An analysis of household water consumption in the City of Cape Town using a panel data set (2016-2020)

Anna Leah Kaplan

A Thesis Submitted for the Degree of a Data Science Masters at
University of Cape Town in 2022
Statistics Department

Under the supervision of:
Dr. Şebnem Er & Prof. Martine Visser

Abstract

Understanding consumer behaviour with respect to water consumption has become an active field of study. This thesis uses a household billing dataset that tracks the quantity of water consumed by households in the City of Cape Town (CoCT) from 2016 to 2020. The household billing data was filtered to include only household observations and then aggregated to the ward level. As a result, the aggregated data is a balanced spatial panel dataset including 20 quarterly observations for each of the 88 wards. Using the billing data set, multiple linear regression models, panel data models as well as spatial panel models were implemented to predict ward level water consumption. Using several visualisations and statistical measures, this thesis found that consumption dropped significantly during the drought period (2016-2018) and also found spatial clusters of water consumption in the CoCT. The data showed that before and after the drought, water consumption exhibited a seasonal pattern which was absent during the drought period. It is also noted that although consumption levels after the drought increase, they do not rise as high as pre-drought levels. The linear models implemented in this thesis resulted in an Adjusted R-squared values of up to 0.85, implying that the independent variables used in the models explain a large amount of variation observed in the dependent variable, quantity of ward level water consumption.

Keywords: spatial panel analysis, water consumption, City of Cape Town, spatial lag model, moran's I

Declaration

This thesis is an original work of my research and contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature: ALK

Print Name: Anna Leah Kaplan

Date: October 9, 2022

Acknowledgements

Throughout the writing of this thesis I received guidance, support and assistance from several individuals and institutions for which I am very grateful.

I would like to start of by thanking the University of Cape Town (UCT) for the funding that I received to complete this Masters.

I would like to continue by thanking my supervisor, Dr. Sebnem Er, whose expertise was invaluable in formulating the research questions and methodology. Your dedication, guidance and support throughout this process encouraged me to work harder and critically evaluate my work.

I would also like to thank all those involved from the Economics department at UCT, in particular my co-supervisor Prof. Martine Visser. Thank you for providing great insight into my topic and dedicating time towards helping me throughout my journey. Another individual who played a pivotal role in the success of my thesis is Johanna Brühl. Thank you for your guidance, patience and helping me understand and explore the data used for this thesis.

This being said, I am appreciative of the City of Cape Town for allowing me to use their data to conduct my research.

In addition, I would like to thank my parents for providing me the opportunity to complete my Masters and for their support throughout the process. My parents along with the rest of my family members have inspired me to complete this thesis and for this I am truly appreciative.

Finally, I could not have completed this thesis without the support of my partner, best friend and friends. Thank you for making me smile, laugh and stimulate discussions throughout my journey.

Contents

Abstract i

Declaration ii

Acknowledgements iii

List of Figures vi

List of Tables vii

Abbreviations viii

1 Introduction 1

2 Literature Review 5

3 Methodology 11

4 Data 20
- 4.1 Variables . . . 21
- 4.2 Missing Data . . . 24
- 4.3 Exploratory Data Analysis . . . 25
- 4.4 Outliers . . . 28
  - 4.4.1 Spatial Analysis . . . 29

5 Implementation and Results 35
- 5.1 Multiple Linear Regression Results . . . 35
- 5.2 Panel Models . . . 41
  - 5.2.1 Hausman Test . . . 41
  - 5.2.2 Panel Model Results . . . 43
- 5.3 Spatial Analysis . . . 49
  - 5.3.1 Moran's I index . . . 49
  - 5.3.2 Local Moran's I . . . 51
  - 5.3.3 Spatial Panel Model . . . 53

6 Conclusion 58

A Figures 60

B Linear Model Results **62**

C Panel Model Results **69**
C.1 Model Outputs 69
C.2 Model Outputs 69
C.3 Best Performing Panel Model results - Model 4 69
C.4 Ward Fixed Effects - Model 4 76

D Spatial Panel Model Results **79**
D.1 Best Performing Spatial Panel Model results - Model 2 79

Bibliography **85**

List of Figures

4.1 Median quarterly water consumption using the raw quarterly data 26
4.2 Average quarterly water consumption using the raw quarterly data 28
4.3 Density Map by Quantity of Water Billed, City of Cape Town 31
4.4 Map of the City of Cape Town, including ward boundaries 32
4.5 Average ward water consumption, by year 33

5.1 Model 4: Residuals vs Fitted Values 40
5.2 Model 4: Histogram of Residuals 40
5.3 Model 4: Q-Q Plot 41
5.4 Visualising in-sample prediction for wards 22, 61, 77 and 88 48
5.6 Local Moran's I, by year & quarter 53
5.7 SAR Model 2 predictions for wards 22, 61, 77 and 88 56

A.1 Average quarterly water consumption (L), by ward 60
A.2 Average quarterly amount billed (ZAR), by ward 61

C.1 Panel Model 4: Actual vs Predicted 70
C.2 Panel Model 4: Actual vs Predicted 71
C.3 Panel Model 4: Actual vs Predicted 72
C.4 Panel Model 4: Actual vs Predicted 73
C.5 Panel Model 4: Actual vs Predicted 74
C.6 Panel Model 4: Actual vs Predicted 75

D.1 Spatial Panel Model 2: Actual vs Predicted 79
D.2 Spatial Panel Model 2: Actual vs Predicted 80
D.3 Spatial Panel Model 2: Actual vs Predicted 81
D.4 Spatial Panel Model 2: Actual vs Predicted 82
D.5 Spatial Panel Model 2: Actual vs Predicted 83
D.6 Spatial Panel Model 2: Actual vs Predicted 84

List of Tables

5.1 Multiple Linear Regression Results 36
5.2 FE Panel Model Results 43
5.3 Fixed Effects - Model 4 45
5.4 Model 5: fixed effects associated with wards 26, 61, 77 and 88 47
5.5 Moran's I values for quarterly ward water consumption 50
5.6 Spatial Panel Model Results using the the Baltagi et al. (2003) methodology 54
5.7 Fixed Effects Spatial Panel Model - Model 2 55

B.1 MLR - Model 1 62
B.2 MLR - Model 2 65
B.3 MLR - Model 3 65
B.4 MLR - Model 4 68

C.1 Fixed Effects - Model 1 69
C.2 Fixed Effects - Model 2 69
C.3 Fixed effects panel model 5, the fixed effects associated with all wards . . 78

Abbreviations

CoCT	**C**ity **o**f **C**ape **T**own
FE	**F**ixed **E**ffects
RE	**R**andom **E**ffects
FD	**F**irst **D**ifference
SAR	**S**patial **L**ag **M**odels
SEM	**S**patial **E**rror **M**odels

Chapter 1

Introduction

Understanding consumer behaviour with respect to water consumption has become an active field of study. Being able to understand trends, patterns and behavioural traits in this domain leads to significant economic and environmental benefits. Accurately predicting water demand or distinguishing spatial clusters, geographical areas sharing a particular trend or pattern, in consumption allows policy makers to plan and implement tariffs where needed, ensuring a reliable water distribution system is in place. Failing to have a sound understanding of water consumption and demand patterns may result in the mismanagement of water supply and control. A classic example of this would be when a natural disaster occurs, such as a drought, resulting in the government needing to control the supply of water to ensure city's dams and other sources of water do not run dry.

Spatial econometrics deals with statistical methods that exploit the interactions among geographical units. Over the past years, this has been an area in econometrics that has been explored rigorously. Acknowledging, in many cases, geographic units that are close to one another are more likely to exhibit similar behavioural traits than those that are further from one another is essential for economists when designing policies and managing water distribution. For example, one would believe that the behaviour of households living near a natural source of water, like a flowing river, would have similar water consumption behaviour. Which would most likely be different from the consumption behaviour of households living in the city, where the main source of water would be the municipal reservoirs.

Having access to historical data allows for a better understanding of consumption behaviours over time. There are several statistical models used for modeling time series data and in particular panel data. Thus allowing for the prediction of future events based on past occurrences in the data. Using such predictive statistical models to understand water consumption demand would allow governments to adequately prepare and manage water supply ensuring water sources are not exhausted and water is efficiently distributed.

The water consumed by households in the CoCT is stored in one of the 14 dams located in different areas in and around Cape Town. The dams in and around Cape Town form part of the Western Cape Water Supply System. Not only do these dams service Cape Town, the system supplies water to towns in the Overberg, Boland, West Coast, and Swartland areas, and provides irrigation water for agriculture. The six major dams supplying Cape Town are: Berg River, Steenbras Lower, Steenbras Upper, Theewaterskloof, Voëlvlei, Wemmershoek. When these dams are full they can store up to 898,221 megaliters (Ml) of water (CoCT 2018*a*). On its way to the city, the water passes through pipelines tunnelling through various terrains and is cleaned at one of the city's 12 treatment plants (CoCT 2018*c*).

Cape Town is located in a water-scarce region vulnerable to droughts and has recently recovered from one of the worst droughts in the city's history. In 2018, Cape Town recorded some of the lowest dam levels ever reported. The six major surface water dams that supply the CoCT dropped from 98.1% filled capacity in 2014 to 25.1% in 2017 (CoCT 2018*b*). It is important to note that the last 10% of the aforementioned dams' water is difficult to use for daily consumption due to the quality of water stored. In early 2018, dam levels dropped even more and the city was at risk of reaching "Day-Zero", the day when the municipal piped system would stop delivering water to homes and businesses. Cape Town almost became the first major city in the world to exhaust its water sources supplying homes and businesses (Bruhl, le Roux, Visser & Köhlin 2020). This is one of the main reasons it is critical to have a clear and sound understanding of water behaviour in the city and its surrounding areas. This historical drought provides a great use case for the application of predictive modeling in this particular industry. Using weather forecasts and being able to predict and understand the demand for water allowed the CoCT to prepare and implement different policies and tariffs, when the supply of water was at risk.

Cape Town is not the only city that has experienced water insecurity as the topic of water demand is well explored across the world (Cheng et al. 2009, Romano & Kapelan 2014, Walker et al. 2015, Adamowski et al. 2012, Bakker et al. 2014). Several researchers have explored water demand in an attempt to help governments understand the demand for water and simultaneously control the supply of water. This thesis is influenced by the work done by South African researchers as well as the methods adopted by researchers across the world. Due to the nature of the data set used in this thesis being one characterised as a spatial panel data set, a spatial analysis using different ward behaviour in the CoCT will be conducted.

Although non-linear methods, models that describes nonlinear relationships in experimental data using non-linear statistical equations, have recently proven to be powerful predictive approaches for dealing with time series forecasting, they often require a large amount of historical data. The data used in this thesis undergoes a data management process, after which there are five years of quarterly data for each of the 88 wards in the CoCT (20 observations per ward and 1760 observations in total). This is not necessarily enough data for some of the more recent non-linear models such as Recurrent Neural Networks (RNN) and Long Short Term Neural Networks (Brownlee 2017). However, the data with the aforementiond dimensions can be used for linear models. Linear models are still powerful tools used by many researchers and arguably allow for better interpretations of model effects.

The primary motivation of this thesis is to perform a thorough spatial analysis of water consumption in the CoCT and implement several linear models, using water consumption as the dependent variable (variable of interest). The intention of this thesis is that the results generated will be able to help in designing/developing effective policies for the management of water services in the CoCT.

There is a vast amount of research on water behaviour in the CoCT and other major cities around the world. Some research projects have influenced the way in which populations consume water by implementing behavioural nudges (Bruhl, Serman & Visser 2020, Brick et al. 2017) and others have been able to guide policy makers through demand management (Bruhl, le Roux, Visser & Köhlin 2020). To my knowledge, no work has been published that explores a spatial analysis of the CoCT's historical (panel) data. This thesis aims to fill this gap and complement prior research on the demand for water

by building a predictive model for water consumption using the spatial panel data of household water billing data that spans over a five year time frame (2016-2020).

The remainder of the thesis is organised as follows: Section 2 discusses the literature on both water consumption and the modeling of spatial panel data. The methodology implemented is described in Section 3. Section 4 explores the data used in this thesis through several visualisations and address the topics of data pre-processing. Discussion and analysis of the results are provided in Section 5 and, finally, the thesis is concluded in Section 6.

Chapter 2

Literature Review

When using historical CoCT water consumption data, one needs to understand and take into consideration the drought experienced from 2016 to 2018. Bruhl & Visser (2021) were able to analyse and focus closely on the drought experienced in Cape Town and how water behaviour changed during this period. The researchers used a municipal utilities database to create a panel data of monthly water usage for approximately 300,000 domestic freestanding households in the city. Using the pre-drought year (June 2014 – May 2015) consumption behaviour, the researchers compared changes in water consumption observed during the city's drought years. Bruhl & Visser (2021) found that as a result of the CoCT repeatedly increasing water tariffs, imposing severe restrictions on water usage, reducing water pressure and using several media campaigns urging people to save water, there was an approximate 50% water usage decrease in less than three years (Bruhl & Visser 2021) in comparison to the June 2014 – May 2015 observations. The large part of the researchers four-year investigation period (June 2014 to May 2018) falls within the same time frame as the data explored in this thesis investigation period (January 2016 to December 2020). Their findings are therefore influential to the results of the current paper.

Several approaches have been developed in order to forecast future events based on past data. This thesis has been influenced and adopts methods found in different papers, across different fields of studies. Apart from work related to forecasting water demand (Tso & Yau 2007), literature focusing on predicting electricity consumption (Shine et al. 2018*b*,*a*, Tso & Yau 2007) and petroleum production have been examined. Comparing

water and electricity consumption can be done due to the similarity in consumption behaviour and the way in which the of data are collected and monitored over time. Just like water consumption is affected by seasonal changes, so too, is electricity consumption. The difference in the consumption patterns of the two commodities is manifested in different consumption behaviour during seasonal changes. In the cold, wet winter months (typically May - August) in Cape Town, one would expect household water consumption to be lower than in the hot summer months (typically November - February) (Bruhl & Visser 2021). However the opposite is true for electricity consumption for which in Cape Town's cold winter months, electricity consumption peaks and is in high demand as households use different methods to keep warm (whether it be boiling hot water, using heaters or underfloor heating) thus increasing individuals' electricity consumption and electricity bills (Eskom 2017). For the reasons stated above, this thesis has been able to learn and implement methods that have been used to predict electricity consumption.

Yea et al. (2018) use tariff data from the National Energy Regulator of South Africa, merged it with the 2010/2011 South African Income and Expenditure Survey in order to examine the standard features of electricity demand, income and price effects. The researchers used monthly household electricity consumption as the dependent variable for their study in an effort to highlight potential avenues for intervention in the residential sector. Using the log of consumption in their second-stage ordinary least squares (OLS) model the researches found that household demand was higher for appliance-rich households in urban areas, especially if there are more household members dwelling in a larger residence. The data used by Yea et al. (2018) was cross-sectional, meaning that time was not considered as a study variable (Yea et al. 2018). Although this type of data only focuses on observations belonging to different households at a single time period, as opposed to the data used in the current thesis using observations from various households at 20 quarterly time periods, the findings in Yea et al. (2018) study are still influential to the research conducted in this thesis. The reason being that there is a similarity between water and electricity consumption and as a result the independent variables used by Yea et al. (2018) could influence the ones used in this thesis.

In 2018, Shine et al. (2018*b*) published their work on multiple linear regression (MLR) modelling for predicting on-farm direct water and electricity consumption for 58 pasture based dairy farms in Ireland, with data taken over a two year time period (2014-2016). MLR models describe the linear relationship between multiple predictor variables in the

prediction of a single dependent variable. As a result the researches do not analyse the study as a time series forecasting problem, rather MLR equations were developed for each month and described the best fit line which minimised the sum of the squared error of the vertical deviations from each observed data point to the line. In total, 15 and 20 dairy farm variables were analysed for their predictive power of monthly electricity and water consumption, respectively. The researchers included variables relating to the number of cows on the farms as well as the ambient temperature levels. If possible, this thesis could aim to use similar features such as the total number of household per ward and features relating to weather over time. The researchers removed outliers from the study due to the inherent nature of the data recorded, where many external factors such as, but not limited to, meter faults, leakage and human error may have an impact on the model weight calculations. The same holds true for this thesis. The researchers suggest that the findings and models developed in their study can be used by governing bodies.

Shine et al. (2018*a*) extend their research by using the same data used in their previous study exploring MLR modelling of on-farm direct water and electricity consumption on pasture based dairy farms (Shine et al. 2018*b*) by implementing multiple machine learning algorithms such as Classification And Regression Trees (CART), random forest ensemble, artificial neural networks (ANN) and a support vector machine. Shine et al. (2018*a*) also implement a variable selection technique known as backward sequential variable selection in an attempt to increase the model prediction performance. The ML algorithm that resulted in the most accurate monthly water and electricity prediction was the random forest and support vector regression, respectively. The researchers found that the developed machine learning models, including the ANN model, improved the prediction accuracy for both electricity and water consumption, when compared to results previously obtained using a MLR approach (Shine et al. 2018*b*). The ANN in the study was out-performed by the two aforementioned ML techniques, this finding is not necessarily surprising as one would expect a classic neural network successor, a RNN, to handle time series better. Although the current thesis does not report on any of the aforementioned ML models, a basic decision tree was implemented but spatial effects were not taken into consideration. This provides an opportunity for future researchers to build on the work conducted in this paper.

The standard approach in most empirical work, dealing with spatial data is to start with a non-spatial linear regression model and then test whether or not the model

needs to be extended with spatial interaction effects. This approach is known as the general-to-specific approach (Elhorst 2010, Millo et al. 2012, Salima et al. 2018) and has influenced the way in which this thesis conducts its analysis. Elhorst (2010) produced work on spatial panel models and adopted the general-to-specific approach (Elhorst 2010). Elhorst (2010) empirical illustration of the models discussed in his paper is the same problem as the one used by Baltagi & Li (2008). The spatial panel data includes 46 states in America in which real per capita sales of cigarettes measured in packs of cigarettes per capita (C_{it}) was regressed on the average retail price of a pack of cigarettes measured in real terms (P_{it}) and on real per capita disposable income (Y_{it}). Elhorst (2010) used 29 observations per each state in the United States (U.S.) (1963-1992) where Baltagi & Li (2008) used the first 25 years of the panel dataset (Baltagi & Li 2008). Elhorst (2010) was able to show that spatial econometric models that include lags of the dependent variable and of the independent variables in both space and time provide a useful tool to quantify the magnitude of direct and indirect effects, both in the short term and in the long term. The current thesis intends on following a similar methodology to Elhorst (2010) and implementing some of the spatial panel models described and explored in his paper.

Later in 2018, Salima et al. (2018) follow a similar approach to Elhorst (2010) and investigate the various models that can be implemented when working with panel data (Salima et al. 2018). The empirical application used in their work pertains to Verdoorn's second law (Verdoorn 1980). This law links, in a linear fashion, labour productivity growth rates with those of output in the manufacturing sector for a range of economies. The authors make use of the statistical computing and graphics packages housed in the R environment (R Core Team 2013). The two main packages, used in the R environment, that are required for the estimation of panel data models are *plm* (Croissant & Millo 2018, 2008, Millo 2017) and *splm* (Bivand et al. 2021, Millo et al. 2012). To select the most appropriate specification for the panel data set, including six periods for 1032 European regions, Salima et al. (2018) start with a model without spatial autocorrelation and implement both the Hausman (Durbin 1954, Wu 1973, Hausman 1978) and Lagrange multiplier test. The Hausman test was used to determine whether a fixed effects or random effects model should be used (Salima et al. 2018). When using the Hausman test for plm data, Random effects are preferred under the null hypothesis due to higher efficiency, while under the alternative Fixed effects is at least as consistent and

thus preferred (Salima et al. 2018). Salima et al. (2018) reject the null hypothesis and consequently the empirical analysis that followed was a fixed effects model. The authors compare the results of various iterations of a fixed effects model with different estimation methods (maximum likelihood versus the (Kelejian & Prucha 2007) methodology). The results generated, using the *splm* package (Millo et al. 2012, Bivand et al. 2021), show a positive and significant autocorrelation coefficient. This thesis intends on following a similar approach and structure to the one displayed by Salima et al. (2018).

Millo et al. (2012) provide a general description of the *splm* R package (Millo et al. 2012). The authors make use of a well known example from Munnell (1990) with productivity data for 48 U.S. states observed over 17 years. Millo et al. (2012) implement both fixed effects and random effects models using the productivity data set with the intention to display the results generated from functions in the *splm* package. The authors explain the mathematics behind the various spatial panel models and discuss the results generated in R. For example, when implementing a fixed effects model they discuss what the outputs and parameters are in the model such as ρ and the coefficients of the the spatially lagged dependent variable. This thesis makes use of the findings and analysis conducted in Millo et al. (2012) by using it as a guide to navigate through the *splm* package.

Kelejian & Prucha (2007) consider a panel data model with error components that are both spatially and time-wise correlated. The researchers analysis is geared towards samples where N is large relative to T (N being the number of observations per cross section time period and T being the number of time periods). Using a Monte Carlo experiment with $N = 100$ and $T = 5$, the authors compare the model that they consider - a linear panel data model that allows for disturbances to be correlated over time and across spatial units (Kelejian & Prucha 2007). The findings and model produced by Kelejian & Prucha (2007) is used in this thesis.

Kiziltan (2021) investigated the effects of water consumption on electricity consumption in Turkey's provinces over the period 2008-2018, using a balanced panel sample, a dataset in which each panel member (i.e., province) is at each time point. The study included 486 observations. Kiziltan (2021) used Moran's I index to detect whether or not spatial autocorrelation existed amongst the observations in the Turkish electricity and water consumption dataset. The preliminary results of the analysis indicated the existence of spatial autocorrelation and dependence amongst the provinces electricity consumption.

Based on these results, Kiziltan (2021) continued his research by the implementation of models that captured these spatial features. Kiziltan (2021) compares the results from a spatial autoregressive model (SAR) and a Spatial Durbin Model (SDM) using the Likelihood-Ratio test and the Akaike Information Criteria (AIC). The model with the lowest AIC was the SDM. Using this model, the results showed that provincial income, population, total water abstraction by municipalities in the provinces, and provincial industrialisation positively affect the provinces' total electricity consumption. It should be noted here that being able to identify which independent variables affect the dependent variable is one of the advantages of using spatial panel models for data that are time dependent. Using non-linear models may improve accuracy but simultaneously makes it harder for the interpretability of model results. This thesis will follow the same methodology as the one found in the work done by Kiziltan (2021), with respect to the methodology implemented.

