Modelling Residential Electricity Usage within the eThekwini Municipal Area

Author:

Samantha Reade

Supervisor:

Professor Temesgen Zewotir

Co-Supervisor:

Professor Delia North

Submitted in fulfilment of the academic requirements for the degree of Master of Science in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Westville Campus

24 November 2014

As the candidate's supervisors we have approved this dissertation for submission.

Supervisor: Signed: _____ Date: _____

Co-supervisor: Signed: _____ Date: _____

Preface and Declarations

The work described in this thesis was carried out at the University of KwaZulu-Natal, School of Mathematics, Statistics and Computer Science, Westville, from **February 2013** until **November 2014**, under the supervision of **Professor T. Zewotir**.

This thesis is entirely, unless specifically contradicted in the text, the work of the candidate, **Samantha Reade**, and has not been previously submitted, in whole or in part, to any other tertiary institution. Where use has been made of the work of others, it is duly acknowledged in the text.

Declaration - Plagiarism

I, _____ declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
    a. Their words have been rewritten but the general information attributed to them has been referenced.
    b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.
5. This thesis does not contain, text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: _____

Abstract

In this study we use linear mixed models to model residential electricity consumption within the eThekwini municipal area. Utilities around South Africa are required to estimate monthly electricity consumptions for each household within their jurisdiction however, little work has been done to find models that may be used to do so. As part of the modelling process we investigate seasonal trends in consumption as well as temporal and spatial variations. Data for the study were obtained from eThekwini Electricity, a subsidiary of eThekwini Municipality. Key findings of the research include confirming the presence of a seasonal pattern in monthly electricity consumption and proving that variations in consumption of different households are not related to the physical distance between them. Models developed in this study also have applications in prediction and may be used to predict future electricity consumption for individual households. Predictions made using the models from this study were found to be closer to the actual value, than that of the customary eThekwini Electricity predicted values.

Keywords

Electricity consumption, generalized linear mixed models, linear mixed models, prediction, residential, spatial variations, temporal variance

Acknowledgements

I am deeply thankful to my supervisor, Professor Zewotir, for his unwavering support, guidance and encouragement.

I am grateful to my co-supervisor, Professor North, for all of her advice and assistance.

The eThekwini Municipality is thanked for their willingness to assist, and provide data for this study. An extra thank you to Mr Pierre Maree and Mr Lee Mackenzie from eThekwini Municipality. Thank you Pierre, for organizing and co-ordinating this process. Thank you Lee, for compiling the data set from the various databases and providing background information regarding the variables.

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and not necessarily to be attributed to the NRF or UKZN.

To my brother, I express my deepest thanks and appreciation for all his technical assistance and guidance.

Finally, I would like to thank my mom for all her support, patience and encouragement throughout the writing of this dissertation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

eThekwini is situated on the East Coast of South Africa, within the province of KwaZulu-Natal. A map depicting eThewkini's precise location is shown in Figure 1.1, eThekwini is the shaded dark area. eThekwini Municipality covers a land area of approximately 2000 square kilometres and encompasses the city of Durban (eThekwini). According to Census 2011, the population of Durban (eThekwini) is about 3.4 million (StatsSA, 2012), accounting for approximately 7% of South Africa's entire population.

Figure 1.1: The location of eThekwini Municipality

Electricity to the eThekwini Municipality is supplied by eThekwini Electricity. eThekwini Electricity estimate that they supply electricity to 655 338 households (eThekwini Electricity, 2013). Of these households, 333 434 are prepaid customers and account for 8% of

all power sold to the residential sector while, 321 904 are credit customers who account for 23% of all power sold to the residential sector. The difference between the two types of customers namely being; prepaid customers purchase their electricity upfront by means of vouchers whereas, credit customers are charged monthly for the amount of electricity they consumed during the previous month.

All credit customers have an electricity meter on their property that is read at approximate 3 month intervals throughout the year. As a result, we expect each household within eThekwini to have 4 meter readings in a 12-month period. eThekwini Electricity use these meter readings to calculate the amount of electricity used, and bill customers accordingly. In the months where meters are not read, customers pay estimated charges (eThekwini Municipality, 2013). Any differences between actual and estimated charges are accounted for in the monthly bills of meter-reading months.

Currently, there are too many credit customers over too wide an area, for eThekwini Electricity to feasibly consider conducting monthly readings and doing away with estimations. eThekwini Electricity estimates consumption by means of a cumulative total of weighted previous actual usage. Whereby, the most recent consumptions carry the highest weightings while weightings of older consumptions decrease exponentially. This method may be summarized by the following formula:

$$E(y_{ij})_{eThekwini} = \sum_{k=1}^{n-1} \left(\frac{1}{2}\right)^k lag_k + 2\left(\frac{1}{2}\right)^n lag_n$$

where n is the total number of measurement periods that each household has; y_{ij} represents the current electricity usage for the i^{th} household at time t_{ij}, $j = 1, \ldots, n$; and previous electricity consumption values for the i^{th} household are denoted using lags whereby, $lag_k = y_{ij-k}$ for $k = 1, \ldots, n-1$ and $lag_n = y_{ij-n}$.

Based on the understanding that each household has 4 meter readings in a 12-month period, it follows that from the current electricity consumption value there would be 4 lags for the 12-month period prior to it. Similarly, we expect there to be 8 lags over a 24-month period, 12 lags over a 36-month period, 16 lags over a 48-month period and so on. From the customary estimation formula given above, we observe that lag_1 carries a weighting of 0.5, lag_2 carries a weighting of 0.25, lag_3 carries a weighting of 0.125 and lag_4 carries

a weighting of 0.0625. Once these weighted lags are summed, it becomes evident that the bulk of the estimate for current electricity usage comes from the first 4 lags that a household has. That is, the most recent previous year's electricity consumption accounts for the majority of the eThekwini estimate.

Clearly, this customary estimation method does not allow for any seasonal or cyclical trends in consumption. Such estimation also does not take into account the individual customers' electricity consumption variability and pattern. Despite dwelling type also being a logical consumption-influencing factor to consider, it too, is not utilized in the customary method. To date, both in South Africa and other countries research into the modelling of monthly household electricity consumption at an individual household level has been limited.

1.1 Literature Review

The national aggregate electricity demand of South Africa and the examination of factors likely to influence demand has been a primary focus of local research. In the 1980s Pouris (1987) used annual data for the period 1950-1983 and an unconstrained distributed lag model to estimate long-run price elasticities of the aggregate electricity demand in South Africa. Modelling demand as a function of price and GDP, Pouris (1987) found the elasticity of electricity demand to be -0.9 while income remained inelastic. More recently, Inglesi (2010) sought to specify variables that could be used to explain aggregate electricity demand within South Africa then, used these variables to produce forecasts up to the year 2030. Inglesi (2010) found that there was a long-run relationship between electricity consumption, electricity price and economic growth. Sigauke & Chikobvu (2011) developed a combination regression-SARIMA-GARCH model to predict short term daily peak aggregate electricity demand in South Africa. The Reg-SARIMA-GARCH model proposed by Sigauke & Chikobvu (2011) captured various short term demand-influencing factors such as days of the week, holidays and temperature. Further to their 2011 study, Chikobvu & Sigauke (2013) employed a piecewise linear regression model and extreme value theory to model the influence of temperature on South Africa's daily average electricity demand. The results of Chikobvu & Sigauke (2013) showed that in South Africa daily electricity demand was highly sensitive to cold temperatures. Inglesi-Lotz & Blignaut (2011) per-

formed a sectoral decomposition of South Africa's electricity consumption and looked at how factors such as increased production, structural changes and efficiency improvements affected consumption in different sectors. Consequently, a sectoral approach to energy policy-making in South Africa was recommended by Inglesi-Lotz & Blignaut (2011) because they found that the identified factors had varying effects in different sectors.

Following a model akin to that of Pouris (1987), Ziramba (2008) examined electricity demand as a function GDP per capita and electricity price in the South African residential sector for the period 1978-2005. Ziramba (2008) found that in the long run, income was the main factor that determined residential electricity demand while the price of electricity was insignificant. A similar study was also carried out using Sri Lankan data for the period 1960-2007 by Athukorala & Wilson (2010). In addition to modelling residential electricity demand as a function of per capita income and electricity price, Athukorala & Wilson (2010) also included the price of alternative energy sources such as kerosene as a demand-influencing factor. The results of Athukorala & Wilson (2010) were somewhat alike to those of Ziramba (2008), in that it was determined that increasing electricity price was not an effective tool to reduce consumption and that increases in household incomes would likely see increases in consumption. Other studies that have also modelled residential electricity demand as a function of per capita income, price as well as other additional factors have been carried out in Australia (Narayan & Smyth, 2005), the United States (Dergiades & Tsoulfidis, 2008), and Taiwan (Holtedahl & Joutz, 2004).

Narayan & Smyth (2005) used a bounds testing approach to cointegration in an autoregressive distributive lag (ARDL) framework to estimate the short and long-run elasticities of residential electricity demand. Income and price of electricity were found to be the main determinants of demand, while temperature was found to be only sometimes important and prices of alternative fuels were irrelevant (Narayan & Smyth, 2005). Dergiades & Tsoulfidis (2008) assumed residential electricity demand to be a function of per capita income, electricity price, weather conditions, prices of alternative sources (e.g. gas or oil) and stock of appliances within dwellings. Using ARDL cointegration, Dergiades & Tsoulfidis (2008) showed that a single cointegration relation existed among these factors.

In addition to studies where the focus has been price elasticity or forecasting electricity demand as a function of GDP, research has also gone into other factors affecting electricity consumption. Firth et al. (2008) identified trends between electricity consumption and appliance usage. Marvuglia & Messineo (2012) used artificial neural networks (ANNs) for short-term electricity forecasting in Italy and focused on how the use of air-conditioning affected electricity consumption. Yohanis et al. (2008) carried out a comprehensive study in Northern Ireland on the patterns of electricity consumption of 27 households. Yohanis et al. (2008) took into account a wide range of household factors such as dwelling type, location, dwelling size, household appliances, attributes of the occupants and income, and found that each factor had an impact on electricity consumption.

Aside from the small study of 27 households by Yohanis et al. (2008), the problem with studies such as those cited, is that they are not able to describe the individual household level of electricity consumption. Though such studies are useful to estimate the national residential electricity usage, the main challenge in the new democratic South Africa is service delivery at the municipal and ward levels. Hence, our study initiates and motivates further research on estimation of electricity consumption of a typical household in a given municipality in South Africa. In this regard our study differs from the works already cited, in that our main focus is at a household-level as opposed to modelling and predicting consumption for the entire residential sector. We also step away from the traditional time series and econometric modelling approaches, choosing instead an applied statistical approach. Accordingly, we use linear mixed models to model household electricity consumption as a function of month of meter reading, dwelling type, a household-specific intercept and lagged consumption.

1.2 Objectives and Significance of the Study

The aim of this study is to model household electricity consumption within the eThekwini Municipality. Presently, we are aware that lagged consumption values are used to estimate current electricity consumption, with the 4 most recent lags contributing the most to the estimate. In this study we will formally assess whether or not lagged consumption values have an effect on current electricity consumption. Furthermore, if an effect is present we

intend to ascertain whether all available lags for a household are significant or not, as well as seek to establish if there is any seasonal pattern occurring within the lags. In addition to studying the effect of lagged consumption, we also strive to address the shortcomings of the customary eThekwini Electricity estimation technique by considering dwelling type and month of meter reading. By including both lags and month of meter reading in the modelling process, we will be able to assess two different factors. The inclusion of the month of meter reading enables us to tentatively gauge if some months have higher or lower electricity consumption than others. The inclusion of lags allows for the assessment of a potential seasonal pattern occurring within households electricity usage. Both temporal and spatial variations in households electricity consumption are also to be investigated in the study. Initially, our focus shall be on temporal variances of electricity usage within households thereafter, we will examine spatial variations in consumption between households. To carry out these modelling procedures, we shall use linear mixed models.

The models developed in this study will also have applications in prediction. We can use our final models to predict monthly electricity consumption for individual households. To validate our final model/s, we shall compare our predictions to estimates derived from the customary eThekwini Electricity estimation method, assessing which method renders results closest to actual monthly consumption values.

Finding a model that enables better prediction of electricity usage is significant, as the outcome could potentially affect over 300 000 credit customers (households). Even though differences in estimated and actual charges get reversed every 3 months, hypothetically, credit customers could be routinely paying too much or too little for electricity on a monthly basis. Indelibly, this would impact on monthly household budgets. Considering the wide ranging economic statuses of residents in eThekwini, this could stand to make a big difference to households on tight budgets.

Of equal importance to improving prediction, is proving whether or not a spatial relationship between electricity consumption of individual households exists. This would enable us to ascertain how feasible it may be, to base a household's predicted electricity consumption, on that of their neighbours. Potentially, this would provide a solution to when meters get skipped during meter-reading months (an event that often occurs if meters are

not accessible and house owners are unavailable). It may also suggest that for prediction purposes, fewer meters need be read every 3 months.

This study is clearly beneficial, both to the energy sector in South Africa and to the municipality who provided the data. To begin the study, we look at the data in Chapter 2, proceeding to the modelling processes in Chapters 3 to 6. In the final chapter, Chapter 7, we provide a broad summary of our results and discuss our findings.

Chapter 2

Data Description

The data required for this study were provided by eThekwini Electricity. Two databases, namely COINS and GIS, were used to compile the data set. COINS stores data relating to every one of eThekwini Electricity's customers, including electricity users in the residential, commercial and industrial sectors. However, for the purposes of this study, only information pertaining to credit residential electricity users was extracted.

eThekwini Electricity continuously collects data from their credit customers (electricity consumers). All credit customers have an electricity meter on their property, that is read by eThekwini Electricity officials at approximate 3 month intervals throughout the year. This is done, so as to determine the amount of consumed electricity to bill the customer for. Meter readings as well as, the dates they were taken on are then stored in the COINS database. As this is an ongoing process, the result is a longitudinal, repeated measures data set, containing rich historical information for individual electricity consumers, spanning several years. With the aid of property identifiers, information regarding the exact spatial location of all these credit electricity consumers could be linked, and included from the GIS database.

The final, repeated measures type data set from eThekwini Electricity contained information for approximately 300 000 properties of individual credit electricity customers. For the purposes of this study and simplicity henceforth, we refer to individual credit electricity customers as households. The data covers an approximate five year period, with most households' meter readings beginning in late 2007 and ending in mid 2013. As most of the information was extracted straight from COINS, aside from meter readings, dates and

GIS co-ordinates, a lot of additional household data was made available to us. We focus further discussions only on variables pertinent to this study.

2.1 Pertinent Identifiers and Variables

Three different identifiers were given in the data; a property key, electrical connection ID and meter ID. The property key identifies one property from another, connection ID identifies each electrical connection from the electricity grid to a dwelling and meter ID is a number unique to each electricity meter.

For the purposes of this study, we use the electrical connection ID as our household identifier. The reason for us choosing this identifier, is that it remains unchanging and allows us to uniquely identify individual electricity consumers. Although we are able to recognize individual properties through the property key, we note that on a single property may exist several dwellings with several different households. A basic example of such a scenario would be a block of flats. The block itself has a single property key but, each individual unit/flat would have its own electrical connection ergo, its own connection ID. Furthermore, we prefer connection ID over meter ID because, meter ID is linked directly to the meter and would change in the event of a meter upgrade or replacement.

Household information includes five years worth of repeated meter readings, reading dates, GIS co-ordinates, suburb and city locations and dwelling type classifications. There are 3 categories of dwelling types; houses, shareblocks ≤ 2 storeys high and shareblocks > 2 storeys high.

We refer to the period between two consecutive readings as a usage period. The amount of electricity consumed in a usage period is measured in kilowatt hours (kWh), and calculated by subtracting the previous meter reading from the current one. eThekwini Electricity uses a built-in algorithm to handle this calculation in the event of a meter roll-over. Meter roll-over is when a meter reaches its maximum possible reading then begins again from zero.

In the business and industrial sector, the difference between readings is multiplied by some

meter coefficient, as their actual consumption is much higher. However, in the residential sector the meter coefficient is 1, allowing us to simply use the difference between two consecutive readings. These differences, coefficients and calculated consumptions (including eThekwini Electricity's algorithm for meter roll-overs) are all stored in the COINS database. As a result, actual electricity consumptions for all usage periods throughout the 5 years are also included as a repeated measure in our household data. Since we expect each household to have 4 meter readings in a 12-month period, it follows that each household should have a minimum of 20 meter readings for the 5-year period. Therefore, it is possible that each household in the data could have at least 19 lagged electricity consumption values, with 4 lags occurring in a 12-month period.

Using the reading dates, we were able to add to our already existing household data. We calculated the length of all the usage periods (in days), and we extracted the months that each meter was read in. An additional variable, t, starting from zero and counting the number of measurement occasions within each household, was also created. This data, combined with the other variables and lags already discussed constitutes the applicable household information contained in the data set available for use in this study.

2.2 Selecting a Sample from the Data

The original data set received from eThekwini Electricity was exceptionally large, containing information for approximately 300 000 households across some 10 million rows. The necessary computing equipment required for processing such large data far exceeds the available computing facilities in the University of KwaZulu-Natal. Since the main focus is modelling electricity consumption, so as to later enable future prediction of it, using limited data serves the purpose. Accordingly, a sample was drawn from the original data set. We model electricity consumption for the sampled data, demonstrating effective methods that could be implemented on a larger scale, with the necessary computing capacity.

Before selecting a cohort, the original data set needed to be sorted and cleaned. The data was sorted by connection ID, then chronologically by reading date. Households that had unpopulated fields for variables such as dwelling type or GIS co-ordinates were removed.

Due to the vast area and number of credit customers, it is not possible for eThekwini Electricity to read every customer's meter on a monthly basis. Instead, eThekwini Electricity ideally try to read the majority of their customers' meters on an ongoing basis, at approximate 3 month intervals. Therefore, in this study we limit our sampling frame to households that have measurement periods that are approximately 3 months in length. To guarantee an approximate 3-month period, we only consider households whose measurement periods are between 80 and 110 days in length. This ensures that we stay within the "ideal" or norm of eThekwini Electricity, enabling us to later compare typical municipal estimations to predictions made using models developed in this study.

In addition to limiting the sampling frame by length of measurement period, we also restricted it by city location. Only households that had a city classification of "Durban", were considered for sampling. The city of Durban encompasses most suburbs within eThekwini's municipal boundaries and incorporates households from wide ranging backgrounds. When we look back to the original data set from eThekwini Electricity, we also see that more than two thirds of the households had a city classification of "Durban". Consequently, we consider our defined sampling frame to be adequate, in regard to ensuring a sample that is more or less representative of typical credit electricity consumers within eThekwini.

The intention of this study is to demonstrate methods of modelling residential electricity consumption. However, we also wish to avoid computational burdens brought about with large data. Therefore, for demonstration purposes and for this study, we select a sample of 1500 households using the method of simple random sampling. To ensure that each dwelling type would be represented in the sample, we sampled 500 households from each type.

2.3 Descriptive Statistics

The specification of the sampling frame ensured no anomalies in regard to the regularity of meter readings/measurement occasions however, no attention was paid to the actual electricity consumption. It is difficult to simply decide if a consumption value may be unnaturally high or low, due to the vastly diverse backgrounds and needs of each electric-

ity consumer within Durban. Therefore, we refrain from looking at any specific expected range of possible consumption values. However, we acknowledge that it would be considered "abnormal", if any usage period had an actual consumption value of zero. It is very difficult (albeit not impossible) for an average household to consume absolutely no electricity within 3 months. Zero usage more likely suggests that a dwelling had no occupant during that time or, measurement error at the time of meter reading or, meter tampering. Whatever the reasons, zero usage in a measurement period is certainly deemed unusual for the majority of households. We identified and removed 22 such households within the sample, leaving us with a sample size of 1478 households. For simplicity purposes, instead of referring to an 8 digit long connection ID throughout the study, households were labeled 1 to 1478.

For this study it is more practical to work with scaled data that will further minimize the burden on computing memory. We scaled the consumption data in such a way so as to not disturb the month of meter reading. The data was scaled by dividing the 3-month consumption values by a factor of 3. These scaled values intuitively created a monthly scale. In a similar fashion, we scaled the number of days in the usage periods down by a factor of 3. The creation of such monthly scaled values will simplify later model interpretation and also ensure the applicability of any results, as we recall that eThekwini Electricity deal with monthly estimates for all of their customers. The interpretation assumes these scaled values are the actual consumptions for the months the meter was read.

