DEPARTMENT OF STATISTICAL SCIENCES
UNIVERSITY OF CAPE TOWN

Energy Modelling

Determining the key factors that differentiate household electricity consumption - A South African study.

Mahlagaume Dinkwanyane

10/25/2010

.

UNIVERSITY OF CAPE TOWN

STATISTICS HONOURS THESIS

Plagiarism Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.

2. Each contribution to, and quotation in, this project from the work(s) of other people has been attributed, and has been cited and referenced.

3. This project is my own work.

4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

Name	Mahlagaume Dinkwanyane
Student number	DNKMAH001
Date	25 October 2010
Signature	_____

ACKNOWLEDGEMENTS

This project topic was proposed by Mr Stephen Davis (Energy Research Centre, University of Cape Town).

I would like to thank my supervisors:

Associate Professor Sugnet Lubbe (Department of Statistical Sciences, University of Cape Town):

> This project would have not been possible without your dedication, patience and guidance.
>
> Thank you.

Mr Stephen Davis and Ms Alison Hughes (Energy Research Centre, University of Cape Town):

> Thank you for the wonderful topic, inputs and the time you put into this project, I hope this report has done it justice.

Lastly, Ms ‘Mamahloko Senatla

> Thank you for helping with the data.

ABSTRACT

This report describes the results obtained from an investigation focused on determining household features that differentiate the amount of electricity consumed within a household. In analysing the data, Classification and Regression Trees (CART) and Generalised Linear Models (GLM) were used. This report also compares the competence of the CART method and the variable selection in GLM in identifying the key variables

The household data used was originally obtained from a survey and from monitoring electric meters per household. The data used had 131 independent variables and only 179 households. CART and GLM methods both showed that the as number of appliances, time length since electrification of household and the number of people per household increase so does the amount of electricity. Households that use of coal for cooking used less electricity. Households from some residential areas used less due to cultural differences and the average socio-economic status in the respective residential areas.

With the structure of the data, CART was more competent in identifying the household features that differentiate household electricity consumption.

Table of Contents

List of illustrations

List of tables

List of acronyms and abbreviations

AIC	Akaike Information Criteria
Amps	Ampere
CART	Classification and Regression Trees
DSM	Demand Side Management
ESKOM	Electricity Supply Commission
GDP	Gross Domestic Product
GLM	Generalised Linear Model
KOE	Kilogram of Oil Equivalent
KWh	Kilowatt hour
NRS	National Regulatory Services
V	Voltage

List of symbols

n	Sample size
n_{ij}	Sample size in the ij^{th} node
k	Number of variables measures per subject
f	Decision rule
O	Outcome set
$\mathbb{R}$	Set of real numbers
X	Design matrix
$\boldsymbol{x_i}$	Covariate vector
x_{ij}	Observation of the j^{th} variable from the i^{th} subject
h_{ij}	Impurity level in j- side node of i^{th} split
w_{ij}	Weighting of ij^{th} node

1. INTRODUCTION

This report describes the results obtained from an investigation focused on determining features that influence amount of electricity consumed within a household. In analysing the data, Classification and Regression Trees (CART) and Generalised Linear Models (GLM) were used. This report also compares the competence of the CART method and the Variable Selection in GLMs in identifying the key variables.

BACKGROUND

South Africa has been characterised as a high-energy intensity economy. This is a quantitative way of claiming that the economy is highly inefficient in converting the energy it uses into GDP or rather energy consumption relative to GDP is too high.

This situation is largely accredited to the energy intensive mining and industrial activities on which the South African economy heavily relies. Even so, all the sectors of the consumer market - households, enterprises and public institutions - are jointly responsible for the high-energy intensity. In this respect, the Department of Energy and ESKOM have embarked on projects and initiatives that focus on reducing the demand for energy and "promoting energy efficiency" (1) in the South African economy.

This study is dedicated to the household sector. Determining and understanding the elements (within households) that influence the amount of energy consumed would give policy makers an idea of what should be targeted when aiming to control the demand for energy.

The objectives of this study are:

- to determine and describe the socio-economic features of a household which are associated with the amount of energy consumed by the household.
- to give a short literature review on the statistical methods used in determining these major features.
- to interpret the results obtained from the analysis by both methods, and
- to compare the competence of the statistical methods used.

The data analysed in this study was obtained from the Energy Research Centre (ERC) of the University of Cape Town (UCT). It was originally obtained from a household energy consumption survey facilitated by the National Regulatory Services (NRS) in conjunction with ESKOM, starting from 1996 to 2006 called the National Load Research programme (NLR). This study focuses on data ranging from December 2004 to February 2006.

PLAN OF DEVELOPMENT

This report begins with a background to the South African energy profile; followed by the methodology which looks at data collection method, description of data (i.e. variables and subjects). It then gives an overview of the statistical methods used: a literature review about CART and a brief overview of variable selection in GLM.

Results from data exploration and analyses are then presented, interpreted and compared between using CART and GLM. Finally, conclusions are made about the results and competence of CART and GLM in reaching the objectives of this study.

2. BACKGROUND

2.1.OVERVIEW

Energy facilitates human activity to a great extent. Most, if not all, activities within households, workplaces, civil institutions (hospitals, schools and prisons) and public infrastructures (roads, dams, power plants and water networks) either rely directly on energy or use services that rely on energy.

The terms energy and electricity are often used interchangeably but they are not synonymous; energy is the umbrella term which incorporates electricity. Two fundamental forms of energy are distinguished, that is "primary energy" and electricity (2). Primary energy covers all the natural resources used to produce electricity; these include substances such as coal, petroleum, natural gases, water and wood. Although these are used to produce electricity, they can be used for several activities before being transformed into electricity. Wood, coal and paraffin are examples of primary energy sources as they can be used directly as opposed to having to be transformed into electricity.

2.2.EFFECTS OF SOURCING, GENERATING AND USING ENERGY

2.2.1. Environmental

Atmospheric and noise pollution, global warming, ozone layer depletion and ecological distortions are some of the environmental repercussions of generating electricity. Recently, using sources that are more environmentally friendly or controlling the amount of pollution posed by current sources is of huge interest.

2.2.2. Socio-Economic

Environmental pollution from energy-manufacturing activities poses health hazards. For example, respiratory diseases from atmospheric pollution, hearing complications from noise pollution, skin diseases induced by overheating due to global warming.

Acid rain and mining activities lead to deteriorations of buildings and general infrastructure. Ecological distortions may have negative economic impacts. Economies directly dependent on aquatic life may be seriously impacted by hydro-electricity infrastructure.

2.3.THE IMPORTANCE OF ENERGY MODELLING

Energy Modelling is a tool that is used to understand factors that influence demand for energy. With an understanding of these factors, Demand Side Management (DSM) can know what to strategise around or what to target, in order to control demand. Why does demand need to be controlled? Perhaps demand is exceeding supply, resulting in issues like load-shedding; as an aid to lower the extent of environmental repercussions from producing energy; or for future planning.

2.3.1. The Current South African Situation

South Africa has been characterised as a "high-energy intensity" economy. Energy intensity of an economy is measured as the ratio of the total energy consumed and the gross domestic product (GDP) of the economy. A high-energy intensity indicates that there is a large amount of energy involved in producing a rand worth of goods and services. This is quantitative way of claiming that the economy is highly inefficient in converting the energy it uses into GDP.

Figure 1 shows the global countries grouped into 6 classes depending on their respective energy intensities. The unit of measurement of energy consumed per country was "kilograms of oil equivalent" (KOE) which is a standardised measure of energy obtainable from a matter expressed in reference to the amount of oil that would give an equivalent amount of energy (3). The countries shaded with the darkest green are those with the highest level of energy intensity, and the level of energy intensity decreases as the shading gets lighter. Thus countries shaded with lightest blue-green are those with the lowest records of energy intensity. It can be observed that South Africa falls in the second darkest green class (out of five classes), which implies that South Africa has an energy intensity that is above average.