From this chapter, it is evident that the current thesis was influenced by a range of prior research. The pre-processing decisions made in this thesis were influenced by research conducted in the past, such as but not limited to Shine et al. (2018*b*,*a*) and (Yea et al. 2018). The general-to-specific approach structure displayed in Chapter 3, as well as the models implemented in Chapter 5 were influenced by the aforementioned work conducted by Elhorst (2010), Millo et al. (2012), Salima et al. (2018), Baltagi et al. (2003) and Kiziltan (2021). Apart from the above mentions, it should be noted that this thesis works very closely with the findings and guidance implemented in Millo, Piras & Bivand (2012). In particular, the research conducted by Millo et al. (2012) helped provide structure and guidance to the current thesis when dealing with both the panel and spatial element present in the data set used for this thesis. This thesis intended on filling the predictive research gap present in water consumption, in the CoCT. As explained above, the thesis leverages off the methods implemented by prior researchers and complements the research by Bruhl & Visser (2021) who focus on water consumption in the CoCT during the same time period as the one investigated in this thesis.

Chapter 3 follows on from the current Chapter by discussing the general-to-specific methodology implemented in this thesis and highlights the statistical models used in this thesis.

Chapter 3

Methodology

When dealing with panel data the most commonly estimated models are fixed effects and random effects models. Several considerations affect which model is most appropriate for a particular use case. This thesis follows the same standard approach taken in most empirical work by starting with a non-spatial linear regression model, ignoring the fact that the data being used is panel in nature. In order to implement the linear models the *lm* function in the R *stats* package (R Core Team 2013) is used. From there, non-spatial panel models are implemented using the *plm* package (Croissant & Millo 2018, 2008, Millo 2017). The residuals are then examined to see whether or not spatial correlation is present, if the residuals are spatially auto-correlated, this indicates that the model is misspecified and consequently the model estimates could include some sort of bias. Signs of spatial autocorrelation in residuals calls for spatial panel models to be implemented. This was implemented using the *splm* package (Millo et al. 2012, Bivand et al. 2021).

The first model implemented in this thesis is one that ignores the panel and spatial element of the dataset. The equation for the linear regression model is as follows (Underhill & Bradfield 2013):

$$y = X\beta + \epsilon \tag{3.1}$$

Where:

$$\epsilon \sim N(0, \sigma^2 I) \tag{3.2}$$

In Equation 3.1:

- y is the n x 1 outcome vector
- X is the n x $(k+1)$ design matrix of independent predictor variables (including a vector of ones corresponding to the intercept)
- β is the k x 1 matrix of parameters

There are three types of regression for panel data: PooledOLS, Fixed Effects (FE) Model and Random Effects (RE) Model. The equation for the PooledOLS model regression is Equation 3.1. The problem with using PooledOLS for panel data is that one of the key assumptions made about the model is often violated - homogeneity. Heterogeneity is the unobserved dependency of independent variables that often leads to biased results. Even if the homogeneity assumption holds true for PooledOLS, the individual error term might have a serial correlation over time and therefore makes it inappropriate to use PooledOLS for panel data as the model estimates may contain bias.

The FE and RE models on the other hand take into consideration both time and individual characteristics. The FE model determines individual effects of unobserved, independent variables as constant over time while RE models determine individual effects of unobserved, independent variables as random variables over time. The fixed effect models assume that the explanatory variable has a fixed or constant relationship with the response variable across all observations. By assuming the above, FE includes a dummy for each individual (in this thesis ward) and consequently assumes each ward has its own intercept but same slope. A random-effects model also assumes that explanatory variables have fixed relationships with the response variable across all observations, but that these fixed effects may vary from one observation to another.

In order to decide between FE or RE the Hausman test can be used. When dealing with panel data, the null hypothesis for the Hausman test favours the RE model while the alternative hypothesis implies that a FE model should be used as the estimators are consistent (Baltagi et al. 2003). The Hausman test was implemented on the data used for this thesis and resulted in a Chi-Squared value of 139.99 with an associated *p-value = 0.00*. The aforementioned results implied the null hypothesis should be rejected and consequently the FE model is preferable and will be used for both the panel and spatial panel models that follow.

The FE model, controlling for individual effects is as follows (Millo et al. 2012):

$$y_{it} = \beta_{it}^{T} x_{it} + \alpha_i + \mu_{it} \tag{3.3}$$

where i=1,...,n is the individual (ward) index, t=1,...,T is the time index.

In Equation 3.3:

- y_{it} is the dependent variable observed for ward i at time t (for this thesis t = quarterly observations)
- X_{it} is the time-variant (1 x k) independent variables regressor vector.
- β is the k x 1 matrix of parameters.
- α_i is the unobserved time-invariant individual effect.
- μ_{it} is the error term with a random disturbance term of mean 0.

Since α_i is not observable, it cannot be directly controlled for. The FE model eliminates α_i by demeaning the variables using the within transformation (Baltagi 2005):

$$y_{it} - \overline{y}_i = \beta(X_{it} - \overline{X}_i) + (\alpha_i - \overline{\alpha}_i) + (\mu_{it} - \overline{\mu}_i) \tag{3.4}$$

Seen as α_i is the time-invariant individual effect and fixed over time, the within transformation equation is:

$$\ddot{y}_{it} = \beta\ddot{X}_{it} + \ddot{\mu}_{it} \tag{3.5}$$

The FE estimators (β) are then obtained by an OLS regression of $\ddot{y}$ on $\ddot{X}$, the demeaned variables.

The main benefit of FE estimations in comparison to PooledOLS is that the potential sources of biases in the estimations are limited. This being said there are several important limitations of FE estimations. One of the important limitations of the FE model, in the current thesis, is the models ability to handle spatial data and the associated spatial correlation.

When spatial effects, are suspected amongst observations, it is essential to continue the analysis taking this into consideration. Spatial effects refer to spatial dependence in empirical data including spatial autocorrelation and spatial heterogeneity. In the context of this thesis, spatial effects are suspected amongst neighbouring wards - this meaning the thesis explored whether or not wards share similar behavioural patterns in water consumption. In order to explore these spatial effects, one needs to define spatial neighbours and create a spatial weights matrix. The structure of the interactions between each pair of spatial units is represented by means of a spatial weights matrix, denoted as **W**. Neighbourhood relationships can be defined in several ways, including but not limited to contiguity weights and distance weights. This thesis makes use of a contiguity-based neighbourhood. A contiguity matrix is based on the fact that two spatial units are simply neighbors to each other based on the mere adjacency between two polygons. In this case two polygons, indexed by S_i and S_j are said to be neighbours if they share a common border. This relationship is studied in various manners. Using Rook Contiguity it can be said that two polygons, S_i and S_j, are neighbours if they share the same border. On the other hand, Bishop Contiguity states that two polygons are spatial neighbours it they meet at one vertex. Queen Contiguity is a combination of Rook and Bishop Contiguity. Identifying neighbourhoods using the Queen Contiguity results in two polygons being classified as neighbours if they share any border (Arbia 2006). This thesis explores different spatial weights matrices and reports on the weights that ultimately perform the best.

The spatial weights matrix, **W**, is then created according to the type of neighbourhood relationship defined above. **W** is a **N**x**N** positive symmetric matrix where the non-zero elements, w_{ij}, indicates whether two locations in the data are neighbors. w_{ij} represents the degree of spatial relationship between the units i and j, where:

$$w_{ij} = \begin{cases} 1, & \text{when } i \neq j \text{ and i, j are neighbours} \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

The weights matrix is generally used in row standardized form, meaning the cell value is divided by the sum of the rows in which it is located (Kiziltan 2021). This is done in order to create proportional weights in cases where features have an unequal number of neighbours. The weights matrix in this thesis is constructed using polygons of ward

shapes in the CoCT, it is recommended to apply row standardization when doing the above (Millo et al. 2012).

Working in the R environment, the classic package used to create weight matrices is *spdep* (Bivand & Wong 2018, Bivand et al. 2013). The *nb2listw* function in *spdep* supplements a neighbours list with spatial weight. This function allows row standardization by setting the style parameter to w (Kelejian & Prucha 2010, Tiefelsdorf et al. 1999, Bivand et al. 2013).

Using the spatial weights matrix defined above, a local Moran's I is calculated as a means of understanding spatial autocorrelation. It was developed by Anselin (1995) as a class of local indicators called Local Indicators of Spatial Association (LISAs) (Anselin 1995). The local Moran's I statistic offers insight into the behaviour of data at local levels, by providing a decomposition of the Moran's I global statistic into the degree of spatial association associated with each observation. The statistic identifies local clusters or local outliers to understand their contribution to the 'global' clustering statistic. A local Moran statistic for an observation i may be defined as (Anselin 1995):

$$I_i = z_i \sum_j w_{ij} z_j \tag{3.7}$$

where:

- z_i, z_j are in deviations from the mean
- the summation over j is such that only neighboring values $j \in \mathrm{J}_i$ are included
- w_{ij} represents the degree of spatial relationship between the cross sectional units i and j

For this spatial statistic this thesis makes use of a queen contiguity-based spatial weights to calculate the statistic. This decision was made as the queen contiguity is a combination of both the Rook and Bishop contiguity matrices thus it is assumed that if any spatial relationships exist using Rook or Bishop contiguity matrices, so too will they be present with queen.

Visualizing the local Moran allows for us to view spatial clusters within the dataset (Anselin 1995). Visualizing the Local Moran's I statistic for each quarter of each year in

the current data set, shows that there is clustering over time amongst the wards. This provides a justification as to why one may not exclude/ignore the spatial element within the data. Ultimately, this thesis continues by implementing spatial models.

The spatial models implemented in this thesis are defined below. Both these models are fitted to the spatial panel data used in this thesis. The model that performs the best, has the lowest root mean square error (RMSE) is then chosen. The RMSE is the standard deviation of the residuals, which are the models prediction errors. RMSE is a measure that tells you how concentrated the data is around the line of best fit.

The first model is the fixed effects spatial lag model (SAR), the version of this model was introduced by Baltagi et al. (2003). The equation for this model can be written in stacked form as follows (Millo et al. 2012):

$$y = \lambda(I_T \otimes W_N)y + (i_T \otimes I_N)\mu + X\beta + \epsilon \tag{3.8}$$

Where:

$$\epsilon \sim N(0, \sigma_\epsilon^2) \tag{3.9}$$

In Equation 3.8 and 3.9:

- λ is the spatial autoregressive coefficient, when equal to zero there is no evidence of spatial dependencies
- W_N is a non-stochastic spatial weights matrix
- i_T is a column vector of ones of dimension T
- I_N is a N x N identity matrix
- X is the (1 x k) independent variables regressor vector
- β is the k x 1 matrix of parameters
- ϵ is a well-behaved error term

The coefficient estimations for these models are done through a maximum likelihood approach where general estimation theory for maximum likelihood resembles the cross-sectional case (Millo et al. 2012). In Equation 3.8 the presence of the spatial lag introduces a form of endogeneity that violates the assumption of standard regression models. In order to maximize the likelihood function Elhorst (2003) suggests transforming the variables in Equation 3.8 by eliminating the time invariant individual effects. The transformation is obtained by subtracting the average for each cross-section over time which consequently removes terms that do not vary over time to be removed from the model (Millo et al. 2012).

This transformation is represented in Equation 3.10

$$y^* = \lambda(I_T \otimes W_N)y^* + X^*\beta + \epsilon^* \tag{3.10}$$

where:

- y^* is equal to $Q_0 y$
- X^* is equal to $Q_0 X$
- ϵ^* is equal to $Q_0 \epsilon$
- $Q_0 = (I_T - \frac{J_T}{I_T}) \otimes I_N$

The log-likelihood function of Equation 3.8 is:

$$L = -\frac{NT}{2} ln(2\pi\sigma_\epsilon^2) + T ln|I_N - \lambda W_N| - \frac{NT}{2\sigma_\epsilon^2}\epsilon^\top \epsilon \tag{3.11}$$

where:

- $\epsilon = y - \lambda(I_T \otimes W_N)y - X\beta$
- $|I_N - \lambda W_N|$ is the Jacobian determinant

A numerical optimisation procedure is needed to obtain the value of λ that maximises Equation 3.15. Finally, estimates for β and σ_ϵ^2 are obtained from the first order conditions of the likelihood function by replacing λ with its estimated value from the maximum likelihood (Millo et al. 2012).

Another popular spatial panel model is the spatial error model (SEM), which once again is the version introduced by Baltagi et al. (2003). Using Baltagi et al. (2003) model equation, the fixed effects SEM can be written as (Millo et al. 2012):

$$y = (i_T \otimes I_N) + X\beta + \mu \tag{3.12}$$

Where:

$$\mu = (I_T \otimes I_N)\mu + \epsilon \tag{3.13}$$

and

$$\epsilon = \rho(I_T \otimes W_N)\epsilon + v \tag{3.14}$$

In Equation 3.12 and 3.13:

- ρ is the spatial autocorrelation coefficient
- W_N a non-stochastic spatial weights matrix
- i_T a column vector of ones of dimension T
- I_N an N x N identity matrix

As done with the SAR model above, a concentrated likelihood approach can be taken but an iterative procedure is needed to estimate the parameters of the spatial error model. According to Millo et al. (2012) the idea is to iterate between maximum likelihood and generalized least squares until a convergence criterion is met.

The log-likelihood function for the model stated in Equation 3.12 can be written as (Millo et al. 2012):

$$L = -\frac{NT}{2}ln(2\pi\sigma_\epsilon^2) + Tln|B_N| - \frac{1}{2\sigma_\epsilon^2}\epsilon^\top[I_T \otimes (\mathbf{B}_N^\top B_N)]\epsilon \tag{3.15}$$

where:

- $\epsilon = y - X\beta$

- $B_N = (I_N - \lambda W)^-1$

Estimators for β and σ_ϵ^2 are then derived from the first order conditions as (Millo et al. 2012):

$$\beta = [X^T(I_T \otimes \mathbf{B}_N^\top B_N)X]^{-1}\mathbf{X}^\top(I_T \otimes \mathbf{B}_N^\top B_N)y \tag{3.16}$$

and

$$\sigma_\epsilon^2 = \frac{\epsilon(\rho)^\top \epsilon(\rho)}{NT} \tag{3.17}$$

where the notation indicates the explicit dependence of the residuals on ρ.

Ultimately the concentrated likelihood and the GLS estimators are alternately computed until convergence. The parameters reported using the *splm* package (Bivand et al. 2021, Millo et al. 2012), for the SEM in this thesis, undergo the aforementioned procedure.

Where the SAR posits that the dependent variable depends on the dependent variable observed in neighbouring units as well as on a set of observed local characteristics the SEM posits that the dependent variable depends on a set of observed local characteristics and that the error terms are correlated across space.

This thesis continues by exploring the data used in this thesis through various visualizations and discussing the data management process implemented. The pre-processing of the data discussed in the proceeding section includes, but is not limited to, the way in which missing data and outlier are accounted for. After which, this thesis proceeds by following a sequential approach of applying the methods explained in this chapter to the data.

Chapter 4

Data

The data set used in this thesis is spatial panel data. Panel data refers to a cross section of observations (groups, households, countries, provinces, wards) repeated over several time periods. In a spatial panel setting, the observations are associated with a particular point in space. Data can be observed either at point location or aggregated over areas (Millo et al. 2012).

Before pre-processing occurs, the data in its raw form, is a collection of repeated household monthly observations. Each household has an associated point location and falls within the perimeters of one of the 116 wards the CoCT is divided into. In this thesis, there were two options, firstly, to analyse the household water consumption at point locations or, secondly, to average consumption at the ward level. The latter of the two was chosen. The main reason the decision was made to analyse and implement a spatial panel model once aggregated up to ward level consumption was to provide a relevant use case for the findings produced from the current thesis. Implementing a spatial panel model at ward level allows ward councillors to get a better understanding of consumption in their respective wards thus allowing the findings of this thesis to influence water supply decisions. In times of crisis, such as during the 2016-2018 drought, having a closer understanding of consumption levels from this perspective is essential when interventions are put in place to adjust consumption behaviour.

In most cases, it is necessary to pre-analyse, as well as pre-process time series data in order to ensure an optimal outcome of the processing. This section describes the data

management process which the raw data went through in order to be used in the final predictive model.

4.1 Variables

To illustrate what was said above, one can think of the example of the number of child births in a city. To create a time series of child births we would normally look at the number of births per month or, more granular, per day. In this cases the data points need to be accumulated or averaged over a time interval to obtain a series. Similarly, the data used in this study has been aggregated on a quarterly basis, such that the variable of interest in this study, water consumption, has been accumulated over time to obtain a quarterly total water consumption value instead of a monthly value. This has been done in order to deal with missing data and to ensure a the panel data used is balanced - this point is explained in further detail in 4.2. Ultimately, the sum of three months water consumption observations have been calculated to obtain one quarterly (total) value.

The original dataset included 12 variables from households' monthly billing data. The list of variables used can be found below:

1. The unique contract hashed value

   Type: Factor

   Description: This attribute is used to identify unique households in the dataset

2. Billing year

   Type: Categorical/factor

   Description: 5 factors, 2016-2020

3. Billing month

   Type: Categorical/factor

   Description: 12 factors, January(1)-December(12)

4. Quantity of litres (L) billed

   Type: Continuous

Description: This is the quantity of water, in liters, that appears on a households' monthly bill

5. Amount of South African Rand (ZAR) billed

   Type: Continuous

   Description: This is the amount required for payment by a household for their monthly water bill

6. Business Area

   Type: Categorical/factor

7. Rate Category

   Type: Categorical/factor

8. Zoning

   Type: Categorical/factor

9. Suburb

   Type: Categorical/factor

   Description: The different suburbs in the CoCT

10. Ward number

    Type: Categorical/factor

    Description: Each household in the dataset falls within the parameters of one of the 116 wards in the CoCT

11. Subcouncil

    Type: Categorical/factor

    Description: A subcouncil is a geographically defined area within the CoCT which is made up of between three and six neighbouring wards. Each household in the dataset falls within the parameters of one of the 24 subcouncils in the CoCT.

12. Geo location of household

Type: Categorical/factor

Description: The longitude and latitude coordinates of a billed household

Adjusted variables

Some of the aforementioned variables were left untouched and some were adjusted in order to get the data set in the desired format. Specifically, the billing month variable was translated into a quarter variable, taking on values 1, 2, 3 and 4, and the Geo location was split by its co-ordinates, creating two variables Latitude and Longitude which allowed for the data to be plotted in R.

Cape Town is divided into 116 areas, called wards. Wards are groups of neighbouring suburbs that are managed together, allowing the City to manage service delivery successfully for all residents (CoCT 2021*b*). The CoCT also make use of Sub-councils which serve as the link between local communities in Cape Town and the City Council. A sub-council is a geographically defined area within the city which is made up of between three and six neighbouring wards. There are a total of 24 sub-councils which make up the CoCT's municipal structure (CoCT 2021*a*) Including these variables will enrich the study by allowing us to observe how different sub-councils and wards behave with respect to water consumption. We may find that belonging to different wards affects water consumption behaviour due to the way in which matters such as controlling water consumption are managed and maintained.

The full billing dataset provided includes billing data for different types of properties and commodities, for this reason the dataset was filtered according to two variables in the dataset; rate category and zoning. Only households with a rate category of WATERFULLD, WATERFULLFA147, WAT-IND135, WAT-IND180, WAT-IND240, WAT-IND300, WAT-INDUNC or WATER-INDI are included in the study. These households must also have a zoning value equal to "Single Residential 1: Conventional Housing". By applying these filters the thesis limited the study to residential houses present in the dataset.

External Variables

As done in previous literature, a weather attribute is included as an external attributes to the raw CoCT data. The attribute included is a drought indicator. It is undoubtedly

true that Cape Town's drought affected water consumption (Bruhl, Serman & Visser 2020, Bruhl, le Roux, Visser & Köhlin 2020, Bruhl & Visser 2021, Brick et al. 2017) and in general one would think that an ongoing drought would affect water behaviour. For this reason a binary variable labelled Drought has been included in the data frame. The variable is set to 1 during a drought period and 0 otherwise. In the context of this study, the drought period is from the start of 2016 to the third quarter of 2018. Although prior research (Bakker et al. 2014) emphasized the importance of including further weather attributes when discussing the current domain, these have been excluded from this study for the following reasons. Firstly, the data being aggregated to quarterly observations per ward. Considering the average quarterly weather temperature (2016-2020) there was little to no variation from one year to the next - this attribute/pattern in the data would be accounted for using the pre-existing quarterly observation (Q1, Q2, Q3, Q4). It is also assumed that all the wards municipal supply of water would be affected similarly by rainfall. This being said, it was not possible to acquire such data for the purpose of this thesis due to the grain of the data explored being on a quarterly, ward level. Data matching this grain was not found and consequently could not be used to enrich the CoCT's dataset.

A political party variable is also included in some of the models implemented in Chapter 5. This variable s set to 1 if a ward is governed by the Democratic Alliance (DA) or the African National Congress (ANC). These political parties were voted into power in 2016 and were the ruling party over the five year time frame considered n this thesis. As a result the attribute is time-invariant and will be dropped from the fixed effects models implemented in Chapter 5.2 and 5.3. The justification for including a variable of this nature in the analysis is to determine the effects each political party may have on the distribution of municipal supplies, such as water.

4.2 Missing Data

The data, in its natural form, is collected and compiled on a monthly basis. This meaning each data point represents a monthly billed value per household. In time series analysis, such as the one conducted in the current thesis, a balanced panel dataset is desirable. This meaning there is a unique observation per household per time point. For the data used in this thesis, some households had missing bills for some months within

a given year. In order to minimise the impact of missing observations in the dataset, the current thesis used total quarterly consumption instead of monthly consumption. Thus, each data point represents an estimated total consumption over three months. As done in past research (Shine et al. 2018*b*,*a*, Yea et al. 2018), all missing data points were removed from the data and as a result 15% of all individuals in the dataset were excluded from the study. Dropping all unique observations with missing quarterly water consumption values resulted in 265,630 unique contracts (households) over the five-year period.

There are however other ways of dealing with missing data such as imputation. The mean, median or most occurring values for household water consumption could have been imputed for missing values. Such methods could have introduced bias into the dataset and reduced variance, for this reason have been avoided in this thesis due to the effect it would have had on the models being used in this thesis. Imputation using k-nearest neighbours (K-NN) was a potential option and is often favoured over the three aforementioned imputation methods as it can be more accurate.