We are aware that eThekwini Electricity's customary estimation method does not account for differences between dwelling types. However, cursory examination of our sampled data at baseline ($t = 0$), reveals the mean monthly electricity consumption to be 734.26 kWh for houses, 564.43 kWh for shareblocks ≤ 2 storeys, and 507.25 kWh for shareblocks > 2 storeys high. This serves as a possible indication that dwelling type does in fact have an effect on electricity consumption, and further investigation is warranted in the modelling process.

In addition to not accounting for differences due to dwelling type, the customary estimation method also does not allow for any cyclical or seasonal effect in electricity usage. The customary estimate is based on a weighted average of previous actual consumption,

with more than 90% of the estimate coming from the 4 most recent lags which equates to the most recent previous year's electricity usage. A focal point of this study includes determining whether lagged consumption values do in fact affect current electricity consumption and if so, establishing how many and which lags in particular effect significance on current consumption. By scrutinizing the lags, we are inherently able to assess the possible presence of a seasonal effect in the electricity consumption data. We begin our investigation by examining a simple profile plot of a few randomly selected households from the sample. The profile plot of 15 randomly selected households is displayed in Figure 2.1.

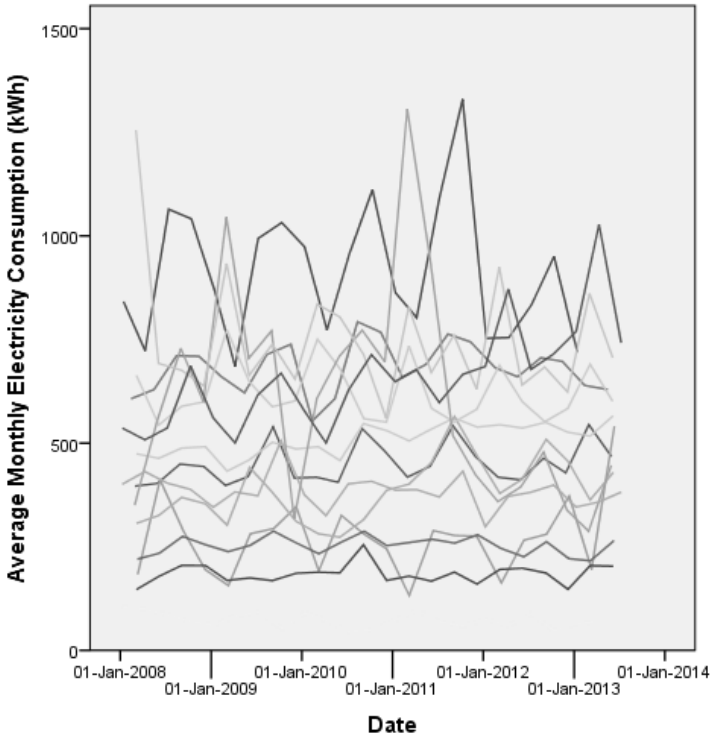

Figure 2.1: Profile plot of monthly electricity consumption for 15 households

Figure 2.1 reveals that there is both between and within household variation. Furthermore, we can clearly see a recurring pattern taking place within each household. It appears as though each household's average monthly electricity usage at the same time in the year, over the 5-year period, are somewhat similar to one another. That is, monthly consumption values that are 12 months apart appear similar, suggesting the presence of a seasonal pattern. Based on this preliminary observation, it is evident that at least some lags will have an effect on current electricity consumption values. Hence, the presence of a seasonal pattern as well as, the number of lags affecting current consumption is investigated

further. This is done in the application section of Chapter 3. However, before proceeding to deeper analysis of the electricity data, we first explore the distributional properties of the electricity consumption values. Figure 2.2 displays a histogram that shows how the monthly electricity consumption values are distributed.

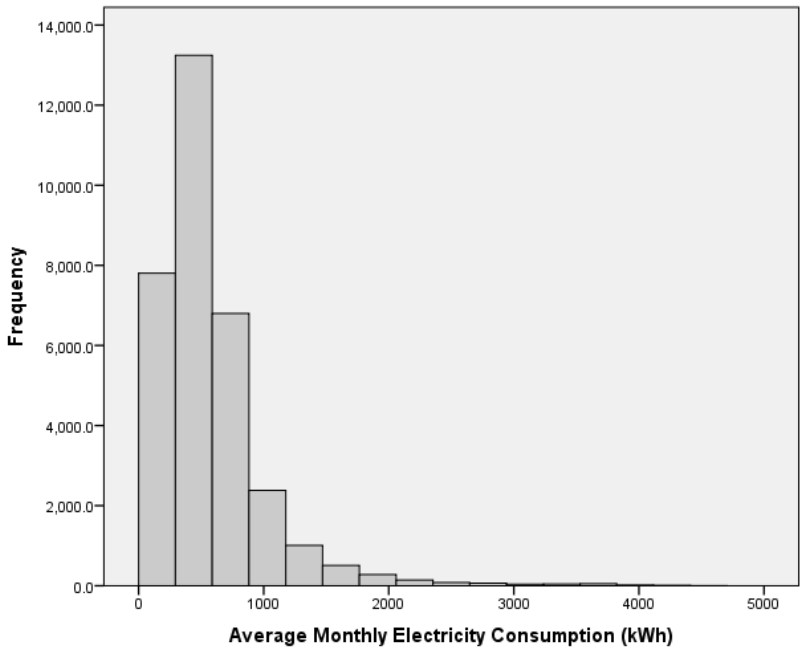

Figure 2.2: Histogram showing the distribution of monthly electricity consumption

From Figure 2.2 we see that the consumption data is skewed to the right. This suggests that the electricity consumption data is unlikely to satisfy normality assumptions. A common remedy for non-normally distributed data is to either apply a transformation to the data or implement a generalized modelling approach. Both avenues are explored in greater detail in the application sections of subsequent chapters.

Following these rudimentary observations, we proceed to deeper analysis of the electricity data, using a mixed modelling approach. Mixed models are well-suited to handling repeated measures data, enabling us to model the monthly electricity consumption of individual households. In addition to modelling both between and within household variations, with mixed models we also have the ability to model either temporal or spatial variations.

Chapter 3

The Linear Mixed Model

In this chapter we look at the linear mixed model and how it may be applied to longitudinal data analysis. Following discussions on repeated measures analysis and mixed models, we proceed to the modelling of our sampled electricity data. The application in this chapter focuses solely on modelling temporal variations within households. To fit linear mixed models to the data we make extensive use of the statistical software, SAS. Inferences made from the final model are discussed at the end of the chapter.

3.1 Mixed Effects Model Definition and Estimation

A linear mixed model (LMM) is so named, as it is a linear model containing both fixed and random effects. It may also be referred to as a linear mixed effects model. A fixed effect is universal to the whole target population. Whereas, a random effect occurs when the factor variable being modelled, constitutes only a random sample of all possible factor levels within the target population. It is a common understanding that we say we *predict* random effects and *estimate* fixed effects (McCulloch et al., 2008). For a detailed historical account of the development of the mixed model refer to Searle et al. (2006) or West et al. (2007).

The general form of the LMM is given by

$$\mathbf{y} \quad = \quad \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{3.1}$$

where $\mathbf{y}$ is a vector of response variables; $\mathbf{X}$ is a $(n \times p)$ known design matrix of fixed

numbers associated with $\boldsymbol{\beta}$; $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown constants, the fixed effects of the model; $\mathbf{Z} = |\mathbf{Z}_1|\mathbf{Z}_2| \ldots |\mathbf{Z}_r|$, where $\mathbf{Z}_i$ is a $(n \times q_i)$ known design matrix of the random effect factor i; $\mathbf{u}' = |\mathbf{u}_1'|\mathbf{u}_2'| \ldots |\mathbf{u}_r'|$, where $\mathbf{u}_i$ is a $(q_i \times 1)$ vector of the random variables; $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of error terms. It is assumed that both the random effect, $\mathbf{u}$, and error term, $\boldsymbol{\varepsilon}$, are of the multivariate normal distribution, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$, where $\mathbf{G}$ is block diagonal with the i^{th} block being $\sigma_i^2 \mathbf{I}_{q_i}$ and $\mathbf{R}$ is positive definite. It is further assumed that $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ are independent. The response variable $\mathbf{y}$ is also of the multivariate normal distribution with $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $var(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$, or, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$.

There are varying opinions and methods as how to best estimate model parameters (see e.g. Harville, 1977; Laird & Ware, 1982; Robinson, 1991; Searle et al., 2006) however, the most frequently employed method of estimation is that of maximum likelihood (ML) and restricted maximum likelihood (REML) (Fitzmaurice et al., 2008).

If we assume that $\mathbf{V}$ is known, and employ either the method of ML or, methods of least squares estimation, we find that the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ with $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}; (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})$. The realized value of $\mathbf{u}$ is given by $\tilde{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $\tilde{\mathbf{u}}$ is known as the best linear unbiased predictor (BLUP) of $\mathbf{u}$.

However, when $\mathbf{V}$ is unknown and not a function of $\boldsymbol{\beta}$, the log likelihood equation needs to be maximized with respect to $\mathbf{V}$. Once the estimate of $\mathbf{V}$, $\hat{\mathbf{V}}$, is found it may be substituted back into the likelihood equation to find the remaining parameter estimates.

Searle et al. (2006) and McCulloch et al. (2008) showed how to find $\hat{\mathbf{V}}$ using the profile likelihood equation and iterative methods such as Newton-Raphson and Fisher Scoring. Alternatively, as first introduced by Patterson & Thompson (1971), $\mathbf{V}$ can also be estimated using the method of restricted maximum likelihood (REML). McCulloch et al. (2008) recognized the growing popularity in the use of REML estimation over ML estimation when estimating $\mathbf{V}$, as the resulting estimates are less biased.

3.2 The Linear Mixed Model in Longitudinal Data Analysis

Longitudinal data occurs when a specific characteristic of a subject is measured or observed repeatedly (on two or more occasions) over a period of time. Subjects are regarded as independent of each other while, observations within subjects are related to one another. Accounting for the correlation of within subject observations as well as, variability between subjects is of particular importance in longitudinal studies. This allows us to gain reliable inference on parameters of interest.

One of the most widely used methods for analyzing continuous, longitudinal data is the linear mixed model (Fitzmaurice et al., 2008). Laird & Ware (1982) were the first authors who illustrated the analysis of longitudinal data using a more flexible class of mixed-effect models. Laird & Ware (1982) used the EM (expectation maximization) algorithm in their iterative mixed model estimation of the variance components and parameters of the mixed model. Later, Jenrich & Schluchter (1986) suggested several other algorithms, among them Fisher Scoring (FS) and Newton-Raphson (NR). Subsequently, several authors employed and built upon using linear mixed models in longitudinal data analysis (see e.g. Laird et al., 1987; Lindstrom & Bates, 1988; Diggle, 1988).

Repeated measures data may be classified as either balanced or unbalanced. A data set is considered balanced when each subject has an equal number of repeated observations that have been measured at the same time points. Conversely and more common in occurrence, unbalanced data occurs when the number of repeated observations per subject are different, and/or have measurements that have been gathered at different time points (Verbeke & Molenberghs, 2000; Fitzmaurice et al., 2004).

Der & Everitt (2006) said that LMMs for repeated measures data formalized the idea that a subject's response pattern would depend on many characteristics of that individual including, some characteristics that may not have been observed. Verbeke & Molenberghs (2000) gave a detailed construction of a two-stage mixed effects model for repeated measures data. Common textbooks such as Verbeke & Molenberghs (2000), Fitzmaurice et al. (2004) or, McCulloch et al. (2008) explained how the two-stage fitting of the mixed effects model allowed for the handling of unbalanced data. By combining the steps in the

two-stage model, one arrives at a similar linear mixed model for longitudinal data, as first proposed by Laird & Ware (1982).

A linear mixed model, adapted for longitudinal data is given by

$$\mathbf{y_i} \quad = \quad \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b_i} + \mathbf{e_i} \tag{3.2}$$

where $\mathbf{y_i}' = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$, if y_{ij} represents the response of the i^{th} individual measured at time t_{ij} for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$; $\mathbf{X_i}$ is an $(n_i \times p)$ matrix of known covariates associated with the fixed effects; $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown regression parameters representing the fixed effects; $\mathbf{Z_i}$ is an $(n_i \times q)$ design matrix associated with the random effects; $\mathbf{b_i}$ is a $(q \times 1)$ vector of random effects, representing the random subject-specific effects, such that $\mathbf{b_i} \sim N(\mathbf{0}, \mathbf{G})$ where $\mathbf{G}$ is block diagonal with the i^{th} block being $\sigma_i^2\mathbf{I}_{q_i}$; $\mathbf{e_i}$ is an $(n_i \times 1)$ vector of residual components where it is assumed $\mathbf{e_i} \sim N(\mathbf{0}, \mathbf{R})$ and $\mathbf{R}$ is positive definite. We assume $\mathbf{y_i}$ is multivariate normal, $\mathbf{y_i} \sim N(\mathbf{X_i}\boldsymbol{\beta}, \mathbf{Z_i}\mathbf{G}\mathbf{Z_i}' + \mathbf{R} = \mathbf{V_i})$, and that $\mathbf{b_i}$ and $\mathbf{e_i}$ are independent.

Similar to the results in the LMM, if $\mathbf{V_i}$ is known, the BLUE of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N}(\mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{X_i}))^{-1}\sum_{i=1}^{N}(\mathbf{X_i}\mathbf{V_i^{-1}}\mathbf{y_i})$ where $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}, (\sum_{i=1}^{N}(\mathbf{X_i'}\mathbf{V_i^{-1}}\mathbf{X_i}))^{-1})$; the best linear unbiased predictor (BLUP) of $\mathbf{b_i}$ is given by $\tilde{\mathbf{b_i}} = \mathbf{G}\mathbf{Z_i}'\mathbf{V_i}^{-1}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta})$. If $\mathbf{V_i}$ is unknown, following the same procedure outlined in Section 3.1, we find $\hat{\mathbf{V}}_\mathbf{i}$ and substitute it back into the likelihood equation to find the remaining parameter estimates.

With the marginal model, $\mathbf{y_i} \sim N(\mathbf{X_i}\boldsymbol{\beta}, \mathbf{Z_i}\mathbf{G}\mathbf{Z_i}' + \mathbf{R} = \mathbf{V_i})$, we notice that the fixed effects, $\boldsymbol{\beta}$, are used for modeling the mean while, the random effects govern the variance-covariance structure. Traditionally in longitudinal analysis, it is often more convenient to proceed with analysis using this marginal model (Wikle & Royle, 2004). Temporal correlation is accounted for in the variance-covariance structure of $\mathbf{y_i}$ however, there is no explicit interest in the underlying process. Of particular importance, is the correlation between pairs of repeated measures on the same subject. Therefore, special attention needs to be given when selecting a covariance structure for $\mathbf{R}$.

The importance of selecting an appropriate covariance model is further highlighted by Littell et al. (2006). Guerin & Stoup (2000) demonstrated the effect of fitting different

covariance structures to a repeated measures data set. Their work agreed with the notion that repeated measures analysis is robust, provided care is taken to ensure the covariance structure is approximately correct.

Variance may be assumed either homogenous or heterogenous. Jenrich & Schluchter (1986) gave general information about covariance structures for longitudinal data. Wolfinger (1996) provided discussion on heterogenous covariance structures suited to repeated measures data. Common covariance structures often fitted to longitudinal data include autoregressive order one (AR(1)), compound symmetry (CS) and Toeplitz as well as, an unstructured form (UN). Detailed discussions on these covariance patterns along with other structures, was provided by common textbooks such as Verbeke & Molenberghs (2000), Fitzmaurice et al. (2004) or West et al. (2007).

When selecting an appropriate covariance structure, several factors should be taken into account. To ensure reliable inferences later, a selected structure should result in a converged solution, with both the $\mathbf{G}$ and Hessian matrix being positive definite. Insight as to why there are occasions when these conditions are not always met was given by Kiernan et al. (2012). Kiernan et al. (2012) also suggested several remedies for the shortcomings.

To select the best model, many authors choose to use information criteria (IC) such as Akaike IC (AIC), corrected Akaike IC (AICC), or Bayesian IC (BIC) (see e.g. Littell et al., 2000; Moser & Macchiavelli, 2002). This is done by fitting all possible covariance structures to the model then, comparing the IC of each of them. IC may either be specified as smaller-is-better or larger-is-better.

AIC was developed by Akaike (1974), with Burnham & Anderson (1998) later coming up with a finite-population corrected form of the AIC. Schwarz (1978) introduced the BIC. Littell et al. (2006) said that models that minimize AIC, AICC or BIC werre preferable and that simpler models were generally a better choice in the interest of selecting a parsimonious model. Keselman et al. (1998) compared AIC and BIC for their potential of selecting the correct covariance model. Guerin & Stoup (2000) conducted a similar study, and found that when emphasis is on Type I error control, it was best to use AIC whereas, if loss of power was more important, it was better to use BIC.

Other factors to be considered include the number of parameters that require estimating in the selected structure as well as, how the selected covariance structure compares to that of the unstructured form. Whereby, a structure similar to the unstructured form is preferable.

Residuals measure the deviation of responses from the fitted model, they allow us to assess the normality assumptions of our model (Zewotir & Galpin, 2004). In a good fitting model, we expect the residuals to have a constant variance and follow a normal distribution.

Following from the model definition of Equation 3.2, the predicted response for the i^{th} subject is given by $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{X_i}\hat{\boldsymbol{\beta}} + \mathbf{Z_i}\tilde{\mathbf{b}_\mathbf{i}}$. The marginal residual, $\mathbf{e_i} = \mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}}$, aids us in identifying potential subject outliers. While, the conditional or subject-specific residual, $\mathbf{e_i} = \mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}} - \mathbf{Z_i}\tilde{\mathbf{b}_\mathbf{i}}$, helps us to detect possible within-subject case outliers. Definitions similar to these are used by many standard linear mixed model textbooks (see e.g. Verbeke & Molenberghs, 2000; Fitzmaurice et al., 2004; Littell et al., 2006; West et al., 2007).

Each residual has its own variance, making raw residuals impossible to compare. To eliminate this difficulty, we standardize the residuals. Common standardized residuals are the studentized residual; $\frac{e_i}{\sqrt{\hat{v}_i}}$ where e_i is a typical raw residual with estimated variance $\hat{v}_i$ and the Pearson residual; $\frac{e_i}{\sqrt{\widehat{Var[Y_i]}}}$ where e_i is a typical raw residual and $\widehat{Var[Y_i]}$ is the estimated variance of the response of the i^{th} subject.

A method that is often employed by authors to assess whether or not residuals satisfy Gaussian assumptions, is that of probability plots (see e.g. Cook & Weisberg, 1982; Zewotir & Galpin, 2004). Ideally, if the assumption of normality is satisfied, one should observe an approximate straight line pattern in the Q-Q (probability) plot.

Other plots commonly used to assess model assumptions include the scatter plot and histogram. For a scatter plot of predicted mean versus residual, one expects a random scattering about zero, as this suggests a constant variance in the error term. As with probability plots, histograms indicate whether or not the assumption of normality is violated. Alternatively, a non-graphical means for assessing normality makes use of the

Shapiro-Wilk (W) statistic (Shapiro & Wilk, 1965). Pearson et al. (1977) showed that the W statistic is a good, omnibus test for normality.

In addition to validating model assumptions, it is also important to assess the robustness of a model. Of particular interest in repeated measures data, is the influence that an individual subject may hold. A robust model should not be significantly affected by either the removal or addition of a particular subject. It should also stand up to being tested with different data. Influence diagnostics provide us with a tool for checking how robust a model is.

Cook & Weisberg (1982) suggested a diagnostic measure based on case deletion. Case deletion involves parameter estimates first being calculated using all the data points then, cases get removed from the data and the model refitted. This enables us to compute statistics based on the change between the full-data and reduced-data estimations, allowing us to quantify local influence. Both case deletion and model perturbation are well established, practical approaches to influence diagnostics in statistical modelling (Lawrance, 1990) and mixed models (Zewotir & Galpin, 2005; Zewotir, 2007).

Applying the concept of local influence in a linear mixed model setting, Beckman et al. (1987) assessed the effect of perturbing the random-effects and error covariances as well as, the effect of perturbing the response vector. Zewotir (2007) studied the influence infinitesimal model perturbations had in the linear mixed model. However, for this study our interest lies in the detection of influential subjects, not influence brought about by model perturbation. Lesaffre & Verbeke (1998) showed how local influence could be used to detect influential subjects within longitudinal data analysis.