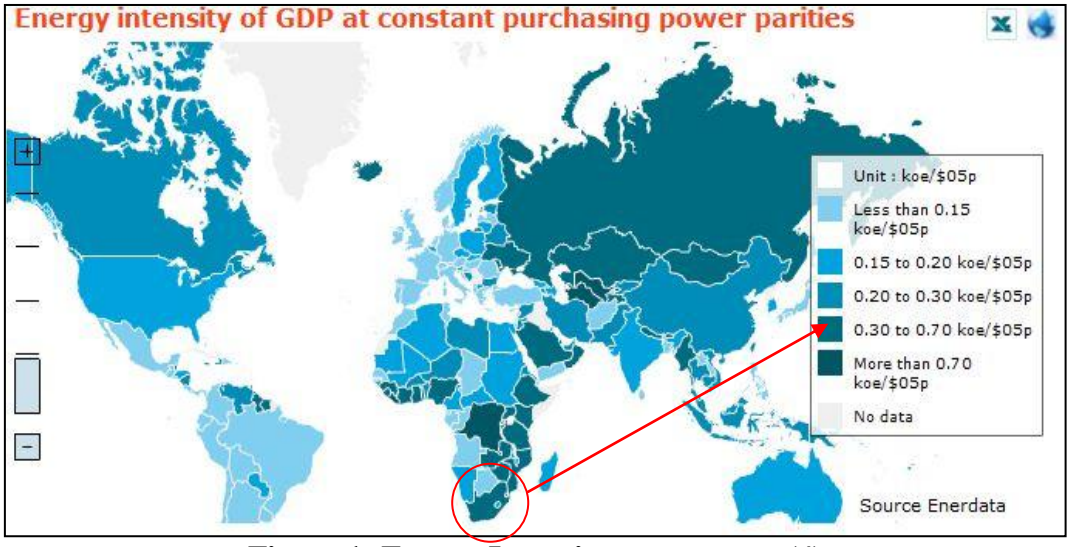


Figure 1: Energy Intensity per country (4)

2.3.2. Domestic Supply of Electricity

South Africa uses a range of sources of primary energy. The following was information was sourced from a publication by the South African Energy Department

Fossil fuels: A high proportion of the energy used in South Africa is produced from domestically produced coal. Furthermore, there is a limited amount of natural gas available and most of the crude oil use is imported.

Uranium: South Africa mines and exports raw uranium, but imports enriched uranium for its nuclear power plant, Koeberg.

Wind: South Africa currently has one wind farm, Klipheuwel, operating near Cape Town. There is moderate potential for wind energy along coastal areas of the Western and Eastern Cape.

Hydro-electricity, Biomass and Solar: South Africa also uses hydropower, most of which is imported. Electricity is also generated from other renewable energy sources mainly biomass and to a lesser extent solar energy.

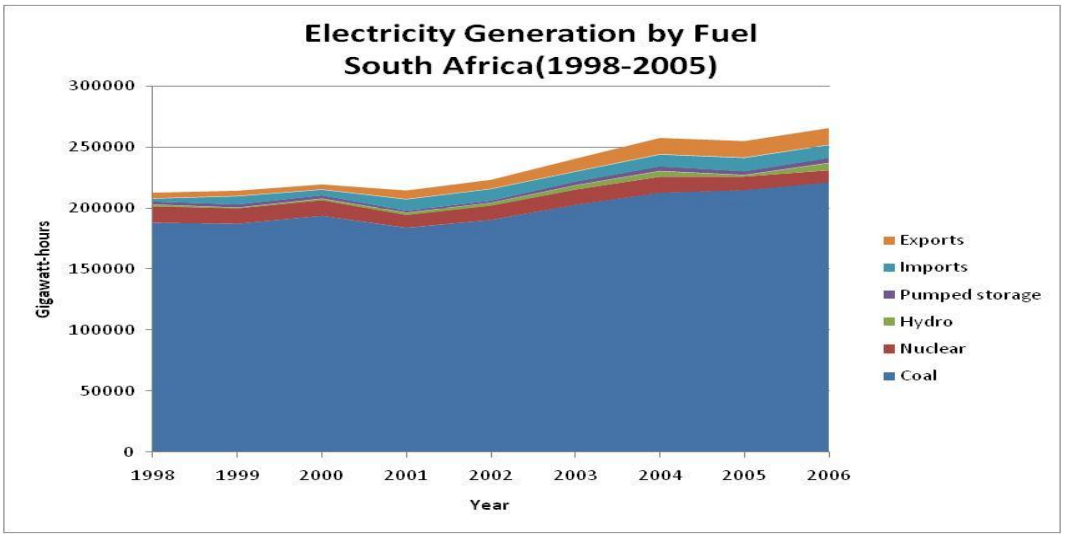


Figure 2: Electricity Generation by Fuel

Figure 2 shows the amount of electricity, in terms of gigawatt-hours, obtained from different sources over years 1998 to 2005. It can be observed that South Africa relies heavily on coal as a source of electricity.

2.4.THE RESIDENTIAL SECTOR

The key sectors in the demand market for electricity are industrial, residential and commerce.

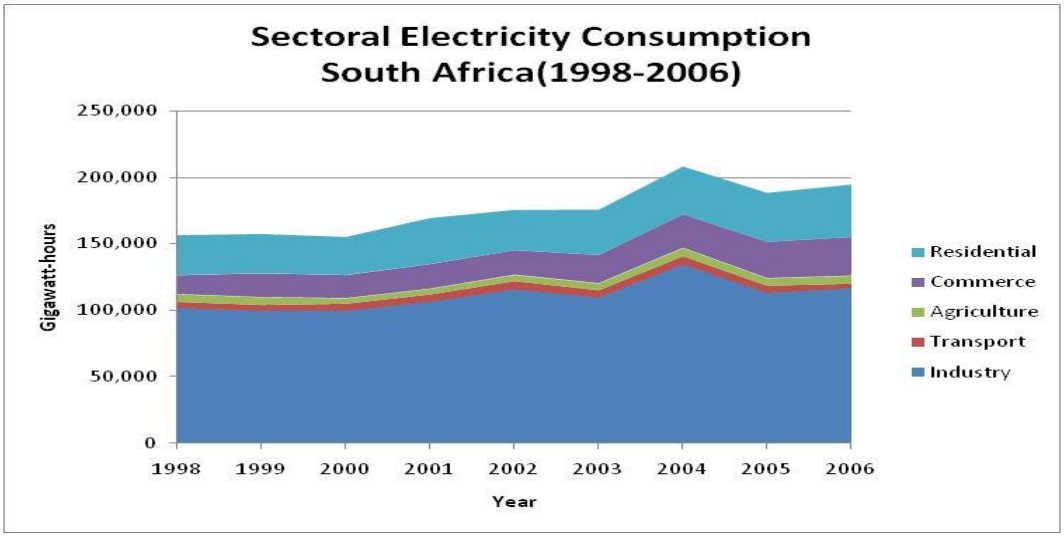


Figure 3: Sectoral Consumption of Electricity

The big question that one could ask is: Why are we interested in studying the factors that influence the amount of energy consumed within a household?

South Africa has experienced overwhelming load shedding problems during the year 2008. This situation was mainly due to demand out running supply. This was evidence for the need of new strategies with regards to both the supply and demand structures of the South African economy. However, given the already high energy consumption in South Africa, it was clear that demand side management has a more rigorous role to play in correcting the situation.

Over the period starting from 1998 to 2006, the residential sector has consumed an annual average of 19.03% of the total South African electricity consumption; coming second after the industrial sector which consumed an annual average of 63.23% over the same period. For DSM, it is important to understand the factors that influence the amount of electricity used within households because understanding these factors would provide aids towards strategising upon the appropriate tools that can be used in managing energy demand of the residential sector (5).

3. METHODOLOGY

3.1 DATA COLLECTION

All the data used in this study was obtained from a database from the ERC (UCT). It was originally obtained through:

- Survey

  Household data was obtained from a survey - the National Load Research programme - conducted by the NRS between 1996 and 2006.

- Electricity Meter monitoring

  ESKOM monitors the electrical current consumed (in terms of amperes) per household on meters. These units were recorded daily, at five minute intervals.

Only the monthly average of this five minute recordings and the number of units read collected per household were accessible, as opposed to the actual recordings.