The next subsection continues by focusing on visualising the variables within the billing data used to conduct this thesis.

4.3 Exploratory Data Analysis

This section focuses on visualising the data used in this thesis. The visuals explored in this section allow for the analysis of the data and provides a way to get a better understanding of the data. The figures that follow reveal trends, patterns, seasonality and clustering present amongst the observations in the data set. A shape file (CoCT 2019), containing the polygons of wards in the CoCt, is merged with the billing data, allowing the data to be visualised The figures in this section are generated before outliers are excluded from the analysis (Figures 4.1 & 4.2).

Figure 4.1, shows the quarterly median household water consumption value over a five year period in the CoCT (2016-2020) while Figure 4.2 represents the average quarterly household consumption. Important to note is that these two visuals are created on a household grain as this analysis took place before the data was aggregated up to a ward level in order to see if the data being used for this thesis conformed with the

findings of (Bruhl, le Roux, Visser & Köhlin 2020). Median consumption was visualised and explored as an alternative to total or average consumption due to the existence of extreme outliers in the data. As previously mentioned, Cape Town experienced a drought from the start of 2017 to the end of 2018 (indicated as the section that falls between the two vertical, dotted red lines in both Figure 4.1 and 4.2).

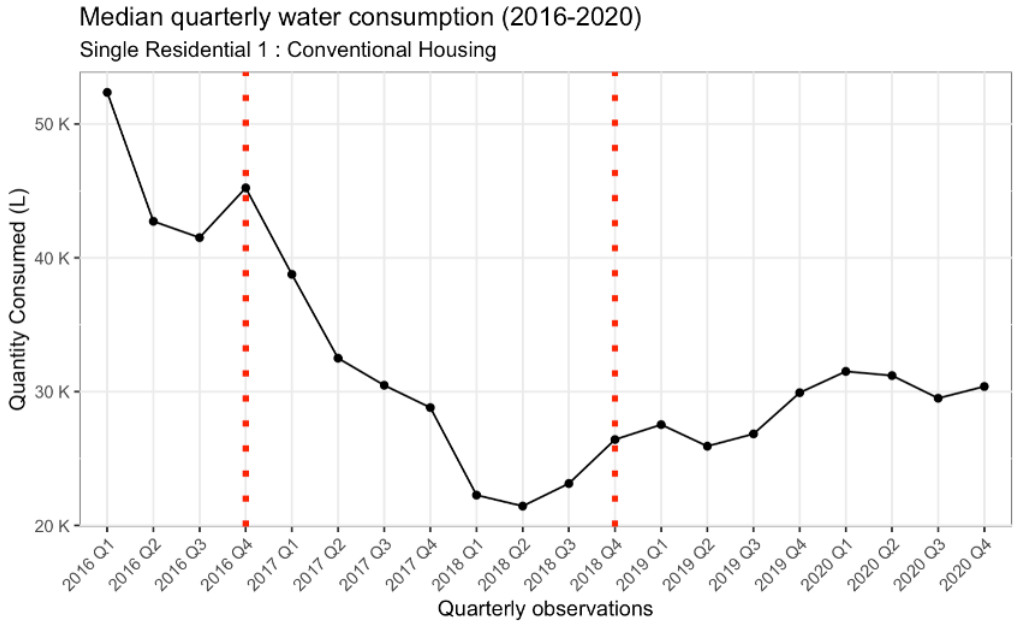


Figure 4.1: Median quarterly water consumption using the raw quarterly data

In Cape Town, the summers are warm, dry, and mostly clear; while the winters are cool, wet, and partly cloudy. In South Africa, roughly speaking, the summer months are December to February, Autumn runs from March to May, Winter is the period from June to August, and Spring is from September to November. Understanding that different seasons are characterised by different weather patterns, one would expect household consumption to fluctuate with seasonal changes. This being said, we expect water consumption to be high in the warm, dry months and lower in the wet months of the year. Therefore we would expect peak, water demand, periods to be in quarters 1 and 4 and low demand in quarters 2 and 3. This assumption was made as one would expect some households to water their gardens and fill up their pools when the weather is hotter and rainfall is scarce. Actions such as these would contribute to a higher water consumption in quarters 1 and 4. On the other hand, during quarters 2 and 3, the frequent rainfall would naturally water gardens, fill up pools and water tanks - resulting in households consuming less municipal supplied water.

The assumptions stated above are confirmed in Figure 4.1, however one ought to remember the effect of the drought when looking at this figure. In 2016 (the pre-drought year) Figure 4.1 shows that seasonality is suggested in the data (high consumption 2016 Q1 & Q4, lower consumption 2016 Q2 & Q3). Although consumption is high in quarter 4 of 2016, it is not as high as quarter 1 2016 - one would image that consumption would be a lot more similar than what is represented in Figure 4.1 however, this difference is caused by the start of the drought in the CoCT.

At the start of 2017, quarter 1 consumption was high but not as high as quarter 1, 2016 or any other quarters in the pre-drought year. This is likely due to households being aware of the drought the city was starting to experience. Water consumption continued to drop until quarter 3, 2018. This pattern of consumption is expected as the government implemented several tariffs, imposed severe restrictions on water usage, reduced water pressure and used several media campaigns urging people to save water during this time period. The measures imposed by the government were purely motivated to decrease household consumption, due to the fear of running out of water in the CoCT.

The lowest median quarterly consumption value was recorded in quarter 2, 2018. This, once again is expected, as the residences of Cape Town were continuously warned about "Day-Zero" which was predicted to be early on in 2018. The concept of "Day-Zero" coupled with low dam levels and other government interventions clearly affected the way in which the Cape Town residences consumed water in their households. However, after a wet rainy season in 2018, dam levels started to rise (CoCT 2018*b*) and so too did water consumption. Restrictions on consumption were eased in the latter part of 2018 as the six major dams collectively rose to 53.3% by July, 2018.

Figure 4.2 suggests something different. While Figure 4.1 shows post-drought levels not reaching the same levels as pre-drought median consumption levels, Figure 4.2 suggests that average post-drought consumption levels return to pre-drought levels in 2020. Although this is likely caused by outliers in the data, it is interesting to see the contradicting messages of Figure 4.1 and 4.2 with respect to pre- and post-drought consumption levels.

The next section of the thesis focuses on the outliers present in the dataset before moving on to introduce the spatial element present in the data set and then to different model implementation.

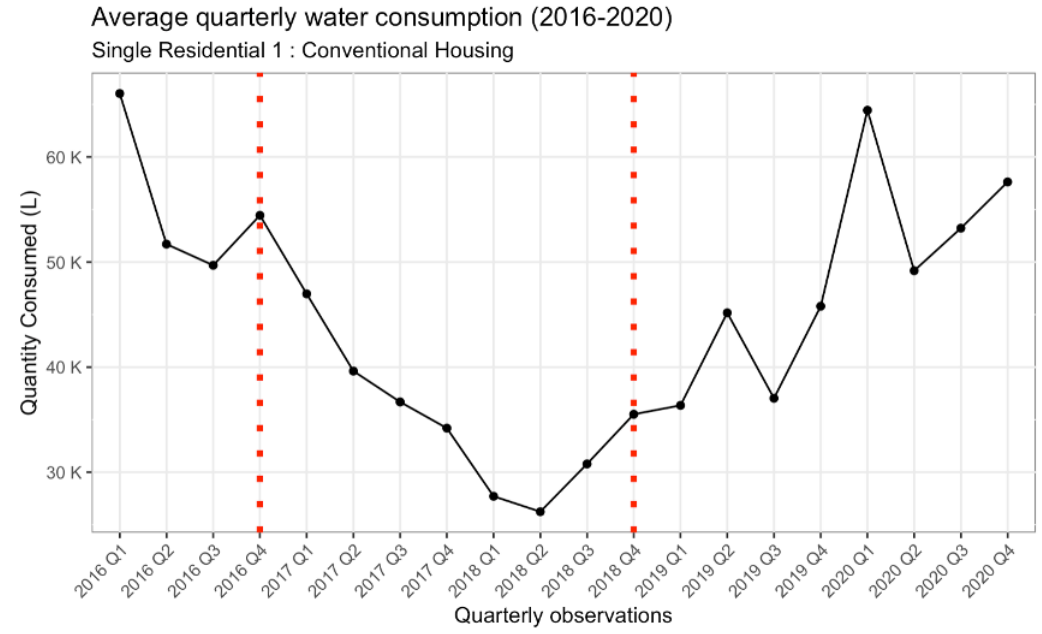


Figure 4.2: Average quarterly water consumption using the raw quarterly data

4.4 Outliers

The analysis identified outlying observations based the *Water Quantity Billed Amount* variable. Due to the nature of water consumption billing, outliers were expected. These outliers could be associated with leaks, misread meters among other reasons. For these reasons, outliers were identified and dealt with appropriately in data.

Outliers were identified using the box-and-whisker diagram constructed by year and quarter. For example the full dataset was filtered by quarter 1, 2016, then the values in the *Water Quantity Billed Amount* attribute were analysed and all outliers were identified. Observations were marked as outliers if they lie beyond the extremes of the whiskers in a box and whiskers diagram. Using this technique, to identify outliers in the dataset, outliers are defined as observations greater than (Underhill & Bradfield 2013):

$$x_m + 6(x_u - x_m) \tag{4.1}$$

or less than:

$$x_m - 6(x_m - x_l) \tag{4.2}$$

where:

- x_m is the median value
- x_l is the lower quantile value
- x_u is the upper quantile value

Outliers were detected at quarterly household consumption level, using the above methodology in R, using the *boxplot.stats* function in the *grDevices* R package (R Core Team 2013). The decision to deal with outliers before aggregating the data to ward level was to ensure that no ward was excluded from the entire data set, rather selected households within the wards were excluded. Doing so resulted in 77,147 unique households with one or more observations classified as an outlier. All observations associated with these households were marked for deletion using a binary indicator. The models and some the visualisations in the proceding sections filter the outlier observations out and work with complete cases. Removing these observations results in the removal of 29.89% of the observations in the dataset. This decision was influenced by past literature that focused on water consumption and used linear models. Past literature justified the removal of outliers due to the nature of water consumption data, whether it be billing data or meter readings. Outliers could be present as a result of, but not limited to, an erroneous bill, burst pipe or broken meter. Linear models are largely impacted by outliers, therefore including them may result in the models' parameters being altered in order to account for these outliers - which in this use case would be undesirable.

The next section focuses on the spatial element that exists the data set. It does so by plotting different maps, using different packages in R, in an attempt to convey different elements and patterns in the data.

4.4.1 Spatial Analysis

As previously mentioned, the dataset used in this thesis is a spatial panel data. The spatial element of the data has allowed for the visualisation of the water quantity billed using the *Leaflet* (Graul 2016) and *tmap* (Tennekes 2018) package in R. Visualising the water billing data on a map allows for us to identify water consumption hot spots (areas with higher than normal consumption levels) and clusters (areas that share a commonality in water consumption). Identifying these hot spots and clusters allows policy makers and ward councils to understand the place-based demand for water and

take informed action when needed. The clustering of ward consumption further allows for similar wards, with regards to consumption behaviour, to adopt similar approaches when wanting to control or alter water consumption.

Figure 4.3 is another visualisation that revealed the presence of outliers in the dataset by highlighting areas in top quantile consumption groups in dark red. The density map 4.3, has shaded different regions of Cape Town based on the total amount of water consumed over the five year period. This visualisation was created using the x and y coordinates associated with each household in the billing dataset. This visualisation has been created on a household grain, before the data set was aggregated to ward level consumption. The output of Figure 4.3 helped identify high consumption areas, areas with a darker red shading, over the five year period using total consumption levels. In order to generate the results in Figure 4.3 the data points were grouped into five quantiles. As previously stated this not only allowed for the visualisation to act as a means to identifying high consumption areas, it has also proven to be a useful visualisation that identified areas with similar consumption based on region shadding.

The thesis continues by an exploration that aggregated household water consumption to the ward level, this meaning that the quantity of water billed on a household grain is not analysed in Sections 4.3 and 5, instead water consumption is analysed on a ward level grain. In order to conduct this type of research a shape file of the CoCT and its boundaries of the wards was necessary. This dataset was acquired from the CoCT open data portal (CoCT 2019).

Figure 4.4 displays the shape file used in this thesis. The visual shows the outline of the CoCT and how it is divided into 116 wards. This shape file acts as the base layer for all other ward level visualisations that follow in this section. Three wards are labeled in Figure 4.4, namely; ward 29, 54 and 62. The reason for doing this is to allow for discussion and to provide examples going forward. The shape file used to create Figure 4.4 was also used to create the weights matrix (W). The spatial weights matrices used to explore spatial relationships between wards in this thesis is both the Queen and Rook contiguity-based weights.

Typically, when working with spatial data, one would expect there to be similarities between neighbouring areas. For example, in this thesis, we would expect neighbouring

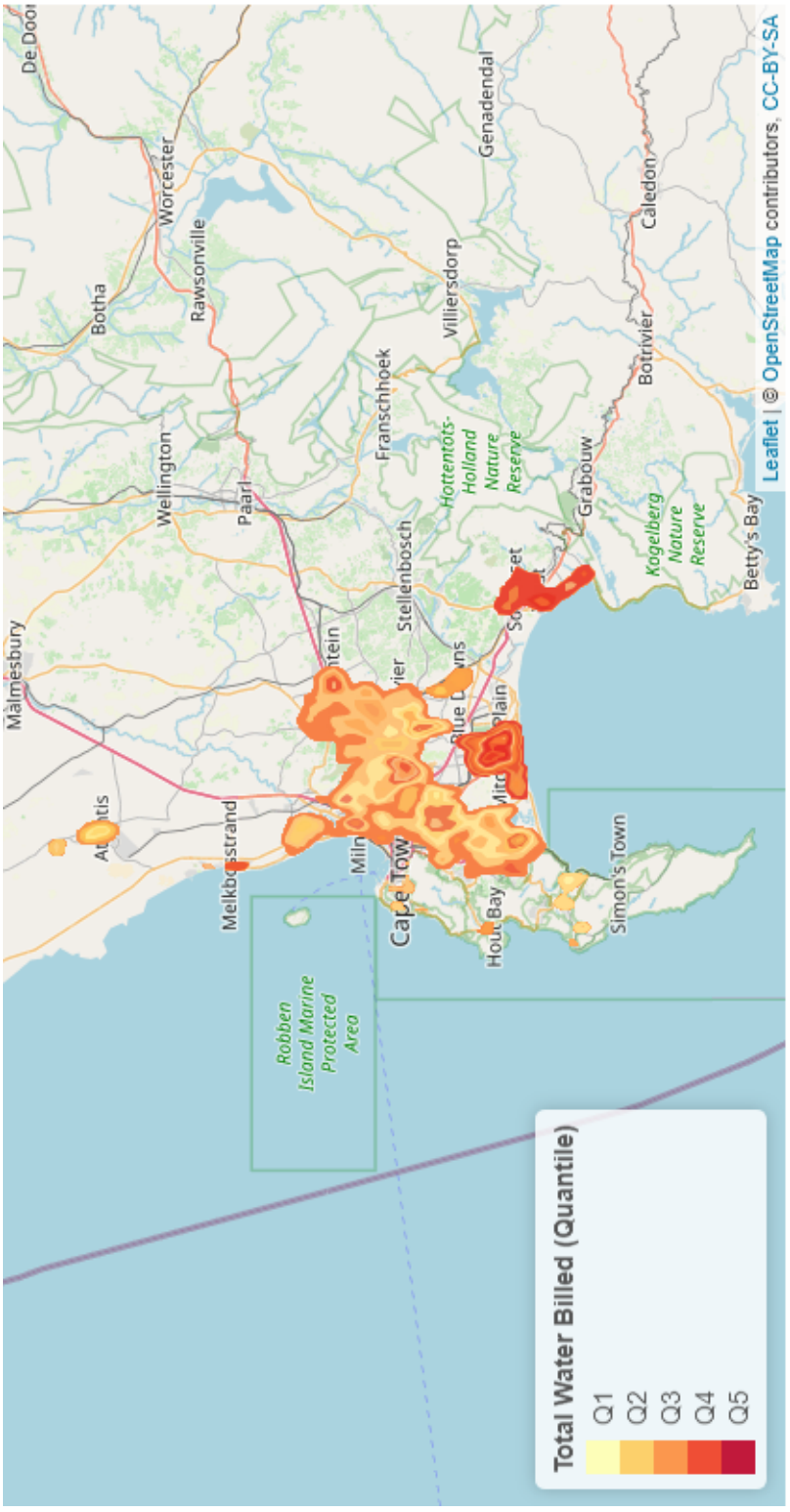
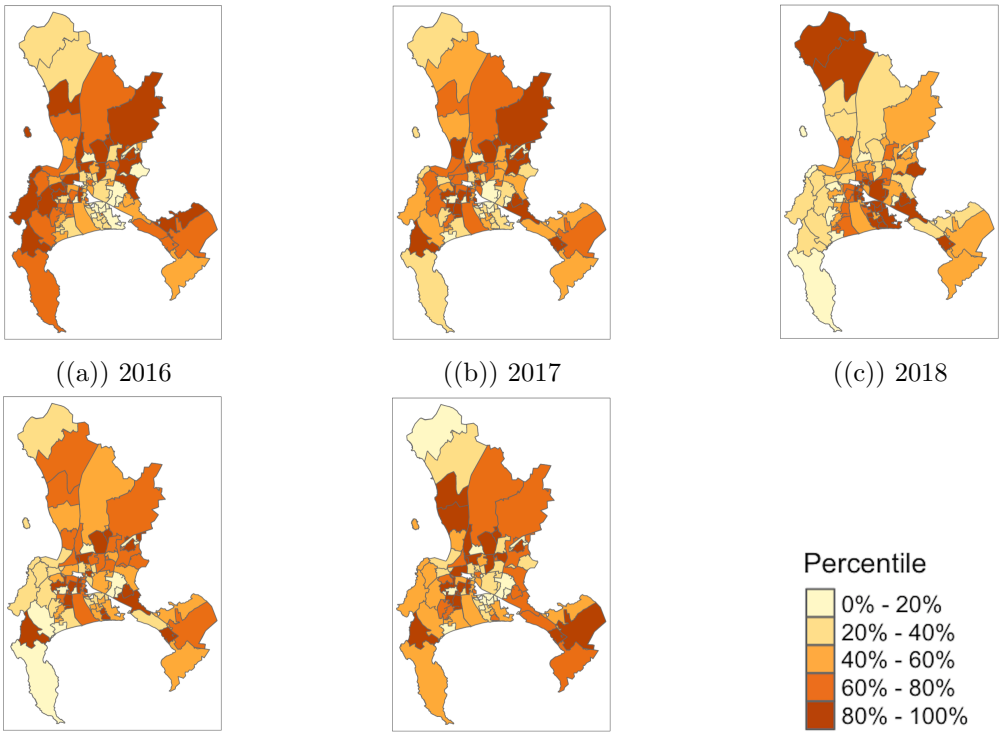


Figure 4.3: Density Map by Quantity of Water Billed, City of Cape Town

Figure 4.4: Map of the City of Cape Town, including ward boundaries

wards to show similarities in water consumption. In the case of this thesis the hypothesize would be that wards that are close together demonstrate a similar behaviour when it comes to water consumption in comparison to wards that are further apart. However in some cases, given the social construct in South Africa, the opposite may hold true. In the CoCT there are wealthy neighbourhoods situated next to informal settlements, in situations such as this we would expect water consumption to be vastly different if the wealthy neighbourhood belongs to a different ward than the informal settlement. For example, Hout Bay has some of the highest value properties in Cape Town but is also home to two large informal settlements. If areas such as these are split into different wards it would be reasonable to hypothesis that consumption between these neighbouring wards would not be similar in nature. Creating effective visualisations of water behaviour, using a map of the CoCT with clear boundaries for the 116 wards, will provide insight into the aforementioned hypotheses. The visualisations that follow attempt to accomplish the above.

Going forward, the visualisations and models implemented in this thesis exclude outliers from the analysis. Figure 4.5 is a choropleth map which visualises mean ward water



((a)) 2016 ((b)) 2017 ((c)) 2018

((d)) 2019 ((e)) 2020

Figure 4.5: Average ward water consumption, by year

consumption, by year (2016-2020). The visualisation divides the water consumption into five percentile groups (0%-20%, 20%-40%, 40%-60%, 60%-80%, 80%-100%). Doing the the above allows for the tracking and observation of ward water consumption relative to other wards in the CoCT. For example, in its entirety, Figure 4.5 allows us to track ward 29 over a 5 year time frame (make use of Figure 4.4 to identify ward 29). Ward 29's water consumption fluctuates over time. In 2016 and 2017, Figure 4.5(a) and 4.5(b), ward 29 was in the percentile group (20% - 40%) but according to Figure 4.5(c), in 2018, the ward was in percentile group (80% - 100%). In the proceeding years (2019 & 2020) the ward moved into lower consumption groups. What one may infer from the story told about ward 29, over the five year period, is that households in ward 29 did not decrease their consumption during the drought period as much as households in neighbouring wards. On the other hand, ward 29's behaviour pattern may not have changed at all but because other wards reduced their consumption, this ward moved up the ladder with respect to their water consumption.

Apart from the above use case for Figure 4.5, this visualisation could also be used to analyse the different behaviours of wards during a period of time when policies were being implemented to alter water consumption. For example, during the drought period it is

clear that some wards changed their consumption behaviour more so than other wards. This could be as a result of different policies, tariffs or behavioural implementations being imposed on different areas during the drought. Taking an example of ward behaviour from Figure 4.5, before the drought, ward 54 and 62 were on average two of the highest consuming wards, falling into the highest quantile group (80%-100%). However, as the CoCT enters a drought in 2017 and 2018 these two wards showed a significant drop in consumption, dropping into lower quantile groups. A possible reason for the drop in water consumption in these wards could be as a result of these factors may include these wards being particularly wealthier wards. Some of the households in these wards may have sourced water from external providers or installed water tanks during the drought. Due to their consumption behaviour in 2016 they may have also been subjected to higher tariffs, which would contribute to their decrease in consumption. Consumption in these wards do increase post-drought however they still remain in the lower quantile groups than before the drought (Figure 4.5(e)).