There are several statistics that may be calculated so as to assess influence in both the fixed effects and covariance parameter estimates. However, it is wise to first gauge a measure of overall influence. This is done by measuring changes in the objective functions, and as Schabenberger (2004) noted, in linear mixed models the objective function is related to ML and REML estimation. Therefore, a common measure of overall influence is the (restricted) likelihood distance (RLD) of (Cook & Weisberg, 1982) or, as it is oft termed, likelihood displacement (Beckman et al., 1987).

A large (restricted) likelihood distance indicates that a subject may indeed be influential but, does not indicate the nature of the influence. It is important to investigate further, so as to determine whether a subject is influencing the fixed effects or covariance parameters (Schabenberger, 2004).

There are many summary statistics that may be employed to further investigate influential subjects in either the fixed effects or covariance parameters, namely; Cook's D (Cook, 1977), multivariate DFFITS (MDFFITS) (Belsley et al., 1980), the PRESS statistic (Allen, 1974), the covariance ratio (CovRatio) and the trace of the covariance matrix (CovTrace) (Christensen et al., 1992). West et al. (2007) and Zewotir & Galpin (2007) gave concise summaries of these statistics, as did authors Schabenberger (2004) and Littell et al. (2006).

Using iterative influence analysis, Cook's D and CovRatio may be recomputed for the covariance parameters. Cook's D indicates the degree to which the parameter estimates may be influenced by a particular subject while, CovRatio indicates how severely precision of the estimates may be affected by a subject. In the case of Cook's D, large values are indicative of an influential subject whereas, the benchmark value for CovRatio is the value one. Values greater than one suggest higher precision for the full-data parameter estimates, while values less than one suggest increased precision for the reduced-data estimates (Zewotir & Galpin, 2005; Littell et al., 2006).

If an identified influential subject is removed, and there is no significant change in the model inference, it indicates that the influence of the subject is not substantial. It also suggests that the subject is in fact a true outlier and not influential hence, it may remain in the data set. However, such identified subjects should be scrutinized for any possible cases presenting anomalies.

3.3 Application

The cohort that was randomly selected for this study consisted of 1478 households. The data contained repeated measures over a 5-year period, with each household having measurement occasions where electricity meters were read. For each meter reading a date and

consumption value was recorded. From each reading date the month of meter reading was extracted so as to include in the modelling process. In addition to the recorded consumption values and month of reading, a dwelling type classification for each household was also provided.

Despite us using a sample of 1478 households, 20 repeated measures for each household still creates a large data set. Kiernan et al. (2012) provided useful insight as to how to best approach large data scenarios when using mixed models in SAS. To fit a mixed model to the electricity data we had a choice of two SAS procedures namely; PROC MIXED and PROC HPMIXED. SAS (2011) highlighted the pros and cons of each. Although PROC HPMIXED is preferable to use in large data scenarios, it limits the choices of covariance structures for $\mathbf{R}$, by always assuming $\mathbf{R} = \mathbf{I}\sigma^2$ (SAS, 2011). For this reason, the use of PROC MIXED was preferred when carrying out modelling and analysis of the electricity data.

We assume that for each repeated measurement on the i^{th} household, $i = 1, \ldots, 1478$, there are n_i measurement occasions and that y_{ij} is the electricity consumption value at time t_{ij}, $j = 1, \ldots, n_i$. As meters are read at 3-month intervals, we expect each household to have a minimum of 20 measurement occasions for the 5-year period. This means that each household should have a minimum of 20 consumption values, 1 most recent or current consumption value and 19 lagged values. We recall that a focal point of this study is to determine whether these lagged values affect the current electricity consumption value and if so, how many and which lags are effecting significance on a household's current electricity usage. Cursory examination of a profile plot in Figure 2.1 indicated that a seasonal pattern was occurring, which in turn confirmed that at least some lags would have an effect on current consumption.

To explore the possibility that every available lag a household has affects the current electricity consumption of that household, we tentatively fit a marginal linear regression model without specifying a specific covariance structure. Current consumption is modelled as a function of month of meter reading, dwelling type and the 19 lagged values. We then study the coefficients and significance of each lag, so as to determine whether there is value in retaining all the lags that a household has in the regression model. The coefficients and

p-values of the 19 lagged consumption values are displayed in Table 3.1.

Table 3.1: Coefficients and p-values of the 19 lags

Lag	Coefficient Estimate	Pr > |t|
1	0.4386	<0.0001
2	0.2230	<0.0001
3	0.06617	0.0001
4	0.1685	<0.0001
5	-0.1026	<0.0001
6	-0.06780	0.0001
7	0.07944	<0.0001
8	0.1309	<0.0001
9	-0.01415	0.3798
10	-0.1098	<0.0001
11	0.004017	0.7813
12	0.1882	<0.0001
13	-0.03844	0.0153
14	0.1119	<0.0001
15	-0.05653	0.0049
16	0.1873	<0.0001
17	-0.1367	<0.0001
18	-0.1454	<0.0001
19	0.05585	<0.0001

From Table 3.1 we see that all the lags except for lags 9 and 11 are significant at a 5% level of significance. However, closer examination of the lag coefficients reveals that after lag_{12}, the remaining lags in fact have little effect on the overall current electricity consumption value of a household. That is, we see that the coefficients of lags 13, 14, 15, 16, 17, 18 and 19 contribute very little, with the positive coefficients of lags 14, 16, and 19 and the negative coefficients of lags 13, 15, 17 and 18 almost canceling each other out. This leads us to conclude that there is little value in retaining more than 12 lagged values in the regression model. Hence, in the interests of parsimony we continue the modelling process using only 12 lags in the model. This equates to all electricity consumption values up to and including 3 years prior to a household's current consumption value being used.

By retaining lags within the model, we are later able to examine their estimated coefficients in the final model which enables us to formally assess whether or not a seasonal pattern is present. In addition to assessing seasonal trends, we are also able to make observations as to how the most recent previous electricity consumption values affect the modelling of current electricity usage. Along with the 12 lagged values, dwelling type and the month of meter reading also remain in the model. By including both lags and month of meter reading in the modelling process, we are able to analyze two different factors. The inclusion of the month of meter reading enables us to tentatively gauge if some months have higher or lower electricity consumption than others while, lags allows for the assessment of a potential seasonal pattern occurring within households electricity usage.

Using month of reading, dwelling type and 12 lagged values we specify the following model

$$\begin{aligned} E(y_{ij}) &= \beta_0 + \alpha_{i0} + \beta_1 DwellingType_{House} + \beta_2 DwellingType_{Shareblock \leq 2\ storeys} + \\ &\quad \beta_3 MonthRead_{Jan} + \beta_4 MonthRead_{Feb} + \ldots + \beta_{13} MonthRead_{Nov} \\ &\quad + \beta_{14}\ y_{ij-1} + \beta_{15}\ y_{ij-2} + \ldots + \beta_{25}\ y_{ij-12} \end{aligned} \tag{3.3}$$

where $E(y_{ij})$ is the expected electricity consumption for the i^{th} household at time t_{ij}, $j = 1, 2, \ldots, n_i$, if n_i is the number of measurement occasions for the i^{th} household. We also fit a household-specific random intercept, α_{i0}, to the model, so as to account for the variability of different households within the variance components.

Model Selection

After specifying the model, we next focus on selecting an appropriate covariance structure. We applied several covariance structures to the data: unstructured (UN), compound symmetric (CS), first-order autoregressive (AR(1)), autoregressive moving-average of order one (ARMA(1,1)), first-order ante-dependance (ANTE(1)) and Toeplitz (Toep). Both the ANTE(1) and Toeplitz models failed to reach convergence and are no longer considered viable covariance structures for our data. Table 3.2 displays the corresponding fit statistics of the structures that converged as well as, the number of iterations until convergence of the Newton-Raphson algorithm was attained.

Table 3.2: Fit statistics for selected covariance structures

Covariance Structure	Iterations	AIC	BIC
UN	4	182924.3	183279.2
CS	3	184957.9	184973.8
AR(1)	2	184896.9	184912.7
ARMA(1,1)	6	184286.0	184307.2

From Table 3.2, we see that there are four possible covariance models from which we may choose. We also note that though these models converge, they converge with a non-positive definite $\mathbf{G}$ matrix. One possible reason for this could be due to model over-specification, possibly due to the inclusion of a random intercept variance component. Nevertheless, we refrain from removing the random intercept from the model. Not only is it possible that an influential household/s could be bringing about such an effect but, we intend to later use our model to predict electricity consumption for individual households and the inclusion of a random intercept improves the model's prediction capabilities.

To select the structure best suited to our data, we go according to the smaller-is-better information criteria. However, when selecting a covariance structure, another factor to consider is the number of parameters that require estimating. Too many parameters make the model less favourable. Based on this, we exclude the unstructured model. Even though it has the smallest AIC value, it has far too many parameters that require estimating. Of the remaining structures we see from Table 3.2, that the ARMA(1,1) model clearly has the lowest fit statistics, having an AIC value of 184286.0. We note that this structure only has one additional parameter than the CS and AR(1) models, which is not enough to exclude it from being the best suited covariance structure. Therefore, according to the smaller-is-better IC, we select the ARMA(1,1) model as the best suited covariance structure and use it to model $\mathbf{R}$.

Having fitted the model with an ARMA(1,1) covariance structure and random intercept, we now assess the validity of the normality assumptions in the model, recalling that the histogram in Figure 2.2 already suggested the that the assumption was violated. To assess normality we now study a scatter and Q-Q plot of the conditional studentized residuals. The plots are given in Figure 3.1.

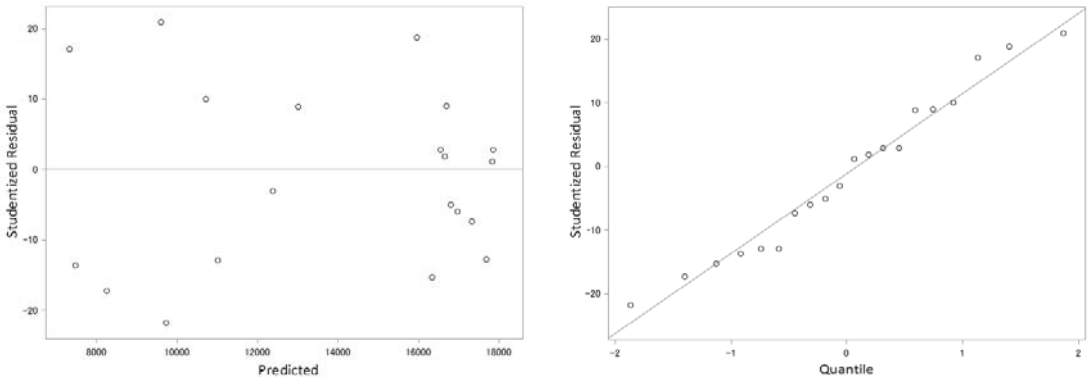

Figure 3.1: Scatter and Q-Q plot of the conditional studentized residuals

In Figure 3.1, we observe a random scattering about zero in the scatter plot, supporting the assumption of constant variance in the error term. From the Q-Q plot in Figure 3.1, we see the majority of points close to and around the straight line, suggesting that the normality assumption of the model is not severely violated. However, we do note that the points fail to form a straight line and that there are also several outliers. This confirms our suspicions from the histogram in Figure 2.2 that is, the electricity consumption data does not adhere to normality assumptions.

When a model or data fail to adequately satisfy normality assumptions, there are two possible remedies a statistician may employ. Either, the model can be refitted using a different approach (often a generalized approach) or, the data may be subjected to a transformation so as to improve normality.

We elect to use the method of transformation in this chapter. An alternate approach is demonstrated in Chapter 4, where we refit the model using a generalized linear mixed model. Various power transformations ($y_{ij}^{(\lambda)}$) were applied to the data, Table 3.3 displays the corresponding fit statistics for each transformation.

From Table 3.3, we see that from the positive power side, as λ approaches zero our model fit improves. That is to say, we observe that the closer λ gets to zero, the faster the model converges and the fit statistics get smaller. Similarly, from the negative power side, as the power gets closer to zero we see that the model fit gets better. This suggests to us, that

Table 3.3: Fit statistics for various data transformations

Transformation	Iterations	AIC	AICC	BIC
λ = -2	8	-32947.9	-32947.9	-32926.7
λ = -1.5	6	-47304.6	-47304.6	-47288.4
λ = -1	5	-59718.0	-59718.0	-59696.8
λ = -0.5	5	-67568.7	-67568.7	-67552.8
λ = 0 (natural log)	4	-14.7	-14.7	6.5
λ = 0.5	3	63350.4	63350.4	63371.6
λ = 1 (no transformation)	6	184286.0	184286.0	184307.2
λ = 1.5	7	311215.2	311215.2	311230.4
λ = 2	19	450911.5	450911.5	450932.7

the best convenient transformation for our data is when $\lambda = 0$. However, by both Tukey's power transformation definition (Tukey, 1957) and the Box-Cox power transformation definition (Box & Cox, 1964), when $\lambda = 0$ we instead take the natural log of y_{ij}. Further support for using the natural log transformed data as opposed to the observed data can be seen when we compare the Q-Q probability plots of each data. Figure 3.2(a) shows the probability plot of the observed untransformed data while, Figure 3.2(b) displays the probability plot of the natural log transformed data.

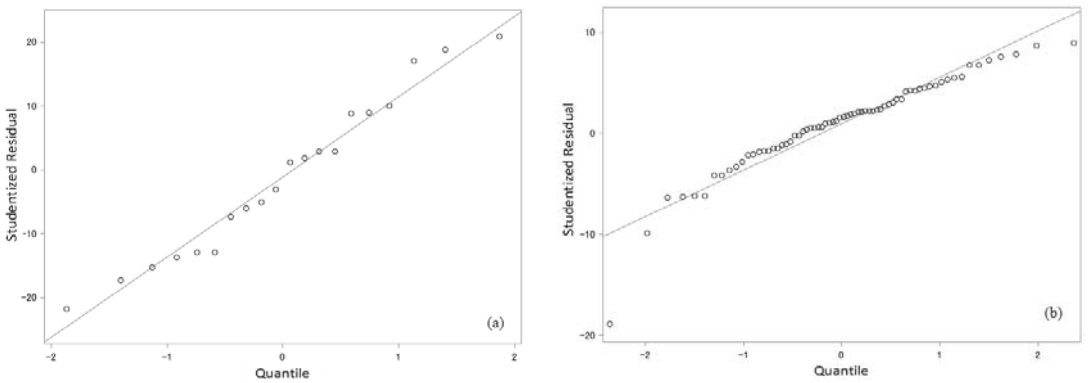

Figure 3.2: Q-Q plots of the conditional studentized residuals for: (a) observed data; (b) natural log transformed data

When we compare the plots of Figure 3.2(a) and 3.2(b), we can clearly see that the log transformed data provides a better fitting model and better satisfies the normality as-

sumption. There is evidence of this both in the improved fit statistics and the fact that other than one or two outliers, most points lie on the line in the Q-Q plot of Figure 3.2(b). Therefore, we proceed with our analysis using the convenient natural log transformed data.

Next, we study the effect any influential households may have on the model. As we expressed earlier, in repeated measures mixed modelling we are particularly interested in the covariance structure as it accounts for the correlation of observations within a household. Thus, it follows that we are most concerned about households that exhibit influence in the covariance parameters.

However, despite this being our interest, we are unable to study the case deletion plots of Cook's D and the covariance ratio (CovRatio) for the covariance parameters. This is because their values were not estimable, due to the variance component of the random intercept being negative.

Instead, we begin by identifying the 3 outliers that draw our attention in the Q-Q plot of Figure 3.2(b). We see that they belong to households 46, 202 and 1205. Next, we study the restricted likelihood (RLD) case deletion plot, looking to see if these same 3 households (as well as any others) stand out as having overall influence in the model. The RLD case deletion plot is given in Figure 3.3.

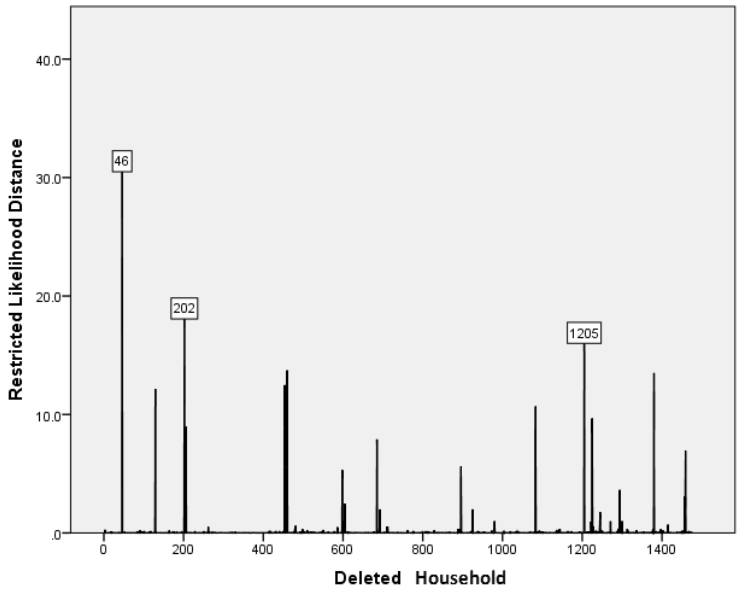

Figure 3.3: Plot displaying RLD per deleted household

From Figure 3.3, we clearly see that in terms of overall influence, households 46, 202 and 1205 stand out as being the most influential. Of all the households in the data, they have the largest RLD values of 33.976, 18.972 and 16.906 respectively.

Having identified households 46, 202 and 1205 as both outliers in the Q-Q plot and potentially influential according to the RLD plot, we remove them from the data and observe the effect (Zewotir & Galpin, 2005). If no significant change is observed, we may classify them as non-influential outliers and keep them in the data set. However, if we observe significant change in the model fit and covariance parameter estimates, we consider the household genuinely influential and remove it from the data set. Table 3.4 displays the fit statistics and the covariance parameter estimates of the full and reduced data models. The relative error between the full-data and reduced-data covariance parameters is expressed as a percentage and also displayed in the table.

Table 3.4: AIC statistics and covariance parameter estimates: full-data and reduced-data models

Data Set	AIC	Intercept	% Rel. Error	$\hat{\rho}$	% Rel. Error	$\hat{\gamma}$	% Rel. Error	Residual	% Rel. Error
Full-data	-14.7	-0.00114	-	0.5308	-	0.2777	-	0.06301	-
46 removed	-679.4	0.000424	137.19	0.4877	-8.11	0.2646	-4.72	0.05891	-6.51
202 removed	-701.4	-0.00101	11.40	0.5393	1.60	0.2924	5.29	0.06060	-3.82
1205 removed	-591.6	-0.00096	15.79	0.5440	2.49	0.2719	-2.09	0.06037	-4.19
46, 202 removed	-1403.0	0.000566	149.65	0.4958	-6.59	0.2797	0.72	0.05647	-10.38
46, 1205 removed	-1289.7	0.000456	140.00	0.5149	-3.00.	0.2572	-7.38	0.05633	-10.60
202, 1205 removed	-1303.3	-0.00087	23.68	0.5567	4.88	0.2869	3.31	0.05798.0	-7.98
46, 202, 1205 removed	-2041.3	0.000541	147.46	0.5314	0.11	0.2724	-1.91	0.05394	-14.39

From Table 3.4, we see that the removal of household 46 brings about an improved AIC value as well as significant change in the model. We clearly see that whenever household

46 is removed, whether on its own or with other identified households, the estimate for the random intercept becomes positive. This shows us that household 46 is very influential in the model, as it alone determines whether or not the model converges with a positive definite $\mathbf{G}$ matrix. Further to this, we note that the largest changes in both $\hat{\rho}$ and $\hat{\gamma}$ are witnessed upon household 46's removal. This leads us to classify household 46 as influential. Hence, we remove it from the data and continue analysis using the reduced-data model. However, the fact that household 46 is influential, suggests to us that it ought to be scrutinized thoroughly and further investigations carried out on its characteristics.

When we re-visit the observed data we see that household 46 differs from the other households, in that it has 3 consecutive measurement periods where the monthly electricity consumption is very low. The monthly electricity consumption for these 3 months is 1.00KWh, 2.33kWh and 0.67kWh. This anomaly is then carried over to the transformed data, where we took the natural log of the observed values. Possible reasons for this anomaly could include errors in the meter reading process or, that the property was vacant during this time.