For each household, the monthly count of units read was not necessarily the same. For this reason, the grand average (as opposed to cumulative amount of recorded units) was considered. Thus the response variable is the grand five-minute average over all the 15 months.

Energy consumption is conventionally expressed in terms of kilowatt-hours (KWh) rather than amperes, thus the amperes were converted to KWh:

$$\text{Amperes} \times \frac{230\text{Volts} \times 24\text{hours} \times 30\text{days}}{1000} = \text{KWh}$$

3.2 SUBJECTS, SCOPE AND LIMITATIONS

The study was restricted to households that were surveyed during 2005 in the areas shown in Table 1. Across the areas shown in Table 1, a total of 179 valid profiles were recorded with the sample size per area as given in Table 1.

Table 1: Residential areas surveyed during 2005

Location	Number of households per location
Peacetown	3
Khayelitsha	12
Matshana	59
Vlaklaagte	45
Driekoppies	12
Greenturf	24
Kabega	24

Figure 1Figure 4 shows the geographical locations of some of the areas shown in Table 1. Peacetown could not be located on the South African map.

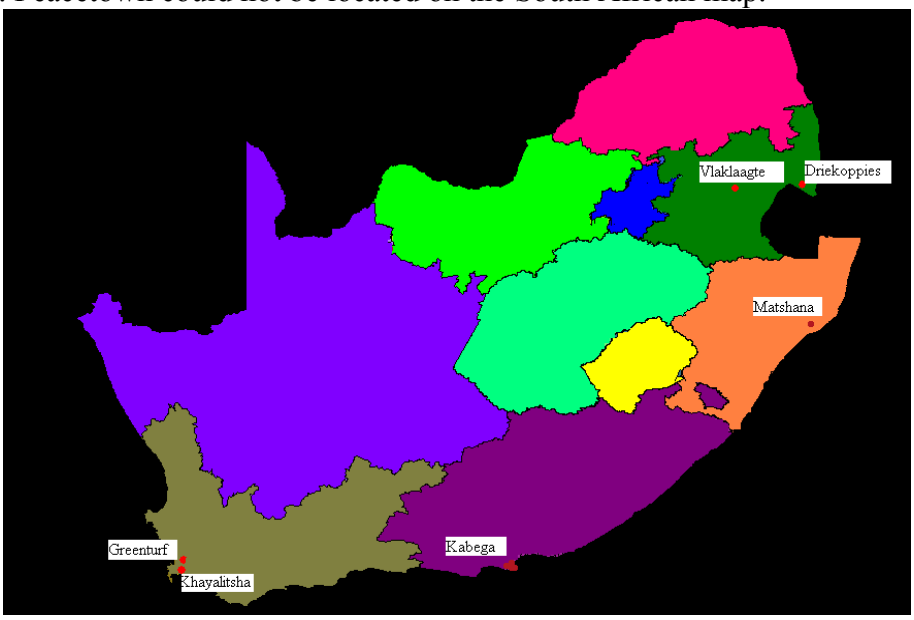


Figure 4: Geographical locations of the areas included in the survey

Site synopsis of the above locations is placed in Appendix 0.

3.3 VARIABLES INCLUDED IN THE STUDY

3.3.1. Outcome of Interest

The primary outcome of interest is the amount of electricity used by a household over a typical month. As outlined in section 3.1, the outcome of interest is the grand average of the recorded amperes over the 15 month period between December 2004 and February 2006. Although this is conventionally expressed in terms of KWh, the dataset used in this study has the amount of electricity used per household in terms of amperes.

3.3.2. Socio-Economic Features of each Household

For each household surveyed, a total of 131 features were recorded. The following is a summary of the features that were collected.

Table 2: Household features included in the study

	Variable name	Description	Type	Values
1	AltFuelCharcoalCook	Does household use charcoal for cooking?	Categorical	0-No 1-Yes
2	AltFuelCharcoalHeat	Does household use charcoal for heating?	Categorical	
3	AltFuelCoalCook	Does household use coal for cooking?	Categorical	
4	AltFuelCoalHeat	Does household use coal for heating?	Categorical	
5	AltFuelGasCook	Does household use gas for cooking?	Categorical	
6	AltFuelGasHeat	Does household use gas for heating?	Categorical	
7	AltFuelParaffinCook	Does household use paraffin for cooking?	Categorical	
8	AltFuelParaffinHeat	Does household use paraffin for heating?	Categorical	

9	AltFuelWoodCook	Does household use wood for cooking?	Categorical	
10	AltFuelWoodHeat	Does household use wood for heating?	Categorical	
11	ApplianceGeyserUsage	How often is a geyser used in household?	Categorical	0-Never 1-Monthly 2-Weekly 3-Daily 4-Unknown
12	ApplianceHeaterUsage	How often is a heater used in household?	Categorical	
13	ApplianceHotplateUsage	How often is a hotplate used in household?	Categorical	
14	ApplianceIronUsage	How often is a iron used in household?	Categorical	
15	ApplianceKettleUsage	How often is a kettle used in household?	Categorical	
16	ApplianceMicrowaveUsage	How often is a microwave used in household?	Categorical	
17	ApplianceOtherUsage	How often are *other appliances* used in household?	Categorical	
18	ApplianceStove3plateUsage	How often is a 3-plate stove used in household?	Categorical	
19	ApplianceStove4plateUsage	How often is a 4-plate stove used in household?	Categorical	
20	ApplianceTumbleDrierUsage	How often is a tumble drier used in household?	Categorical	
21	ApplianceWashingMachineUsage	How often is a washing machine used in household?	Categorical	
22	Ceiling	Does house have ceiling?	Categorical	0-No 1-Yes
23	Insulation	Does house have insulation?	Categorical	
24	IncomeAggriculturalYesNo	Are there who receive income from an agricultural business?	Categorical	
25	IncomeSmallBusinessYesNo	Are the members who receive income from a small business or child grant?	Categorical	
26	IncomeExternalYesNo	Are the members who receive income from other external sources?	Categorical	
27	OwnDwelling	Is this house rented or own dwelling	Categorical	
28	SexHouseholdHeadMale	Is sex of household head male?	Categorical	
29	IncomeRefuse	Did house refuse to disclose level of income earned by adults?	Categorical	
30	SmallBusiness	Does house own a small business?	Categorical	
31	SupplyOfOutBuildings	Does house supply electricity to any external buildings?	Categorical	
32	GROUPID	See Table 1	Categorical	See Table 1
33	Hotwater Source	How do household members obtain hot water?	Categorical	Coal Wood Paraffin Electric kettle Hotplate Electric stove Geyser
34	Spouse' Highest Education	Highest level of education obtained by spouse?	Categorical	0-No 1-Yes
35	Head's Highest Education	Highest level of education obtained by head?	Categorical	
36	EmploySpouse	Employment status of spouse?	Categorical	1-Full time 2-Part time 3-Self-employed 4-Pension 5-Unemployed
37	EmployHead	Employment status of head?	Categorical	
38	LanguageID	Home Language	Categorical	1-Zulu 2-Sotho 3-Xhosa 4-Afrikaans 5-English 6-Sepedi 7-Tswana 8-Ndebele