A visualisation such as the one being discussed provides valuable insight to policy makers as it clearly shows how some wards changed their consumption behaviour more/less than others when interventions were being imposed. Looking at this visualisation in conjunction with intervention timelines and details, such as the one found in Bruhl and Visser's paper (Bruhl, le Roux, Visser & Köhlin 2020), policy makers will be able infer which types of policies were more successful than others.

Figure 4.5 also justifies the inclusion of a drought indicator in the analysis of water behaviour in the CoCT. This assumption is made due to the change in water consumption behaviour displayed in Figure 4.5 during the CoCT's drought period. For this reason, a binary (1 or 0) drought indicator is included as an independent variable when implementing all the models in Section 5.

Chapter 5

Implementation and Results

This chapter explores the several models implemented in this thesis. The flow of this chapter takes after the methodology where it begins with the implementation and assessment of the goodness of fit of several multiple linear regression models. This is then followed by the implementation and assessment of panel models and spatial panel models. The results from each model explored was published and assessed in this section. The code used to create these models can be found on a GitHub page including all the Rmd files for this thesis (Github 2022).

The dependent variable used in all the models in this section is the log of average quarterly ward consumption value (*ln_consump*). This decision was made in order to reduce the skewness of the original data and appropriately deal with the target variable which is, by nature, a positive value. The predicted values from the linear regression using the unadjusted average quarterly ward consumption value may be negative. The predicted values from a log-transformed regression can never be negative.

5.1 Multiple Linear Regression Results

To begin, several multiple linear regression models were implemented. The results of the four models can be found in Table 5.1. The coefficients along with their associated *p-values* for the models that are not deemed can be found in the Appendix A.

The implementation of a multiple linear regressions ignores the fact that the data in this thesis is panel data. As a result it is the foundation phase to building a model that includes the underlying characteristics of the data.

Model	No. Estimates	Adjusted R^2	RMSE
1	6	0.0731	0.2577
2	93	0.2174	0.2567
3	10	0.6907	0.1367
4	97	0.8349	0.1367

Table 5.1: Multiple Linear Regression Results

R^2 is a statistical measure that represents the proportion of the variation in a dependent variable that is explained by an independent variable or variables in a regression model. The reported measure, for each respective model, in Table 5.1 is the Adjusted R^2 value. This statistical measure is a modified version of the R^2 value that takes into account the number of predictors in the model. It would therefore be expected that the Adjusted R^2 increases only when a new attribute improves the model and decreases when a predictor improves the model by less than expected. The Adjusted R^2 will always be less than the R^2 due to the penalty it imposes when including more attributes into the model (Underhill & Bradfield 2013).

The root mean square error (RMSE) statistic reported in Table 5.1 has been calculated using k-fold cross validation. Cross validation (CV) is a re-sampling method that uses different portions of the data to test and train a model on different iterations. For this thesis "leave-one-out" cross validation (LOOCV) is implemented, this is because, after filtering the dataset, there are 88 wards in the data set and ideally we would want all observations per ward included when the ward is selected to be apart of the training set. During the CV process we train our models on 87 of the folds (87 wards) and validate on the last fold (1 ward). Implementing the above resulted in the values seen in Table 5.1, model 3 and 4 have the lowest error on average and if one were to choose a model based on the RMSE value it would be justifiable to choose either one of these models as it is assumed that these models are the best predictor of out-of-sample data.

The sections that follow explain the models and associated results.

Model 1:

The independent variables in this model are:

1. Drought (Binary)
2. Quarter (Factor - quarter 1 is the base category and is excluded to avoid multicollinearity)
3. Political Party (Binary)

As seen in Table 5.1 the Adjusted R^2 value was particularly small, implying this model is a poor fit for predicting the dependent variable. As a result, additional independent variables were added to this model in an attempt to improve the model's predictive accuracy.

Model 2: Following on from Model 1, above, the independent variables used in this model include all those listed in model 1 as well as:

4. Ward (87 binary ward indicators)

To avoid the occurrence of high inter-correlations among two or more independent variables in a multiple regression model (multicollinearity) one of the wards is excluded from the model and used as a base year. Consequently, 87 binary ward indicators were included (reminder: there are 88 wards in the data frame).

Doing the above resulted in an increase in the Adjusted R^2 value, in comparison to Model 1. This result clearly indicates that including the ward as an explanatory variable is helpful in predicting the average quarterly water billed amount. However, the Adjusted R^2 value for this model, 0.2174, is still low for the model indicating that the independent variables used in this model are not highly predictive of the dependent variable. In comparison to the other models in this sub section, the RMSE statistic for the model is the second highest. If one were to choose one of the models in this section based on their associated RMSE values, this model would not be regarded as the best model due to its higher RMSE value. As a result of the Adjusted R^2 and RMSE value associated with this model it is not selected as the best performing model for this section.

Model 3: In an attempt to find a model that best fits the data a fourth independent variable is added to Model 1, that also differs from Model 2. The independent variables used in this model include all those listed in model 1 and in addition:

4. Year (4 binary coefficients - 2016 is used as a base category)

In order to improve the fit of model 1 and model 2, the year attribute in the dataset was included. To avoid multicollinearity in the model, 4 binary ward indicators were included for the year attribute (*2017*, *2018*, *2019*, *2020*).

Including year as an explanatory variable resulted in a large increase in the Adjusted R^2 value. This result clearly indicates that including the year is helpful in predicting the average quarterly water billed amount. If we were to select a model based on the RMSE values associated with the models, model 3 or 4 would be chosen as they both have the same, lowest, RMSE value in comparison to all the other models in this section. As the Adjusted R^2 value is present for each model, the selection of the model that best fits the data can be made using both performance metrics. As a result, model 3 is not selected as the best performing model in this section as the Adjusted R^2 value for model 3 is smaller than that of model 4, 0.69 and 0.83 respectively.

Model 4: The independent variables used in this model include all those listed in model 1 as well as both the:

4. Year (4 binary estimates)
5. Ward (87 binary estimates)

The final linear model fitted to the data, in this section, builds on from model 1 by including both the year and ward number attribute. Excluding one factor level from each of the aforementioned attributes ensure multicollinearity is avoided in the model.

As previously mentioned, this model has the highest Adjusted R^2 value amongst all four models. To some extent, these results are expected as one would believe that consumption values associated with particular wards are similar in nature and that consumption varies by year. As a result, including both attributes in the model would most likely increase the models predictive power. It also evident that including ward without year, and vice versa, the Adjusted R^2 value increases in comparison to model 1, where both of these independent variables are excluded. We could therefore presume that including both these attributes would lead to a higher Adjusted R^2 value.

Having the highest Adjusted R^2 value (0.8349) and the a shared lowest RMSE value (0.1367) this model is seen as the best model amongst the other multiple linear regression

models in this section. In order to determine whether or not this model can be or should be used going forward, the residuals need to be explored.

Now that the best performing model in this section has been identified it is essential for the model to be analyzed and checked for some of the flaws that may exist. Some of these flaws include, but are not limited to; non-linearity of the response predictor relationships, correlation of error terms, non-constant variance of error terms, high-leverage points and collinearity may be present in the model. The discussion that follows explores some of the aforementioned problems, with respect to Model 4.

1. Non-linearity in the data

The linear regression model assumes a straight-line relationship between the independent and dependent variables. It is however possible that a linear relationship may not exist between all predictors and the response and failing to acknowledge this will result in the conclusions we draw from the model being unreliable (Tibshirani et al. 2013). In order to examine non-linearity in the model one can analyse the results show in Figure 5.1, where the model's residuals are plotted against the fitted values generated by the model. The red line in Figure 5.1 shows a smooth fit to the residuals, which helps detect any trends. A red line sitting exactly on the dotted line would suggest no evidence of non-linearity. Looking at Figure 5.1 (from the readers left to right), the red line initially diverges from the dotted line suggesting that the assumption of linearity does not hold. This result encourages and motivates the need to explore an alternative model.

2. Residuals follow a Normal Distribution with a mean of 0 and a constant variance

In order to check whether or not this property holds true for Model 4, Figure 5.2 and 5.3 were plotted. Plotting a histogram of model 4's residuals, Figure 5.2, suggest that the residuals are symmetrical around a mean of 0. The Q-Q plot, Figure 5.3, shows evidence that suggests the residuals deviate from normality. If the residuals were to follow a normal distribution, all the points in Figure 5.3 would fall directly on the dotted line. In the Q-Q plot associated with model 4, the residuals in the bottom left corner and top right corner diverge from the straight line and the normality assumption does not hold. This being said, the large majority of the residuals fall along the dotted

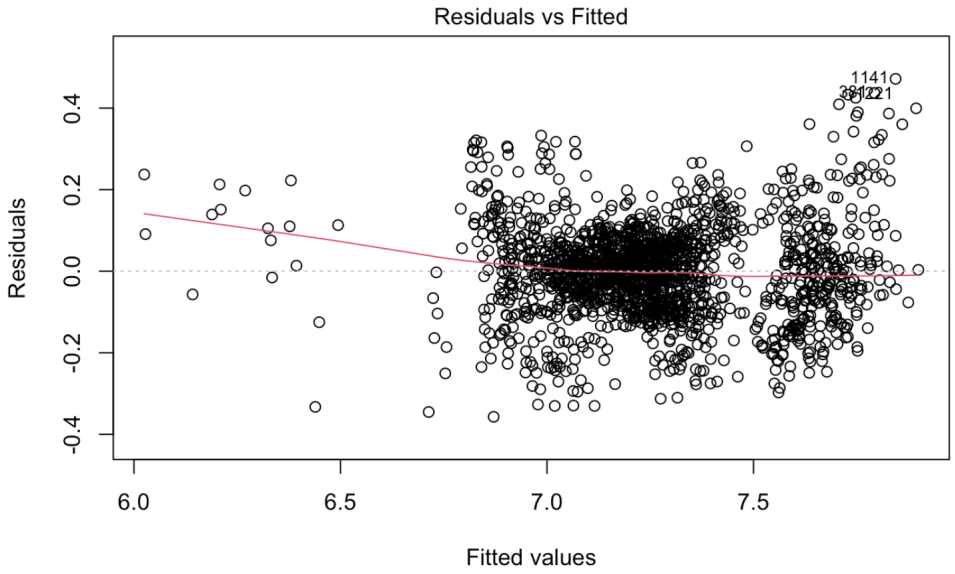


Figure 5.1: Model 4: Residuals vs Fitted Values

line and the bell shaped histogram in Figure 5.2 provides evidence that the normality assumption is not necessarily violated.

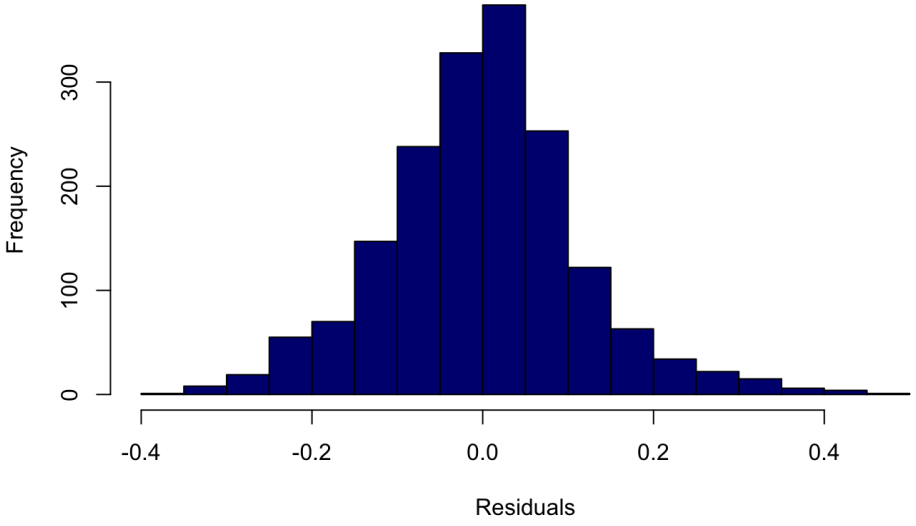


Figure 5.2: Model 4: Histogram of Residuals

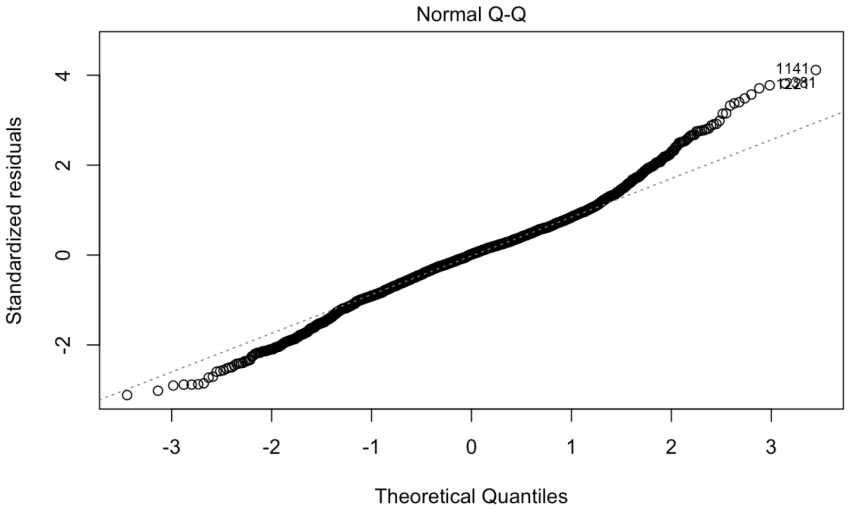


Figure 5.3: Model 4: Q-Q Plot

5.2 Panel Models

This section follows on from Section 5.1 by exploring different panel models in an attempt to find a model that is most appropriate for the data used in this thesis. As discussed in Section 5.1, the models including both the ward and year attributes was the bed performing model. The individual effects in the panel dataset is the associated ward that a particular observation belongs to while the time feature is a combination of year and quarter. Panel models are designed to take these characteristics into account and as a result are explored in this section of the thesis.

As done with the multiple linear regression models, the dependent variable used for all the panel models is the log of average quarterly ward consumption value.

5.2.1 Hausman Test

As suggested and implemented in prior research (Baltagi et al. 2003, Baltagi & Li 2008, Millo et al. 2012), the choice between using a fixed or random effects model is based on the Hausman-type test results.

The R function; *phtest* in the *plm* package (Croissant & Millo 2018, 2008, Millo 2017) computes the Hausman test which is based on the comparison of two sets of estimates. Essentially the test checks how similar the beta coefficients generated in the fixed effects model are to the beta coefficients generated in the random effects model. The main

arguments of the Hausman test is two panel model objects. In order to compute this test two panel models are computed, a fixed effects and a random effects model. The abbreviated terms for the attributes used in the models, for this test and all models going forward, are as follows:

1. log(Consumption) - ln_consump
2. Drought Indicator - drought_ind
3. Political Party Indicator - pp_ind
4. Ward Number - ward

The formula used for both the fixed and random effects models was as follows:

$$ln\_consump = drought\_ind + quarter + pp\_ind \tag{5.1}$$

The null and alternative hypothesis for the Hausman test are stated below (Millo et al. 2012):

H_0: Fixed effect coefficients are not significantly different from the random effects coefficients

H_A: Fixed effect coefficients are not significantly different from the random effects coefficients

Using the above formula in both a fixed and random effects model, produced a *p-value* < 0.001 for the Hausman test. Based on these results, we reject H_0 and conclude that there is enough statistical evidence in favour of the alternative. The reason for this is that in statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct and a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. Therefore in the context of this test, the fixed effect model is implemented in this thesis.

5.2.2 Panel Model Results

Based on the outcome of the Hausman test, this section of the thesis focuses on fixed effects panel models. The results from the various fixed effects panel models are displayed in Table 5.2.

Model	No. Estimates	Effect	Adjusted R^2	RMSE
1	5	Individual	0.0192	0.2495
2	2	Time	0.0419	0.2609
3	89	Time	0.5611	0.2820
4	9	Individual	0.7926	0.1146

Table 5.2: FE Panel Model Results

For the models that have not been selected as the best performing models in this section, the outputs can be found in the Appendix C. The model that performs the best, by having the highest Adjusted R^2 value and lowest RMSE, is Model 4. As a result, this model is discussed extensively in this section. Before doing so, a brief discussion on the three other panel models implemented is given. It should be noted here that the RMSE reported in Table 5.2 is the RMSE generated from using the full dataset and not the RMSE generated using cross validation and testing on unseen data or a validation set. This is primarily because of a software limitation and leaves space for future researches to implement predictions on unseen data using panel models. Consequently a disclaimer should be noted that one may be skeptical as to whether or not these models are over fitting.

Model 1: The first panel model implemented in this section of the report is fairly basic and uses the following formula:

$$ln\_consump = drought\_ind + quarter \tag{5.2}$$

The effects parameter was set to *individual* for this model, this indicates that we were isolating the individual effects over time. As a result, time-invariant attributes are excluded from the model such a as the *pp_ind* attribute. Although all the beta coefficients in this model were significant. Results found in Appendix A, the model has a low Adjusted R^2 (0.0192) indicating that the independent variables used in this model do not explain much of the variation in the dependent variable.

Model 2: Due to the poor performance of model 1, model 2 sets the effect parameter in the *plm* function to *time*. As a result, there is only one attribute included in the model (*pp_ind*).

The formula used for Model 2 is as follows:

$$ln\_consump = drought\_ind + quarter + pp\_ind \tag{5.3}$$

In an attempt to see the effects of time on the model, the effects parameter was set to *time*. Setting the effect parameter to *time* for a FE model returns a vector/matrix containing the values in deviation from time means (Croissant & Millo 2008). Due to the aforementioned parameter being set to *time*, the Drought Indicator and Quarter attribute were excluded from the model. Evidently, when looking at the Adjusted R^2 and the training RMSE, this model did not perform as well as other models implemented in this Section (Models 4 and 5) as the Adjusted R^2 is very low implying little to no predictive power while the RMSE value is the highest amongst all models in this section (0.2609).

Model 3: Learning from the Section 5.1, where the ward attribute was included in the model as an independent variable, Model 3 includes *ward* as an explanatory variable resulting in an increase in the models performance. In the context of this thesis, the ward attribute represents the individual element of the panel dataset while the combination of quarter and year create the time element in the dataset. As a result, the inclusion of *ward* as an attribute in the FE model with the effect parameter set to *time* returns the demeaned data as a vector/matrix containing the values in deviation from the time means.

The formula used for model 3:

$$ln\_consump = drought\_ind + quarter + pp\_ind + ward \tag{5.4}$$

Using this formula for the FE model, there are 89 β coefficients estimated. There are 87 ward coefficients, ward 1 is excluded and used as the base category and there is one coefficient for *pp_ind*. Both the *drought_ind* and *quarter* attribute were dropped from the model. Doing the above improves the models performance with respect to the Adjusted

R^2 value, however the RMSE is the highest out of all the models. Model 3 shows an increase in the Adjusted R^2 value, 0.5611, in comparison to model 1 and 2 in this section this statistic improves.

The model with the lowest RMSE amongst the Multiple Linear Regression models (Section 5.1), is the model that includes year as an independent variable. Model 4 below learns from the findings in the previous section by including year as an independent variable in an attempt to improve the predictive accuracy of the basic fixed effects model generated in Model 1 of the current section.

Model 4: The formula used for FE panel model, model 4 is as follows:

$$ln\_consump = drought\_ind + quarter + year \tag{5.5}$$

This model performs the best out of all 4 panel models implemented for in this thesis. As a results the estimates generated from running this model are shown in Table 5.3.

	Estimate	Std. Error	t value	Pr(>|t|)
Drought	-0.1021	0.0102	-9.9562	0.0000
Q2	-0.1142	0.0079	-14.3645	0.0000
Q3	-0.1172	0.0079	-14.7413	0.0000
Q4	-0.0551	0.0079	-6.9376	0.0000
2017	-0.2745	0.0118	-23.3542	0.0000
2018	-0.6267	0.0103	-61.0788	0.0000
2019	-0.5464	0.0092	-59.0878	0.0000
2020	-0.4223	0.0092	-45.6612	0.0000

Table 5.3: Fixed Effects - Model 4

Based on the results in Table 5.3 all the β coefficients in this model are significant. The aforementioned statement is justified by the small p-values values seen in column 5 ($Pr(>|t|)$). The *p-value* for each attribute is associated with a test where the null hypothesis (H_0) states that the respective coefficient is equal to zero, this indicating that the independent variable has no effect on the dependent variable. The low p-values (*p-value* $<<$ 0.001) associated with each independent variable in Table 5.3 provides enough evidence to reject the null hypothesis and conclude that the independent variables in this model are meaningful as changes in the predictor's value are related to changes in the response variable.

All the β estimates in the model have a negative effects on the dependent variable. This observation can be made by looking at the negative sign associated with each β estimate reported in Table 5.3. This is expected as the base category excluded from the *quarter* and *year* attributes are Q1 and 2016, respectively. One would expect water consumption to be lower during a drought period and highest in the first quarter of the year. As previously mention, the exclusion of these factor levels ensure that multicollinearity is avoided in the model. When interpreting the results of Q2, Q3 and Q4, this is done in comparison to the base category Q1. This holds true for the year estimates reported, where 2016 is the base category.

One of the advantages of using a linear model is being able to interpret the results in a way that can be understood and explained easily. This being said the models in this thesis take the log of average quarterly water consumption as the dependent variable and therefore interpreting the β estimates is not as simple as it would be if average quarterly water consumption was used as the dependent variable. However an easy transformation can be done to make these estimates easy to understand. To interpret the β estimates in a meaningful way one can exponentiate the coefficient reported in the model, subtract one from this number, and multiply by 100. This gives the percent increase or decrease in the response for every one-unit increase in the independent variable.

The drought indicator shown in the output of Table 5.3 is associated with the indicator being set to 1. Using the transformation technique described above, this estimator indicates that if the observation is taken during a drought period, the model predicts water consumption to decrease by 10.21%, all else held constant. Looking at the Q3, β coefficient reported Table 5.3, one can say that consumption is approximately 11.72% lower in quarter 3 of the year than quarter 1, holding all else constant. The other estimates in the model can be interpreted in a similar way to the two aforementioned estimates.

