially be used to create accurate predictions. However, they provide minimal insight into how the model is learning and the function that is creating the outputs from the inputs. In contrast, using time-series models such as SARIMAX has the advantage of being easy to interpret the structure of the model, the function creating the outputs, and the results.

5.5 Summary

In chapter 5, two SARIMAX models were developed for two buildings on the UBC campus using the modified Box-Jenkins methodology introduced in chapter 4. The method used daily average temperature and humidity as external variables and predicted daily average electricity consumption and peak daily demand in the buildings. The outcome of each step of the methodology was presented. The results were interpreted, and the underlying factors behind the

varying levels of success in the two buildings were discussed. The results demonstrated that the SARIMAX method produced more accurate results in the case with a more established weekly pattern and a more significant presence of process loads. It was pointed out that the strict weekly pattern in the Pharmaceutical Building was a result of the composition of the building as it had classrooms, labs, and office spaces. It was also noted that the significant presence of process loads in the Pharmaceutical Building was a result of housing several labs and a data centre.

The models can be used in several ways by building energy managers to operate the buildings more energy and cost-effectively. The applications include detecting abnormalities in the building's electricity consumption trend, quantifying energy-saving measures, making informed decisions regarding renovation activities, and quantifying abrupt changes in the system or the behaviour of occupants. The application of the models is discussed in more detail in the next chapter.

Chapter 6: Conclusion

6.1 Applications

The models developed in this work have several impactful applications. They can be used by building energy managers for the following applications:

1. Detect abnormalities in consumption trends.

   The models can be used to calculate what the consumption values and patterns should be in the coming year. As the year starts, data can be compared to the model, and any significant deviations can be flagged as abnormal consumption.

2. Quantify energy and cost-saving measures.

   The models can be used to create forecasts of energy consumption under the current conditions. Once the energy measures take effect, the values under new conditions can be compared to that of the model, and the differences can be easily calculated.

3. Make informed decisions about the renovation or recommissioning activities in the building.

   The models can be used to forecast future values. If the consumption levels rise to a critical level in the forecast, then decisions can be made to recommission or renovate relevant parts of the building.

4. Quantify the effects of sudden changes or disruptions in the system or the way occupants behave.

   An example of such a scenario is the state of lockdown in 2020 caused by the COVID19 pandemic. The models can be used to forecast consumption in 2020. The impact of the lockdown situation can be calculated as the difference between the actuals in 2020 and the models.

Also, the methodology used in this thesis is flexible enough to be expanded to modelling other time-series data generated in buildings, such as the energy consumption of chillers.

6.2 Limitations

Some of the limitations of this research and the modelling approach are:

- The models are specific to each building and cannot be generalized to predict values in other buildings (although the methodology can be generalized to similar models for other buildings).
- The models did not consider the effects of holidays and special days (days close to holidays).
- The models did not use the number of occupants as an external variable.
- SARIMAX models do not allow for multiple seasonality in the data, for instance, both weekly and yearly trends.
- SARIMAX models do not work well with long seasons, e.g. s = 365.

6.3 Summary

Forecasting electricity demand and consumption in buildings has gained attention recently due to its potential in helping building energy managers operate buildings more efficiently and cost-effectively [4]. Time-series models use the historical and present values of a series to predict its future values. However, traditional time-series modelling techniques such as ARMA or ARIMA cannot consider external variables. In the context of buildings, several external factors such as temperature and humidity highly influence electricity consumption. As such, traditional techniques may not be able to predict future values accurately. SARIMAX is a class of time series models that considers external variables and automatically deals with seasonality in data.

The goal of this thesis was to apply the SARIMAX technique to predict electricity demand and consumption in university buildings and evaluate their performance. This was done through a modified Box-Jenkins methodology [40]. Figure 6-1 shows the steps of the modified methodology, as it was shown in chapter 4.