				9-Tsonga 13-Swati
39	MainSwitch	Ampere reading on main switch	Categorical	20-20 Amps 60-60 Amps
40	PQGoesOff	How often does the township (in which the house is) experience and general power failure in the entire?	Categorical	0-Never 1-Monthly 2-Weekly 3-Daily 4-Unknown
41	PQLightsDim	How often do lights get dim at night but power not tripping completely?	Categorical	
42	PQTrips	How often does main switch trip or circuit break in the house?	Categorical	
43	RoofMaterialCode	Ty pe of roof for house	Categorical	
44	IncomeCode	Income class code	Categorical	0: R0-R500 1:R500-R1000 2-R1000-R1500 … 34-R17000-R17500
45	TotalIncomeCode	Total income class code	Categorical	0: R0-R500 1:R500-R1000 2-R1000-R1500 … 34-R17000-R17500
46	WallMaterialCode	Wall material code	Categorical	
47	WatersourceCode	Water source code	Categorical	
Number of units per type of appliance				
48	ApplianceDeepFreezeNumber	Deep Freeze	Continuous	Non-negative counts.
49	ApplianceFridgeFreezerNumber	Fridge Freezer	Continuous	
50	ApplianceGeyserBroken	Geyser	Continuous	
51	ApplianceGeyserNumber	Geyser	Continuous	
52	ApplianceHeaterBroken	Heater	Continuous	
53	ApplianceHeaterNumber	Heater	Continuous	
54	ApplianceHiFiRadioNumber	Hi-Fi Radio	Continuous	
55	ApplianceHotplateBroken	Hotplate	Continuous	
56	ApplianceHotplateNumber	Hotplate	Continuous	
57	ApplianceIronBroken	Iron	Continuous	
58	ApplianceIronNumber	Iron	Continuous	
59	ApplianceKettleBroken	Kettle	Continuous	
60	ApplianceKettleNumber	Kettle	Continuous	
61	ApplianceLightsNumber	Lights	Continuous	
62	ApplianceMicrowaveBroken	Microwave	Continuous	
63	ApplianceMicrowaveNumber	Microwave	Continuous	
64	ApplianceOtherBroken	Other	Continuous	
65	ApplianceOtherNumber	Other	Continuous	
66	ApplianceStove3plateBroken	Stove3plate	Continuous	
67	ApplianceStove3plateNumber	Stove3plate	Continuous	
68	ApplianceStove4plateBroken	Stove4plate	Continuous	
69	ApplianceStove4plateNumber	Stove4plate	Continuous	
70	ApplianceTumbleDrierBroken	Tumble drier	Continuous	
71	ApplianceTumbleDrierNumber	Tumble drier	Continuous	
72	ApplianceTVNumber	TV	Continuous	
73	ApplianceWashingMachineBroken	Washing machine	Continuous	

74	ApplianceWashingMachineNumber	WashingMachine	Continuous	
How many meals of this type are prepared for this type of household members on this day? See arrows for example?				
75	CookAdultsSaturdayBreakfast		Continuous	
76	CookAdultsSaturdayDinner		Continuous	
77	CookAdultsSaturdayLunch		Continuous	
78	CookAdultsSundayBreakfast		Continuous	
79	CookAdultsSundayDinner		Continuous	
80	CookAdultsSundayLunch		Continuous	
81	CookAdultsWeekDayBreakFast		Continuous	
82	CookAdultsWeekDayDinner		Continuous	
83	CookAdultsWeekDayLunch	AdultsSaturdayBreakfast	Continuous	Non-negative counts.
84	CookChildrenSaturdayBreakfast		Continuous	
85	CookChildrenSaturdayDinner		Continuous	
86	CookChildrenSaturdayLunch		Continuous	
87	CookChildrenSundayBreakfast		Continuous	
88	CookChildrenSundayDinner		Continuous	
89	CookChildrenSundayLunch		Continuous	
90	CookChildrenWeekDayBreakfast		Continuous	
91	CookChildrenWeekDayDinner		Continuous	
92	CookChildrenWeekDayLunch		Continuous	
93	EmployOlder16Full	Number of people older than 16 who are employed on full time basis.	Continuous	Non-negative counts.
94	EmployOlder16part	Number of people older than 16 who are employed on part time basis.	Continuous	
95	EmployOlder16Pension	Number of people older than 16 who earn a pension.	Continuous	
96	EmployOlder16Self	Number of people older than 16 who are self-employed.	Continuous	
97	EmployOlder16Unemployed	Number of people older than 16 who are unemployed.	Continuous	
98	EmployUnder16Unemployed	Number of people UNDER than 16 who are unemployed.	Continuous	
99	Females16To24	Number of females aged 16To24 in household	Continuous	Non-negative counts.
100	Females25To34	Number of females aged 25To34 in household	Continuous	
101	Females35To49	Number of females aged 35To49 in household	Continuous	
102	FemalesOlder50	Number of females aged over 50 in household	Continuous	
103	FemalesYoungerThan16	Number of females aged under 16 in household	Continuous	
104	FloorArea	Floor area	Continuous	All positive numbers
105	IncomeAdults	Amount of income earned by adults	Continuous	All positive numbers
106	IncomeAggricultural	Amount of income from agricultural business	Continuous	
107	IncomeExternal	Amount of income from external sources	Continuous	
108	IncomeSmallBusiness	Amount of income from child grant or small business	Continuous	
109	Males16To24		Continuous	Non-negative counts.
110	Males25To34	Number of males aged 25To34 in household	Continuous	
111	Males35To49	Number of males aged 35To49 in household	Continuous	
112	MalesOlderThan50	Number of males aged over 50 in household	Continuous	

113	MalesYoungerThan16	Number of males aged under 16 in household	Continuous	
114	TotalNumber Of People In Household	Total number of people in household	Continuous	Non-negative counts.
115	NumberAdultsEarningSalary	Number of adults earning a salary	Continuous	Non-negative counts.
116	OccupationYears	Occupation years	Continuous	Non-negative counts.
117	OtherAppliances	Other appliances	Continuous	List of other appliances
118	Rooms	Rooms	Continuous	Non-negative counts.
119	TimeWithElectricity	Time length for which household had electricity	Continuous	Non-negative counts.
120	Total Females	Total number of females	Continuous	Non-negative counts.
121	TotalEducatedPeopleOlder16	Total number of educated people older than 16	Continuous	Non-negative counts.
122	TotalEducatedPeopleUnder16	Total number of educated people under than 16	Continuous	Non-negative counts.
123	TotalIncomeInHousehold	Total income in household	Continuous	All positive numbers
124	TotalNumbeOlder16Employed	TotalNumbeOlder16Employed	Continuous	Non-negative counts.
125	TotalNumberOfAppliances	Total number of appliances	Continuous	Non-negative counts.
126	TotalNumberOfAppliancesExclLights	Total number of appliances excluding lights	Continuous	
127	TotalNumberOfBrokenAppliancesExclLights	Total number of broken appliances excluding lights	Continuous	
128	TotalNumberOfMales	Total number of males	Continuous	
129	TotalNumberOfNonBrokenAppliancesExclLights	Total number of non- broken appliances excluding lights	Continuous	
130	TotalWeekdayAdultMeals	Total number of weekday meals for adults	Continuous	
131	TotalWeekdayChildMeals	Total number of weekday meals for children	Continuous	

The full data exploration will be done in Chapter 5.

4. STATISTICAL METHODS

Variable selection is a method of investigating relationships between the outcome of interest and the independent variables. Variable selection in multiple linear regression can be considered for this data, but this methods assumes that

- the response variable follows a normal distribution
- there are linear association between the independent variables and the response

In cases where the normality assumption does not hold, GLM are used to overcome the distribution problem.

At times, the linear association method does not hold either; CART is a more flexible method for investigating the relationships without the linearity assumption. GLM might struggle in large data sets like the one analysed in this study as collinearities between the independent variables are more likely, thus we might encounter singularity problems in the parameter estimation process. CART triumphs GLM in this aspect.

4.1 VARIABLE SELECTION AND MODELLING BY GLM

Generalised linear modelling is a statistical modelling tool, used to quantify linear relationships between a set of independent variables and a transformation of the expected value of the outcome of interest.