By including the year attribute in the model, the effects parameter is set to *individual*. We are therefore able to extract the fixed effects that are associated with each ward, which do not vary over time. The fixed effects are something specific to a ward that make consumption higher or lower in that particular ward. Selecting four random wards (22, 61, 77, 88), Table 5.4 looks at the fixed effects associated with each ward respectively. The results in Table 5.4 suggest that ward 22, 61 and 88 have lower consumption than the

average consumption value whereas ward 77 has a higher consumption than the average consumption value. Based on the associated *p-values* in Table 5.4, the estimates for wards 61 and 88 are significant as the associated are 0.00 and 0.01, respectively. Low p-values such as the ones previously mention indicate that the fixed effects estimates in column 2 of Table (*Estimate*) 5.4 did not occur by chance. It should however be noted that ward 22 and 77 have p-values of 0.68 and 0.65, respectively, and as a result the associated estimate is insignificant. This meaning that we can not draw much evidence from these particular estimates. The fixed effects associated with all the wards in the dataset are included in the Appendix C, in Table C.3.

Ward	Estimate	Std. Error	t-value	Pr(>|t|)
22	-0.01	0.03	-0.41	0.68
61	-0.17	0.03	-6.08	0.00
77	0.01	0.03	0.46	0.65
88	-0.07	0.03	-2.57	0.01

Table 5.4: Model 5: fixed effects associated with wards 26, 61, 77 and 88

Using the same four, randomly selected wards, Model 4's in-sample prediction is visualised in Figure 5.4. The plots in Figure 5.4 show the actual (y) values, the predicted ($\hat{y}$) values and a 95% confidence interval band that is generated when using Model 4. The black line in Figure 5.4 represents the actual response variable (*the log of average quarterly water consumption*), the red line indicates the predicted response variable while the dotted grey lines represent the upper and lower boundary of the 95% confidence interval. Visualising the results in this way allows the reader to see the difference in actual and predicted values when using Model 4.

From the visualisation (Figure 5.4) the black and red line graph follow a similar trend. The gap between the two curves, representing the error/mismatch between predicted and actual is not large. For wards 26, 61 and 77 the actual values fall within the 95% confidence interval which is a favourable result. The 95% confidence interval displayed in the visualisation highlights a range of values that one can be 95% confident contains the true prediction. It should however be noted that for ward 88 the actual values around time point 12 and 13 sit outside the 95% confidence interval band. Looking at the actual values for ward 88 at these time points suggest that there was a spontaneous increase in consumption which the model does not do a great job at predicting. Nonetheless, the above observations suggest that the model does a good job at predicting the dependent variable. However it should be noted that, as mentioned at the start of this chapter,

these results are generated using the full sample and not a training set. The model is therefore not tested on an unseen, testing dataset. This is one of the limitations of this section and could be a section that can be focused on in future research.

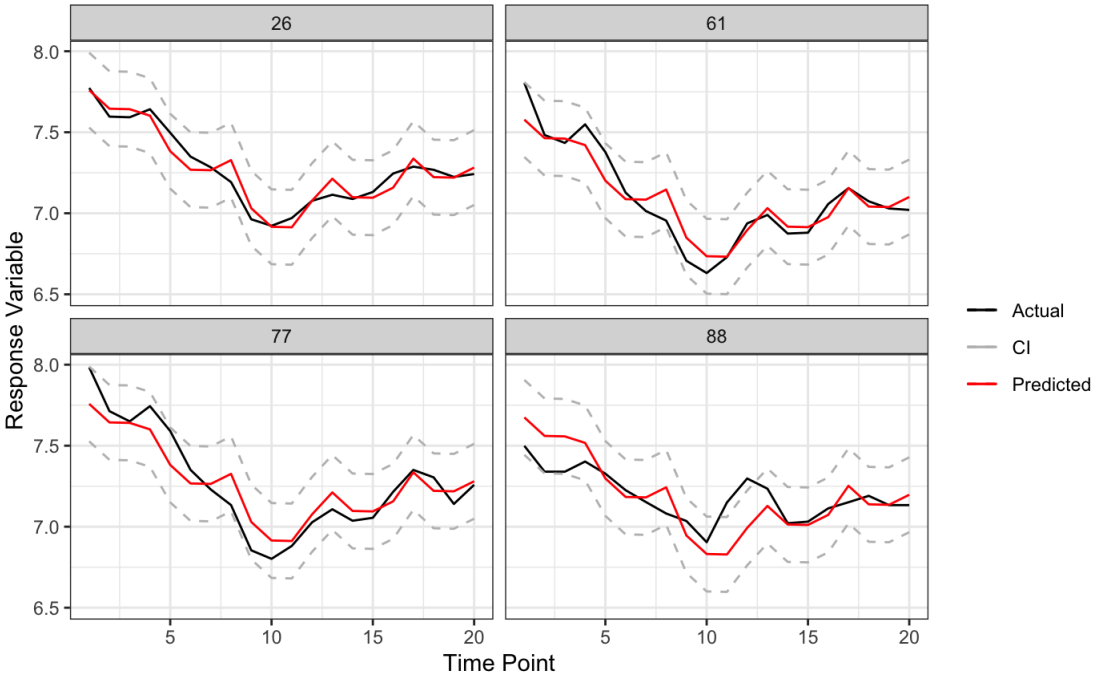


Figure 5.4: Visualising in-sample prediction for wards 22, 61, 77 and 88

5.3 Spatial Analysis

There are several studies that have focused on testing for cross-sectional dependence in spatial econometrics literature. This thesis focus on the work done by (Baltagi et al. 2003, 2007) who explore spatial dependencies in panel data. Baltagi et al. (2007) derive a joint Lagrange Multiplier (LM) test for the existence of spatial error correlation as well as random region effects in a panel data regression model.

The Baltagi, Heun Song, Cheol Jung & Koh (2007) LM tests for spatial error correlation in panel models is implemented using the *bsjktest* function the *splm* package (Bivand et al. 2021, Millo et al. 2012). The null and alternative hypothesis for this test is as follows:

H_0: No spatial dependence in error terms

H_A: Spatial dependence in error terms

The results from the test produced a *p-value* < 0.001, leading to the rejection of null hypothesis and concluding that the errors in the best performing panel model are spatially autocorrelated. As a result, the exploration of the spatial element in the panel data set is justified and necessary.

5.3.1 Moran's I index

As done in prior literature, this thesis makes use of the Moran's I index to detect spatial autocorrelation in the data (Bivand & Wong 2018, Kiziltan 2021, Bivand et al. 2013). This index measures spatial autocorrelation using a spatial weights matrix in weights list form. As discussed in Section 3, this thesis uses contiguity matrices for the weights matrix and for the Moran's I index only the queen contiguity matrix is used. Given a set of features and an associated attribute, Moran's I evaluates whether the pattern expressed is clustered, dispersed, or random (Bivand & Wong 2018). The correlation coefficient takes on values ranging from -1 to 1. A positive and statistically significant value of the Moran's I statistic indicates spatial clustering (1 indicates perfect clustering of similar values), while a negative and statistically significant Moran's I value shows spatial spillover (-1 is perfect clustering of dissimilar values). If Moran's I statistic is equal to zero, this means no spatial autocorrelation (Kiziltan 2021).

One of the limitations to the Moran's I statistic is that the test gives a single and average correlation for all wards included in the sample. This meaning that some wards may have a negative spatial autocorrelation, while others have a positive spatial autocorrelation (Kiziltan 2021). We would expect the above to be true for the observations found in the dataset used in this thesis given the socioeconomic structure of Cape Town. Nonetheless, results show positive and significant relationship.

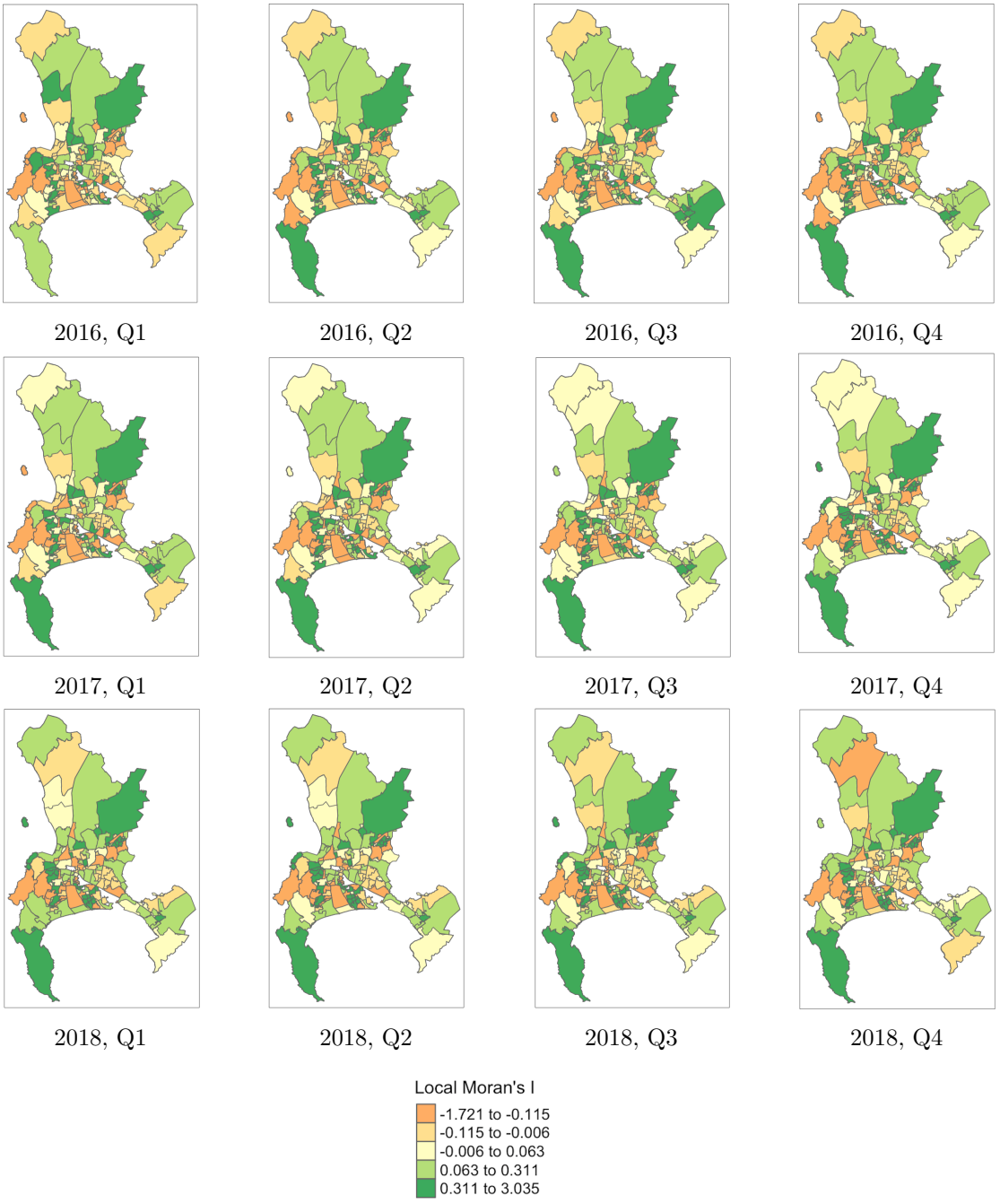
Quarter	Moran's I	z	p value
Year = 2016			
1	0.102	4.763	0.000
2	0.109	5.072	0.000
3	0.114	5.288	0.000
4	0.111	5.134	0.000
Year = 2017			
1	0.110	5.111	0.000
2	0.108	5.014	0.000
3	0.107	4.979	0.000
4	0.109	5.083	0.000
Year = 2018			
1	0.104	4.854	0.000
2	0.105	4.875	0.000
3	0.102	4.734	0.000
4	0.098	5.579	0.000
Year = 2019			
1	0.095	4.435	0.000
2	0.103	4.812	0.000
3	0.102	4.761	0.000
4	0.101	4.723	0.000
Year = 2020			
1	0.096	4.513	0.000
2	0.105	4.893	0.000
3	0.107	4.966	0.000
4	0.105	4.872	0.000

Table 5.5: Moran's I values for quarterly ward water consumption

5.3.2 Local Moran's I

As mentioned in the methodology section, the Local Moran's I statistic is calculated and visualised in an attempt to disclose the spatial clusters within the dataset.

In Figure 5.6, the Local Moran test is performed using *spdep* package in R (Bivand & Wong 2018, Bivand et al. 2013). The figure shows possible clustering of wards quarterly, over the five years. Although some wards change categories (change colour) over time, on average the same wards appear to cluster together. This is a valuable insight into ward behaviour with respect to water consumption. LISA allows us to see which wards behave in similar ways with regards to their water consumption behaviour and as a result groups wards into various clusters. Clustering wards based on their behaviour will allow ward councils to partner up and enforce similar policies for wards belonging to the same clusters. The insights from Figure 5.6 enrich the findings from Figure 4.5, using these two visuals in conjunction with one another prove to be a valuable source of insight into ward behaviour over time.


2016, Q1
2016, Q2
2016, Q3
2016, Q4
2017, Q1
2017, Q2
2017, Q3
2017, Q4
2018, Q1
2018, Q2
2018, Q3
2018, Q4
Local Moran's I
-1.721 to -0.115
-0.115 to -0.006
-0.006 to 0.063
0.063 to 0.311
0.311 to 3.035

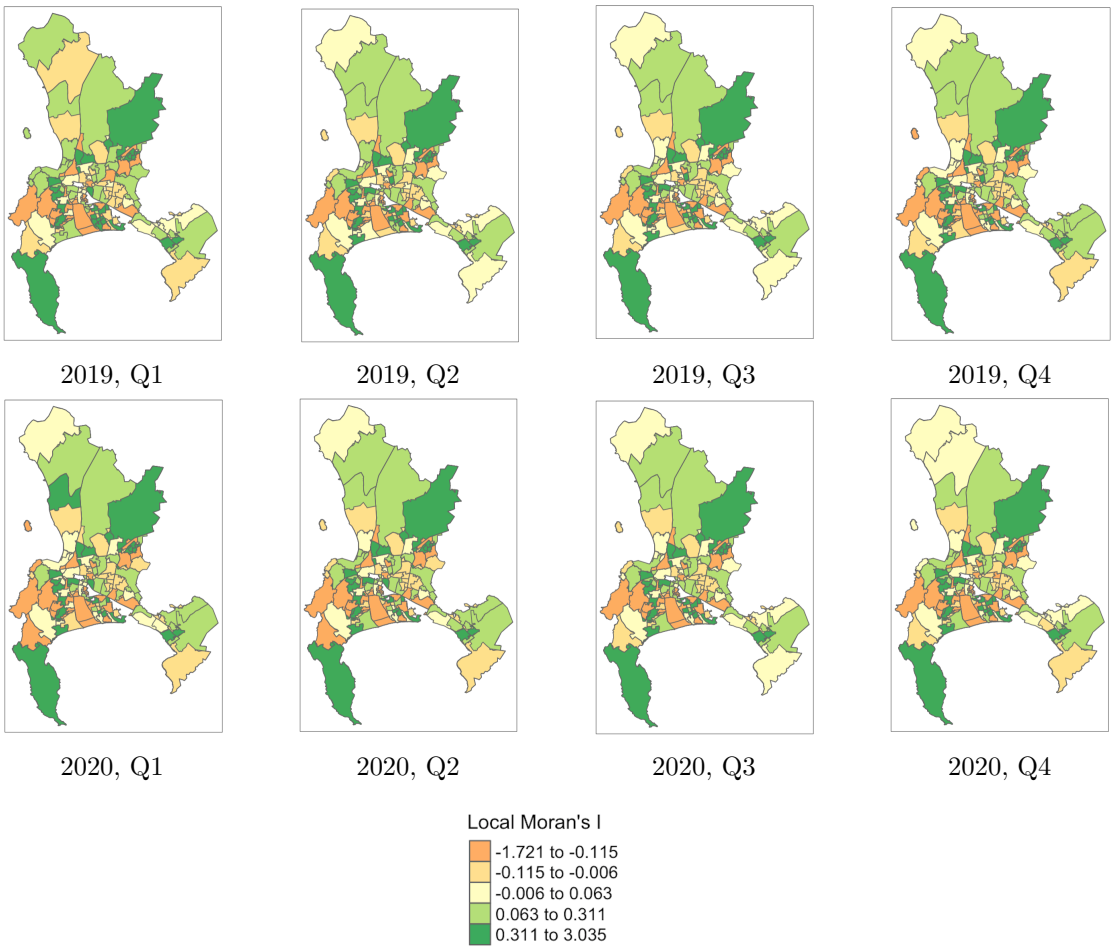


Figure 5.6: Local Moran's I, by year & quarter

5.3.3 Spatial Panel Model

The same attributes used in the best performing model in sub-section 5.2.2 is used in the spatial panel models implemented in this section of the thesis. The formula being:

$$ln\_consump = drought\_ind + quarter + year \tag{5.6}$$

The results from the various fixed effects spatial panel models are displayed in Table 5.6. Due to the software used to create the spatial models reported in this section, it was not possible to test the performance of these models on unseen data. As a result the RMSE values reported in this section are those generated from the full training set. This is once again a limitation of the results reported in this thesis and the low RMSE

values may be a result of over fitting. Future research can focus on using these models to predict consumption based on unseen data, doing this will give a true representation on the predictive power of these models.

Model	Effect	Spatial Lag	Spatial.Error	Weights	RMSE
1 SAR	Individual	TRUE	b	Queen	0.0942
2 SAR	Individual	TRUE	b	Rook	0.0941
3 SEM	Individual	FALSE	b	Queen	0.1147
4 SEM	Individual	FALSE	b	Rook	0.1147

Table 5.6: Spatial Panel Model Results using the the Baltagi et al. (2003) methodology

The approach taken throughout this paper has been to select the best performing model based on its RMSE value and Adjusted R^2 value. Leading on from this logic, one would choose the spatial panel models over the panel models that do not take spatial dependencies into account as the associated RMSE values for both the SAR and SEM are smaller than those of the standard fixed effects panel models. Setting the *Spatial.Error* parameter in the splm models to *'b'* ensures the Baltagi et al. (2003) methodology and equations displayed in Section 3 is adopted for both the SAR and SE models. Doing the above means that the idiosyncratic errors are spatially auto correlated and follow a spatial autoregressive process seen in Equation 3.14.

Spatial Error Models

Although the SE models produce very low RMSE values, it is mentioned above that the SAR model's outperform the SE model's by having lower associated RMSE values. This being said it should be noted all the β coefficients as well as the ρ coefficient is significant in both the SE models. This implies that including spatial dependencies in the model is crucial, failing to ignore it would introduce bias into a panel model. The outputs for all the SE models can be found in the code associated with this thesis.

Spatial Lag Models

Both SAR models implemented in this section have small RMSE values in comparison to all other models used and reported on in this thesis. The difference between the two SAR models is the weights matrix used. Both models use contiguity based spatial weights except, model 1 uses queen contiguity while model 2 uses Rook contiguity. Using the Rook contiguity evidently produces the smallest RMSE value, the difference is however minuscule.

The model coefficients for the SAR model using the Rook contiguity weights matrix can be seen in Table 5.7.

	Estimate	Std. Error	t value	Pr(>|t|)
lambda	0.8122	0.0195	41.672	0.0000
Drought	-0.0193	0.0050	-3.8374	0.0000
Q2	-0.0213	0.0042	-5.0763	0.0000
Q3	-0.0217	0.0042	-5.1426	0.0000
Q4	-0.0102	0.0037	-2.7384	0.0000
2017	-0.0498	0.0074	-6.7428	0.0000
2018	-0.1135	0.0126	-8.9747	0.0000
2019	-0.1003	0.0112	-8.9505	0.0000
2020	-0.0775	0.0090	-8.5622	0.0000

Table 5.7: Fixed Effects Spatial Panel Model - Model 2

In Table the small *p-value* associated with the lambda coefficient implies that the null hypothesis, $\lambda = 0$, can be rejected and evidently conclude that λ differs from 0. This meaning that spatial effects do in fact play a role in predicting the dependent variable. Although not reported here the lambda coefficient in SAR model 1 is also significant. Based on these findings, it can be assumed that leaving these spatial effects out of the model would result in bias estimates for a panel model.

Comparing the outputs from the panel model in the previous section, in Table 5.3, to the result of the SAR model in Table 5.7 it is noted that the magnitude of the β estimates decrease once spatial effects are accounted for. The nature of the relationship between the independent and dependent variables do not change from one model to the next, with negative relationships between all the independent variables in both the model outputs in Table 5.3 and Table 5.7.

The estimates reported in Table 5.7 suggest that during the first quarter of the year consumption is higher than all other quarters. For example, holding all else constant, if a consumption observation is taken in the third quarter of the year (Q3) the the water consumption value will be approximately 5.35% less than quarter one (Q1). From the model results, it is also clear that the pre-drought year, 2016, had the highest consumption values and evidently consumption has not necessarily returned to pre-drought levels. The estimates in Table 5.7 clearly show that during a drought, holding all else constant, consumption is approximately 4.71% lower implying that consumers behave differently during a drought period. Knowing whether this behaviour is influenced by

heightened tariffs or behavioural nudges is unclear in this model but this is something that can be further researched in the future.

As done for the best performing panel model, using four randomly selected wards, the actual versus the predicted dependent variable are plotted using the fixed effects SAR Model 2 in Figure 5.7. A 95% confidence interval is also included for the predicted values in Figure 5.7. Comparing Figure 5.4 and 5.7 it is clear that the predicted values estimated using a spatial panel model are closer to the actual values - this is justified by the fact that the red line depicting predicted values is closer to the black line representing actual values. One would expect this to be the case as the RMSE associated with the spatial model is smaller than that of the standard fixed effects panel model. It should also be noted that all the actual and predicted values fall within the 95% confidence interval, even ward 88 which experienced the large jump in consumption from time point 10 to 11. When comparing Figure 5.4 and 5.7, the results presented in Figure 5.7 help promote the spatial model implemented in this thesis.