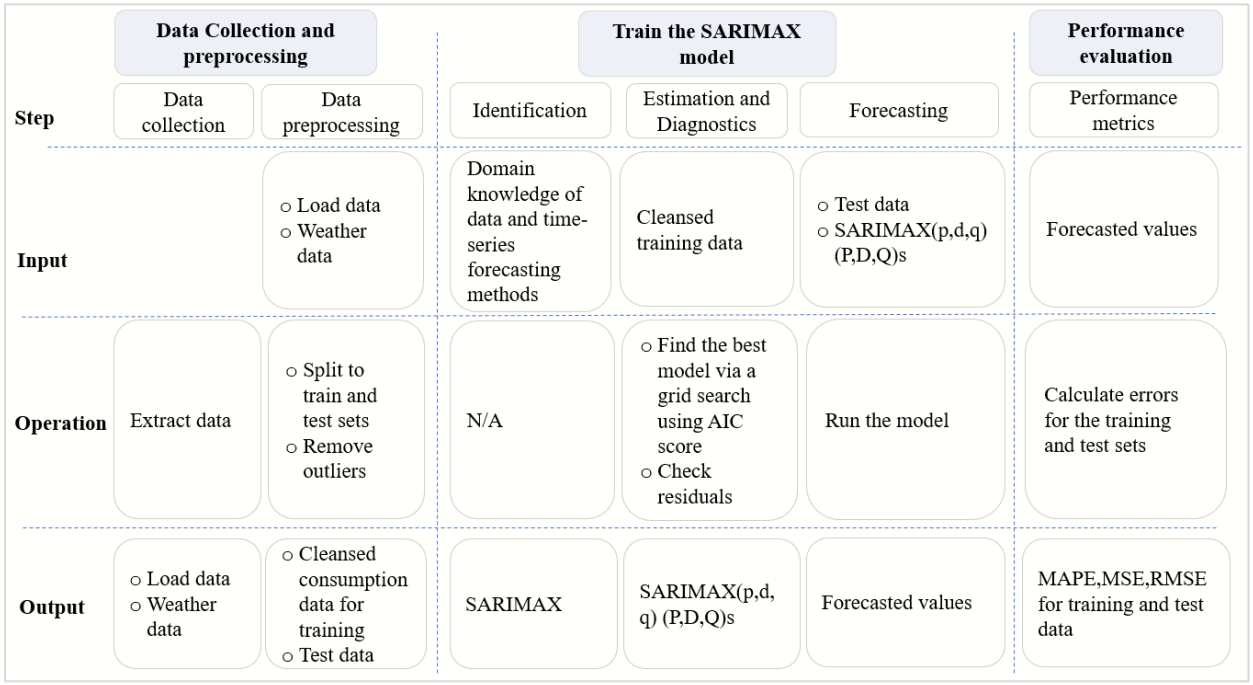

Figure 6-1 Modified Box-Jenkins methodology used in the thesis

Electricity consumption data and weather data were collected from 2017 to 2019 for two university buildings. After replacing the outliers and resampling the data, SARIMAX was identified as a suitable class of models due to seasonality in data and the effect of external variables such as temperature on electricity consumption. Next, using the 2017 and 2018 electricity consumption and weather data, a grid search was conducted for each building to find the best model from 60 candidate models. Candidate models were created from different combinations of seasonal (P, D, Q) and non-seasonal parameters (p, d, q). Each parameter was set to take a value of either zero or one. The model with the least AIC was selected. In the next stage, diagnostic tests were done on the residuals of the selected model to determine if the modelling assumptions were

satisfied. Then the model was used to predict values in 2019 using average daily temperature and humidity in 2019 as external variables. Finally, the predicted values were compared with the actual values in 2019 and forecast accuracy measures, including MAPE, MSE, and RMSE were calculated. The performance of each model is presented in the summary Table 6-1.

Table 6-1 Summary of the performance of the models for the two buildings studied

Building	Model	MAPE	MSE	RMSE
Pharmaceutical Sciences Building	SARIMAX (1,0,0) $(0,1,1)_7$	4.1%	7428.1	86.2
Robert H. Lee Alumni Centre	SARIMAX (1,0,1) $(0,1,1)_7$	12.8%	288.0	17.0

6.4 Contributions

Traditional time-series forecasting methods may not be able to accurately predict electricity consumption since this type of data is seasonal and highly affected by external variables such as temperature. This thesis applied the SARIMAX technique with two external variables, namely daily average temperature and humidity, to predict the daily peak electricity demand and the daily average electricity consumption in two university buildings.

The specific contributions of this thesis are:

1. A methodology for modelling energy consumption in buildings using time-series forecasting methods.

   The methodology was built using the Box-Jenkins methodology and outlined the steps to collect and process data, train a forecasting model, and evaluate the performance of the model.

2. Forecasting models of energy consumption in two university buildings.

   The models used temperature and humidity data as external variables and predicted daily average electricity consumption and daily peak demand in two buildings on the UBC campus. The model in the former case was substantially more accurate.