Before using GLMs, the following properties must be verified:

a) Exponential family
GLMs require that the underlying distribution followed by the outcome of interest, Y , be a member of the exponential family. That is, the density function should be of the form:

$$f(y|\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

b) A set of coefficients
As per the purpose of GLMs, from a dataset that has $(Y_i, X_{i1}, \dots, X_{ik})$ for each subject, we wish to attain estimates for values $\beta_0, \beta_1, \dots, \beta_k$ such that

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

is the best possible **linear predictor** (η_i) of some pre-specified transformation of the expected value of the outcome of interest.

c) Link function
The linear predictor (η_i) is the best estimate for a transformation of the expected outcome, $g(E[\boldsymbol{Y}])$. $g(x)$ has to be a differentiable, monotonic

function and its form is suggested by the form $a(y)$. The end result is to obtain:

$$g(E[\boldsymbol{Y_i}]) = \boldsymbol{X^T\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

a) Procedure

The idea in GLM variable selection is to consider **all the possible subsets** of the independent variables, and select the subset that results in the minimum AIC[1]. This means that a total of models that have to be considered is

$$\binom{k}{1} + \binom{k}{2} + \binom{k}{3} + \cdots + \binom{k}{r} + \cdots + \binom{k}{k-1} + \binom{k}{k}$$

which becomes an enormous amount as k gets larger.
The process of selecting variables by GLM is called stepwise variable selection and it can either be forward, backward or in both directions. Consider a study that has k independent variables $X_1, X_2, X_3, \ldots, X_k$. With the forward selection, the algorithm

- starts with the minimal model[2] and evaluate the corresponding AIC,
- Add each of the independent variables one at a time and evaluate the corresponding AIC,
- choose that variable that results in the lowest AIC (compared to the other variables and the minimal model)
    - if none of the variables result in an AIC lower than that of the minimal model, then that is evidence that none of the variables can be used to explain the variation in the outcome of interest. Thus the algorithm terminates.
- After finding the first variable, consider the remaining variables. Adding each of the variables one at a time to the model that already contains the first variable. The second best variable would be the one that results in the lowest AIC, which is also lower than the AIC corresponding to the first variable.
- The process continues onto the selecting more variables for as long as the AIC is decreasing, and halts when AIC the stops decreasing.

The backward selection follows a similar algorithm, except it starts with the maximal model[3], excluding one variable at a time for as long as the AIC is still decreasing.

At maximum, the total number of models that have to be considered in these 'directed' algorithms is:

$$\binom{k}{1} + \binom{k}{2} + \binom{k}{3} + \cdots + \binom{k}{r} + \cdots + \binom{k}{k-1} + \binom{k}{k}$$

before interaction terms!

[1] There are other criteria that could be used besides the AIC but in this study we will only use the AIC.
[2] Minimal model is that which contains no independent variables.
[3] Maximal model is that which contains all the independent variables in the study.

The undirected selection method can start with either the minimal and maximal model. The key difference with this method is that, it can add and exclude variables in the process in search of the subset that results in minimal AIC.

4.2 A SHORT LITERATURE REVIEW ON CART

4.2.1. Introduction

Classification and Regression Trees (CART) are a statistical technique used for identifying features (independent variables) of subjects from a sample that have radical influence on the outcome of interest (6), either categorical or continuous. We shall refer to these variables as classifiers. The process is exploratory in nature and is often part of the data mining process.

4.2.2. Overview and Origin of the Classification Problem

Using trees to group subjects is a relatively well known method in several fields of study. This method was initiated at the University of California, San Diego Medical Centre where the objective was to classify heart attack patients, with regards to their likeliness of living for more than 30 days or not, since admission (7). During admission, 19 features that were considered to well-abridge the patient's medical and physiological profile were measured. Age, systolic blood pressure and the presence of sinus tachycardia were some of those features; and the patients were classified into either category based on their profiles.

For this grouping task to be *accurate*, it is imperative for the facilitator to understand which of the all the 19 features had fundamental influence on the outcome of interest. A "*learning sample"* was used to identify which features seemed to have radical influence on the outcome of interest using this tree method, and future patients were grouped - solely- based on these radical features.

From the learning sample; only age, systolic blood pressure and the presence of sinus tachycardia emerged as having strong association with the patient surviving for more than 30 days (or not) since admission. The results were presented as follows:

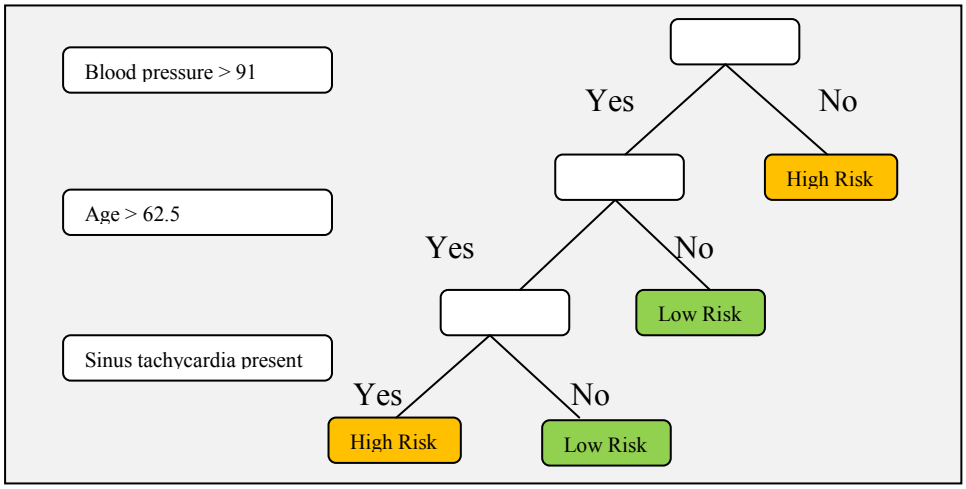


Figure 5: Results from the San Diego Medical Centre (6)

The above problem entailed classifying subjects into categories, and thus the self-explanatory term classification tree. Although for this case the outcome only had two classes, the technique can be generalised to an outcome with more than two categories. For instance, the facilitator above could have been interested in relatively

finer partitions of the survival groups and classified the subjects as those who survived

- for more than 30 days
- to the 20th day but died before the 30th day
- to the 10th day but died before the 20th day
- for less than 10 days

Thus the problem would expand to grouping across four categories.

The group prior to splitting is referred to as the parent node while the emerged - after splitting - subgroups are referred to as child nodes.

4.2.3. Description of Classification Trees

a) Setting

Assume a study that aims to classify an $n-$sized sample of subjects into a set of outcome classes. Let $\boldsymbol{O}$ represent the set of all possible outcomes, this may be a set of categories (high risk or low risk from the above example); further assume that k variables are measured from each of the subjects studied, and let $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \dots x_{ik})$ represent the covariate $k-$ vector for the i^{th} subject.

CART is concerned with coming up with a rule f, that accurately groups the subjects into outcome classes based solely on the covariate vector. That is:

$$f(\boldsymbol{x}) \in \boldsymbol{O}$$

The rule is a set of sequential "yes-no" questions, thus the alternative term "decision tree". The process is aimed at selecting variables (and specific values of these variables - splitting points) that are strongly associated with the outcome and uses these to group the data such that minimal impurity (or the well-known equivalent term "maximum homogeneity") within the terminal subgroups is achieved.

b) Splitting Process

The splitting process is the core part of the CART technique. For each of the variables measured, the process *considers* splits at all possible points of the variable and then evaluates the extent of impurity (see more below) that results in the emerged subgroups. The variable with a splitting point that resulted in minimum impurity is then taken as a possible significant classifier for the outcome. In the San Diego Medical Centre case, when considering systolic blood pressure, splits were done at all reasonable blood pressure points and 91units was the point that resulted in the minimum impurity.

As a general idea, for each splitting point, the extent of *impurity* in the child nodes must be evaluated and the lower the resulting (child nodal) *impurity* the better.
How to evaluate this impurity? How low should the impurity be to validate a splitting point as a definite classifier? This will always depend on the **splitting criteria** used.

c) Splitting Criteria, Impurity and Impurity Functions

The term impurity is used in a similar sense as in a linguistic sense. It refers to the extent of variation (however that may be measured) of the outcome of interest within a node. For instance, if the outcome of interest was continuous and exactly the same value was observed for all the subjects in the node, then the node has no impurity whatsoever.

Splitting criteria encompasses the function rule used to evaluate the level of impurity within a node and guidelines on deciding whether a splitting point is a valid one. Whichever criterion adopted would be used when considering all splitting points in all independent variables. Gini index, Twoing rule, Entropy rule, Chi-Square rule and the minimum error rule are some of the many splitting criteria that have been proposed, with Gini and Twoing being the most widely used (8).

i. Gini Diversity Index

Gini diversity index is an example of what is often referred to as an impurity function. Several impurity functions have been proposed but, the imperative properties that any impurity function should have are:

a. It must be a symmetrical function of the class probabilities $p(1), p(2), \dots p(\omega)$;
b. be convex;
c. attain its maximum value when all proportions of outcomes from each of the outcome classes are uniform. This is an imperative requirement since if the outcome values within a node are uniformly distributed across the outcome classes, then that splitting rule has not assisted in classifying the subjects in anyway;
d. attain its minimum value when all outcomes in a node come from only one class (thus $p(k) = 1$ for some $k \in \boldsymbol{O}$) and there are none from the other classes (thus $p(j) = 0$ for some $j \in \boldsymbol{O}$, $j \neq k$). Implying that there is no impurity within the node.