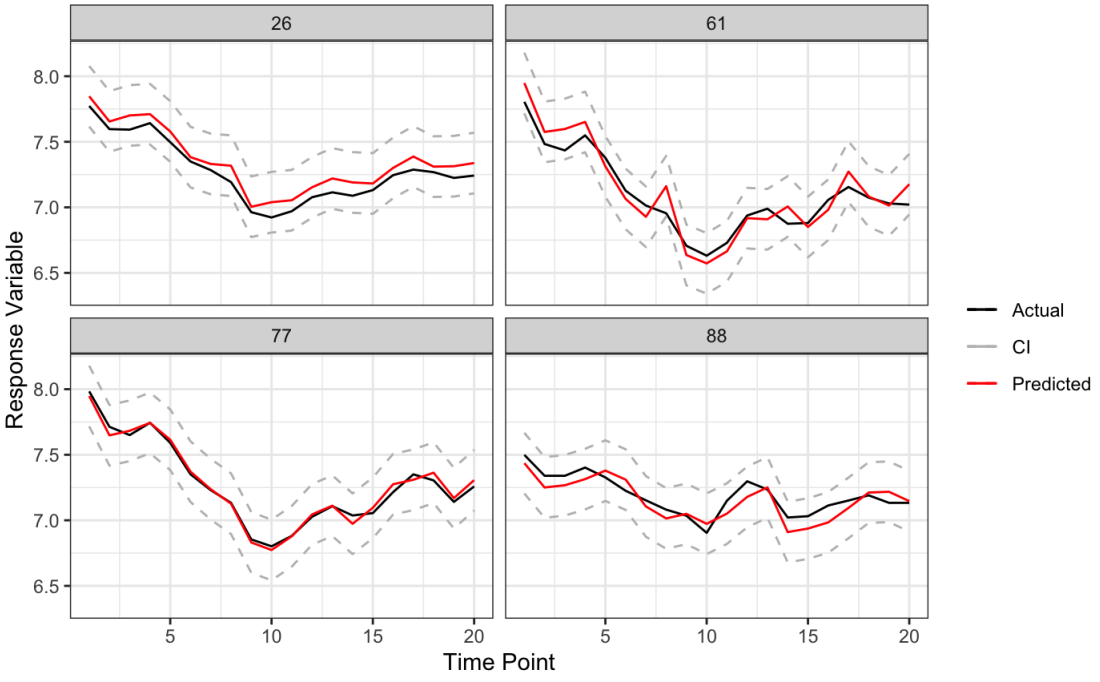


Figure 5.7: SAR Model 2 predictions for wards 22, 61, 77 and 88

As done with the panel model predictions, all other wards actual versus predicted graphs can be found in Appendix D. The actual versus predicted graphs found in the Appendix, for the remaining wards, show similar results to the ones represented in Figure 5.7. The plotted results indicate that, for the large majority of the wards, all the actual and

predicted values fall within the 95% confidence interval. The model is able to predict consumption levels for both wards with low consumption levels (for example ward 6) and high consumption levels (for example ward 7). The model was also able to capture spikes in ward consumption levels, which is favourable to see given the nature of water consumption behaviour.

Chapter 6

Conclusion

This thesis concludes by reinforcing that spatial effects prevail when conducting research on household water consumption in the CoCT. This meaning that the models and other spatial analysis tools implemented in this thesis provide evidence that ward consumption is affected by spatial features. We would therefore expect consumption levels to be similar amongst wards that are closer to one another.

Finding significant spatial effects in the dataset may be advantageous to policy makers as different ward councils may adopt same or similar strategies when it comes to water supply interventions. Instead of applying the same water restrictions or intervention methods to the whole of the CoCT, the various ward councils can justifiably implement ward specific intervention methods such as, but not limited to, ward specific tariffs and behavioural nudge campaigns. This is backed by the idea that the data on household ward consumption over the five year period of this study show spatial clusters in ward consumption.

The results of this thesis also reveal that wards with similar consumption behaviour during time periods where water is not scarce show similar consumption behaviour during time periods where water is scarce. This remark is justified as the data examined in the current thesis looks at observations between 2016 and 2020, during this time frame a drought occurred in the CoCT.

This thesis has explored a billing consumption dataset in the CoCT and the findings of this thesis can be built on by future researchers. Future research topics can focus on using the spatial panel model (fixed effects SAR model with a rook contiguity based

weights matrix) to forecast future ward water consumption in the CoCT. The models included in this thesis have proven to be good predictors of water consumption and being able to use them as a forecasting tool would help the CoCT predict and plan how to control the supply of water in the future. By including a drought indicator in the models allows for further use cases for those using these model to predict water consumption and to plan accordingly when water sources are scarce.

Apart from this, future work can focus on implementing non-linear models to predict water consumption. Initially the intention of this thesis was to implement a LSTM Recurrent Neural Network however after a thorough investigation it was deemed that the data used in this thesis was not appropriate for this model. Firstly if one wanted to include a dummy variable for the unique household identifier attribute, there would be too many factor levels in the attribute to be included in a LSTM. Even when broadening the grain of the research to ward level, this too still had too many factor levels to track over time for a LSTM. Another reason as to why a LSTM was not implemented in this thesis was motivated by the missing observations present in the dataset. Although the dataset contains a lot of observations, these observations are all associated with unique households in the dataset. Neural Networks, in general, perform best when there is a large amount of data being processed. If a household did not have any missing observations there would only be 60 observations per household and as a result each household does not have enough observations for the LSTM to perform optimally, especially once the observations are aggregated from monthly to quarterly consumption values.

Although the two aforementioned reasons discuss why a LSTM is not implemented in this thesis, further research can be done to determine an approach that finds an alternative method to using a dummy variable for household unique identifiers. Future research can look into an embedding layer being incorporated into the LSTM and using a household identifier as a latent variable in the model. It should also be noted that, naturally, as the data set grows with time the use of LSTM and other machine learning methods may be preferable. This is because these types of statistical models perform best with large amounts of data.

Appendix A

Figures

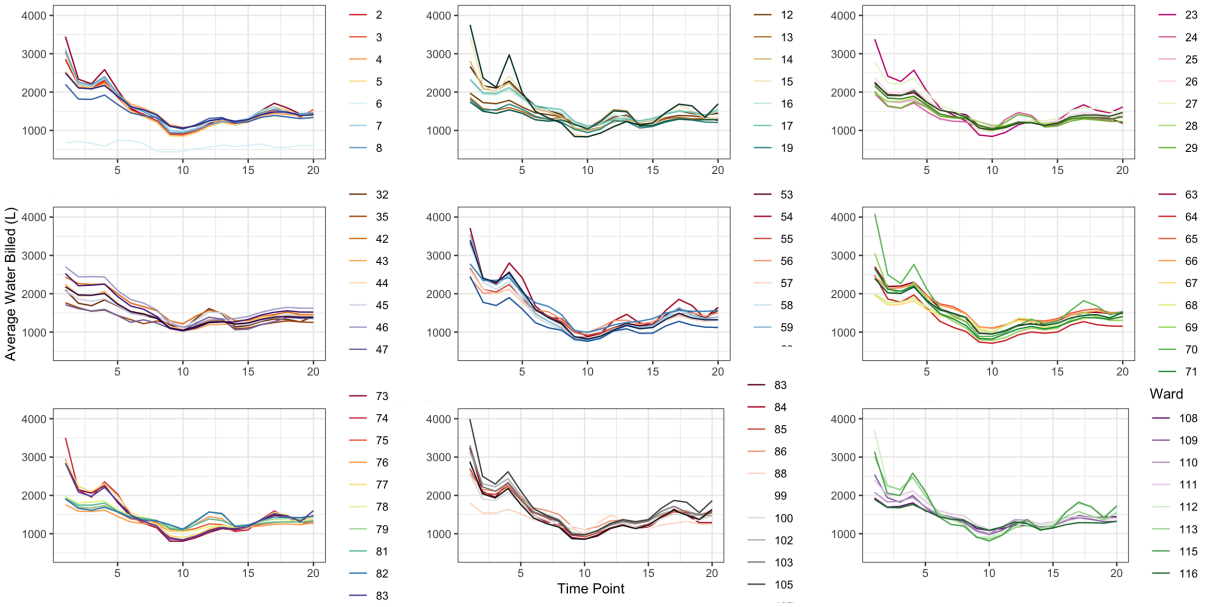


Figure A.1: Average quarterly water consumption (L), by ward

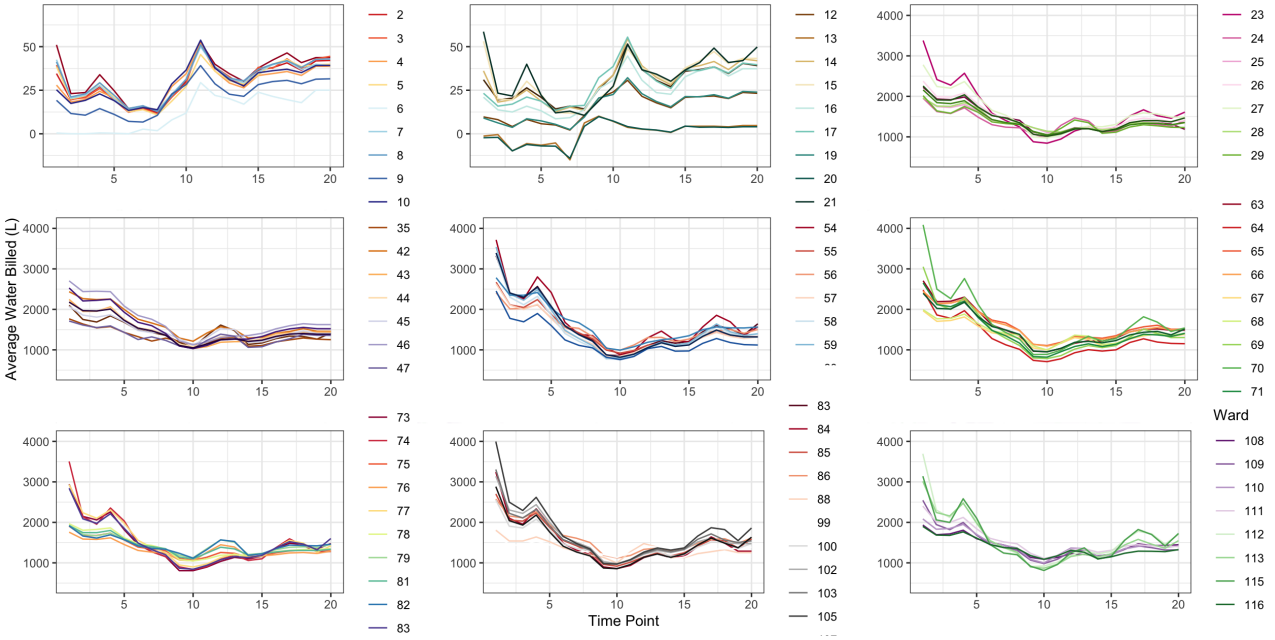


Figure A.2: Average quarterly amount billed (ZAR), by ward

Appendix B

Linear Model Results

Linear model 1 results can be found in Table B.1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.2636	0.0248	292.50	0.0000
Drought	-0.0993	0.0132	-7.50	0.0000
Q2	-0.1142	0.0183	-6.22	0.0000
Q3	-0.1172	0.0183	-6.39	0.0000
Q4	-0.0551	0.0183	-3.01	0.0027
pp_ind	0.1171	0.0226	5.19	0.0000

Table B.1: MLR - Model 1

Linear model 2 results can be found in Table B.2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.3533	0.0585	125.70	0.0000
Drought	-0.0993	0.0125	-7.96	0.0000
Q2	-0.1142	0.0173	-6.61	0.0000
Q3	-0.1172	0.0173	-6.78	0.0000
Q4	-0.0551	0.0173	-3.19	0.0014
pp_ind	0.0823	0.0811	1.02	0.3102
ward_2	-0.0459	0.0811	-0.57	0.5713
ward_3	-0.0504	0.0811	-0.62	0.5339
ward_4	-0.0114	0.0811	-0.14	0.8877
ward_5	-0.0544	0.0811	-0.67	0.5024
ward_6	-0.8576	0.0811	-10.58	0.0000
ward_7	0.0009	0.0811	0.01	0.9913
ward_8	-0.0142	0.0811	-0.17	0.8613
ward_9	-0.0746	0.0811	-0.92	0.3577
ward_10	-0.0116	0.0811	-0.14	0.8862
ward_11	-0.0099	0.0811	-0.12	0.9027
ward_12	-0.0858	0.0811	-1.06	0.2897
ward_13	-0.1073	0.0811	-1.32	0.1857
ward_14	-0.0189	0.0811	-0.23	0.8154
ward_15	-0.0580	0.0811	-0.72	0.4743
ward_16	-0.0046	0.0811	-0.06	0.9545
ward_17	0.0028	0.0811	0.03	0.9722
ward_19	-0.1736	0.0811	-2.14	0.0324
ward_20	-0.1430	0.0811	-1.76	0.0778
ward_21	-0.0180	0.0811	-0.22	0.8245
ward_22	-0.0769	0.0811	-0.95	0.3430
ward_23	-0.0077	0.0811	-0.09	0.9246
ward_24	-0.1235	0.0811	-1.52	0.1277
ward_25	-0.0981	0.0811	-1.21	0.2264
ward_26	-0.0512	0.0811	-0.63	0.5281
ward_27	0.0053	0.0811	0.07	0.9479
ward_28	-0.0912	0.0811	-1.13	0.2605
ward_29	-0.1228	0.0811	-1.51	0.1300
ward_30	-0.0971	0.0811	-1.20	0.2310
ward_31	-0.0690	0.0811	-0.85	0.3949
ward_32	-0.0612	0.0811	-0.76	0.4502
ward_35	-0.0667	0.0811	-0.82	0.4108

	Estimate	Std. Error	t value	Pr(>|t|)
ward_42	0.1464	0.0811	1.81	0.0712
ward_43	-0.0489	0.0811	-0.60	0.5467
ward_44	-0.0365	0.0811	-0.45	0.6523
ward_45	-0.0344	0.0811	-0.42	0.6713
ward_46	0.0884	0.0811	1.09	0.2756
ward_47	-0.1447	0.0811	-1.78	0.0744
ward_48	0.0125	0.0811	0.15	0.8775
ward_49	-0.0532	0.0811	-0.66	0.5115
ward_53	-0.0185	0.0811	-0.23	0.8193
ward_54	0.0495	0.0811	0.61	0.5414
ward_55	-0.0734	0.0811	-0.91	0.3651
ward_56	-0.0290	0.0811	-0.36	0.7208
ward_57	-0.1171	0.0811	-1.45	0.1486
ward_58	-0.1038	0.0811	-1.28	0.2005
ward_59	-0.0631	0.0811	-0.78	0.4361
ward_60	0.0334	0.0811	0.41	0.6807
ward_61	-0.2328	0.0811	-2.87	0.0041
ward_62	-0.0618	0.0811	-0.76	0.4458
ward_63	-0.0171	0.0811	-0.21	0.8334
ward_64	-0.2405	0.0811	-2.97	0.0030
ward_65	0.0344	0.0811	0.42	0.6710
ward_66	0.0196	0.0811	0.24	0.8088
ward_67	-0.1097	0.0811	-1.35	0.1760
ward_68	-0.0831	0.0811	-1.03	0.3054
ward_69	-0.1171	0.0811	-1.45	0.1486
ward_70	0.0332	0.0811	0.41	0.6820
ward_71	-0.1073	0.0811	-1.32	0.1859
ward_72	-0.0516	0.0811	-0.64	0.5247
ward_73	-0.0983	0.0811	-1.21	0.2255
ward_74	-0.0817	0.0811	-1.01	0.3136
ward_75	-0.1221	0.0811	-1.51	0.1322
ward_76	-0.1414	0.0811	-1.74	0.0814
ward_77	-0.0529	0.0811	-0.65	0.5140
ward_78	-0.0993	0.0811	-1.22	0.2208
ward_79	-0.0597	0.0811	-0.74	0.4618

	Estimate	Std. Error	t value	Pr(>|t|)
ward_81	-0.0867	0.0811	-1.07	0.2850
ward_82	-0.0588	0.0811	-0.72	0.4686
ward_83	-0.0797	0.0811	-0.98	0.3257
ward_84	-0.0686	0.0811	-0.85	0.3976
ward_85	0.0421	0.0811	0.52	0.6037
ward_86	0.1228	0.0811	1.52	0.1299
ward_88	-0.0539	0.0811	-0.67	0.5059
ward_99	-0.0506	0.0811	-0.62	0.5322
ward_100	-0.0560	0.0811	-0.69	0.4900
ward_102	0.0189	0.0811	0.23	0.8160
ward_103	0.0174	0.0811	0.22	0.8296
ward_105	0.0830	0.0811	1.02	0.3060
ward_107	-0.0719	0.0811	-0.89	0.3749
ward_109	-0.0715	0.0811	-0.88	0.3782
ward_110	-0.0608	0.0811	-0.75	0.4530
ward_111	-0.0055	0.0811	-0.07	0.9457
ward_112	0.0170	0.0811	0.21	0.8336
ward_113	-0.0362	0.0811	-0.45	0.6556
ward_115	-0.0198	0.0811	-0.24	0.8073
ward_116	-0.1104	0.0811	-1.36	0.1733

Table B.2: MLR - Model 2

Linear model 3 results can be found in Table B.3.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.6387	0.0163	469.34	0.0000
Drought	-0.1021	0.0137	-7.46	0.0000
Q2	-0.1142	0.0106	-10.76	0.0000
quarter3	-0.1172	0.0106	-11.04	0.0000
Q3	-0.0551	0.0106	-5.20	0.0000
pp_ind	0.1171	0.0130	8.98	0.0000
2017	-0.2745	0.0157	-17.50	0.0000
2018	-0.6267	0.0137	-45.76	0.0000
2019	-0.5464	0.0123	-44.27	0.0000
2020	-0.4223	0.0123	-34.21	0.0000

Table B.3: MLR - Model 3

Linear model 4 results can be found in Table B.4.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.7284	0.0275	280.97	0.0000
Drought	-0.1021	0.0103	-9.96	0.0000
Q2	-0.1142	0.0079	-14.36	0.0000
Q3	-0.1172	0.0079	-14.74	0.0000
Q4	-0.0551	0.0079	-6.94	0.0000
pp_ind	0.0823	0.0373	2.21	0.0274
2017	-0.2745	0.0118	-23.35	0.0000
2018	-0.6267	0.0103	-61.08	0.0000
2019	-0.5464	0.0092	-59.09	0.0000
2020	-0.4223	0.0092	-45.66	0.0000
ward_2	-0.0459	0.0373	-1.23	0.2184
ward_3	-0.0504	0.0373	-1.35	0.1762
ward_4	-0.0114	0.0373	-0.31	0.7589
ward_5	-0.0544	0.0373	-1.46	0.1448
ward_6	-0.8576	0.0373	-23.01	0.0000
ward_7	0.0009	0.0373	0.02	0.9811
ward_8	-0.0142	0.0373	-0.38	0.7039
ward_9	-0.0746	0.0373	-2.00	0.0456
ward_10	-0.0116	0.0373	-0.31	0.7556
ward_11	-0.0099	0.0373	-0.27	0.7903
ward_12	-0.0858	0.0373	-2.30	0.0214
ward_13	-0.1073	0.0373	-2.88	0.0040
ward_14	-0.0189	0.0373	-0.51	0.6117
ward_15	-0.0580	0.0373	-1.56	0.1199
ward_16	-0.0046	0.0373	-0.12	0.9012
ward_17	0.0028	0.0373	0.08	0.9396
ward_19	-0.1736	0.0373	-4.66	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
ward_20	-0.1430	0.0373	-3.84	0.0001
ward_21	-0.0180	0.0373	-0.48	0.6297
ward_22	-0.0769	0.0373	-2.06	0.0393
ward_23	-0.0077	0.0373	-0.21	0.8369
ward_24	-0.1235	0.0373	-3.31	0.0009
ward_25	-0.0981	0.0373	-2.63	0.0086
ward_26	-0.0512	0.0373	-1.37	0.1702
ward_27	0.0053	0.0373	0.14	0.8870
ward_28	-0.0912	0.0373	-2.45	0.0145
ward_29	-0.1228	0.0373	-3.29	0.0010
ward_30	-0.0971	0.0373	-2.61	0.0092
ward_31	-0.0690	0.0373	-1.85	0.0644
ward_32	-0.0612	0.0373	-1.64	0.1007
ward_35	-0.0667	0.0373	-1.79	0.0738
ward_42	0.1464	0.0373	3.93	0.0001
ward_43	-0.0489	0.0373	-1.31	0.1901
ward_44	-0.0365	0.0373	-0.98	0.3273
ward_45	-0.0344	0.0373	-0.92	0.3562
ward_46	0.0884	0.0373	2.37	0.0178
ward_47	-0.1447	0.0373	-3.88	0.0001
ward_48	0.0125	0.0373	0.34	0.7375
ward_49	-0.0532	0.0373	-1.43	0.1535
ward_53	-0.0185	0.0373	-0.50	0.6194
ward_54	0.0495	0.0373	1.33	0.1842
ward_55	-0.0734	0.0373	-1.97	0.0490
ward_56	-0.0290	0.0373	-0.78	0.4371
ward_57	-0.1171	0.0373	-3.14	0.0017
ward_58	-0.1038	0.0373	-2.79	0.0054
ward_59	-0.0631	0.0373	-1.69	0.0905
ward_60	0.0334	0.0373	0.89	0.3709
ward_61	-0.2328	0.0373	-6.24	0.0000
ward_62	-0.0618	0.0373	-1.66	0.0974
ward_63	-0.0171	0.0373	-0.46	0.6473
ward_64	-0.2405	0.0373	-6.45	0.0000
ward_65	0.0344	0.0373	0.92	0.3556
ward_66	0.0196	0.0373	0.53	0.5987
ward_67	-0.1097	0.0373	-2.94	0.0033
ward_68	-0.0831	0.0373	-2.23	0.0259
ward_69	-0.1171	0.0373	-3.14	0.0017
ward_70	0.0332	0.0373	0.89	0.3729
ward_71	-0.1073	0.0373	-2.88	0.0041
ward_72	-0.0516	0.0373	-1.38	0.1667
ward_73	-0.0983	0.0373	-2.64	0.0085
ward_74	-0.0817	0.0373	-2.19	0.0285
ward_75	-0.1221	0.0373	-3.27	0.0011
ward_76	-0.1414	0.0373	-3.79	0.0002
ward_77	-0.0529	0.0373	-1.42	0.1559
ward_78	-0.0993	0.0373	-2.66	0.0078
ward_79	-0.0597	0.0373	-1.60	0.1097

	Estimate	Std. Error	t value	Pr(>|t|)
ward_81	-0.0867	0.0373	-2.33	0.0201
ward_82	-0.0588	0.0373	-1.58	0.1151
ward_83	-0.0797	0.0373	-2.14	0.0327
ward_84	-0.0686	0.0373	-1.84	0.0659
ward_85	0.0421	0.0373	1.13	0.2591
ward_86	0.1228	0.0373	3.30	0.0010
ward_88	-0.0539	0.0373	-1.45	0.1481
ward_99	-0.0506	0.0373	-1.36	0.1744
ward_100	-0.0560	0.0373	-1.50	0.1334
ward_102	0.0189	0.0373	0.51	0.6128
ward_103	0.0174	0.0373	0.47	0.6399
ward_105	0.0830	0.0373	2.23	0.0261
ward_107	-0.0719	0.0373	-1.93	0.0538
ward_109	-0.0715	0.0373	-1.92	0.0554
ward_110	-0.0608	0.0373	-1.63	0.1028
ward_111	-0.0055	0.0373	-0.15	0.8823
ward_112	0.0170	0.0373	0.46	0.6478
ward_113	-0.0362	0.0373	-0.97	0.3322
ward_115	-0.0198	0.0373	-0.53	0.5959
ward_116	-0.1104	0.0373	-2.96	0.0031

Table B.4: MLR - Model 4

Appendix C

Panel Model Results

PLM - Model 1 results shown in Table C.1

C.1 Model Outputs

	Estimate	Std. Error	t value	Pr(>|t|)
Drought	-0.0993	0.0102	-7.9593	0.0000
Q2	-0.1142	0.0173	-6.6057	0.0000
Q3	-0.1172	0.0173	-6.7790	0.0000
Q4	-0.0551	0.0172	-3.1903	0.0000

Table C.1: Fixed Effects - Model 1

PLM - Model 2 results shown in Table C.2

C.2 Model Outputs

	Estimate	Std. Error	t value	Pr(>|t|)
pp_ind	0.1171	0.0119	9.849	0.0000

Table C.2: Fixed Effects - Model 2

C.3 Best Performing Panel Model results - Model 4

Using the PLM model that performs the best, Figures C.1, C.2, C.3, C.4, C.5 and C.6 depict the actual versus predicted response variable, by ward.