Next, we look at the other two households identified in Table 3.4, households 202 and 1205. It is clear that their removal does not affect the model as significantly as the removal of household 46. That is to say, regardless of their inclusion or exclusion the model still does not converge with a positive definite $\mathbf{G}$ matrix. Nevertheless, we do note a significant improvement in the AIC values upon their removal. However, this is to be expected as we noted them appearing as outliers in the Q-Q plot. Though the removal of outliers undoubtedly improves the fit statistics of a model, it does not necessarily mean that these households exercise influence. As we do not see hugely significant changes in the covariance parameters upon the removal of household 202 and 1205, we classify them as non-influential outliers in the model and retain them in the data. When we re-visit the data, we see that a possible reason for them appearing as outliers is that they both have between 3 and 5 consecutive measurement occasions where their monthly electricity consumption is very low, compared to those of their other monthly averages.

As we have decided to remove household 46 from the data and re-fit the model, it is prudent that we re-examine the reduced-data probability plot. We expect to see some

changes in it as the reduced-data model now converges with a positive definite Hessian and **G** matrix. The reduced-data Q-Q plot is given in Figure 3.4.

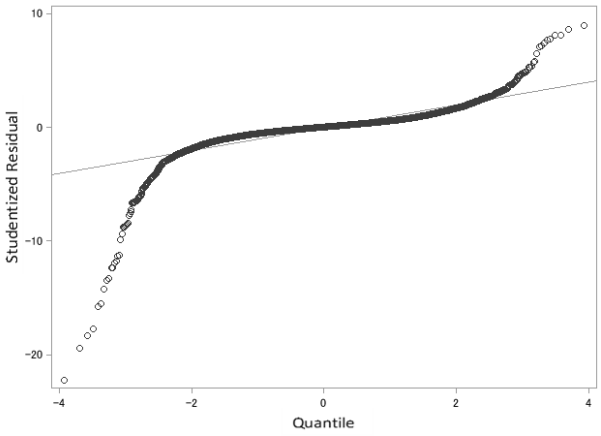

Figure 3.4: Q-Q plot of studentized residuals for reduced-data model

From Figure 3.4 we see that the majority of points lie on the straight line while the tails in the plot depart from the line. As the data that we are dealing with are large, we are satisfied with the fact that clearly most of the points adhere to the normality assumption. The tails that depart from the line are due to a few outlying households in the data. So long as these outliers have proven not to be influential we retain them in the data, such is the case for households 202 and 1205. Having thoroughly studied the 3 identified households and the reduced-data Q-Q plot, we proceed to making model inferences. Model inferences are based on the reduced-data model where household 46 has been removed.

Model Inference

We begin our analysis by examining the type 3 tests and solutions for the fixed effects. Before we look at tables of our results we recall that we are using the convenient natural log transformation of the response. Thus, it follows that the lagged values in the model are also logged. That is, using the reduced-data model with $y^*_{ij} = ln(y_{ij})$ and $y^*_{ij-1}, \ldots, y^*_{ij-12} = ln(y_{ij-1}), \ldots, ln(y_{ij-12})$, our inferences are according to the following model:

$$\begin{aligned} E(y^*_{ij}) &= \beta_0 + \alpha_{i0} + \beta_1 DwellingType_{House} + \beta_2 DwellingType_{Shareblock \leq 2\ storeys} + \\ &\quad \beta_3 MonthRead_{Jan} + \beta_4 MonthRead_{Feb} + \ldots + \beta_{13} MonthRead_{Nov} \\ &\quad + \beta_{14}\ y^*_{ij-1} + \beta_{15}\ y^*_{ij-2} + \ldots + \beta_{25}\ y^*_{ij-12} \end{aligned} \tag{3.4}$$

We note for all further inferences, due to ease of explanation, we refer to $y^*_{ij-1}, \ldots, y^*_{ij-12}$ as $lag^*_1, \ldots, lag^*_{12}$. We begin with the type 3 test results for dwelling type and month of meter reading, the results are displayed in Table 3.5.

Table 3.5: Type 3 test results for dwelling type and month read

Effect	Num DF	Den DF	F Value	Pr > F
Dwelling Type	2	13303	0.30	0.6919
Month Read	11	13303	12.04	<0.0001

From Table 3.5, we see that dwelling type does not appear to be significant in the model. At a 5% level of significance we fail to reject the type 3 test null hypothesis, H_0: $\beta_1 = \beta_2$, as we observe a large p-value of 0.6919, indicating that dwelling type does not play an important role in the model. Further evidence of this is presented in Table 3.6, where it is clearly seen that relative to shareblocks > 2 storeys high, houses and shareblocks $\leq$ 2 storeys high are not significant at a 5% level of significance, having p-values of 0.8012 and 0.4047 respectively.

Table 3.6: Solutions for fixed effects: dwelling type

Dwelling Type [Ref: Shareblocks > 2 storeys]	Estimate	Standard Error	DF	t Value	Pr > |t|
House	-0.00178	0.007053	13303	-0.25	0.8012
Shareblocks $\leq$ 2 storeys	-0.00565	0.006783	13303	-0.83	0.4047

Next from Table 3.5, we consider month of meter reading (month read). At a 5% level of significance we reject the null hypothesis, H_0: $\beta_3 = \beta_5 = \ldots = \beta_{13}$, as we observe a p-value of <0.0001. This suggests that at least one month is significant in our model and that the month in which a meter is read does play a role in modelling the current electricity usage. Thus, further examination of individual estimates is required. The estimates and significance of each month are displayed in Table 3.7.

Before we start studying these estimates, we first recall that meters are read at approximate 3 month intervals throughout eThekwini. This implies, that the month that a meter is read in not only refers to the average consumption of that month, but also two months prior to it. Therefore, any inferences made ought to also include two months prior to the month of reading.

Table 3.7: Solutions for fixed effects: month read

Month Read [Ref: Dec]	Estimate	Standard Error	DF	t Value	Pr > |t|
Jan	0.008710	0.01021	13303	0.85	0.3936
Feb	0.006739	0.009328	13303	0.72	0.4700
Mar	0.01886	0.007096	13303	2.66	0.0079
Apr	0.01463	0.01056	13303	1.38	0.1661
May	-0.00951	0.009171	13303	-1.04	0.2997
Jun	0.02471	0.007431	13303	3.32	0.0009
Jul	0.02059	0.01157	13303	1.78	0.0750
Aug	0.04700	0.01006	13303	4.67	<0.0001
Sep	0.06388	0.007659	13303	8.34	<0.0001
Oct	0.05999	0.01185	13303	5.06	<0.0001
Nov	0.01060	0.01052	13303	1.01	0.3134

From Table 3.7 we are able to see that relative to December, the months August, September and October are highly significant, having p-values that are <0.0001. We also observe that these months have the highest coefficient estimates, suggesting that these months may have higher electricity consumption compared to the reference month, December. However, our inference ought to also include two months prior to these months. That is, if August appears to have a high average monthly consumption, we must acknowledge that this measurement period also included the months June and July. By this logic, as we found the months August, September and October to be significant and have large coefficients, we infer that the period from approximately June through to October, relative to December, is likely to have higher monthly electricity consumption. We note that this period encompasses approximately winter in South Africa, possibly suggesting that electricity usage increases during winter.

Lastly, we remind the reader that "month read" is only approximate, in the sense that the 3 month actual usage is distributed to months by means of scaling the 3-month value by a factor of 3. eThekwini Electricity also do not read meters according to calendar months, but rather approximate 90 day periods. So, although we gleam interesting inferences based on the months meters are read in, caution should be exercised when doing so. The final

fixed effects that we examine are the intercept and lags, their corresponding estimates and p-values are displayed in Table 3.8.

Table 3.8: Solutions for fixed effects: intercept and lags

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	0.04818	0.02513	1472	1.92	0.0554
lag_1^*	0.3316	0.008167	13303	40.60	<0.0001
lag_2^*	0.1134	0.008354	13303	13.58	<0.0001
lag_3^*	0.1466	0.008618	13303	17.01	<0.0001
lag_4^*	0.1939	0.008740	13303	22.19	<0.0001
lag_5^*	-0.02743	0.008802	13303	-3.12	0.0018
lag_6^*	-0.04052	0.009002	13303	-4.50	<0.0001
lag_7^*	0.03767	0.009231	13303	4.08	0.0001
lag_8^*	0.2128	0.009519	13303	22.35	<0.0001
lag_9^*	-0.07722	0.009512	13303	-8.12	<0.0001
lag_{10}^*	-0.02966	0.009611	13303	-3.09	0.0020
lag_{11}^*	-0.00992	0.009872	13303	-1.00	0.3151
lag_{12}^*	0.1348	0.009783	13303	13.78	<0.0001

First from Table 3.8, we see that at a 5% level of significance the model intercept, β_0, is not significant in the model. We also see that the only lag not significant at a 5% level of significance is lag_{11}^*, as it has a p-value of 0.3151. Nevertheless, this is not where our interest lies. Our main focus is the coefficients of the remaining lags, as we use them to assess the presence of any seasonal trends.

We see that lag_1^* carries the highest coefficient estimate, 0.3316. This tells us that the most recent consumption value contributes more to the model than other older consumption values. Additionally, we observe that lag_2^*, lag_3^* and lag_4^* also have high coefficient estimates (0.1134, 0.1466 and 0.1939 respectively). This information, combined with the coefficient value of lag_1^*, indicates that the most recent previous year's electricity consumption contributes considerably to the model.

We also note that lag_4^* has a larger coefficient than that of lag_2^* and lag_3^*. We find this interesting because, lag_4^* in fact corresponds to 12 months prior to the current electricity

consumption value, y^*_{ij}. Similarly, we observe large coefficients for lag^*_8 and lag^*_{12} (0.2128 and 0.1348 respectively). Both of these lags also refer to the same period in time as y^*_{ij}, only 24 and 36 months previously. Therefore, the large and significant coefficients of lags 4, 8 and 12 provide substantial evidence of there being a seasonal pattern present in monthly electricity consumptions. It shows that when predicating electricity usage for a particular month, consumption from the same month in previous years is just as important to the most recent previous year's usage. Hence, a seasonal pattern is clearly present and ought to be taken into account when modelling current household electricity consumption.

Finally, we look at the covariance parameter estimates, recalling that we fitted a first-order autoregressive moving average (ARMA(1,1)) structure. This allowed us to model the correlation of monthly electricity consumption at different measurement occasions within the same household. The ARMA(1,1) model ensures that the more time that passes between measurement occasions, the less correlated consumption values are. Similarly, electricity usages at consecutive (adjacent) measurement occasions or, measurement occasions close together in time, are highly correlated. We also recall that a household-specific random intercept was included in the model. The covariance parameter estimates are displayed in Table 3.9.

Table 3.9: Covariance parameter estimates

Covariance Parameter	Estimate	Standard Error	Z Value	Pr Z
Intercept	0.000424	0.000960	0.44	0.6584
$\hat{\rho}$	0.4877	0.06814	7.16	<0.0001
$\hat{\gamma}$	0.2646	0.02619	10.10	<0.0001
Residual	0.05891	0.001358	43.43	<0.0001

The first observation we make from Table 3.9 is that in this reduced-data model, the random intercept appears to no longer be significant. Nevertheless, much as we mentioned at the beginning of this application, we retain the random intercept in the model as it improves the model's prediction capabilities.

The covariance parameter $\hat{\gamma}$, estimates the lag_1 correlation while, remaining consecutive correlation is partially estimated by $\hat{\rho}$. From Table 3.8, we see that the estimates for γ and ρ are 0.2646 and 0.4877 respectively. Following from the ARMA(1,1) structure, we

see that lag_1 correlation is constant and that its corresponding covariance function may be estimated by $(0.000424) + (0.05891)(0.2646)$, where the random intercept is accounted for by 0.000424. Subsequent (lag_2 and onwards) correlations decrease with the amount of time that passes between measurement occasions, that is the covariance becomes a function of the lag and is estimated by $(0.000424) + (0.05891)(0.2646)(0.4877)^{lag}$.

Summary

The modelling process began with us investigating whether or not it was necessary to utilize all previous consumption values of a household. We determined that it was only of value to retain up to and including only 3 years of prior consumption values. That is, we found it was beneficial to retain 12 lags within the regression model. Following this discovery, we proceeded to fit a linear mixed model to the electricity data. We included month of reading, dwelling type, 12 lags and a household-specific random intercept in the model. An ARMA(1,1) covariance structure was used to model $\mathbf{R}$. Residual analysis showed that for the fitted model, the majority of data appeared to satisfy the assumptions of normality. However, it was clear that the normality assumption was not entirely satisfied. To improve upon this, we elected to use the method of transformations. Upon testing several power transformations, we found the best convenient transformation to be the natural log transformation. Influence diagnostics revealed that household 46 was influential enough to warrant its removal from the data. While the other identified households; 202 and 1205 were confirmed as non-influential outliers and retained in the data.

Essentially there were 3 components that we dealt with in our model. We had dwelling type classification, the months that meters were read in and we had the 12 lags. We observed that dwelling type did not play a significant role in predicting electricity consumption. The inclusion of 12 lags allowed us to model and assess how the previous consumption of a household affected their current electricity usage. To this end we found that the most recent previous year's electricity consumption, as well as that which was 12, 24 and 36 months prior to the current consumption was also important. However, we note that each household had their meter read at different months in the year, for example some may follow a reading pattern of February, May, August, November while, some may follow a pattern of January, April, July, October. Though the pattern in the lags for these households remains, using the months we were able to further assess if there were some

months that had higher consumption than others. We found that the months July, August and September likely had higher electricity consumption, relative to the reference month December. We observed that these months also constituted an approximate winter period in South Africa.

The transformation approach we used improved the fit. However, the normality plot did not clearly favour the normality of the transformed data. Perhaps directly searching for the type of distribution might be the solution. In the next chapter we will explore this in detail.

Chapter 4

The Generalized Linear Mixed Model Approach

In the previous chapter, we recall that in order to satisfy the Gaussian assumptions of the model, we performed a log transformation on the response variable. Such a transformation approach is often employed in various scenarios (see e.g. Box & Cox, 1982; Gurka et al., 2006). In a mixed model setup Zewotir & Galpin (2004) recommended such transformations if the error structure had a constant variance. Interpretation of the parameters on the transformed scale and the uniqueness of the transformation are some of the concerns of the transformation approach (see e.g Box & Cox, 1982; Games, 1984; Manning, 1998). Tabachnick & Fidell (2013) cautioned that although often used to remedy violations in normality and outliers, transformations should not be universally recommended. Analysis should be interpreted from the variables that are in it and the interpretability of the parameter estimates should be taken into consideration.

Though the natural log transformation used in Chapter 3 is a commonly used scale transformation and interpretation is not getting sophisticated by the transformation, it is not a direct modelling of the observed monthly electricity consumption. A viable alternative to using transformations is that of generalized linear models. Generalized linear models is a general approach that is not restricted by the assumption of normality (Montgomery et al., 2012). It assumes a more general family of probability distributions called the exponential family of distributions, which include the normal, gamma and many other distributions.

In this chapter, we look at pertinent theory of the generalized linear mixed model (GLMM) and use this generalized approach to refit the municipal data.

4.1 The Generalized Linear Model (GLM)

The overall idea of generalized linear models is to specify the model

$$g(E(\mathbf{y})) = \mathbf{X}\boldsymbol{\beta}$$

in terms of the distribution of the data and mean function and to express the mean in terms of a transformation of $\mathbf{X}\boldsymbol{\beta}$ (Littell et al., 2006).

Nelder & Wedderburn (1972) are widely credited for their work that unified a broad class of existing models into a GLM definition. They did this, by proving that so long as the response could be classified into the exponential family, the maximum likelihood estimators for all of the models could be obtained using the same algorithm.

Fundamentally, there are three parts to specifying a GLM: a distributional assumption, a systematic component and a linear predictor (Fitzmaurice et al., 2004).

Under the distributional assumption, we assume that independent observations, Y_i for $i = 1, \ldots, n$, have a probability distribution that belongs to the exponential family. For the systematic component, we relate the effect of the covariates $X_{i1}, \ldots, X_{ik}$ on the expected value of Y_i by means of a linear predictor. The linear predictor is given by $g(E(Y_i)) = \eta_i = X_i\beta$, with maximum likelihood methods being used to estimate the parameter β. Lastly, the link function, $g(.)$, is a monotonic differentiable function that describes the relationship between the linear predictor, η_i, and the expected value of Y_i. If the expected value of Y_i is given by μ_i, then we may express the link function as $g(\mu_i) = \eta_i$.

Having briefly examined the constituents of a GLM, we may now move on to look at the generalized form of the linear mixed model. For a more detailed, mathematical breakdown of generalized linear models, refer to common textbooks such as McCulloch & Nelder (1989), Fitzmaurice et al. (2004), McCulloch et al. (2008) or Montgomery et al. (2012).

4.2 The Generalized Linear Mixed Model (GLMM)

Fitting a GLMM allows us to extend the GLM by enabling us to incorporate correlations among responses (Schabenberger, 2005). There are two ways in which we may model correlation namely, either by adding a random effect inside the linear predictor or, by modelling the correlations directly within the data itself. It has become commonplace to refer to the afore mentioned, respectively, as modelling the G-side and R-side random effects. These terms stem from the definition of the linear mixed model variance components given in Chapter 3. A GLMM may contain either a G-side or R-side random effect, or both, or neither. The modelling in the GLMM assumes the specified covariance structures of $\mathbf{G}$ and/or $\mathbf{R}$.

As with a GLM, a GLMM also has a linear predictor, $\boldsymbol{\eta}$, a monotonic differentiable link function, $g(.)$, and the assumption that the distribution of Y as well as, the conditional distribution, belong to the exponential family. If we wish to fit a GLMM that contains a G-side random effect, we specify the conditional model

$$g(E(\mathbf{y}|\mathbf{u})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$$

where we assume $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, select a covariance structure for $\mathbf{G}$, and specify the distribution of $\mathbf{y}|\mathbf{u}$. We may also note, in our model specification we have defined the linear predictor as $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ (i.e. we have added a random effect into our linear predictor of GLM).

However, if we would rather fit a GLMM that only has R-side random effects, we find we specify a model that resembles that of a GLM

$$g(E(\mathbf{y})) = \mathbf{X}\boldsymbol{\beta}$$

where we may select the covariance structure of $\mathbf{R}$. Our linear predictor is the same as it was in a GLM, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. With no random effect in the linear predictor, we have nothing to condition our response on. We may obtain the mean by using the inverse link function, i.e. $E(\mathbf{y}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$. While, the variance is given by $Var(\mathbf{y}) = \mathbf{A}^{\frac{1}{2}}\mathbf{R}\mathbf{A}^{\frac{1}{2}}$ where $\mathbf{A}$ is a diagonal matrix that contains the variance functions.

Of course, it is also possible to specify a GLMM that contains both G-side and R-side random effects. To do so, we would select the covariance structures for $\mathbf{G}$ and $\mathbf{R}$, then the conditional mean and variance would be given by:

$$E(\mathbf{y}|\mathbf{u}) = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})$$

$$Var(\mathbf{y}|\mathbf{u}) = \mathbf{A}^{\frac{1}{2}}\mathbf{R}\mathbf{A}^{\frac{1}{2}}$$

We recall from Chapter 3, that the method of maximum likelihood or restricted maximum likelihood was used to estimate the model parameters in the LMM. That is to say, estimations of parameters were found by maximizing the objective (likelihood) function. Similar estimation approaches also apply to the GLMM whereby, estimation is based on the likelihood principle (Schabenberger, 2005). However, in GLMMs the procedure of obtaining the objective (likelihood) function is somewhat more complex than that of LMMs.

If we define $p(\mathbf{u})$ as the distribution of the random effects and assume it to be multivariate normal, and define $f(\mathbf{y}|\mathbf{u})$ as the conditional distribution of the data and assume it to belong to the exponential family then, we define the marginal likelihood function as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y}) = \int f(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u}.$$

We already know that in a linear setting it is easy for us to maximize $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y})$, however, if the random effects are nonlinear in form, it becomes difficult for us to obtain the above integral. Various solutions have been used and different recommendations have been made, as to how to get around this estimation problem (see e.g. Kachman, 1998; Fitzmaurice et al., 2004; Demidenko, 2004; McCulloch et al., 2008). We may group these solutions into two broad categories, namely, integral approximation and model linearization (Schabenberger, 2005).

Integral approximation provides us with an actual objective function we can maximize, by using a variety of numerical methods to approximate $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{y})$. Some of the different techniques used for integral approximation include; Laplace methods (see e.g. Breslow & Clayton, 1993; Breslow & Lin, 1995; Raudenbush et al., 2000; Demidenko, 2004), quadrature methods (see e.g. Pan & Thompson, 2003; Fitzmaurice et al., 2004) and Monte Carlo integration methods (see e.g. Pan & Thompson, 2007). For a concise summary of the afore

mentioned methodologies refer to Demidenko (2004).