6.5 Future work

The number of occupants in the building can be used in the SARIMAX models as an external variable. However, accessing high-quality and reliable occupancy data can be quite challenging. Cooling degree days can be used instead of outside air temperature as an external variable to correlate temperature and energy consumption better. Fourier terms and TBATS models can be used to model electricity consumption with multiple seasonality. Also, considering factors such as holidays, special days (days close to holidays or long weekends), and the late December period, all of which have similar characteristics to consumption trends on weekends, can improve the accuracy of forecasts. Other avenues to explore are to apply different modelling and machine learning techniques such as neural networks and clustering to the context of this thesis and compare the results. Clustering techniques can also be used in conjunction with time-series modelling to make specific models for different clusters of consumption and possibly improve prediction accuracy.

Bibliography

[1] P. R. S. Jota, V. R. B. Silva, and F. G. Jota, “Building load management using cluster and statistical analyses,” *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 8, pp. 1498–1505, 2011.

[2] E. Kyriakides and M. Polycarpou, “Short Term Electric Load Forecasting: A Tutorial,” in *Trends in Neural Computation*, K. Chen and L. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 391–418.

[3] US Energy Information Administration, “US Energy Flow.” [Online]. Available: https://www.eia.gov/totalenergy/data/monthly/pdf/flow/total_energy.pdf. [Accessed: 21-May-2020].

[4] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*, vol. 0035, no. March. 2002.

[5] J. K. Larkin, D. Thomas, and E. Jerome, “Notes from the Field : Identifying Demand-Response Measures Customers Will Actually Do,” pp. 153–164.

[6] M. Cai, M. Pipattanasomporn, and S. Rahman, “Day-ahead building-level load forecasts using deep learning vs . traditional time-series techniques,” *Appl. Energy*, vol. 236, no. November 2018, pp. 1078–1088, 2019.

[7] P. Price, “Methods for Analyzing Electric Load Shape and its Variability,” no. May, 2010.

[8] G. J. Rios-Moreno, M. Trejo-Perea, R. Castaneda-Miranda, V. M. Hernandez-Guzman, and G. Herrera-Ruiz, “Modelling temperature in intelligent buildings by means of autoregressive models,” vol. 16, pp. 713–722, 2007.

[9] C. Hor, S. J. W. Member, and S. Majithia, “Daily Load Forecasting and Maximum Demand Estimation using ARIMA and GARCH,” no. July 2006, 2014.

[10] E. S. Handbook, “6.4.4.2. Stationarity.” [Online]. Available:

https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc442.htm. [Accessed: 25-Apr-2020].

[11] N. Citroen and M. Ouassaid, “Long Term Electricity Demand Forecasting Using Autoregressive Integrated Moving Average Model : Case Study of Morocco,” pp. 1–6, 2015.

[12] M. De Felice, A. Alessandri, and P. M. Ruti, “Electricity demand forecasting over Italy : Potential benefits using numerical weather prediction models,” *Electr. Power Syst. Res.*, vol. 104, pp. 71–79, 2013.

[13] M. Cools, E. Moons, and G. Wets, “Investigating the Variability in Daily Traffic Counts Through Use of ARIMAX and SARIMAX Models Assessing the Effect of Holidays on Two Site Locations,” pp. 57–66, 2005.

[14] G. Papaioannou, C. Dikaiakos, A. Dramountanis, and P. Papaioannou, “Analysis and Modeling for Short- to Medium-Term Load Forecasting Using a Hybrid Manifold Learning Classical Statistical Models (SARIMAX , Exponential.”

[15] F. Tas and N. Tutkun, “Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods Fatih Tas,” vol. 56, pp. 23–31, 2013.

[16] R. B. Fuller and J. McHale, “WORLD DESIGN SCIENCE,” 1975.

[17] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[18] A. Burdick, “Strategy Guideline : Accurate Heating and Cooling Load Calculations,” no. June, 2011.

[19] L. Pérez-Lombard, J. Ortiz, and C. Pout, “A review on buildings energy consumption

information," *Energy Build.*, vol. 40, no. 3, pp. 394–398, 2008.

[20] H. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," *Renew. Sustain. Energy Rev.*, vol. 16, no. 6, pp. 3586–3592, 2012.

[21] H. Hahn, S. Meyer-nieberg, and S. Pickl, "Electric load forecasting methods : Tools for decision making," *Eur. J. Oper. Res.*, vol. 199, no. 3, pp. 902–907, 2009.

[22] I. Moghram and S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques," vol. 4, no. 4, 1989.

[23] D. Srinivasan and C. S. C. A. C. Liew, "Demand Forecasting Using Fuzzy Neural Computation , With Special Emphasis On Weekend And Public Holiday Forecasting," vol. 10, no. 4, pp. 1897–1903, 1995.

[24] S. Rahman and R. Bhatnagar, "An Expert Based Algorithm for Short Term Load Forecast," vol. 3, no. 2, 1988.