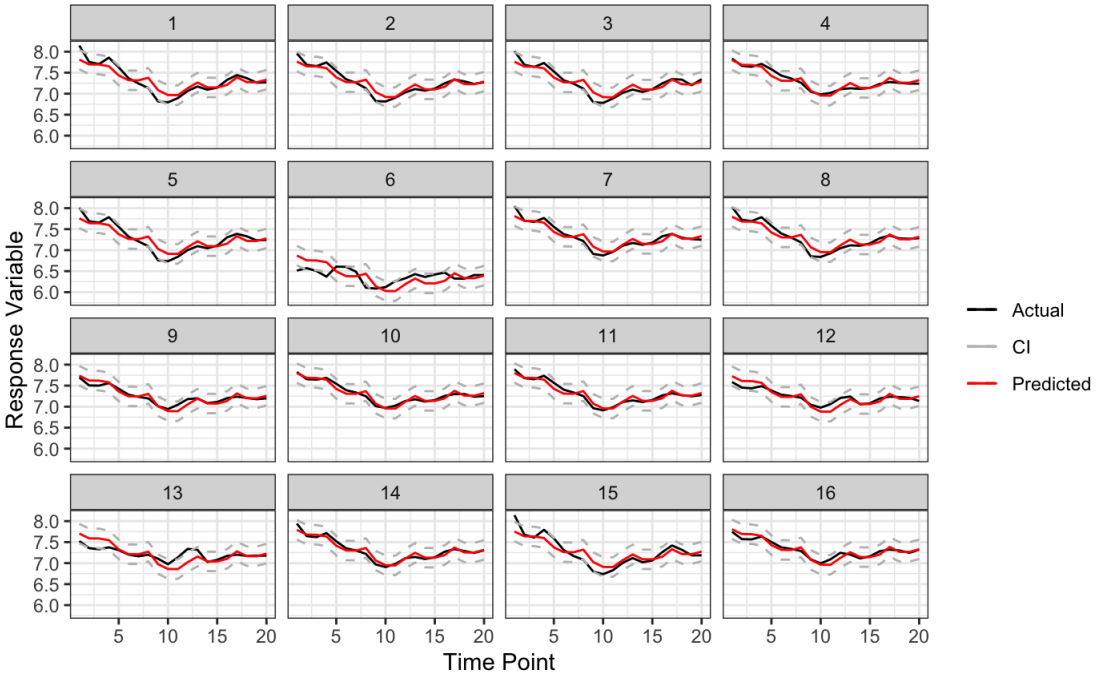


Figure C.1: Panel Model 4: Actual vs Predicted

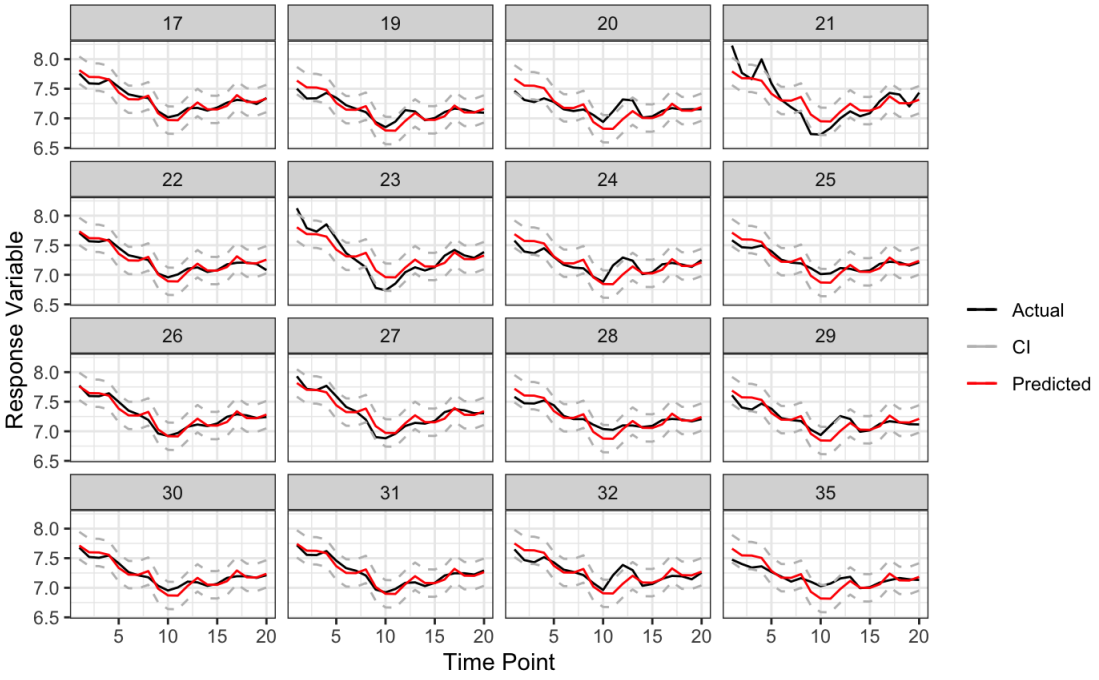


Figure C.2: Panel Model 4: Actual vs Predicted

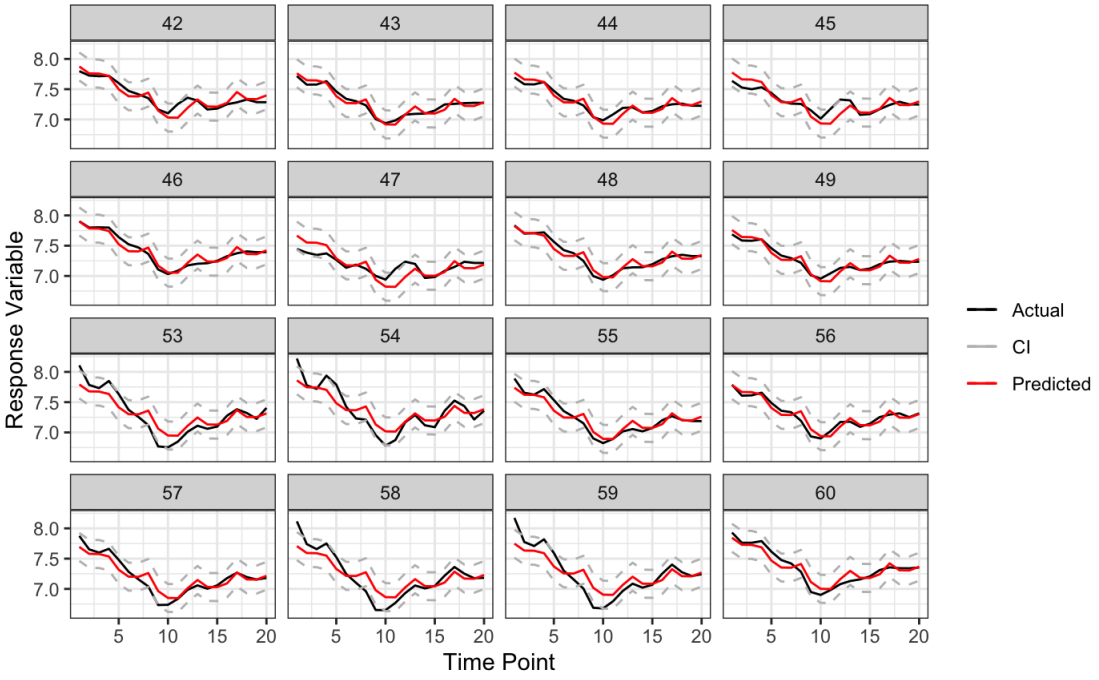


Figure C.3: Panel Model 4: Actual vs Predicted

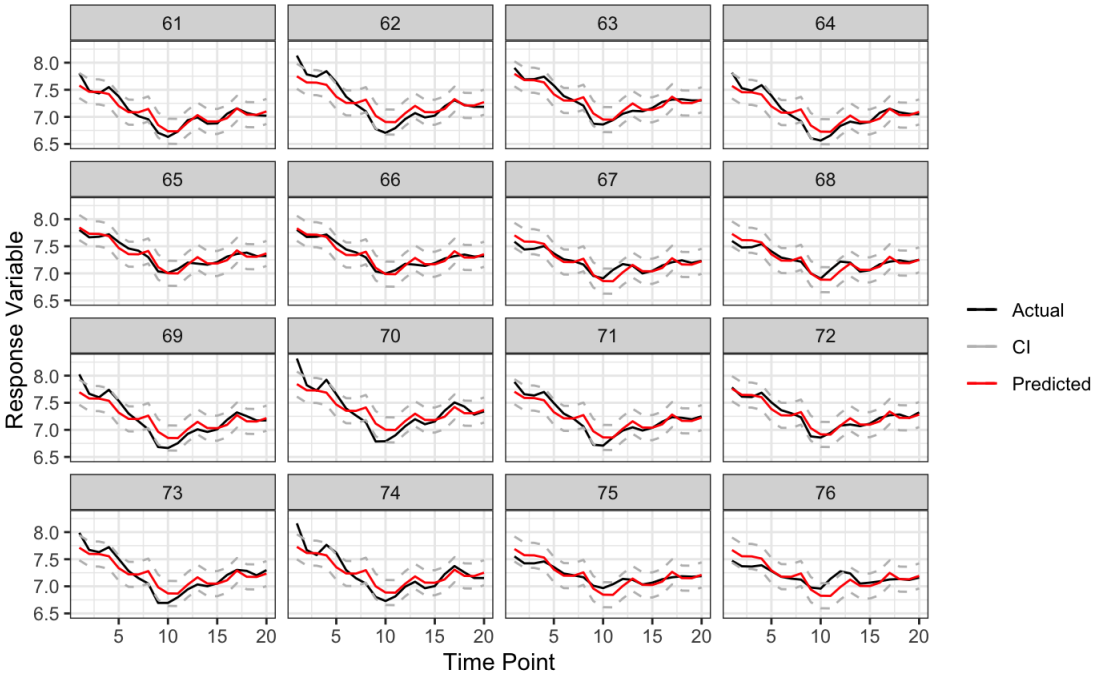


Figure C.4: Panel Model 4: Actual vs Predicted

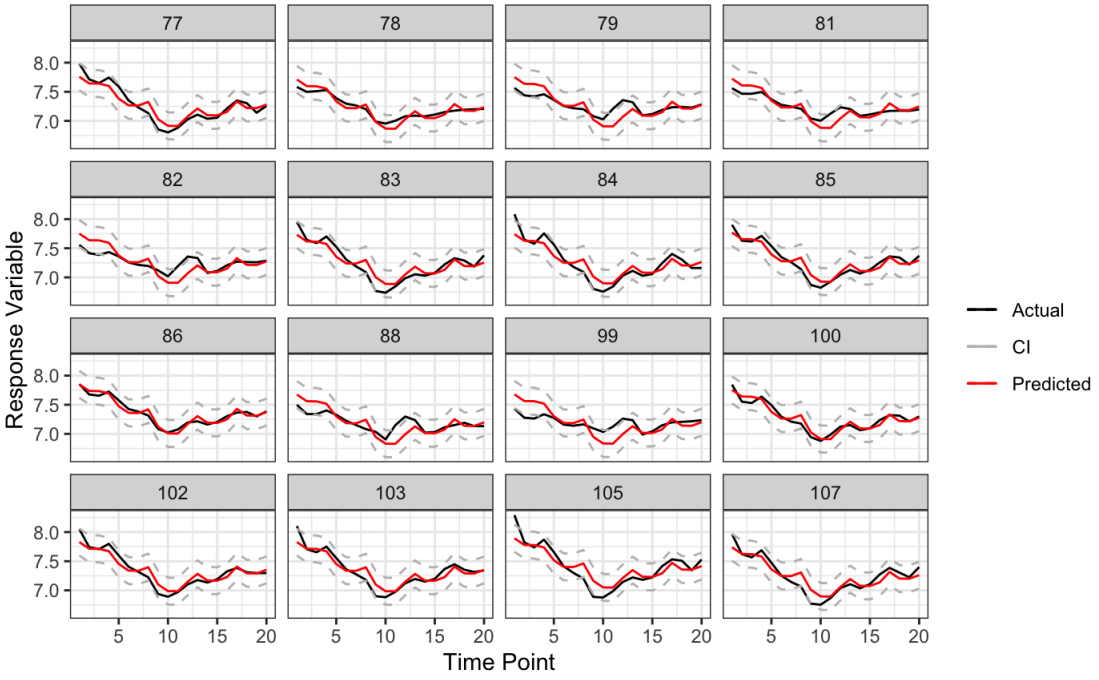


Figure C.5: Panel Model 4: Actual vs Predicted

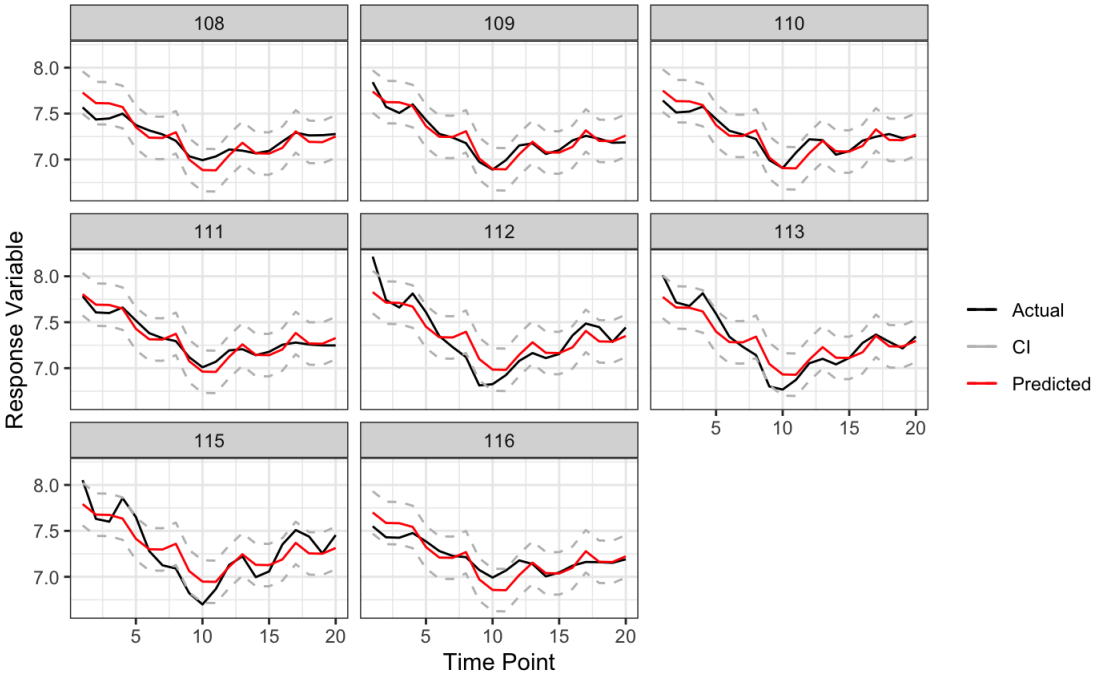


Figure C.6: Panel Model 4: Actual vs Predicted

C.4 Ward Fixed Effects - Model 4

Ward	Estimate	Std. Error	t-value	Pr(>|t|)
1	0.07	0.03	2.38	0.02
2	0.02	0.03	0.71	0.48
3	0.02	0.03	0.55	0.58
4	0.05	0.03	1.97	0.05
5	0.01	0.03	0.40	0.69
6	-0.87	0.03	-31.79	0.00
7	0.07	0.03	2.41	0.02
8	0.05	0.03	1.87	0.06
9	-0.01	0.03	-0.33	0.74
10	0.05	0.03	1.96	0.05
11	0.06	0.03	2.02	0.04
12	-0.02	0.03	-0.74	0.46
13	-0.04	0.03	-1.52	0.13
14	0.05	0.03	1.69	0.09
15	0.01	0.03	0.27	0.79
16	0.06	0.03	2.21	0.03
17	0.07	0.03	2.48	0.01
19	-0.11	0.03	-3.93	0.00
20	-0.08	0.03	-2.82	0.00
21	0.05	0.03	1.73	0.08
22	-0.01	0.03	-0.41	0.68
23	0.06	0.03	2.10	0.04
24	-0.06	0.03	-2.11	0.04
25	-0.03	0.03	-1.18	0.24
26	0.01	0.03	0.52	0.60
27	0.07	0.03	2.57	0.01
28	-0.03	0.03	-0.94	0.35
29	-0.06	0.03	-2.08	0.04
30	-0.03	0.03	-1.15	0.25
31	-0.00	0.03	-0.13	0.90
32	0.00	0.03	0.16	0.88
35	-0.08	0.03	-3.03	0.00

Ward	Estimate	Std. Error	t-value	Pr(>|t|)
42	0.13	0.03	4.71	0.00
43	0.02	0.03	0.61	0.55
44	0.03	0.03	1.05	0.29
45	0.03	0.03	1.13	0.26
46	0.15	0.03	5.60	0.00
47	-0.08	0.03	-2.88	0.00
48	0.08	0.03	2.84	0.00
49	0.01	0.03	0.45	0.66
53	0.05	0.03	1.71	0.09
54	0.12	0.03	4.18	0.00
55	-0.01	0.03	-0.29	0.77
56	0.04	0.03	1.33	0.18
57	-0.05	0.03	-1.88	0.06
58	-0.04	0.03	-1.39	0.16
59	0.00	0.03	0.09	0.93
60	0.10	0.03	3.59	0.00
61	-0.17	0.03	-6.08	0.00
62	0.00	0.03	0.13	0.89
63	0.05	0.03	1.76	0.08
64	-0.18	0.03	-6.36	0.00
65	0.10	0.03	3.63	0.00
66	0.09	0.03	3.09	0.00
67	-0.04	0.03	-1.61	0.11
68	-0.02	0.03	-0.64	0.52
69	-0.05	0.03	-1.88	0.06
70	0.10	0.03	3.59	0.00
71	-0.04	0.03	-1.52	0.13
72	0.01	0.03	0.51	0.61
73	-0.03	0.03	-1.19	0.23
74	-0.02	0.03	-0.59	0.56
75	-0.06	0.03	-2.06	0.04
76	-0.08	0.03	-2.76	0.01
77	0.01	0.03	0.46	0.65
78	-0.03	0.03	-1.23	0.22
79	0.01	0.03	0.21	0.83

Ward	Estimate	Std. Error	t-value	Pr(>|t|)
81	-0.02	0.03	-0.77	0.44
82	0.01	0.03	0.25	0.81
83	-0.01	0.03	-0.52	0.61
84	-0.00	0.03	-0.11	0.91
85	0.03	0.03	0.92	0.36
86	0.11	0.03	3.86	0.00
88	-0.07	0.03	-2.57	0.01
99	-0.07	0.03	-2.45	0.01
100	0.01	0.03	0.35	0.73
102	0.08	0.03	3.07	0.00
103	0.08	0.03	3.02	0.00
105	0.15	0.03	5.40	0.00
107	-0.01	0.03	-0.23	0.82
108	-0.02	0.03	-0.61	0.54
109	-0.01	0.03	-0.22	0.83
110	0.00	0.03	0.17	0.87
111	0.06	0.03	2.18	0.03
112	0.08	0.03	3.00	0.00
113	0.03	0.03	1.07	0.29
115	0.05	0.03	1.66	0.10
116	-0.04	0.03	-1.63	0.10

Table C.3: Fixed effects panel model 5, the fixed effects associated with all wards

Appendix D

Spatial Panel Model Results

D.1 Best Performing Spatial Panel Model results - Model 2

Using the SAR model that performs the best, Figures D.1, D.2, D.3, D.4, D.5 and D.6 depict the actual versus predicted response variable, by ward.