The main benefit of these methods is that by having an approximated objective function, we may compare nested models using true likelihood ratio tests. However, it is not always appropriate to use integral approximation. Schabenberger (2005) explained how integral approximation methods were ill-suited to handle complex R-side covariance structures, multiple household effects and large numbers of random effects.

While integral approximation focuses on approximating the objective function, linearization aims to approximate the model. Linearization is an iteratively intense process, whereby, pseudo-data (that contains fewer nonlinear components) is used to approximate the GLMM using an LMM with current parameter estimates. The LMM is fit, and upon convergence the new parameter estimates update the linearization, leaving a new linear mixed model. The process continues until the change in estimates, of two consecutive iterations, reaches a specified level of tolerance (Schabenberger, 2005).

Linearization resolves the setbacks of integral approximation, while also allowing for the use of REML estimation. The main disadvantage of linearization is that we do not have an actual objective function, only one based on the pseudo-data after each linearization, which could lead to biased estimates.

The SAS procedure, GLIMMIX, performs the generalized linear mixed models analyses. When there are random effects in the model, the default method employed in the procedure is a pseudo-likelihood linearization approach (Schabenberger, 2005). Both SAS (2011) and Schabenberger (2005) provided a comprehensive mathematical breakdown of the pseudo-likelihood linearization method.

Under the distributional assumptions of a GLMM, we assume that the conditional distribution of the response variable is a member of the exponential family.

Generally speaking, there are three types of responses one may expect: continuous responses, count data or binary outcomes. We find that there are common distributions that often may be applied to each type of response. For example, we find that in contin-

uous responses, Y is often normally distributed, a Poisson distribution is used for count data and the Bernoulli distribution is applicable to binary outcomes. Of course, there are many other distributions in the exponential family that may be used, however, we shall not discuss all of them. Instead, as our electricity data is continuous in nature, we shall focus on distributions that may be applied to continuous data.

We considered five distributions that belong to the exponential family which may be applied to a continuous response. These are the normal (Gaussian) distribution, the log-normal distribution, the gamma distribution, the exponential distribution and the Inverse Gaussian distribution. The general form of the exponential family probability density function is

$$f(y) = exp\left\{\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y, \phi)\right\}$$

for some functions $a(.)$, $b(.)$ and $c(.)$, where θ is the natural (canonical) parameter and ϕ is the scale (dispersion) parameter.

Each distribution (including the five mentioned above) have their own natural (canonical) link functions. For example, the Normal and Lognormal distributions have the identity link as their canonical link function, while, the Gamma and Exponential distributions have a reciprocal link as their canonical link function.

Although each distribution has its own natural link function, we are not obligated to treat it as the only possible link function we can use. We may choose to fit other link functions to distributions, so as to attain a better model fit. By default the statistical software SAS generally uses canonical links for most distributions. However, with the Gamma and Exponential distributions, it uses the log link to fit a GLMM, while, the reciprocal link is in fact the natural link function (SAS, 2011). Using various links and the distributions mentioned, we shall fit a GLMM to our data using available procedures in SAS.

4.3 Application

The explanatory variables we include in the generalized linear model are dwelling type, month of meter reading and lagged electricity consumption values. Following from our Chapter 3 finding where we determined that it was unnecessary to include more than 12

lagged values in the regression model, we limit the generalized linear model to also only include 12 lags. This leads to the following GLMM being specified:

$$\begin{aligned} g[E(y_{ij})] \;=\; \eta_{ij} \;=\; & \beta_0 + \alpha_{i0} + \beta_1 DwellingType_{House} + \beta_2 DwellingType_{Shareblock \leq 2\ storeys} + \\ & \beta_3 MonthRead_{Jan} + \beta_4 MonthRead_{Feb} + \ldots + \beta_{13} MonthRead_{Nov} \\ & + \beta_{14}\ y_{ij-1} + \beta_{15}\ y_{ij-2} + \ldots + \beta_{25}\ y_{ij-12} \end{aligned} \tag{4.1}$$

where $g(.)$ is the link function, $E(y_{ij})$ is the expected response (monthly electricity consumption) for the i^{th} household, $i = 1, \ldots, 1478$ at time $j = 1, 2, \ldots, n_i$, if n_i is the number of measurement occasions for household i; α_{i0} represents a household-specific random intercept and; η_{ij} represents the linear predictor.

We do not require that the data adheres to the normality assumption as we, ourselves, specify both the distribution and link function. To carry out model fitting we use SAS PROC GLIMMIX. PROC GLIMMIX allows us to specify various link functions and distributions, as well as different covariance structures for **R** and **G**. An unfortunate aspect of PROC GLIMMIX is that it does not allow for advanced diagnostics on the model. To get around this problem we will rely on the identification of case outliers and the households that they belong to. These households may then in turn be removed from the data, the model re-fitted and the resulting effect observed. Before commencing with the model selection procedure, we remind the reader that all the information criteria (IC) are interpreted as smaller-is-better.

Model Selection

To begin model selection we fit five distributions suited to continuous data, along with various link functions. Namely, these are the normal, gamma, lognormal, inverse Gaussian and exponential distributions. Using the resulting AIC values, we select the best fitting distribution. In PROC GLIMMIX we invoke the method of REML estimation, using linearization and a Dual Quasi-Newton algorithm. Instead of searching again for the best suited covariance structure, we use our finding from Chapter 3 and fit an ARMA(1,1) structure to **R**. The full-data results for the various fitted generalized models are displayed in Table 4.1.

Table 4.1: The goodness of fit by distribution and link

Distribution	Link	Convergence	Iterations	AIC	BIC
Normal	identity	Yes (**G** matrix not p.d)	6	184306.1	184322.0
Normal	power ($\lambda = -2$)	No	.	.	.
Normal	power ($\lambda = -1.5$)	No	.	.	.
Normal	power ($\lambda = -1$)	No	.	.	.
Normal	power ($\lambda = -0.5$)	No	.	.	.
Normal	log {power ($\lambda = 0$)}	No	.	.	.
Normal	power ($\lambda = 0.5$)	No	.	.	.
Normal	power ($\lambda = 1.5$)	No	.	.	.
Normal	power ($\lambda = 2$)	No	.	.	.
Gamma	log	No	.	.	.
Gamma	inverse	No	.	.	.
Lognormal	identity	Yes	8	3815.57	3836.76
Inverse Gaussian	inverse squared	No	.	.	.
Exponential	log	No	.	.	.
Exponential	inverse	No	.	.	.

From Table 4.1, we are able to clearly see that there are two possible GLMMs to consider: the normal distribution with the identity link function and the lognormal distribution with the identity link function. These are the only two distributions to successfully reach convergence. However, we discount the normal distribution, as according to the smaller-is-better AIC, the lognormal distribution has the better fit. Further support for the choice of the lognormal distribution is the fact that it is a distribution which naturally only takes non-negative values, making it compatible with electricity consumption data which can also only take on non-negative values. Therefore, we proceed using the lognormal distribution with the identity link function as our GLMM.

Having selected our GLMM, we next seek to confirm the necessity of retaining a household-specific intercept in the model. To do so we apply the Wald test to the covariance parameters. The parameter estimates and p-values from the Wald test are given in Table 4.2.

Table 4.2: Covariance parameter estimates and Z and P-values from the Wald test

Covariance Parameter	Estimate	Standard Error	Z Value	Pr Z
Intercept	0.2607	0.01081	24.11	<0.0001
$\hat{\rho}$	0.6516	0.02783	23.41	<0.0001
$\hat{\gamma}$	0.5302	0.01649	32.15	<0.0001
Residual	0.07834	0.002650	29.57	<0.0001

Based on the p-values in Table 4.2, we see that all the covariance parameters are highly significant, having a p-values <0.0001. Hence, the necessity of retaining a subject-specific intercept and accounting for its corresponding variance component is illustrated. Another reason for retaining a random intercept in the model, regardless of its significance, is that it improves the model's prediction capabilities at an individual level.

Next, we perform residual analysis so as to assess the goodness of the underlying distribution and linearity assumptions. Figure 4.1 contains a scatter plot of the studentized conditional residuals. For the majority of cases there is an evident random scattering about zero. This supports the non-existence of any systematic pattern not accounted for by the model. Though we do also observe several cases that drift away from zero, we reason that relative to the large data these are in fact but a few. Hence, we classify these cases as outliers.

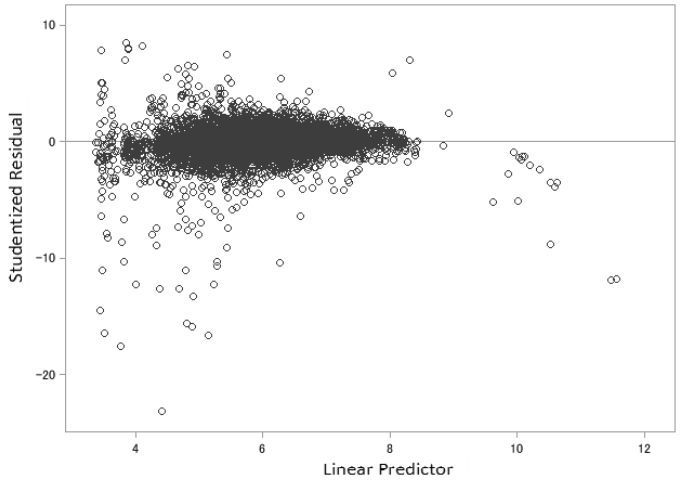

Figure 4.1: Scatter plot of conditional studentized residuals

We use the probability plot in Figure 4.2(a), to examine the goodness of the underlying distribution assumptions. It is clear from this Q-Q plot that most of the points lie approximately on the straight line, favouring the goodness of the lognormal distribution. The tails that depart from the line are due to a few outlying households in the data.

Based on the plots in Figures 4.1 and 4.2(a) and considering how large our data is (repeated measures for 1478 households), we feel comfortable in concluding that the distributional and linearity assumptions have been adequately satisfied, except for a few outliers.

The case outliers that catch our attention in the full-data Q-Q plot belong to households 202 and 1205. We now study the effect of these two outlying households. To do this we first run the full-data model then, compare it to subsequent reduced-data models. Influence is measured by the amount of change in covariance parameter estimates as well as, the observed effect on the AIC values and reduced-data Q-Q plots. Table 4.3 displays the reduced-data covariance parameter estimates, as well as their change relative to the full-data estimates, expressed as a percentage. The plots contained in Figure 4.2(a-d) show both the full-data and reduced-data Q-Q plots.

Table 4.3: AIC statistics and covariance parameter estimates: full-data and reduced-data models

Data Set	AIC	Intercept	% Rel. Error	$\hat{\rho}$	% Rel. Error	$\hat{\gamma}$	% Rel. Error	Residual	% Rel. Error
Full-data	3815.57	0.2607	-	0.6516	-	0.5302	-	0.07834	-
202 removed	3174.22	0.2595	-0.46%	0.6569	0.81%	0.5414	2.11%	0.07613	-0.03%
1205 removed	3299.46	0.2605	-0.08%	0.6523	0.11%	0.5080	-4.19%	0.07302	-0.07%
202 & 1205 removed	2636.95	0.2591	-0.61%	0.6604	1.35%	0.5197	-1.98%	0.07087	-0.10%

From Table 4.3, we see that the individual removal of households 202 and 1205, as well as the simultaneous removal of both households, result in smaller AIC values, suggesting an improved fit. However, this is what we expect to see when outliers are removed and the model re-fitted. Therefore, we do not decide whether or not to remove households 202 and 1205 from the model based only on the improved fit statistics. From Table 4.3,

we also see that upon the removal of households 202 and 1205 only negligible changes in the covariance parameter estimates occur. This tells us that we should not consider these households as influential.

Further evidence supporting that households 202 and 1205 are not influential is found when we first study then, compare the full and reduced-data Q-Q plots in Figure 4.2. In each of the Q-Q plots in Figure 4.2 (a-d), we see that the majority of points lie on the straight lines while, the tails in the plots depart from the lines. As the data that we are dealing with are large, we are satisfied with the fact that clearly most of the points in each plot adhere to the distributional assumption. The tails that depart from the lines are due to a few outlying households in the data. So long as these outliers prove not to be influential we may retain them in the data.

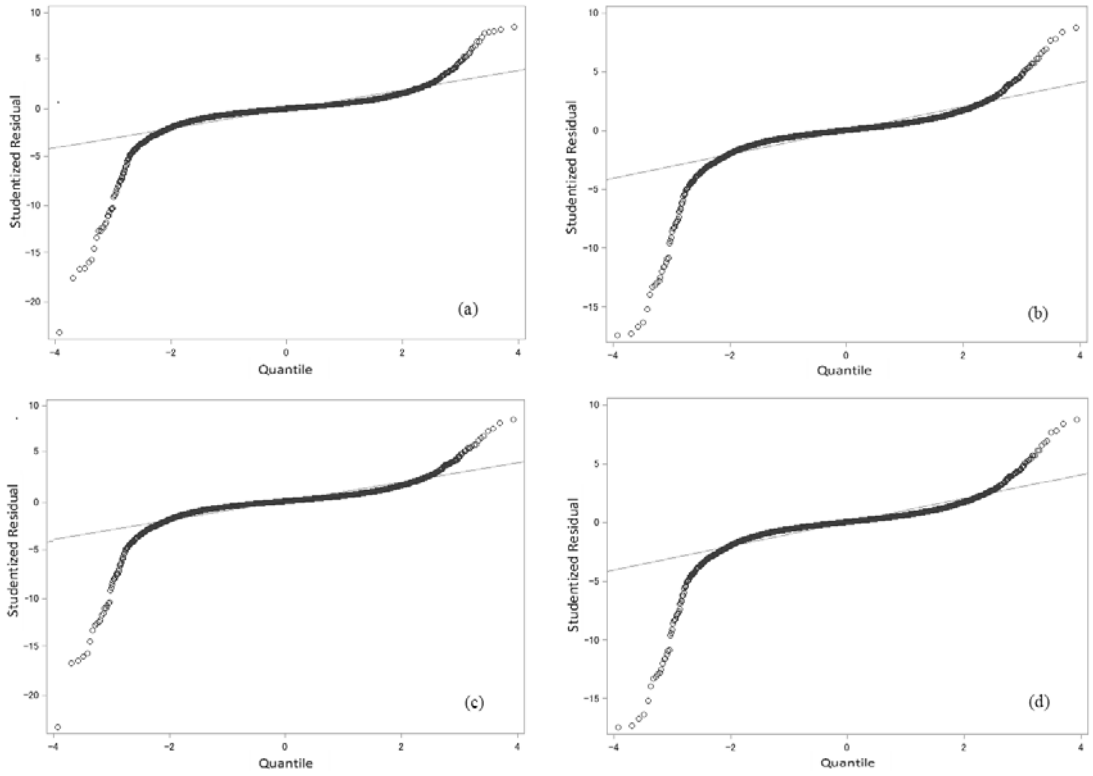

Figure 4.2: Q-Q Plots of conditional studentized residuals when: (a) No households removed (full-data model); (b) Household 202 removed; (c) Household 1205 removed; (d) Households 202 and 1205 removed

Let us now compare the observed changes in the Q-Q plots when households 202, 1205 or both are removed from the data. In Figures 4.2(b) and (d), we see that the removal of household 202 results in a plot containing more outliers than the full-data plot (see Figure 4.2(a)). This indicates to us that household 202 is better left in the model, as its absence is adversely affecting the distributional assumption. Second, we see in Figure 4.2(c), when household 1205 is removed, the resulting Q-Q plot is very similar to that of the full-data plot in Figure 4.2(a). We also see that the plots in Figure 4.2(b) and 4.2(d) are much alike. This suggests to us that we may retain household 1205 in the model, as its absence is neither improving nor adversely affecting the distributional assumption.

If we consider all these factors as a whole, we are able to conclusively confirm that households 202 and 1205 are not effecting influence in the model and are merely model outliers. When we go back to the data to examine these two households, we find a possible reason for them showing up as outliers. Both households have a few consecutive usage periods where the electricity consumption is very low, compared to the majority of consumption values in the other usage periods within the same household. Having classified 202 and 1205 as outliers and confirmed they are not influential, we choose to retain them in the model and proceed with our analysis using the full-data model.

Model Inference

We begin by looking at the type 3 test results for dwelling type and month read. From the results in Table 4.4, we see that both appear to be significant in our model, at a 5% level of significance. We reject both null hypotheses; H_0: $\beta_1 = \beta_2$ and H_0: $\beta_3 = \beta_5 = \ldots = \beta_{13}$, which tells us that at least one of the dwelling types and one of the meter reading months are significant in our model. Further examination of the individual estimates is required. In Table 4.5 we look at the individual estimates for dwelling type, followed by the individual estimates for the month of meter reading in Table 4.6.

Table 4.4: Type 3 test results for dwelling type and month read

Effect	Num DF	Den DF	F Value	Pr > F
Dwelling Type	2	13303	62.19	<0.0001
Month Read	11	13303	25.46	<0.0001

We recall that the 3 dwelling type classifications are: houses; share blocks $\leq$ 2 storeys high and; share blocks $>$ 2 storeys high. With share blocks $>$ 2 storeys as the reference category, from Table 4.5 we see that houses have a parameter estimate of 0.3821 while share blocks $\leq$ 2 storeys have a smaller parameter estimate of 0.1790. This suggests that of the 3 possible dwelling types, households falling under the classification of "house", are likely to have a higher electricity consumption, provided that the reference category remains the same.

Table 4.5: Solutions for fixed effects: dwelling type

Dwelling Type [Ref: Shareblocks > 2 storeys]	Estimate	Standard Error	DF	t Value	Pr > |t|
House	0.3821	0.03429	13303	11.15	<0.0001
Share blocks $\leq$ 2 storeys	0.1790	0.03413	13303	5.24	<0.0001

We next look at the parameter estimates for the months that meters are read in. From Table 4.6, we see that relative to December, the months May and September are highly significant (having p-values of <0.0001) in the model. We also note, that out of all the months, September has the highest positive parameter coefficient estimate of 0.06824 while, May has the highest negative coefficient of -0.05560. This suggests that September likely sees a rise in electricity consumption while, May likely sees a decrease in electricity usage, compared to the reference month, December.

However, we recall the nature of eThekwini Electricity's meter reading procedure whereby, meters are read at approximate 3 month intervals. Therefore, the month of meter reading not only refers to the electricity consumption of that month, but also includes the consumption two months prior to it. By this reckoning, our results suggest that we are likely to see a decrease in electricity usage during the months of March, April and May. While, an increase in electricity consumption is likely to be observed in the months July, August and September. It is also worth acknowledging the fact that these latter months constitute an approximate winter period in South Africa.

Lastly, we remind the reader that "month read" is only approximate, as the 3-month actual usage is distributed to months by means of scaling the 3-month value by factor of 3. eThekwini Electricity also do not read meters according to calendar months, but rather

Table 4.6: Solutions for fixed effects: month read

Month Read [Ref: Dec]	Estimate	Standard Error	DF	t Value	Pr > |t|
Jan	-0.02419	0.02009	13303	-1.20	0.2286
Feb	-0.04400	0.01255	13303	-3.51	0.0005
Mar	-0.01110	0.006453	13303	-1.72	0.0853
Apr	-0.04079	0.02059	13303	-1.98	0.0476
May	-0.05560	0.01221	13303	-4.55	<0.0001
Jun	0.01689	0.006829	13303	2.47	0.0134
Jul	-0.01358	0.02082	13303	-0.65	0.5144
Aug	0.01609	0.01246	13303	1.29	0.1967
Sep	0.06824	0.006832	13303	9.99	<0.0001
Oct	0.03689	0.02103	13303	1.75	0.0794
Nov	-0.01239	0.01325	13303	-0.93	0.3499

approximate 90 day periods. So, although we gleam interesting inferences based on the estimated coefficients for the month a meter is read in, caution should be exercised when doing so.

Following dwelling type and month of meter reading, we move on to examine the lag coefficients, looking out for any seasonal patterns that may be occurring. The coefficient estimates and p-values for both the lags and model intercept are contained in Table 4.7.

The first observation we make from Table 4.7, is that the model intercept has a p-value of <0.0001 and an estimate of 5.5410. This significance and size tells us that not only is a starting value universal to all households necessary but, also shows how a household-specific random intercept is likely to improve the model's prediction capabilities for individual households. The next observation we make from Table 4.7 is regarding the lags. We see that at a 5% level of significance only half the lags appear to be significant in the model. Lags 3, 5, 7, 9, 10 and 11 are not significant as they all have p-values >0.05 while, lags 1, 2, 4, 6, 8 and 12 are significant. We focus our attention on these, the significant lags.