[25] N. Amjady, "Short-Term Hourly Load Forecasting Using Time-Series Modeling with Peak Load Estimation," vol. 16, no. 3, pp. 498–505, 2001.

[26] A. Rahman, V. Srikumar, and A. D. Smith, "Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks," *Appl. Energy*, vol. 212, no. October 2017, pp. 372–385, 2018.

[27] A. Kimbara, S. Kurosu, R. Endo, K. Kamimura, T. Matsuba, and A. Yamada, "On-line prediction for load profile of an air-conditioning system," 1995.

[28] F. Cavallaro, "Electric load analysis using an artificial neural network," no. February, pp. 377–392, 2005.

[29] C. Anners, K. Kraus, and J. Lyhagen, "FORECASTING ENERGY USAGE IN THE INDUSTRIAL SECTOR IN SWEDEN USING SARIMA AND DYNAMIC

REGRESSION Submitted by Supervisors," 2017.

[30] P. L. Bartlett, "Introduction to Time Series Analysis . Lecture 15 .," pp. 1–15, 2010.

[31] T. A. Reddy, "A Fourier Series IVIodel to Predict Hourly Heating and Cooling Energy Use in Commercial Buildings With Outdoor Temperature as the Only Weather Variable," vol. 121, no. February, pp. 47–53, 1999.

[32] G. R. Newsham and B. J. Birt, "Building-level Occupancy Data to Improve ARIMA-based Electricity Use Forecasts," pp. 13–18, 2010.

[33] C. Chatfield, *Time-series Forecasting*. Chapman & Hall/CRC, 2001.

[34] R. Shumway and D. Stoffer, *Time Series: A Data Analysis Approach Using R*. CRC Press, Taylor & Francis Group, 2019.

[35] O. Risicobepaling, "Time Series Models for Road Safety Accident Prediction," 2012.

[36] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," 1960.

[37] R. H. Shumway and D. S. Stoffer, "Time Series Analysis and Its Applications," 2016.

[38] H. Tong, *Threshold Models in non-linear Time Series Analysis*. 1983.

[39] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods : Empirical comparisons *," vol. 08, pp. 69–80, 1992.

[40] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis*. 2008.

[41] S. Makridakis, "ARMA Models and the Box-Jenkins Methodology," vol. 16, no. May 1995, pp. 147–163, 1997.

[42] H. Liu and M. Cocea, "Semi-random partitioning of data into training and test sets in granular computing context," *Granul. Comput.*, vol. 2, no. 4, pp. 357–386, 2017.

[43] D. O. Observations, S. Author, F. E. G. Source, A. Society, and Q. S. Url, "Procedures for Detecting Outlying Observations in Samples," vol. 11, no. 1, pp. 1–21, 2011.

[44] R. K. Pearson, Y. Neuvo, J. Astola, and M. Gabbouj, “Generalized Hampel Filters,” *EURASIP J. Adv. Signal Process.*, 2016.

[45] F. Perez and B. Granger, “Jupyter Notebook,” 2020. [Online]. Available: https://jupyter.org/. [Accessed: 28-Apr-2020].

[46] J. Brownlee, “How to Grid Search ARIMA Model Hyperparameters with Python,” 2019. [Online]. Available: https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/. [Accessed: 28-Apr-2020].

[47] K. P. Burnham, D. R. Anderson, and D. R. Anderson, “Multimodel inference: Understanding AIC and BIC in Model Selection,” 2007.

[48] C. Fulton, “State space diagnostics.” [Online]. Available: http://www.chadfulton.com/topics/state_space_diagnostics.html. [Accessed: 21-May-2020].

[49] Statsmodel, “Co2 Sample dataset.” [Online]. Available: https://www.statsmodels.org/devel/datasets/generated/co2.html. [Accessed: 21-May-2020].

[50] “Leadership in Energy and Environmental Design.” [Online]. Available: https://www.usgbc.org/help/what-leed. [Accessed: 03-May-2020].

[51] “Pharmaceutical Sciences Buildings.” [Online]. Available: https://pharmsci.ubc.ca/about/facilities/about-our-building. [Accessed: 03-May-2020].

[52] D. J. Rumsey, “What a p-Value Tells You about Statistical Data.” [Online]. Available: https://www.dummies.com/education/math/statistics/what-a-p-value-tells-you-about-statistical-data/. [Accessed: 10-May-2020].

[53] “Robert H. Lee Alumni Center.” [Online]. Available: https://news.ubc.ca/2014/05/23/ubc-

reveals-plans-for-18-5-m-robert-h-lee-alumni-centre/. [Accessed: 03-May-2020].