When using these impurity functions as splitting criteria, for each splitting point considered, the level of impurity within the parent node is compared to the weighted[4] average of the impurity levels in the child nodes. The larger the difference in this two figures (with that of the parent node being the largest) the more credible is that point as a classifier.

More formally, let h_{ij} denote the impurity within the j-side (j only indicating left or right) node of the i^{th} split, with the uppermost group being the 0th generation. Let w_{ij}be the weighting held by the j-side node during the i^{th} split and h_{i-1} be the impurity level in the immediately preceding parent node. That is, at a typical i^{th} split on the tree:

[4] The weight used for each subgroup is its size (in terms of number of subjects) relative to the size of the immediately preceding parent node.

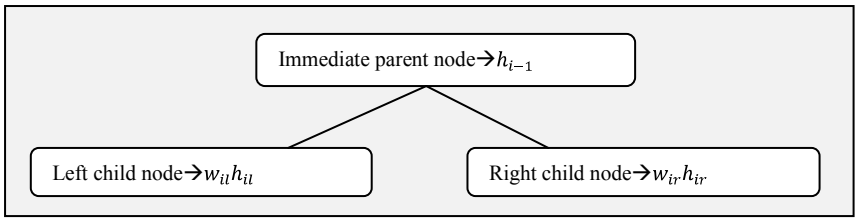


Figure 6: Using impurity as a splitting criteria

An index suitably called an "improvement measure"

$$\Delta h_i = h_{i-1} - w_{il}h_{il} - w_{ir}h_{ir}$$

has been formally employed to indicate the difference in (im-)purity between parent and child nodes.

When using the Gini diversity index method, h_{ij} is calculated as follows

$$h_{ij} = \sum_{\alpha \neq \beta} p(\alpha|i,j)p(\beta|i,j)$$

where $p(\alpha|i,j)$ is the probability of outcome α given that we are at j-side node from the i^{th} split. From the learning sample, these conditional probabilities are calculated as the proportion of outcomes that belong to that respective class. Then the splitting point that results in the maximum value for Δh_i is adopted for that variable.

How large the improvement should be to credit the splitting point will differ across researchers, objectives and context.

ii. Twoing Rule

Twoing is another method of deciding on splitting points. Unlike the impurity functions, Twoing uses the spread of proportions of the outcomes between the two child nodes to evaluate the efficiency of a splitting point. The principle behind this criterion is:

If a splitting point results in outcomes from the same class being equal spread between the child nodes (i.e. 50%-50%), then that splitting point is not associated with that outcome level as the outcome level is *indifferent* to either side of the splitting point. Yet at the other extreme, if all values from the same outcome class end up in the same node (and none in the other) after a split then that is evidence for a strong association between that splitting point and the outcome level.
In order to integrate all the levels of the outcome and size (in terms of number of subjects) of both child nodes, then the Twoing index is calculated as follows:

$$\frac{w_{il}w_{ir}}{4}\left[\sum_{\alpha \in \boldsymbol{O}} |p(\alpha|i,l) - p(\alpha|i,r)|\right]^2$$

with w_{il} and $p(\alpha|i,l)$ as above.

A large Twoing index is taken as evidence for association between the splitting point and the outcome as a whole. There are variations in calculating this index, but the integral idea is that the index should encompass the difference in the proportions per outcome level and the (sample size) weights.

Which criterion to use? Past investigations have shown that the optimal tree does not depend much on the criterion used. (7)

d) Surrogate Splits and Missing Values

Missing values are a frequent encounter in datasets. At times, the design of the study results in structurally missing data. For example, in the NRS data, the questionnaire asked the respondents about often they use a tumble drier; this question will automatically not have an answer for households that did not have a tumble drier, thus a structurally missing value.

In CART, surrogates splits are used when encountering missing values. Surrogate splits are splits that can serve as auxiliary splits to the best split if the best split cannot be applied due to missing values. That is, they are the rules that can predict the results of the best split as closely as possible.

In constructing a tree, to select the best splitting point on a variable, the algorithm only uses the subjects which have a value for that variable.

In the task of splitting new samples, some cases might have missing values for variables on which an optimal split was found. The surrogate split will be used on the cases that have missing values for such variables. Sometimes case might have missing values for both the variable that have the optimal split and that on which the surrogate is based. What then? Statistical programs are orientated to select a sequence of potential surrogate split in order of their merit ranked according to the accuracy at which the splits can predict the results of the main best split. In the situation that the best surrogate split is not applicable, the next best one will be used. (7)

e) Stopping Rules

At times, investigations might involve a large number of variables, and one might feel the urge to control or limit the tree size. Because the tree algorithm *decides* on splits in an ordinal manner - i.e. splitting points/questions that emerge first are the most radical ones and those that come later in the tree are the least influential - one could stop the tree from growing to the end, if there are too many variables.

The main joint disadvantage of using stopping rules is that the tree-growth might be stopped too early and some potentially important splitting questions might not be observed.
To control over-growth (or over fitting) of the tree one can pre-specify

- <u>The minimum number of subjects that may remain in a node after a split.</u>
  This prevents the algorithm from deciding on a splitting point based on too small a sample. The benefit of this rule is that results are likely to be reliable due to the (nodal) sample size at each point being (at least) as large as the facilitator wishes.

- The maximum number of splits that are allowed.
  This prevents the tree from growing beyond a certain point, and the main benefit is that the facilitator can control the size of the tree.
- Maximum value of the splitting criteria used.
  This entails commanding the algorithm to only split if the value of the splitting criteria chosen is larger than some benchmark value.

4.2.4. Regression Trees

Overview

In some situations, one might wish to do a task similar to the classification tree but with a continuous outcome variable; in this case the task is then referred to as a regression tree.

The methodology of splitting discussed in part b) of section 4.2.3 is still applicable, except now there are no classes among which the subjects are classified but just a continuum of values. Now because all splitting criteria discussed in part c) of section 4.2.3 use the concept of *probability of an outcome belonging to a certain class of the whole set of outcome classes*, these functions will not be applicable for evaluating the level of variability within a node in a regression tree problem.

a) Description

The key idea is still to group the subjects such that the extent of variability within the child nodes is minimal, relative to that in the immediately preceding parent node. When dealing with a continuous outcome, deviance (defined below) (9) is the sensible way of evaluating the level of variability within a node. Let y_{ijk} represent the k^{th} outcome in the ij^{th} node, with $k = 1, 2, \dots, n_{ij}$. Recall, this refers to the j-side node resulting from the i^{th} split. And further let $\bar{y}_{ij}$ be the average value in the ij^{th} node; then the observed variance in the ij^{th} node – s_{ij}^2 – is calculated as:

$$s_{ij}^2 = \frac{\sum_{k=1}^{n_{ij}} \left[y_{ijk} - \bar{y}_{ij} \right]^2}{n_{ij}}$$

the maximum likelihood estimate, or

$$s_{ij}^2 = \frac{\sum_{k=1}^{n_{ij}} \left[y_{ijk} - \bar{y}_{ij} \right]^2}{n_{ij} - 1}$$

the unbiased estimate.
For comparative purposes, either estimates will lead to similar results, but if one is interested in obtaining the true variance, the unbiased estimate is more appropriate.