Table 4.7: Solutions for fixed effects: intercept and lags

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	5.5410	0.02706	1473	204.78	<0.0001
lag_1	0.000046	0.000016	13303	2.93	0.0034
lag_2	0.000081	0.000014	13303	5.77	<0.0001
lag_3	0.000023	0.000014	13303	1.65	0.0982
lag_4	0.000201	0.000014	13303	14.69	<0.0001
lag_5	-0.00002	0.000012	13303	-1.75	0.0808
lag_6	-0.00004	0.000013	13303	-3.48	0.0005
lag_7	1.853E-6	0.000014	13303	0.14	0.8924
lag_8	0.000121	0.000014	13303	8.64	<0.0001
lag_9	-0.00002	0.000014	13303	-1.82	0.0692
lag_{10}	0.000012	0.000013	13303	0.92	0.3580
lag_{11}	2.392E-6	0.000013	13303	0.18	0.8568
lag_{12}	0.000202	0.000014	13303	14.41	<0.0001

The fact that lag_1 and lag_2 are significant tells us that the two most recent consumption value are important in modelling electricity consumption. Next, we note that of the significant lags, it's lags 4, 8 and 12 that have the highest coefficient estimates, 0.000201, 0.000121 and 0.000202, respectively. Showing to us, that these lags are contributing a lot to the model. This in itself, is very interesting to us. We recall that we expect 4 lags to occur in a 12-month period. Therefore, we may deduce that lags 4, 8 and 12 all correspond to the same point in time for previous years. That is to say, lag_4 corresponds to 12 months prior to the electricity consumption we are currently modelling. Lag_8 corresponds to 12 months prior to lag_4 or, 24 months prior to the electricity consumption we are currently modelling. This same pattern also applies to lag_{12}. For us, the fact that lags 4, 8 and 12 have large parameter estimates and are all highly significant in the model, suggests that a seasonal pattern is indeed occurring. In a similar fashion to lags 4, 8 and 12, we see that the last significant lag, lag_6, corresponds to 12 months prior to lag_2. This tells us that not only is the second most recent consumption value important but, so too is the value that is most likely to be similar to it from a seasonal perspective.

Based on our observations made from the lag coefficients, we may conclude that a seasonal effect is present in household electricity consumption. Furthermore, we learned that most recent consumption values also play a role when modelling household electricity consumption.

To conclude the application in this chapter, we revisit the covariance parameter estimates, recalling that we included a household-specific random intercept and fitted a first-order autoregressive moving average (ARMA(1,1)) structure. As we are using the full-data model, the significance and estimates for the covariance parameters can be found in Table 4.2. Following the ARMA(1,1) structure, we see that lag_1 correlation is constant, with, the corresponding covariance function being estimated by $(0.2607) + (0.07834)(0.5302)$, where the household-specific random intercept is accounted for by 0.2607. Subsequent (lag_2 and onwards) correlations decrease with the amount of time that passes between measurement occasions, that is the covariance becomes a function of the lag, and is estimated by $(0.2607) + (0.07834)(0.5302)(0.6516)^{lag}$.

Summary

This chapter focused on re-modelling the data from the cohort used in this study, by fitting a generalized linear mixed model. The reason behind this being, we wanted to fit a model that did not rely on transformations in order to satisfy the assumption of normality. We fitted various distributions using different link functions, and found the lognormal distribution to be the best fit. We used SAS PROC GLIMMIX to fit our GLMM which, modelled current household electricity consumption as a funtion of dwelling type, month of meter reading and 12 lags. A household-specific random intercept was also included and an ARMA(1,1) covariance structure was used to model $\mathbf{R}$. Unfortunately, PROC GLIMMIX did not allow for advanced diagnostics of the model. However, the residual plots showed that fewer households were outliers. We also noted that the distributional assumption was not violated.

Upon drawing model inferences, we observed some similarities and some differences compared to our findings from the LMM. We found that dwelling type was significant and that the model intercept value was particularly important. There was also evidence to suggest that electricity usage may increase during the winter months. Using the lag coefficients,

we confirmed that a clear seasonal pattern was present. We also established that the most two most recent electricity consumptions were significant when modelling electricity usage.

As similar key inferences can be drawn from both the fitted LMM and GLMM (namely regarding recent consumption and seasonal trends) and both models were shown to satisfy model assumptions and robustness, we deem both models viable. We refrain from discounting either model and resolve to carry out any further applications to both modelling approaches. The temporal correlation structure is accounted for in both the LMM and GLMM. However, the spatial correlation is not assessed and included in the modelling and analysis. This is done in the following chapter.

Chapter 5

Accounting for Spatial Variability

In the previous two chapters, our main focus was on modelling the temporal variance within each household. We did not take into account possible spatial variations between households. This chapter provides us with the opportunity to now do so. Here, we investigate if any spatial relationships exist and seek to determine whether or not households in the same neighbourhood have similar electricity consumptions. It is in fact a common practice for utilities to use a neighbourhood estimate if, the household meter reading is not done for whatever reason.

We look at relevant theory necessary for modelling spatial variations and apply the discussed methodology to the electricity data. Most discussions are focused on how spatial data analysis is carried out using linear mixed models however, similar theory applies in a generalized setting. We use the statistical software, SAS, to carry out spatial applications for both our fitted linear mixed model and generalized linear mixed model. Findings relating to both models are discussed in Section 5.3.

5.1 Introduction

Spatial variation occurs when responses are separated/dispersed over space. It is a fundamental geographic principal, that responses/observations in close proximity to one another are more alike, compared to responses/observations that are further apart from one another (Tobler, 1970). We quantify this understanding as *spatial autocorrelation* (Johnston et al., 2003).

Spatially dependent data may be classified as either isotropic or, anisotropic. Isotropy is defined as being a property of either a process or data, where spatial autocorrelation changes only with the distance between two locations (Johnston et al., 2003). Whereas, anisotropy is defined as being a property of either a process or data, when spatial autocorrelation depends on both the direction and distance between two locations (Johnston et al., 2003). For the purposes of this study we assume our data isotropic. This is because our main interest lies in establishing if or how the distance between households relates to their electricity consumption. In essence our aim is to determine if households in the same neighbourhood have similar electricity consumptions.

Spatial data can be further subdivided into three broad categories: geostatistical data, lattice data and point patterns (Cressie, 1991). Geostatistical data are collected over a continuous space, often from randomly selected sites. Lattice data pertain to equally spaced locations and point patterns involve the locations of events of interest. As we would classify our municipal data as geostatistical, we focus all further discussions based on this type of data.

There are two main approaches to dealing with spatially dependent data. Either, we may characterize the spatial variance whereby, we find the covariance parameters and describe the nature of spatial correlation or, we adjust for the presence of spatial variation (Littell et al., 2006). The emphasis of this chapter is the former, to characterize any relationship that exists between variance and physical distance between households.

Most statistical methods for dealing with spatial data were developed independently, in the field of geostatistics (Wikle & Royle, 2004). Diggle et al. (1998) credited much of the early developments in geostatistical methods to Matheron (see e.g. Matheron, 1963, 1969, 1971). A common application within the realm of spatial data analysis is that of kriging. A geostatistical term first introduced by Matheron (1963), kriging is the method of optimally predicting values or responses at unmeasured locations. Cressie (1991) detailed several variants of kriging, as well as the conditions and assumptions under which they may be used for prediction. Among those listed are ordinary kriging (Matheron, 1971), robust kriging (Hawkins & Cressie, 1984) and universal kriging (Matheron, 1969; Huijbregts & Matheron, 1971). Nevertheless, for this study our focus lies in characterizing any spatial

variation that exists between households, we have no particular interest in using our data for kriging purposes. Therefore, we refrain from discussing kriging any further. For a detailed discussion on kriging and various methods of it, refer to the listed references or Cressie (1991).

Many authors acknowledge how natural it is to view spatial statistical models from a linear mixed model perspective (see e.g. Wikle & Royle, 2004; Schabenberger & Gotway, 2005). This approach also gives the added benefit of allowing the option of employing a generalized approach (see e.g. Zhang, 2002).

As we have already laid a solid theoretical mixed model foundation for longitudinal (repeated measures) data, in both a standard and generalized form, the remainder of this chapter focuses on how we may extend these models to allow for spatial variation. For a deeper, more mathematical discussion on spatial data analysis refer to Cressie (1991). A comprehensive historical account of the numerous developments and methodologies of spatial data analysis was also given by Cressie (1991).

5.2 Spatial Data Analysis using Linear Mixed Models

Spatial data may be thought of as a special case of longitudinal type data (Wikle & Royle, 2004). Therefore, when attempting to quantify spatial autocorrelation, it is logical to consider a similar modelling approach to that of other types of longitudinal analysis.

Common textbooks, such as Schabenberger & Gotway (2005) or Littell et al. (2006), highlighted similarities between modelling temporal variances in repeated measures data, and modelling variance in spatial longitudinal data. We employ a similar approach, using linear mixed models to model spatial dependance.

First, we recall the general form of a linear mixed model (LMM):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Let us now consider the LMM in a spatial setting. We assume that the response vector, $\mathbf{y}$ is spatially indexed, such that $\mathbf{y} = [y(\mathbf{s_1}), \ldots, y(\mathbf{s_n})]$ for spatial locations $\mathbf{s_i}$, $i = 1, \ldots, n$

where it is further assumed $\mathbf{s} \in D$, if D is some d-dimensional Euclidean space; $\mathbf{X}$ is an $(n \times p)$ known design matrix of fixed numbers associated with $\boldsymbol{\beta}$; $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown constants, the fixed effects of the model; $\mathbf{Z}$ is a known $(n \times q)$ design matrix of indicator variables; $\mathbf{u}$ is a $(q \times 1)$ vector of random spatial effects; $\boldsymbol{\varepsilon}$ is a $(n \times 1)$ vector of error terms. It is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{R})$ and $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, where $\mathbf{R}$ is positive definite and $\mathbf{G}$ is a $(q \times q)$ spatial covariance matrix, with q being the number of spatial locations. We further assume that $\boldsymbol{\varepsilon}$ and $\mathbf{u}$ are independent.

We recall the importance of selecting an appropriate covariance structure in longitudinal data analysis, as this ensured adequate modelling of temporal autocorrelation. Similarly, finding an appropriate covariance function to account for spatial dependance (autocorrelation) is just as important.

Spatial correlation relates to the spatial dependance between pairs of responses at different locations. Modelling spatial correlation enables us to quantify the spatial relationship between $y(\mathbf{s}_i)$ and $y(\mathbf{s}_j)$ for $i \neq j$ with $\mathbf{s}_i, \mathbf{s}_j \in D$.

The general understanding in spatial statistics, is that spatial autocorrelation does not depend on the location of the responses $y(\mathbf{s}_i)$ and $y(\mathbf{s}_j)$ but, rather the distance between them (Cressie, 1991; Johnston et al., 2003). Extending this understanding, we observe that the covariance between responses at different sites, $cov(y(\mathbf{s}_i),\ y(\mathbf{s}_j))$, is some function, $c(.)$, of the distance, h, between the locations. There are two underlying assumptions of $cov(y(\mathbf{s}_i),\ y(\mathbf{s}_j)) = c(h)$, namely; a constant mean ($E[y(\mathbf{s}_i)] = E[y(\mathbf{s}_j)]$) and second order stationarity.

Various spatial covariance structures exist and may be fitted to model the R-side covariance. Common spatial covariance models include the exponential, Gaussian, Matérn, power and spherical structures. The method to select the best possible spatial covariance structure is much the same as that in Chapter 3 whereby, we use the information criteria.

Model estimation is carried out using methods of maximum likelihood. To select the best fitting model, we rely on information criteria. We use residual analysis and case deletion diagnostics for model checking and methods of best linear unbiased prediction if we wish

to use the model for prediction purposes.

Before fitting spatial covariance structures to a linear mixed model, it is wise to first examine the empirical semivariogram. The semivariogram measures spatial variability as a function of distance between two locations, it may be defined as one-half the variance of the difference between two observations made at different locations (Littell et al., 2006). The empirical semivariogram provides a visual depiction of the actual spatial variability of the data (Johnston et al., 2003; Kolovos, 2010). The basic elements of a semivariogram are the nugget, sill and range. Littell et al. (2006) provided a clear illustration of what an ideal semivariogram would look like, which in turn helped us to clearly define its key features.

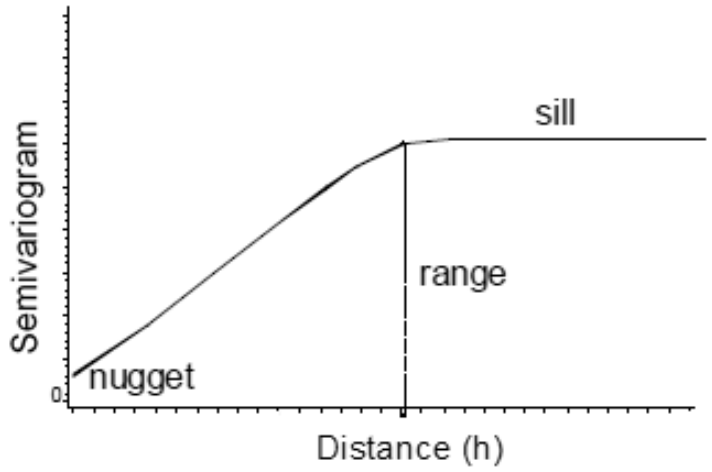

Figure 5.1: Idealized depiction of the semivariogram (Littell et al., 2006)

From Figure 5.1, it is easy for us to define the nugget as the intercept of the semivariogram. As this is the point where $h = 0$, the nugget effect is understood to show errors that are spatially independent or, the variance occurring at a particular location. The sill is the value that the semivariogram tends towards for large values of h. At large distances variables cease to be correlated (Johnston et al., 2003). Therefore, we may infer that when h is very large, the sill in fact corresponds to the variance of an observation (Littell et al., 2006). From Figure 5.1, we can also clearly see that the range can be defined as the value of h at which the semivariogram reaches the sill. Intuitively, this tells us that for all distances less than this value of h, observations are spatially correlated. While,

at distances greater than or equal to this value of h, observations are no longer spatially correlated. For more insight on the semivariogram in the context of linear mixed models, refer to Schabenberger & Gotway (2005) or Littell et al. (2006).

5.3 Application

In this application we explore the possibility of a spatial relationship existing between the electricity consumptions of different households. We are particularly interested in determining if differences between electricity consumptions of different households can be attributed to the distance between them. We recall that to model our electricity data we implemented two different modelling approaches. First, we applied a natural log transformation to the data and fitted an LMM to it. Then, we tried a generalized approach whereby, we fitted a GLMM to the observed data using a lognormal distribution and identity link function. We now seek to establish if spatial correlation is present and if so, to include it in our modelling procedure by means of fitting a spatial covariance structure to the LMM and GLMM specified in Equations 3.4 and 4.1 respectively. We remind the reader that these models were respectively specified as follows:

$$\begin{aligned} E(y^*_{ij}) &= \beta_0 + \alpha_{i0} + \beta_1 DwellingType_{House} + \beta_2 DwellingType_{Shareblock \leq 2\ storeys} + \\ &\quad \beta_3 MonthRead_{Jan} + \beta_4 MonthRead_{Feb} + \ldots + \beta_{13} MonthRead_{Nov} \\ &\quad + \beta_{14}\ y^*_{ij-1} + \beta_{15}\ y^*_{ij-2} + \ldots + \beta_{25}\ y^*_{ij-12} \end{aligned}$$

where $y^*_{ij} = ln(y_{ij})$, $y^*_{ij-1}, \ldots, y^*_{ij-12} = ln(y_{ij-1}), \ldots, ln(y_{ij-12})$; $E(y^*_{ij})$ is the expected response for the i^{th} household at time t_{ij}, $j = 1, 2, \ldots, n_i$, if n_i is the number of measurement occasions for household i; α_{i0} represents a random household-specific intercept.

$$\begin{aligned} g[E(y_{ij})] = \eta_{ij} &= \beta_0 + \alpha_{i0} + \beta_1 DwellingType_{House} + \beta_2 DwellingType_{Shareblock \leq 2\ storeys} + \\ &\quad \beta_3 MonthRead_{Jan} + \beta_4 MonthRead_{Feb} + \ldots + \beta_{13} MonthRead_{Nov} \\ &\quad + \beta_{14}\ y_{ij-1} + \beta_{15}\ y_{ij-2} + \ldots + \beta_{25}\ y_{ij-12} \end{aligned}$$

where $g(.)$ is the link function, $E(y_{ij})$ is the expected response (monthly electricity consumption) for the i^{th} household, $i = 1, \ldots, 1478$ at time $j = 1, 2, \ldots, n_i$, if n_i is the number of measurement occasions for household i; α_{i0} represents a household-specific random intercept and; η_{ij} represents the linear predictor.

To begin, we first establish whether or not there is any spatial correlation present in the data. We do this by constructing and studying the empirical semivariogram. As we made use of both transformed and observed data, we study the empirical semivariogram for evidence of spatial correlation in both data. For both data we use the GIS co-ordinates of each household as a measure of distance between properties. To begin the modelling procedure we first look at the empirical semivariogram of the transformed data.

Before we can start constructing the empirical semivariogram, we need to group locations into classes. Pairs of locations can be grouped into classes according to the common distance between them. To construct the empirical semivariogram, PROC VARIOGRAM requires that we specify both the size of the lag class (i.e. the common distance) and the maximum number of lag classes to include (SAS, 2011).

To make our specifications, we examined the pairwise distribution of our data using a variety of class numbers, then followed guidelines suggested by Journal & Huijbregts (1978). Journal & Huijbregts (1978) recommended that lag classes be specified such that each class contained a minimum of 30 location pairs and only lags up to approximately half of the extreme distance between points be considered. We elected to group our locations across 50 classes. Based on the resulting pairwise distribution, the pairs information we obtained is presented in Table 5.1.

Table 5.1: Pairs information from 50 classes

Number of Lags	51
Lag Distance	7027.22
Maximum Data Distance in Latitude	285473.00
Maximum Data Distance in Longitude	204841.00
Maximum Data Distance	351361.17

Using the information in Table 5.1, to construct the empirical semivariogram we specify a

common lag distance of 7000 and the maximum number of classes to be 25 ($max\ class = (350000/2) \div 7000$). The constructed empirical semivariogram is depicted in Figure 5.2.

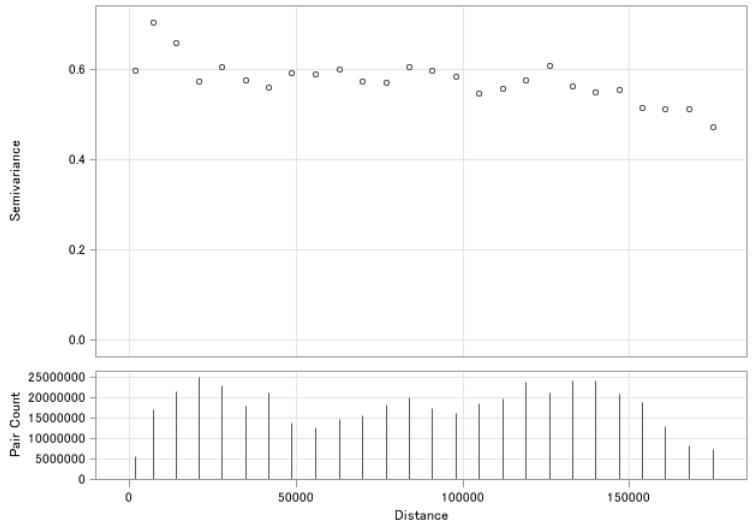

Figure 5.2: Empirical semivariogram for the natural logarithm of monthly electricity consumption

Upon studying the empirical semivariogram in Figure 5.2, we observe that a nugget effect is present. However, this is expected as among other factors, each location is measured repeatedly resulting in within-household variability. From Figure 5.2, we also see that as the distance between properties is increasing, the semivariogram remains more or less constant. This suggests that there is no spatial correlation present between properties, implying that variance is not a function of distance between dwellings.

Though the empirical semivariogram provides evidence that clearly shows no spatial correlation is present, we elect to continue with the modelling process. That is to find and fit a spatial covariance structure to the LMM given in Equation 3.4. However, before doing so we emphasize this to be a demonstrative exercise. The exercise is carried out so as to both further show the ill-suited nature of a spatial covariance structure and to demonstrate how we would proceed with the modelling process, had the empirical semivariogram indicated that spatial correlation was present.