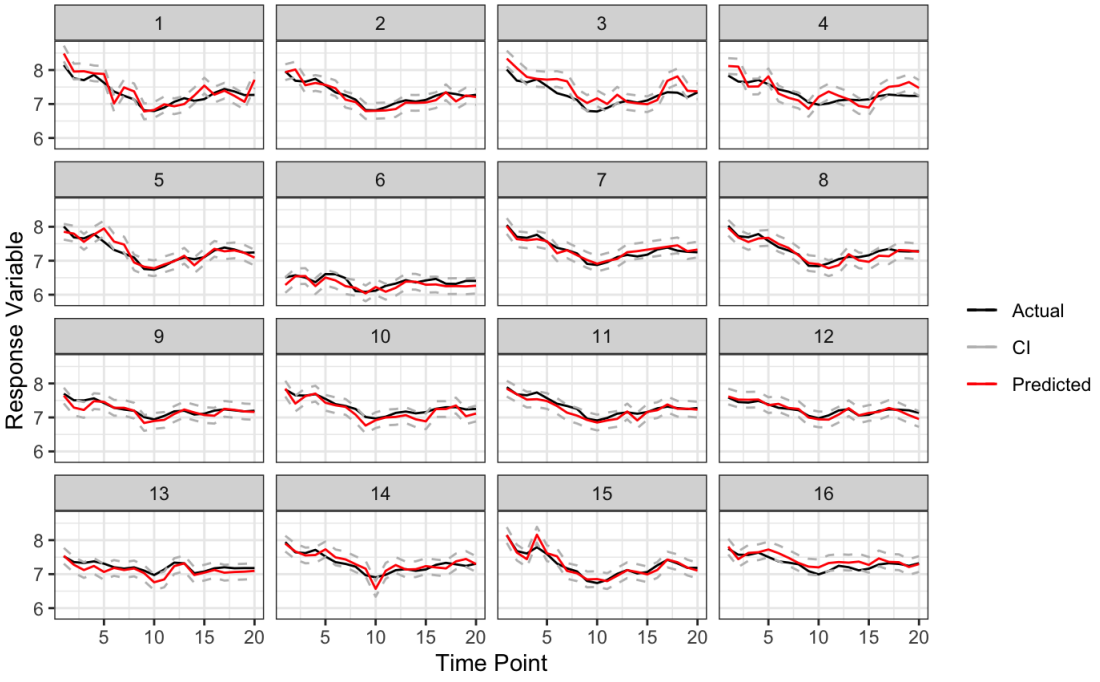


Figure D.1: Spatial Panel Model 2: Actual vs Predicted

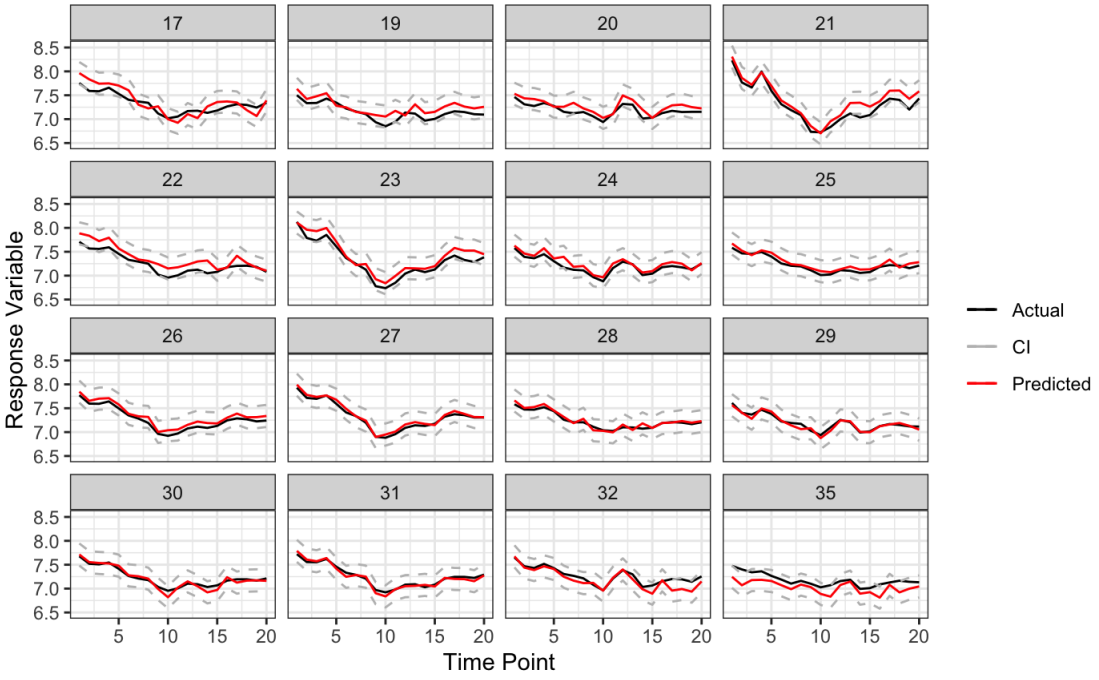


Figure D.2: Spatial Panel Model 2: Actual vs Predicted

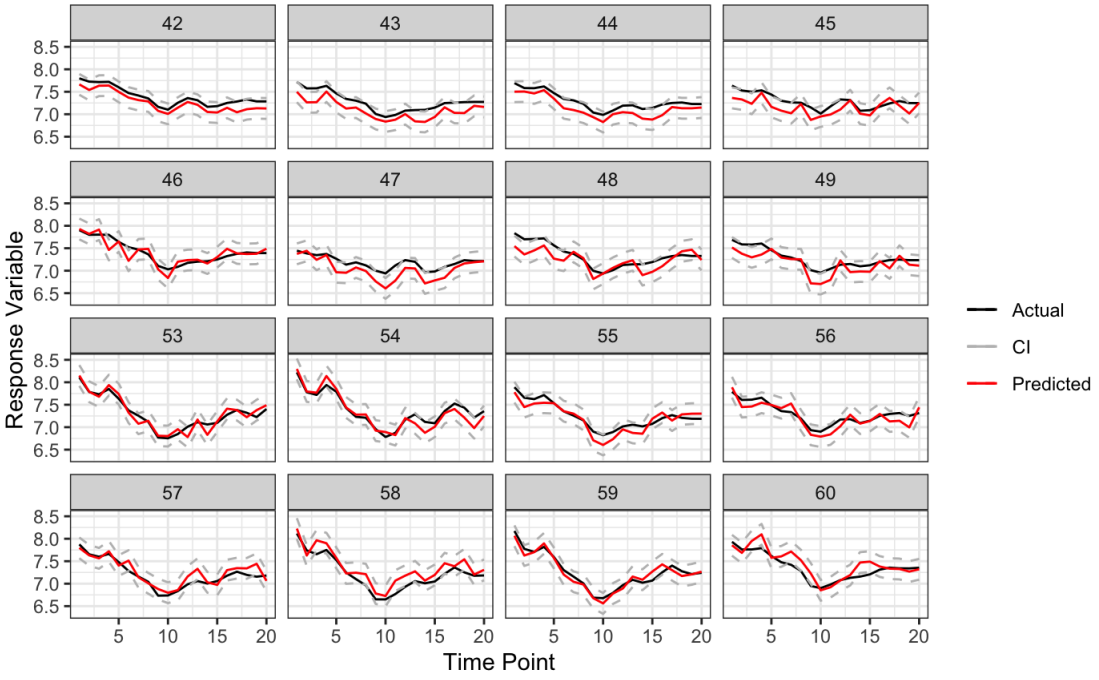


Figure D.3: Spatial Panel Model 2: Actual vs Predicted

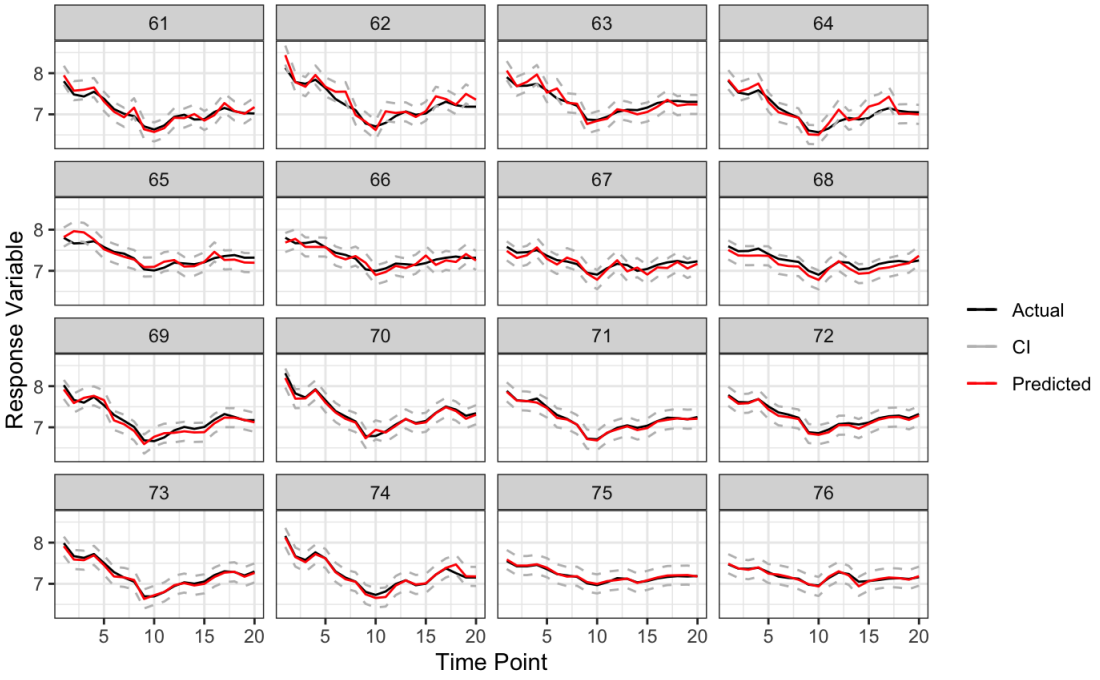


Figure D.4: Spatial Panel Model 2: Actual vs Predicted

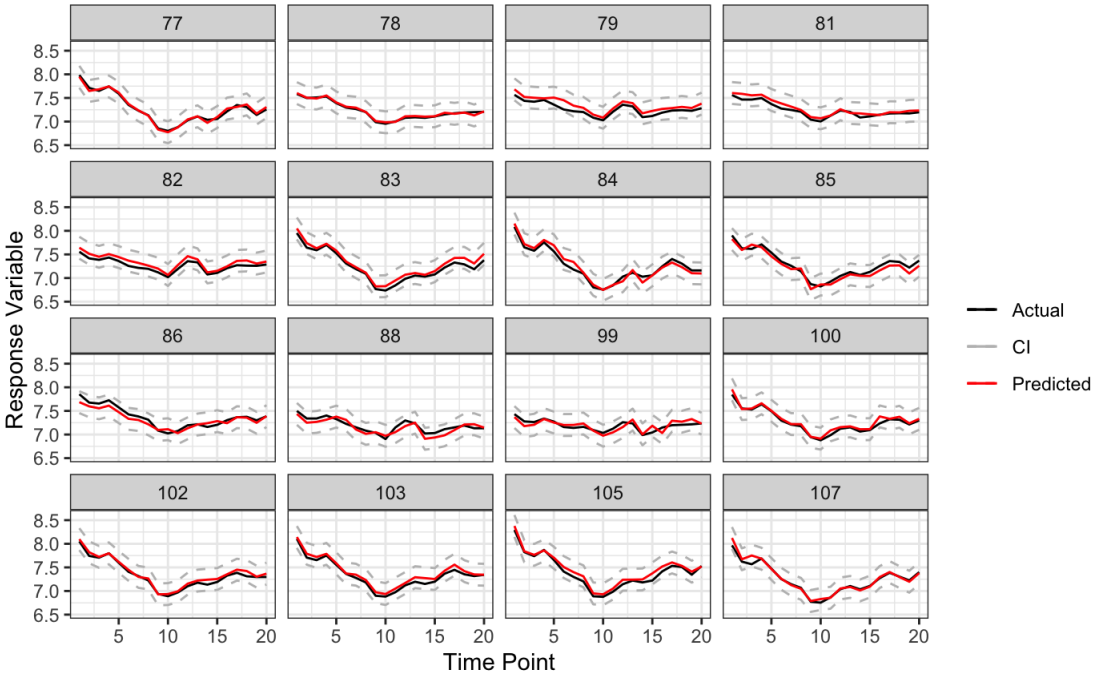


Figure D.5: Spatial Panel Model 2: Actual vs Predicted

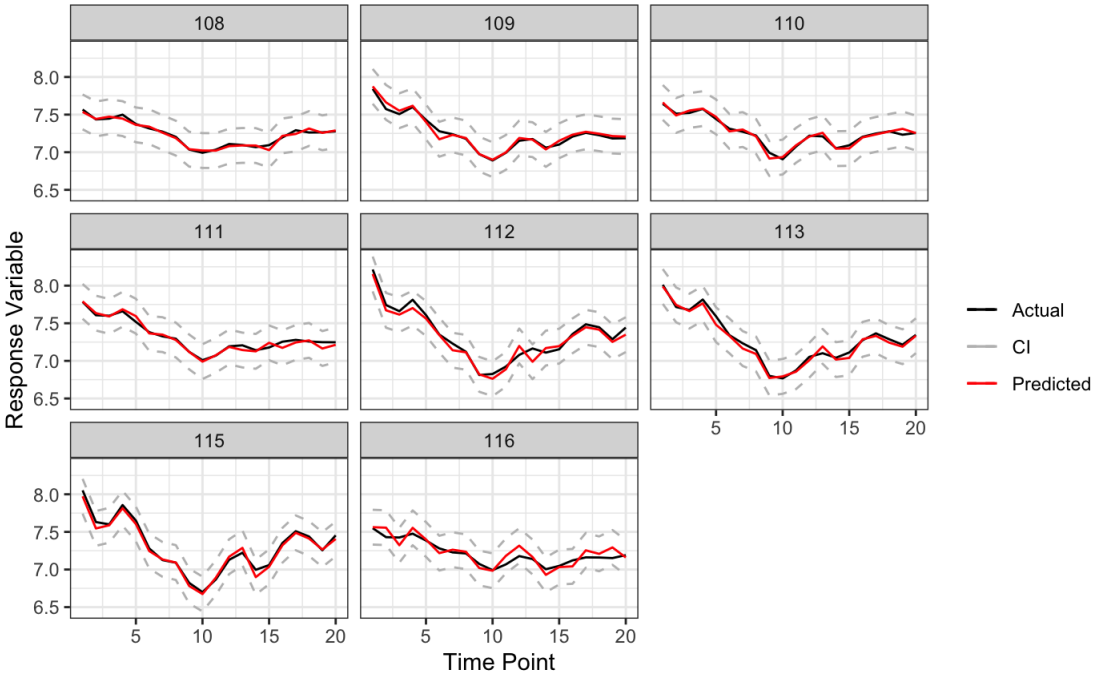


Figure D.6: Spatial Panel Model 2: Actual vs Predicted

Bibliography

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. (2012), 'Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada', *Advancing Earth and Space Science* **48**(1), 1–14.

Anselin, L. (1995), 'Local indicators of spatial association—lisa', *Geographical Analysis* **27**(2), 93–115.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1995.tb00338.x*

Arbia, G. (2006), *Spatial Econometrics*, Springer, Vialle Pindaro, Italy.

Bakker, M., van Duist, H., van Schagen, K., Vreeburg, J. & Rietveld, L. (2014), 'Improving the performance of water demand forecasting models by using weather input', *ScienceDirect* **70**, 93–102.
URL: *https://www.sciencedirect.com/science/article/pii/S1877705814000149*

Baltagi, B. H. (2005), *Econometric Analysis of Panel Data*, John Wiley & Sons Ltd, England.

Baltagi, B. H., Heun Song, S., Cheol Jung, B. & Koh, W. (2007), 'Testing for serial correlation, spatial autocorrelation and random effects using panel data', *Journal of Econometrics* **140**, 5–51.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407606002223*

Baltagi, B. H. & Li, D. (2008), *Prediction in the Panel Data Model with Spatial Correlation*, Springer, Berlin, Heidelberg.

Baltagi, B. H., Song, S. H. & Koh, W. (2003), 'Testing panel data regression models with spatial error correlation', *Journal of Econometrics* **117**(1), 123–150.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407603001209*

Bivand, R., Millo, G. & Piras, G. (2021), 'A review of software for spatial econometrics in R', *Mathematics* **9**(11).
URL: *https://www.mdpi.com/2227-7390/9/11/1276*

Bivand, R. S., Pebesma, E. & Gomez-Rubio, V. (2013), *Applied spatial data analysis with R, Second edition*, Springer, NY.
URL: *https://asdar-book.org/*

Bivand, R. S. & Wong, D. W. S. (2018), 'Comparing implementations of global and local indicators of spatial association', *TEST* **27**, 716–748.

Brick, K., DeMartino, S. & Visser, M. (2017), 'Behavioural nudges for water conservation: Experimental evidence from cape town', *ResearchGate* .

Brownlee, J. (2017), 'How much training data is required for machine learning?'.
URL: *https://machinelearningmastery.com/much-training-data-required-machine-learning/*

Bruhl, J., le Roux, L., Visser, M. & Köhlin, G. (2020), 'The costs of a crisis: Lessons from the cape town drought for urban water policy', *Oxford Encyclopedia of Water Resources Management and Policy* .

Bruhl, J., Serman, K. & Visser, M. (2020), Motivating high water users to conserve water during time of crisis: Evidence from drought- stricken cape town.

Bruhl, J. & Visser, M. (2021), 'The cape town drought: A study of the combined effectiveness of measures implemented to prevent "day zero"', *ScienceDirect* **34**.

Cheng, H., Hu, Y. & Zhao, J. (2009), 'Meeting china's water shortage crisis: Current practices and challenges', *ACS Publications* **43**(2), 240–244.

CoCT (2018*a*), 'Dam levels'.
URL: *https://www.capetown.gov.za*

CoCT (2018*b*), 'Dam levels report'.

CoCT (2018*c*), 'Where does my water come from?'.

URL: *https://www.capetown.gov.za/Family%20and%20home/Residential-utility-services/Residential-water-and-sanitation-services/where-does-my-water-come-from*

CoCT (2019).

CoCT (2021*a*), 'Subcouncils'.

URL: *https://www.capetown.gov.za/Family%20and%20home/Meet-the-City/city-council/subcouncils*

CoCT (2021*b*), 'Wards'.

URL: *https://www.capetown.gov.za/Family%20and%20home/Meet-the-City/city-council/wards*

Croissant, Y. & Millo, G. (2008), 'Panel data econometrics in R: The plm package', *Journal of Statistical Software* **27**(2), 1–43.

URL: *https://cran.r-project.org/web/packages/plm/vignettes/A_plmPackage.html*

Croissant, Y. & Millo, G. (2018), *Panel Data Econometrics with R*, Wiley.

Durbin, J. (1954), 'Errors in variables', *JSTOR* **22**, 23–32.

URL: *https://www.jstor.org/stable/1401917*

Elhorst, J. P. (2003), Spatial panel data models, *in* M. Fischer & A. Getis, eds, 'Handbook of Applied Spatial Analysis', Springer, Berlin, Heidelberg, pp. 377–407.

Elhorst, J. P. (2010), 'Applied spatial econometrics: raising the bar', *Spatial Economic Analysis* **5**(1), 9–28.

Eskom (2017), 'An overview of electricity consumption and pricing in south africa', *Deloitte Consulting Pty (Ltd)* .

Github, A. K. (2022), 'Minor masters thesis model implementation'.

URL: *https://github.com/annakaplan101/Thesis/blob/main/anna_22012022_models.Rmd*

Graul, C. (2016), *leafletR: Interactive Web-Maps Based on the Leaflet JavaScript Library*. R package version 0.4-0.

URL: *http://cran.r-project.org/package=leafletR*

Hausman, J. A. (1978), 'Alternative tests of independence between stochastic regressors and disturbances', *Econometrica* **46**(6), 1251–1271.
URL: *doi:10.2307/1913827*

Kelejian, H. H. & Prucha, I. R. (2010), 'Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances', *Journal of Econometrics* **157**, 53–167.

Kelejian, M. K. H. H. & Prucha, I. R. (2007), 'Panel data models with spatially correlated error components', *ScienceDirect* **140**, 97–130.

Kiziltan, M. (2021), 'Water-energy nexus of turkey's municipalities: Evidence from spatial panel data analysis', *ScienceDirect* **226**, 1–12.

Millo, G. (2017), 'Robust standard error estimators for panel models: A unifying approach', *Journal of Statistical Software* **82**(3), 1–27.

Millo, G., Piras, G. & Bivand, R. (2012), 'splm: Spatial panel data models in r', *Journal of Statistical Software* **47**, 1–38.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Romano, M. & Kapelan, Z. (2014), 'Adaptive water demand forecasting for near real-time management of smart water distribution systems', *Environmental Modelling Software* **60**, 265–276.
URL: *https://www.sciencedirect.com/science/article/pii/S1364815214001819*

Salima, B. A., Julie, L. G. & Lionel, V. (2018), Spatial econometrics on panel data, *in* 'Handbook of spatial analysis — Theory and practical application with R', INSEE Eurosta, pp. 179–2003.

Shine, P., Murphy, M. D., Upton, J. & Scully, T. (2018*a*), 'Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms', *ScienceDirect* **150**, 74–87.
URL: *https://www.sciencedirect.com/science/article/pii/S0168169917315259*

Shine, P., Murphy, M. D., Upton, J. & Scully, T. (2018*b*), 'Multiple linear regression modelling of on-farm direct water and electricity consumption on pasture based dairy farms', *ScienceDirect* **148**(2), 337—-346.

Tennekes, M. (2018), 'tmap: Thematic maps in r', *Journal of Statistical Software* **84**, 1–39.
URL: *https://doi.org/10.18637/jss.v084.i06*

Tibshirani, R., Witten, D., Hastie, T. & James, G. (2013), *An Introduction to Statistical Learning, with Applications in R*, Springer, New York.

Tiefelsdorf, M., Griffith, D. & Boots, B. (1999), 'A variance stabilizing coding scheme for spatial link matrices', *SAGE journals* **31**(1), 165–180.

Tso, G. & Yau, K. (2007), 'Predicting electricity energy consumption: A comparison of regression analysis decision trees and neural networks', *ScienceDirect* **32**(9), 1761–1768.

Underhill, L. & Bradfield, D. (2013), *Introstat*, University of Cape Town, South Africa Cape Town.

Verdoorn, P. J. (1980), 'Verdoorn's law in retrospect: A comment', *The Economic Journal* **90**, 382–385.

Walker, D., Creaco, E., Vamvakeridou-Lyroudia, L., Farmani, R., Kapelan, Z. & Savić, D. (2015), 'Forecasting domestic water consumption from smart meter readings using statistical methods and artificial neural networks', *Procedia Engineering* **119**, 1419–1428.
URL: *https://www.sciencedirect.com/science/article/pii/S1877705815026727*

Wu, D. (1973), 'Alternative tests of independence between stochastic regressors and disturbances', *Econometrica* **41**(4), 733–750.
URL: *https://doi.org/10.2307/1914093*

Yea, Y., Kocha, S. F. & Zhang, J. (2018), 'Determinants of household electricity consumption in south africa', *Energy Economics* **75**, 120–133.
URL: *https://www.sciencedirect.com/science/article/pii/S0140988318303013*