Then for the i^{th} split considered, the improvement (Δs_i^2) measure is calculated as:

$$\Delta s_i^2 = s_{i-1}^2 - w_{il} s_{il}^2 - w_{ir} s_{ir}^2$$

where s_{i-1}^2 is the observed variance within the immediately preceding parent node. That is, at the i^{th} split, we have:

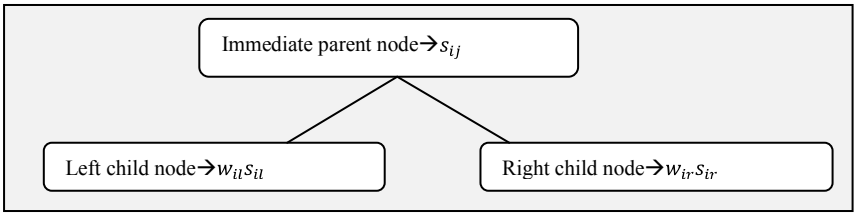


Figure 7: Using deviance as a splitting criteria in regression trees

Again how large should the difference be to validate the splitting as a radical splitting point would differ across researchers, objective and context.

b) Regression Tree Example

Consider a fictional investigation whose aim is to identify the features that are associated with level of income earned by a participant of a specific sector in the labour market. Assume that features collected were:

- Age: continuous variable;
- Gender: categorical,2 levels;
- Location area type: categorical, looking at cities, rural areas, near-city townships and farms;
- Highest level of educational qualification: categorical, considering None, grade 1, up to advanced degree - thus a total of 16 levels.

Income level is essentially a continuous variable, so the regression tree method is applicable.

Assume the following results were obtained:

- Highest level of qualification: Although this variable had 17 levels (regardless of the other features), there was a marked difference between the cluster of income levels for individuals who attained their matric (or beyond) and those whose highest education level was below matric.
- Age: Within the cluster of individuals who had - at least - obtained matric, income levels of those in the cluster of individuals aged below 30 years were notably lower than those aged over 30 years. Yet for individuals who had not attained their matric, a similar cluster difference was apparent at the age of 35.
- Gender: Within the group of individuals who are aged over 30 years and have at least attained their matric, the males exhibited some notably higher income levels compared to that of females.
- Work location (area type): There was somewhat of an income difference between the group of individuals whose workplaces are located in cities and those whose workplaces were in other area types (like rural, near-city townships, farms). This was apparent in the cluster of females aged over 30 and had attained their matric or beyond AND that of individuals aged under 35 and had not attained their matric.

The above results can then be shown in a tree as follows:

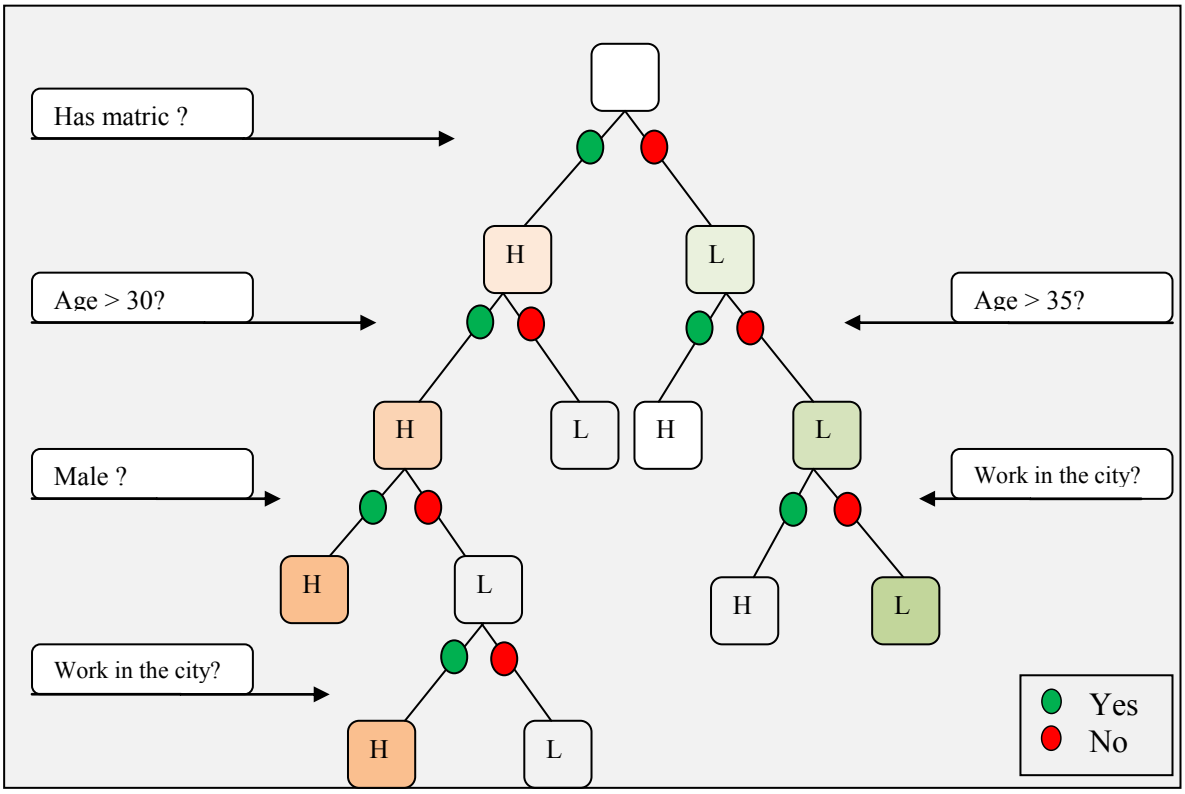


Figure 8: Results from the income level study

With H denoting that the average level of income is higher (in that respective node) than in the other child node that emerged from the same parent (which would have symbol L).

As a final note, these regression trees can also applied to count outcomes or survival outcomes.

4.2.5. Objective Variable Selection

As discussed under part e) of section 4.2.3, with no constraints, the tree might have too many splitting points than is desired or useful for predictive purposes. This is referred to as over-fitting. In this section we look at selecting a subset of variables that are useful for predictive purposes. This process is referred to as variable selection.

Although this sounds similar to applying stopping rules, this process focus on identifying variables that have fundamental influence on the outcome of interest while stopping focus primarily of managing the tree size. Again stopping rules are based on researcher's subjective intuition based on past experience; this section is dedicated to the proposed objective methods for getting the "right-sized tree".

a) Cost-Complexity

Cost-complexity is the most widely used pruning method. The key idea behind this method is to keep on growing the tree for as long as the decrease of rate of errors (or

mean residual deviance in a regression problem) is still significant. These pruning criteria are used in order to assess the extent to which the observed outcomes differ from the predicted value in a node. In a classification problem, an error within a node is the subject $(x_{i1}, x_{i2}, \dots, x_{ik}, Y_i)$ for which

$$f(x_{i1}, x_{i2}, \dots, x_{ik}) \neq Y_i$$

That is the case whose observed outcome is different from the predicted one. Thus, the number of errors in any node is calculated as

$$\sum_i I[f(x_{i1}, x_{i2}, \dots, x_{ik}) \neq Y_i]$$

summing across the subjects in the node, where I is an indicator function,

$$I = \begin{cases} 1 & f(x_{i1}, x_{i2}, \dots, x_{ik}) \neq Y_i \\ 0 & f(x_{i1}, x_{i2}, \dots, x_{ik}) = Y_i \end{cases}$$

In a regression problem the residual deviance in a node is defined as

$$\sum_i [y_i - \bar{y}]^2$$

again summing across the subjects in the node. In a regression problem, the average value in a node, $\bar{y}$, is the predicted value for that node.

Let R_i represent the number of errors(or residual deviance in regression problem) in the i^{th} leave (or terminal node), and R be a the error rate (or mean residual deviance in a regression case) for the whole tree, thus calculated as

$$R = \frac{\sum_{\text{all nodes}} R_i}{n - \text{size}}$$

The cost-complexity method is based on applying a penalty(called the complexity parameter) for each extra leave added onto the tree. If $\alpha \geq 0$ denotes the complexity parameter, then the cost-complexity measure is evaluated as

$$R_\alpha = R + \alpha(\text{size})$$

where "size" is the number of leaves (or terminal nodes) on the tree.