We begin by selecting a spatial covariance to fit to the LMM. To assist with this process

we use PROC VARIOGRAM. Whereby, selected covariance structures supported by both PROC MIXED and PROC VARIOGRAM are fitted using PROC VARIOGRAM. The corresponding fit statistics are presented in Table 5.2.

Table 5.2: Fit Statistics for selected spatial covariance structures also available in PROC MIXED

Spatial Model	Exponential	Gaussian	Matérn	Power	Spherical
AIC	**285.23592**	285.23599	287.23598	285.23628	285.23595

From Table 5.2, we see that the fit statistics for the various models are all very similar, suggesting that no particular structure is better than another. We note that this agrees with our initial observations from the empirical semivariogram, by emphasizing the fact that if variance is not dependent on distance, then no spatial structure will make a significant difference. Yet according to the smaller-is-better information criteria, theoretically if we were to select the best suited structure, it would be the exponential model. However, we note that when comparing a typical semivariogram of the exponential structure (shown in Figure 5.3) to our empirical semivariogram in Figure 5.2, we clearly see that the exponential structure is inappropriate for our data.

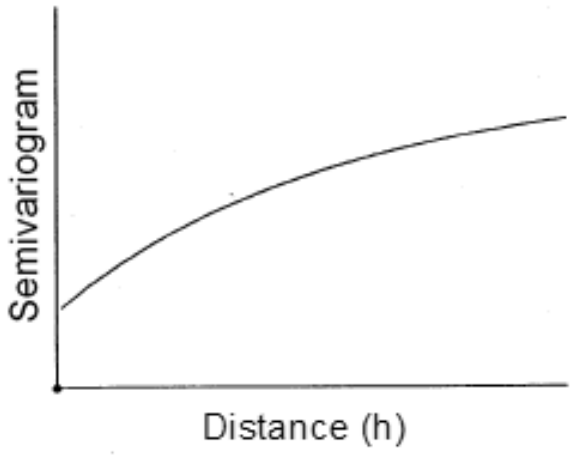

Figure 5.3: Typical semivariogram of the exponential structure (Cressie, 1991)

Nevertheless we continue with our demonstrative exercise, showing how modelling would proceed had spatial correlation been present. We now employ PROC MIXED to fit our linear mixed model, including a nugget effect and using the selected exponential type spa-

tial covariance structure to model **R**. The resulting covariance parameter estimates are displayed in Table 5.3.

Table 5.3: Covariance parameter estimates for the spatial mixed model

Covariance Parameter	Estimate	Standard Error	Z Value	Pr Z
Intercept	0.001559	0.000488	3.19	0.0014
Variance (nugget effect)	0.000867	0.000012	75.35	<0.0001
Sp(exp)	0	.	.	.
Residual	0.05353	0.000710	75.35	<0.0001

From Table 5.3, we see that a p-value for the parameter, sp(exp), was not estimable due to the final Hessian matrix not being positive definite. This again shows the ill-suited nature of a spatial covariance structure. Thus ends our demonstrative exercise, having provided further evidence to support the absence of spatial correlation and having shown how the modelling process would have proceeded.

As there is no evidence to suggest that spatial correlation is present, it is clearly of little value to fit a spatial covariance structure to model **R**. Instead, it is better to model temporal variations in the data. We did this in Chapter 3 by fitting an ARMA(1,1) structure to **R** in the LMM. However, it is important that we also account for the nugget effect that is occurring. We understand in our data, that a nugget effect is in fact referring to the variance occurring at a particular property.

Before we deal with the nugget effect, let us recall that we fitted a household-specific random intercept in the model. From this we understand that in the covariance parameters we are accounting for the variance of an individual household. We further recall that a household was defined such that it identified an individual household or dwelling, noting that some dwellings (such as those of shareblocks like flats or simplexes) shared property identifiers and GIS co-ordinates. Nevertheless, due to us randomly selecting a sample from 300 000 possible properties, it is highly improbable that we select many households that happen to share common GIS co-ordinates. Hence, we infer that for almost all the data, the variance of an individual household is one in the same as the variance at a particular location. Therefore, by including a household-specific random intercept we are in fact

already accounting for the observed nugget effect. Adding a nugget effect into the model will result in the variance being accounted for twice. Therefore, we conclude that for the transformed data, as no spatial correlation is present, it is best to use an LMM fitted with a random intercept and ARMA(1,1) covariance structure.

Having assessed spatial variability in the transformed data, we move on to explore spatial correlation in the observed data. We start by constructing the empirical semivariogram. To do this we first examine the pairwise distribution of the observed data using a variety of class numbers. This helps us determine the specifications required to construct the empirical semivariogram. Table 5.4 displays the pairs information for 100 classes.

Table 5.4: Pairs information from 100 classes

Number of Lags	101
Lag Distance	3513.61
Maximum Data Distance in Latitude	285473.00
Maximum Data Distance in Longitude	204841.00
Maximum Data Distance	351361.17

As we are dealing with GIS coordinates, and not actual units of distances (e.g. kilometers or miles), we observe the same pairs information for maximum distances as given in Table 5.1. The empirical semivariogram is plotted, specifying a common lag distance of 3000 and the maximum number of classes to be 55.

Upon examination of the empirical semivariogram in Figure 5.4, we see that a nugget effect is present and that the semivariogram remains more or less constant even as distances between two properties increase. This is sufficient evidence to suggest an absence of spatial correlation, telling us that variance is unlikely to be a function of distance. Unlike in the case of the transformed data we refrain from continuing with the modelling process as it is clearly of little value to fit a spatial covariance structure to the GLMM. Instead, it is better to model temporal variance. This was what we did in Chapter 4, where we fitted a GLMM to the observed data, and used an ARMA(1,1) structure to model $\mathbf{R}$.

However, it is important that we account for the nugget effect that's observed in the em-

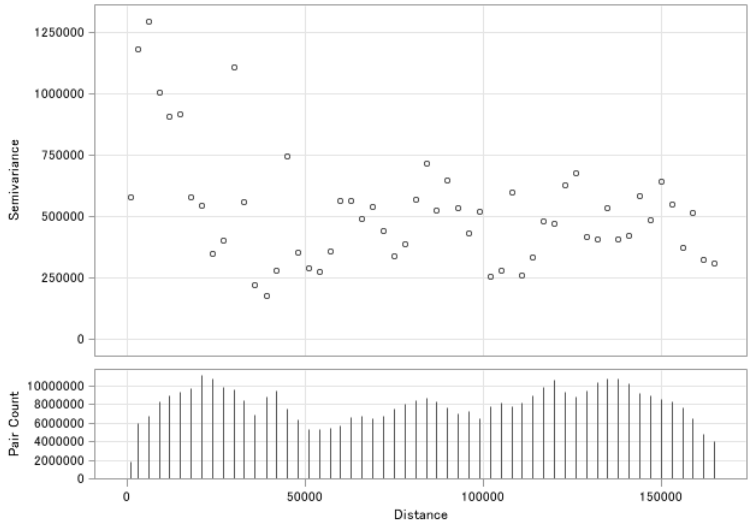

Figure 5.4: Empirical semivariogram for the observed data

pirical semivariogram. From our data, we understand that the nugget effect is the variance occurring at a particular location. We also understand, that the inclusion of a random intercept in the model accounts for the variance of a particular household. As our data was randomly selected from a large population, we expect the majority of households to have a GIS co-ordinate unique to them. Therefore, we recognize that for almost all the data, the variance of an individual household is one in the same as the variance at a particular location. From this understanding, we see that by including a household-specific random intercept, we have already accounted for the nugget effect.

The overall inference we gain from the spatial analysis of both the transformed and observed data, is that regardless of how far apart or close together households may be, the differences in their electricity consumption cannot be attributed to distance between them. A possible explanation for this might be the co-existence of high and low electricity consumers in close proximity to one another.

Summary

We observed that spatial correlation is absent in both the transformed and observed data, though both data have a nugget effect occurring. However, we also established that this effect is already accounted for in both the LMM and GLMM in the form of a household-specific random intercept. Therefore, we conclude that it is of more value to use models

that model temporal variance and include a random intercept.

Though the LMM and GLMM we have fitted so far account for temporal variation and contain random intercepts, they do not take into account that the number of days in each measurement period varies slightly. To resolve this oversight and clear any doubts the reader may have regarding the validity of the LMM and GLMM, we revisit both models and add weightings to them. This is done in the next chapter.

Chapter 6

Weighted Model Parameter Estimates

In this chapter we focus on re-estimating our LMM and GLMM. We revisit these models because, when initially estimating the model parameters, we did not account for the varying number of days in each measurement period.

Despite the sample data being carefully selected, so as to ensure approximately evenly spaced measurement occasions, a small amount of variation was inevitable. We recall that we only considered sampling households whose measurement periods were between 80 and 110 days then, approximated this to a monthly period by scaling these values by a factor of 3. Essentially, this equated to the number of days in a measurement period varying between 26 and 37 days. As we would expect a one month period to be approximately 30 days, we agree that this variation is negligible enough to still use methods for evenly spaced data. However, this small amount of variation may result in the parameter estimates no longer being efficient. A possible solution to this problem is to add weightings to the estimation procedure. Whereby, our estimates are weighted by the length of time (in days) of an approximate monthly measurement period.

We recall the ML estimate of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X_i'V_i^{-1}X_i})^{-1}(\mathbf{X_iV_i^{-1}y_i})$, where $\mathbf{V_i} = \mathbf{Z_iGZ_i'}+\mathbf{R}$. Now, if we add weightings into this estimator, we find that variance-covariance matrix $\mathbf{V_i}$ becomes $\mathbf{V_i} = \mathbf{Z_iGZ_i'}+\mathbf{W}^{-\frac{1}{2}}\mathbf{RW}^{-\frac{1}{2}}$, where $\mathbf{W}$ is a diagonal matrix containing known weightings (SAS, 2011). This new $\mathbf{V_i}$ is then used to calculate $\hat{\boldsymbol{\beta}}$.

Following this idea, we re-estimate both our LMM and GLMM, by adding weightings to the estimators. The weighting is based on the number of days in a monthly measurement period.

6.1 Weighted Estimates for the LMM

In this section we re-estimate the LMM that was fitted to the transformed data. To fit the LMM we performed a natural log transformation on the data and included the variables dwelling type, month of meter reading, and 12 lagged consumption values. The final model used an ARMA(1,1) structure to model $\mathbf{R}$, contained a random intercept and was a reduced-data model where household 46 had been removed. Now, we add weightings by number of days in a period and refit our model. According to the smaller-is-better information criterion, the weighted LMM with an AIC value of -715.0 renders a better fitting model than that of the LMM which, had an AIC value of -679.4. The improved model fit indicates that we should retain the weighted LMM. As our most important findings from the LMM were centered around the lag coefficients, we focus our attention on studying the lag coefficients of the weighted model. Weighted estimates for the lags are displayed in Table 6.1.

From Table 6.1 we see that the only lag not significant to the model is lag^*_{11}, as it has a p-value of 0.2782. Though the remaining lags are all significant at a 5% level of significance, lags 1, 2, 3, 4, 8 and 12 draw our attention as they also have the highest coefficient estimates. The fact that the coefficients for lags 1, 2, 3 and 4 are all significant and large, tells us that the most recent previous year's electricity consumption is both important and necessary when modelling electricity usage. The fact that lags 4, 8 and 12 are both large and significant suggests that a seasonal pattern is present. This is because these lags respectively correspond to 12, 24 and 36 months prior to the current electricity consumption value.

Subsequent to the lag coefficient estimates, we next look at the covariance parameter weighted estimates. We recall that to model temporal variance we fitted an ARMA(1,1) structure to $\mathbf{R}$ and to allow for better prediction for individual properties we included

Table 6.1: Weighted estimates of the lag coefficients for the transformed data

Effects	Weighted Estimate	Pr > |t|
lag_1^*	**0.3260**	<0.0001
lag_2^*	**0.1140**	<0.0001
lag_3^*	**0.1481**	<0.0001
lag_4^*	**0.1047**	<0.0001
lag_5^*	-0.02576	0.0033
lag_6^*	-0.03988	<0.0001
lag_7^*	0.03717	<0.0001
lag_8^*	**0.2123**	<0.0001
lag_9^*	-0.07631	<0.0001
lag_{10}^*	-0.02921	0.0023
lag_{11}^*	-0.01069	0.2782
lag_{12}^*	**0.1355**	<0.0001

a random household-specific intercept. The weighted covariance parameter estimates are displayed in Table 6.2.

Table 6.2: Weighted covariance parameter estimates for the transformed data

Covariance Parameter	Weighted Estimate	Pr Z
Intercept	0.000333	0.7344
$\hat{\rho}$	0.4911	<0.0001
$\hat{\gamma}$	0.2715	<0.0001
residual	1.7973	<0.0001

From Table 6.2 we see that the random intercept variance component appears not to be significant however, we retain the random intercept in the model so as improve prediction later. Following the ARMA(1,1) structure, we see that lag_1 correlation is constant with the corresponding covariance function being estimated by $(0.000333) + (1.7973)(0.2715)$, where the household-specific random intercept is accounted for by 0.000333. Subsequent (lag_2 and onwards) correlations decrease with the amount of time that passes between measurement occasions, that is the covariance becomes a function of the lag and is estimated by $(0.000333) + (1.7973)(0.2715)(0.4911)^{lag}$.

It's clear that the addition of weightings to the LMM do not greatly alter our main model findings. However, as the varying number of days in a measurement period is now accounted for, the new parameter estimates are more efficient and likely to yield better predictions. We now look to add weightings to our GLMM, where the observed data was modelled.

6.2 Weighted Estimates for the GLMM

Following the re-estimation of the LMM, we next look at how weightings may improve the GLMM estimates. We recall that to fit the GLMM we used a lognormal distribution and identity link function. We also included a household-specific random intercept and used a ARMA(1,1) covariance structure to fit the final full-data model. Now, we include weightings by number of days in a period and refit the model. According to the smaller-is-better information criterion, the weighted GLMM with an AIC value of 3781.15 renders a better fitting model than that of the LMM which, had an AIC value of 3815.57. The improved model fit indicates that we should retain the weighted GLMM.

Having established that the GLMM with weighted estimates is better, we proceed much as how we did for the LMM with weighted estimates. First, we look for any patterns in the lag estimates then, study the covariance parameter estimates. The lag coefficient estimates are given in Table 6.3.

From Table 6.3, we see that lags 1, 2, 4, 6, 8 and 12 are all significant at a 5% level of significance. However, of these lags we observe that lags 1, 2, 4, 8 and 12 have the largest coefficient estimates. The fact that lags 1 and 2 have high estimates tells us that the two most recent consumption values are important. The large coefficients of lags 4, 8 and 12 show that a seasonal effect is present, as these lags correspond respectively to 12, 24 and 36 months prior to the current electricity consumption. Following the lags, we next look at the weighted covariance parameter estimates that are displayed in Table 6.4.

It is clear from Table 6.4 that all variance components are significant at a 5% level of significance. We recall that to model temporal variance we fitted an ARMA(1,1) structure to $\mathbf{R}$ and to allow for better prediction of individual properties we included a random household-

Table 6.3: Weighted estimates of the lag coefficients for the observed data

Effects	Weighted Estimate	Pr > |t|
lag_1	**0.000044**	0.0045
lag_2	**0.000081**	<0.0001
lag_3	0.000025	0.0786
lag_4	**0.000202**	<0.0001
lag_5	-0.00002	0.0940
lag_6	-0.00004	0.0005
lag_7	1.578E-6	0.9081
lag_8	**0.000120**	<0.0001
lag_9	-0.00002	0.0717
lag_{10}	0.000012	0.3679
lag_{11}	1.841E-6	0.8892
lag_{12}	**0.000203**	<0.0001

Table 6.4: Weighted covariance parameter estimates for the observed data

Covariance Parameter	Weighted Estimate	Pr Z
Intercept	0.2607	<0.0001
$\hat{\rho}$	0.6506	<0.0001
$\hat{\gamma}$	0.5306	<0.0001
residual	2.3799	<0.0001

specific intercept. From the ARMA(1,1) structure, we see that lag_1 correlation is constant with the corresponding covariance function being estimated by $(0.2607)+(2.3799)(0.5306)$, where the household-specific random intercept is accounted for by 0.2607. Subsequent (lag_2 and onwards) correlations decrease with the amount of time that passes between measurement occasions, that is the covariance becomes a function of the lag and is estimated by $(0.2607) + (2.3799)(0.5306)(0.6506)^{lag}$.

Overall, in both the LMM and GLMM, we see that accounting for the varying number of days in each measurement period by means of a weighting, improves model fit by attaining efficient estimates. Our original model inferences remain much the same as those in Chapters 3 and 4 whereby, we observe that the most recent electricity consumption

values are important and that a seasonal pattern is present. As the weighted models are clearly better, we take them to be our final models that we use to make predictions. Predictions made using both the weighted LMM and GLMM are presented in the Section 6.3.

6.3 Using the Weighted LMM and Weighted GLMM for Prediction

To conclude the study, we illustrate the effectiveness of our fitted models while demonstrating their potential for use in prediction. We compare predictions made using the LMM, GLMM and customary eThekwini method, to actual monthly values that we have in the data. To carry out the comparison, we randomly select some households from across the 3 dwelling types then, we remove the most recent electricity consumption value that each selected household has from the data. Thus, treating this value as that for which we need to predict. However, we in fact now have an observed value that we may compare our predictions to.

To make predictions from our LMM and GLMM, we use the weighted estimates of $\beta_0, \beta_1, \ldots, \beta_{25}$ as well as, that of the household-specific estimate of α_{i0}. To use the eThekwini method of prediction, we first need to recall their methodology. eThekwini Electricity estimate consumption by means of a cumulative total of weighted actual electricity usage whereby, the most recent consumptions carry the highest weights while weightings of older consumptions decrease exponentially. Therefore, to make predictions using eThekwini's method, we make use of the following formula:

$$E(y_{ij})_{eThekwini} = \sum_{k=1}^{n-1} \left(\frac{1}{2}\right)^k lag_k + 2\left(\frac{1}{2}\right)^n lag_n$$

where n is the total number of measurement periods that each household has; $E(y_{ij})$ represents the electricity usage that is being predicted for the i^{th} household; and previous electricity consumption values for the i^{th} household are denoted using lags whereby, $lag_k = y_{ij-k}$ for $k = 1, \ldots, n-1$ and $lag_n = y_{ij-n}$. However, to distribute the prediction to a monthly estimate, we divide our answer by 3.

Predictions made using the weighted LMM, weighted GLMM and the eThekwini method

are displayed in Table 6.5 as well as, the actual observed monthly value. The relative errors of each method for each household are also given in Table 6.5, where they are expressed as a percentage. This helps us to see which method performs best.

Table 6.5: Comparison of actual values with various predictions

		Weighted LMM		Weighted GLMM		eThekwini	
Household	Actual Value	Prediction	Error	Prediction	Error	Prediction	Error
2	467.33	**479.24**	2.55%	437.06	-6.48%	491.00	5.06%
148	565.33	516.22	-8.69%	**532.48**	-5.81%	526.28	-6.91%
298	710.00	**718.0**	1.13%	743.07	4.66%	808.36	13.85%
371	551.67	505.13	-8.44%	**531.01**	-3.74%	518.85	-5.95%
513	629.33	625.91	-0.54%	**632.37**	0.48%	667.66	6.09%
660	429.33	392.91	-8.48%	**430.64**	0.31%	411.86	-4.07%
686	601.00	614.10	2.18%	**608.62**	1.27%	637.28	6.04%
812	446.33	348.44	-21.93%	**466.51**	4.52%	342.21	-23.33%
908	203.00	191.71	-5.56%	**193.80**	-4.53%	186.19	-8.28%
1128	382.00	**372.13**	-2.58%	370.64	-2.97%	361.75	-5.30%
1182	706.00	740.55	4.89%	**678.34**	-3.92%	761.85	7.91%
1262	742.00	831.31	12.04%	**723.62**	-2.48%	886.08	19.42%
1378	69.33	**69.25**	-0.12%	74.82	7.92%	63.30	-8.70%
1430	264.67	228.30	-13.74%	**249.88**	-5.59%	226.56	-14.40%

For each household in Table 6.5, according to the smallest absolute value of the relative errors, the predictions closest to the actual monthly consumptions are indicated in bold. From this we can see that the mixed models consistently provide better predictions than those made using the customary eThekwini method. In addition to this, we also see that the weighted GLMM appears to be outperforming the weighted LMM, having the most number of predictions closest to the actual value.

To further illustrate which predictions are closest to the actual monthly values, we construct a spider graph. The spider graph is depicted in Figure 6.1 and shows the absolute value of the relative errors of each method for each household.

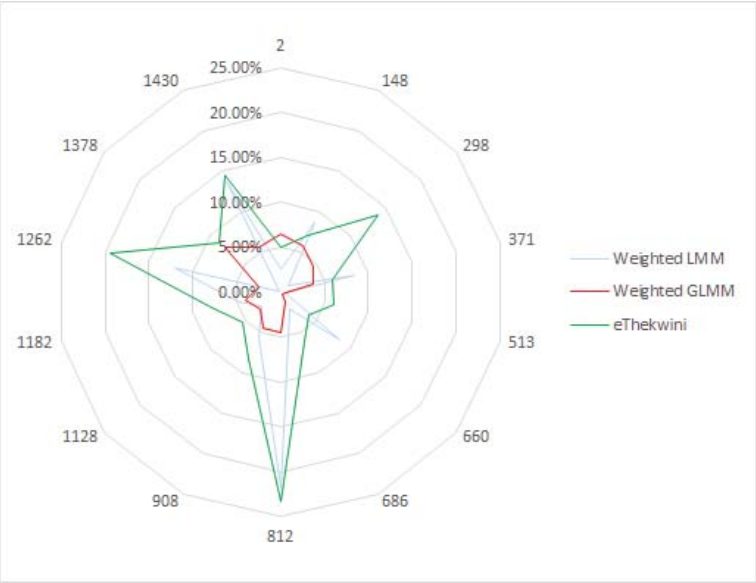

Figure 6.1: Spider graph showing |relative errors| of each prediction method

In Figure 6.1, we see the absolute value of the errors are represented by concentric circles and that each household is shown along the outside of the outer most circle. The innermost circle represents the smallest errors, while subsequent outer circles show larger errors. In the spider graph, we clearly see that the weighted GLMM is the method that is most closely centered in and around the innermost circle. This shows us that the weighted GLMM is the best performing model, as across all the households it has the most number of smallest errors. Meaning, that the weighted GLMM has predictions that are very close to the actual monthly consumption values, for most of the households. After the weighted GLMM, from Figure 6.1 we see that the next best performing model is the weighted LMM. While, the method that results in the highest errors between predicted and observed values is that of eThekwini.

Based on our observations from Table 6.5 and Figure 6.1, we conclude that models that account for seasonal patterns in monthly household electricity consumption are clearly preferable when predicting future electricity consumption at a household level. Having demonstrated the effectiveness of the weighted LMM and GLMM as well as their predicative capabilities, we proceed to the conclusion of our study in the next chapter.

Chapter 7

Conclusion

The aim of this research was to model household electricity consumption within the eThekwini municipal area. The intention being that later, the developed model/s be used to predict future electricity consumption values of individual households. As part of the modelling process, we were interested in determining if or how lagged consumption values affect current electricity consumption values of a household. Of particular interest to us, was establishing whether or not a seasonal pattern existed within households electricity consumption. In addition to seasonality, we also wanted to find out if patterns of spatial variability between households could be identified.

Data was obtained from eThekwini Electricity, a subsidiary of eThekwini Municipality, and official supplier of electricity to the municipality. The data consisted of repeated electricity meter readings, reading dates and electricity consumption values for an approximate five year period. Additional household information contained in the data included GIS co-ordinates, both the suburb and city the household was located in as well as, classifications of dwelling type. There were 3 categories of dwelling types; houses, shareblocks $\leq$ 2 storeys high and shareblocks $>$ 2 storeys high.

For the purposes of this project we worked with only a sample of the whole data set. After data cleaning and sorting, a random sample of 1500 households was selected whereby, 500 households were randomly selected from each of the 3 dwelling types. The sampling frame consisted of households that had a city classification of “Durban” and whose measurement periods were between 80 and 110 days.

Since we wanted to model 'typical', regular residential electricity consumers within eThekwini, the selected sample was checked for anomalies before beginning the modelling process. A total of 22 households were removed from the sample, as they were found to have some measurement periods where their electricity consumption was zero, which was deemed atypical for regular consumers. The final sample consisted of 1478 households.

As we were required to model repeated measures data, it was a natural decision to employ linear mixed models (LMMs). First, a standard LMM was fitted to the observed data, modelling monthly electricity consumption for each household and accounting for temporal variation. We found that lags that extended further than 3 years prior to the current value had little effect when modelling current electricity consumption. Based on this finding and the principal of parsimony, we limited our models to include 12 lagged values. This was equivalent to including all the lags that a household had, up to and including 3 years prior to the current value. We modelled current household consumption as a function of dwelling type, month of meter reading, 12 lags and a household-specific intercept. The inclusion of 12 lagged values enabled us to later assess the presence of seasonality while, the inclusion of a random intercept enhanced the prediction capabilities of the model.

The best LMM was found to have an ARMA(1,1) covariance structure. However, we observed that the assumption of normality was not entirely satisfied. This led us to implement two different modelling approaches. For the first approach we used the method of transformations. We applied various power transformations to the data, and witnessed an improvement in the model fit. We found that the natural log transformation was the best convenient transformation. Upon residual analysis of our model, we observed 3 outlying households, households 46, 202 and 1205. With the use of case deletion diagnostics we classified household 46 as influential and removed it from the data. Households 202 and 1205 were confirmed to be non-influential outliers within the model, hence we retained them in the data.

Several inferences were gleamed from our final reduced-data LMM. We observed that dwelling type did not appear to play a significant role in predicting electricity consumption while, the months that meters were read in, did. We also saw that the most recent

previous year's consumption played an important role in the modelling process. Finally, upon closer examination of the coefficients of the lagged responses, we observed that lags 4, 8 and 12 had large estimates. We took this as evidence to support the presence of a seasonal pattern, as lags 4, 8 and 12 each respectively correspond to 12, 24 and 36 months prior to the current electricity consumption.

Though we saw that the transformation approach improved the model fit, the normality plot did not clearly favour the normality of the transformed data. This prompted us to implement our second modelling approach, that of generalized linear mixed models (GLMMs). The generalized modelling approach allowed us to directly search for the type of distribution that could best satisfy normality assumptions. We again modelled current household electricity consumption as a function of dwelling type, month of meter reading, 12 lags and a household-specific intercept. The best GLMM was found to have a lognormal distribution, used the identity link function and had an ARMA(1,1) covariance structure modelling the temporal variance. Residual analysis of the GLMM identified two outliers, households 202 and 1205. However, case deletion diagnostics proved them not influential therefore, the full-data model was used to fit the GLMM.

Inferences drawn from the GLMM were in some regards similar to those of the LMM, particularly the finding of a seasonal pattern. In the GLMM we saw that dwelling type, month of meter reading and the two most recent consumption values were important in the model. Evidence of a seasonal effect was found upon observing the large coefficient estimates of lags 4, 8 and 12.

Following the use of standard and generalized approaches to model temporal variations in monthly electricity consumption, we investigated the possibility of spatial variability between households. However, examination of the empirical semivariograms for both the observed and transformed data revealed that, variances in electricity consumptions between households could not be attributed to physical distance between them. Consequently, we refrained from using any mixed models fitted with spatial covariance structures in further applications. The empirical semivariograms also showed that a nugget effect was present. However, we found that we had already accounted for this effect in both the LMM and GLMM, by means of a random intercept.

To complete the modelling process we revisited our LMM and GLMM, and added weightings to the parameter estimates. We did this because, we realized that our models had not taken into account the varying number of days in the measurement periods. Weights were based on the length (in days) of the electricity consumption period. We found that weighted estimates for both the LMM and GLMM improved model fit, without drastically changing the initial model inferences.

The study was concluded by illustrating the effectiveness of the weighted LMM and weighted GLMM, as well as showcasing their potential for use in prediction. Several households were randomly selected and their most recent electricity consumption value removed from the data. This provided us with actual observed values from which to compare the predicted values. Predictions were made using the weighted LMM, weighted GLMM and the customary eThekwini method. When comparisons between the predicted and observed values were drawn, the effectiveness of the models developed in this study became evident. Overall, predictions made using the weighted GLMM were closest in value to the observed values. The next best performing model was the weighted LMM while, the method of eThekwini resulted in the highest errors of predicted versus observed values. This led us to conclude that models that allow for seasonal patterns within monthly household electricity consumption are preferable when predicting future consumption at a household level.

Though the results of this study have applications in prediction, the main objective was to find a way to model electricity usage at a household level, according to ideal measurement circumstances. Further studies are required, to find models better suited to prediction. This includes developing methods that can adapt to less than ideal measurement circumstances, such as exceedingly long or, unevenly spaced usage periods.

Other approaches to improving prediction models, could include finding ways to incorporate additional household information. One possibility that warrants further investigation, is the linkage of information from StatsSA household surveys to existing household information, with the GIS co-ordinates serving as the linking factor.

Chapter 8

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*, 16, 125–127.

Athukorala, P. & Wilson, C. (2010). Estimating short and long-term residential demand for electricity: New evidence from Sri Lanka. *Energy Economics*, 32, S34–S40.

Beckman, R. J., Nachtsheim, J. C., & Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, 413–426.

Belsley, D., Kuh, E., & Welsch, R. (1980). *Regression Diagnostics; Identifying Influential Data and Sources of Collinearity.* New York: John Wiley & Sons.

Box, G. & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26(2), 211–252.

Box, G. & Cox, D. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77(377), 209–210.

Breslow, N. & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9–25.

Breslow, N. & Lin, X. (1995). Bias correction in generalized linear mixed models with a single componenet of dispersion. *Biometrika*, 82(1), 81–91.

Burnham, K. P. & Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag.

Chikobvu, D. & Sigauke, C. (2013). Modelling influence of temperature on daily peak electricty demand in South Africa. *Journal of Energy in Southern Africa*, 24(4), 63–70.

Christensen, R., Pearson, L., & Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34, 38–45.

Cook, R. (1977). Detection of influential observations in linear regression. *Technometrics*, 34, 38–45.

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New Jersey: John Wiley and Sons.

Der, G. & Everitt, B. S. (2006). *Statistical Analysis of Medical Data using SAS*.

Dergiades, T. & Tsoulfidis, L. (2008). Estimating residential demand for electricity in the United States, 1965-2006. *Energy Economics*, 30, 2722–2730.

Diggle, P., Tawn, J., & Moyeed, R. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society, Series C*, 47, 299–350.

Diggle, P. J. (1988). An approach to the analysis of repeated measures. *Biometrics*, 44, 959–971.

eThekwini Electricity (2013). 2011/2012 Annual Report. [www.durban.gov.za].

eThekwini Municipality (2013). 2013/2014 eThekwini Electricty Tariffs. [www.durban.gov.za].

Firth, S., Lomas, K., Wright, A., & Wall, R. (2008). Identifying trends in the use of domestic appliances from household electricity consumption measurements. *Energy and Buildings*, 40, 926–936.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longituinal Data Analysis*, chapter 1. Advances in longitudinal data analysis: An historical perspective, (pp. 3–27). Chapman and Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied Longitudinal Analysis*.

Games, P. (1984). Data transformations, power and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin*, 95, 345–347.

Guerin, L. & Stoup, W. (2000). A simulation study to evaluate proc mixed analysis of repeated measures data. In *Proceedings of the Twelfth Annual Conference on Applied Statistics in Agriculture.* Manhatten: Kansas State University.

Gurka, M., Edwards, L., Muller, K., & Kupper, L. (2006). Extending the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society. Series A.*, 169(2), 273–288.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320–338.

Hawkins, D. & Cressie, N. (1984). Robust kriging - a proposal. *Journal of the International Association for Mathematical Geology*, 16, 3–18.

Holtedahl, P. & Joutz, F. (2004). Residential electricity demand in Taiwan. *Energy Economics*, 26, 201–224.

Huijbregts, C. & Matheron, G. (1971). Universal kriging (an optimal method for estimating and contouring in trend surface analysis). In *Proceedings of Ninth International Symposium on Techniques for Decision-Making in the Mineral Industry*.

Inglesi, R. (2010). Aggregate electricity demand in South Africa: Conditional forecasts to 2030. *Applied Energy*, 87, 197–204.

Inglesi-Lotz, R. & Blignaut, J. N. (2011). South Africa's electricity consumption: A sectoral decomposition analysis. *Applied Energy*, 88, 4779–4784.

Jenrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805–820.

Johnston, K., Ver Hoef, J., Krivoruchko, K., & Lucas, N. (2003). *ArcGIS 9: Using ArcGIS Geostatistical Analyst*. Redlands, California: Enviromental Systems Research Institute (ESRI).

Journal, A. & Huijbregts, C. (1978). *Mining Geostatistics*. New York: Academic Press.

Kachman, S. (1998). An introduction to generalized linear mixed models. Department of Biometry, University of Nebraska-Lincoln.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measures. *Communications in Statistics: Simulation and Computation*, 27(3), 591–604.

Kiernan, K., Tao, J., & Gibbs, P. (2012). *SAS Global Forum 2012 Paper 332-2012 : Tips and Strategies for Mixed Modeling with SAS/STAT Procedures.* SAS Institute Inc.

Kolovos, A. (2010). *SAS Global Forum 2010 Paper 337-2010 : Everything in its Place: Efficient Geostatistical Analysis with* SAS/STAT *Spatial Procedures*. Cary, NC: SAS Institute Inc.

Laird, N. M., Lange, N., & Stram, D. O. (1987). Maximum likelihood computations with repeated measures: Application of the em algorithm. *Journal of the American Statistical Association*, 82, 97–105.

Laird, N. M. & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963–974.

Lawrance, A. (1990). *Directions in Robust Statistics and Diagnostics*, chapter Local and Deletion diagnostics, (pp. 141–157). Springer.

Lesaffre, E. & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54, 570–582.

Lindstrom, M. & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of the American Statistical Association*, 83, 1014–1022.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Institute Inc, second edition.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19, 17931819.

Manning, W. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics*, 17, 283–295.

Marvuglia, A. & Messineo, A. (2012). Using recurrent artificial neural networks to forecast household electricty consumtion. *Energy Procedia*, 14, 45–55.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58, 1246–1266.

Matheron, G. (1969). Le krigeage universel. *Cashiers du Centre de Morphologie Mathematique*, (1).

Matheron, G. (1971). The theory of regionalized variables and its applications. *Cashiers du Centre de Morphologie Mathematique*, (5).

McCulloch, C., Searle, S., & Neuhaus, J. (2008). *Generalized, Linear and Mixed Models.*

McCulloch, P. & Nelder, J. (1989). *Generalized Linear Models.* London: Chapman and Hall, second edition.

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis.* John Wiley & Sons, fifth edition.

Moser, E. & Macchiavelli, R. (2002). Model selection techniques for repeated measures covariance structures. In *Proceedings of Fourteenth Annual Kansas State University Conference on Applied Statistics in Agriculture* (pp. 17–31).

Narayan, P. & Smyth, R. (2005). The residential demand for electricity in Australia: an application of the bounds testing approach to cointegration. *Energy Policy*, 33, 467–474.

Nelder, J. & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, 135, 370–384.

Pan, J. & Thompson, R. (2003). Gauss-hermite quadrature approximation for estimation in generalized linear mixed models. *Computational Statistics*, 18, 57–78.

Pan, J. & Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics and Data Analysis*, 51, 5765–5775.

Patterson, H. & Thompson, R. (1971). Recovery of inter block information when block sizes are unequal. *Biometrika*, 58, 545–554.

Pearson, E., D'Agostino, R., & Bowman, K. (1977). Tests for the departure from normality: Comparison of powers. *Biometrika*, 64, 231–246.

Pouris, A. (1987). The price elasticity of electricity demand in South Africa. *Applied Economics*, 19, 1269–1277.

Raudenbush, S., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9, 141157.

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6, 15–51.

SAS (2011). *SAS/STAT 9.3 User's Guide*. SAS Institute Inc, Cary, NC: SAS Institute Inc.

Schabenberger, O. (2004). *SUGI 29 Paper 189-29 : Mixed Model Influence Diagnostics*. SAS Institute Inc.

Schabenberger, O. (2005). *SUGI 30 Paper 196-30 : Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*. SAS Institute Inc.

Schabenberger, O. & Gotway, C. (2005). *Statistical Methods for Spatial Data Analysis*. CRC Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992 and 2006). *Variance Components*. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc.

Shapiro, S. & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Sigauke, C. & Chikobvu, D. (2011). Prediction of daily peak electricity demand in South Africa using volatility forecasting models. *Energy Economics*, 33, 882–888.

StatsSA (2012). Pretoria: Statistics South Africa Census Publication. [Accessed through http://www.statssa.gov.za].

Tabachnick, B. & Fidell, L. (2013). *Using Multivariate Statistics*. Pearson, sixth edition.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2), 234–240.

Tukey, J. (1957). The comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28, 602–632.

Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*.

West, B. T., Welch, K. B., & Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall / CRC.

Wikle, C. & Royle, J. (2004). Spatial statistical modeling in biology. In *Encyclopedia of Life Support Systems (EOLSS)*. Oxford, UK: Developed under the Auspices of UNESCO, EOLSS Publishers. [http://www.eolss.net].

Wolfinger, R. D. (1996). Heterogeneous variance covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(2), 205–230.

Yohanis, Y., Mondol, J., Wright, A., & Norton, B. (2008). Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use. *Energy and Buildings*, 40, 1053–1059.

Zewotir, T. (2007). Infinitesimal model perturbation influence in the linear mixed model. *South African Statistical Journal*, 41(2), 105–126.

Zewotir, T. & Galpin, J. (2004). The behaviour of normality under non-normality for mixed models. *South African Statistical Journal*, 38, 115–138.

Zewotir, T. & Galpin, J. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science*, 3, 153–177.

Zewotir, T. & Galpin, J. (2007). A unified approach on residuals, leverages and outliers in the linear mixed model. *Test*, 16, 58–75.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58, 129–136.

Ziramba, E. (2008). The demand for residential electricity in South Africa. *Energy Policy*, 36, 3460–3466.

Appendix A

SAS Codes

The SAS codes provided in this appendix are for the final LMM fitted in Chapter 3, the final GLMM fitted in Chapter 4, both the pairs information and the semivariograms constructed in Chapter 5, and the final weighted models of Chapter 6.

A.1 Coding for the LMM

```
proc mixed data=durban covtest ic;
   class id month_read dwelling_type;
   model approx_monthly_avg = dwelling_type month_read lag1-lag12 / s residual;
   random int / sub=id s;
   repeated / sub=id type=arma(1,1) R Rcorr;
run;
```

A.2 Coding for the GLMM

```
proc glimmix data=durban plots=studentpanel(unpack);
   class id month_read dwelling_type;
   model approx_monthly_avg = dwelling_type month_read lag1-lag12
                              / s dist=lognormal link=identity;
   random int / sub=id s;
   random_residual_ / sub=id type=arma(1,1);
   covtest / wald;
run;
```

A.3 Coding for the pairs information and empirical semivariogram: Transformed data

```
proc variogram data=durban plots=pairs(mid);
   compute novariogram nhc=50;
   coordinates xc=latitude yc=longitude;
   var approx_monthly_avg;
run;

proc variogram data=durban plots(only)=semivar;
   compute lagd=7000 maxlag=25;
   coordinates xc=latitude yc=longitude;
   model form=auto(mlist=(exp,gau,mat,sph,pow) nest=1 to 2);
   var approx_monthly_avg;
run;
```

A.4 Coding for the pairs information and empirical semivariogram: Observed data

```
proc variogram data=durban plots=pairs(mid);
   compute novariogram nhc=100;
   coordinates xc=latitude yc=longitude;
   var approx_monthly_avg;
run;

proc variogram data=durban plots(only)=semivar;
   compute lagd=3000 maxlag=55;
   coordinates xc=latitude yc=longitude;
   var approx_monthly_avg;
run;
```

A.5 Coding for the Weighted LMM

```
proc mixed data=durban covtest ic;
   class id month_read dwelling_type;
   model approx_monthly_avg = dwelling_type month_read lag1-lag12 / s residual;
   random int / sub=id s;
   repeated / sub=id type=arma(1,1) R Rcorr;
   weight monthly_weighting;
run;
```

A.6 Coding for the Weighted GLMM

```
proc glimmix data=durban plots=studentpanel(unpack);
   class id month_read dwelling_type;
   model approx_monthly_avg = dwelling_type month_read lag1-lag12
                              / s dist=lognormal link=identity;
   random int / sub=id s;
   random_residual_ / sub=id type=arma(1,1);
   covtest / wald;
   weight monthly_weighting;
run;
```