The procedure:

- obtain values of R for all tree sizes between the tree with root only and the maximum tree by cross validation (described below)
- for each α (starting with 0), calculate the corresponding R_α for each tree size
- select sub-tree that minimises R_α
- as α increases, the optimal tree-size gets smaller due to the increasing penalty
- stop the procedure (i.e. increasing α) as soon as the best tree for an α is the tree with the root only
- finally, there will be a sequence of nested sub-trees for each α-value

- the best tree size can be chosen directly or by selecting the *best* value of α.
  Selecting the best value of α:
  - in classification: $\alpha = 2(||\boldsymbol{O}|| - 1)$, where $||\boldsymbol{O}||$ is the number of classes in the outcome space
  - in regression: $\alpha = 2\hat{\sigma}^2$, where $\hat{\sigma}^2$ is the residual mean deviance from the maximum model

    Having obtained a value for α, the best tree size is the one that minimises R_α. The tree obtained from using this rule is the one that has minimum AIC. This was derived from the idea that AIC (as a model selection criteria) penalises the number of parameters (therefore the number of variables) included in the model by adding 2 times (the number of parameters) to -2 times log-likelihood. (9). This rule, however, tends to overfit.

  Directly:
  - alternatively, one could use the "1-SE rule", which states that the best tree is that tree whose R value is the largest within 1 standard error of the minimum R obtained. (9)

b) k fold Cross Validation

Cross-validation is a method used to objectively estimate the true error misclassification error or residual deviance of a tree (R). The procedure consists of:

- randomly divide the learning sample into k roughly equal subgroups
- from the k subgroups, exclude one subgroup at a time and use $k - 1$ to grow the maximum tree
- then use the k^{th} dataset to calculate the residual deviance for each value of te complexity parameter.
- from the k different R estimates obtained (for each tree size), take the average R value for each tree size and the average is the objective estimate of the true misclassification or residual deviance.

Because cross-validation involves random division of the learning sample, different R estimates would be obtained every time the procedure is performed.

c) Balancing Accuracy and Operability

Variable selection by CART is largely a subjective task. Developing the decision tree involves obtaining a *fair* balance between simplicity and accuracy.

5. DATA EXPLORATION

5.1 OUTCOME OF INTEREST

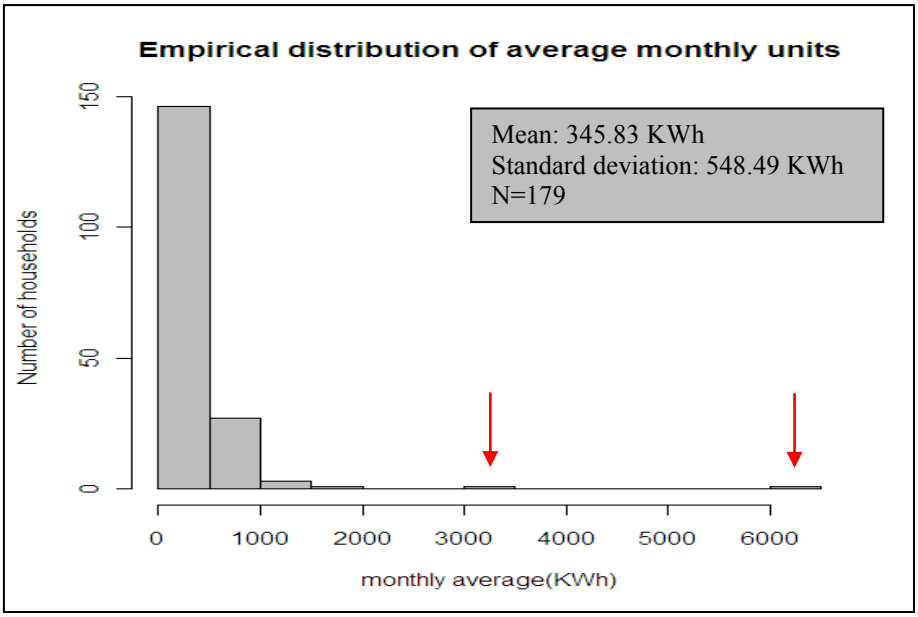


Figure 9: Empirical distribution of average monthly units (in KWh)

Outlying observations were removed (with monthly consumption of 6059.92KWh & 3135.033KWh); these observations were regarded as extremely unlikely for a household.

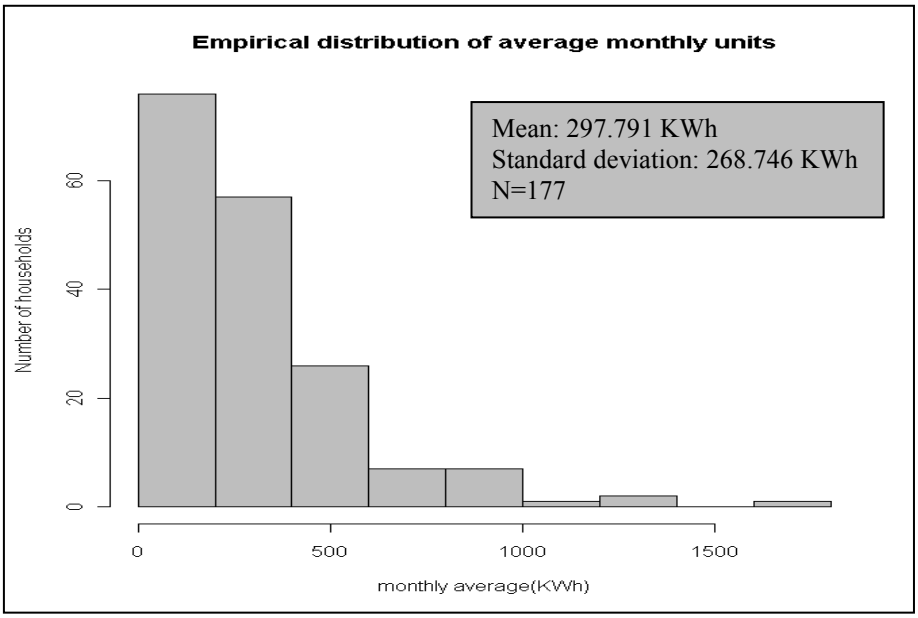


Figure 10: Empirical distribution of average monthly units excluding outliers (in KWh)

Other summary statistics (excluding outliers):

%	0%	25%	50%	75%	99%	100%
Percentile	6.004	105.545	234.440	396.207	1364.278	1648.516

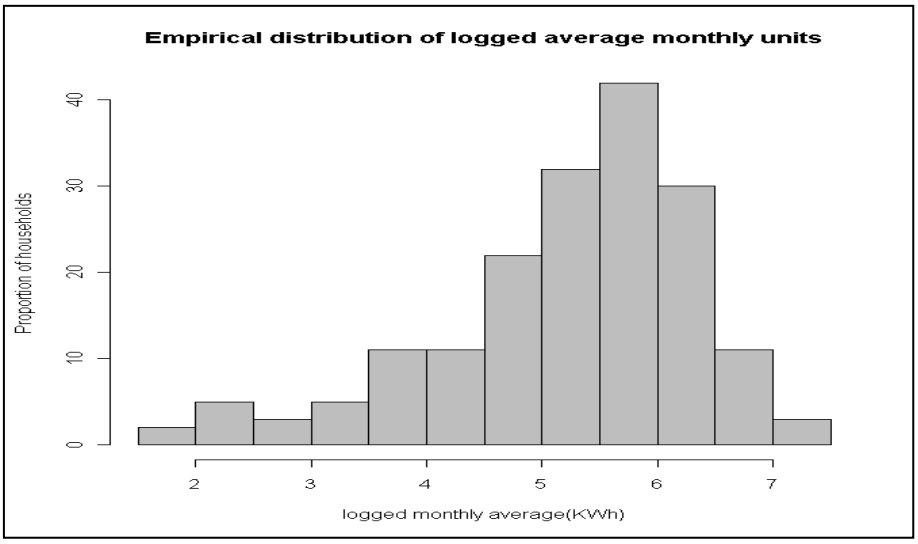


Figure 11: Empirical distribution of transformed average monthly units excluding outliers (in KWh)

Figure 11 shows that the logged average monthly units are not as skew as the original average monthly units but have a tail to the right.

Recall that the Figure 10 shows the averaged figures from the recordings made per month, Figure 12 shows the actual monthly recording (converted into KWh) for a few selected households.

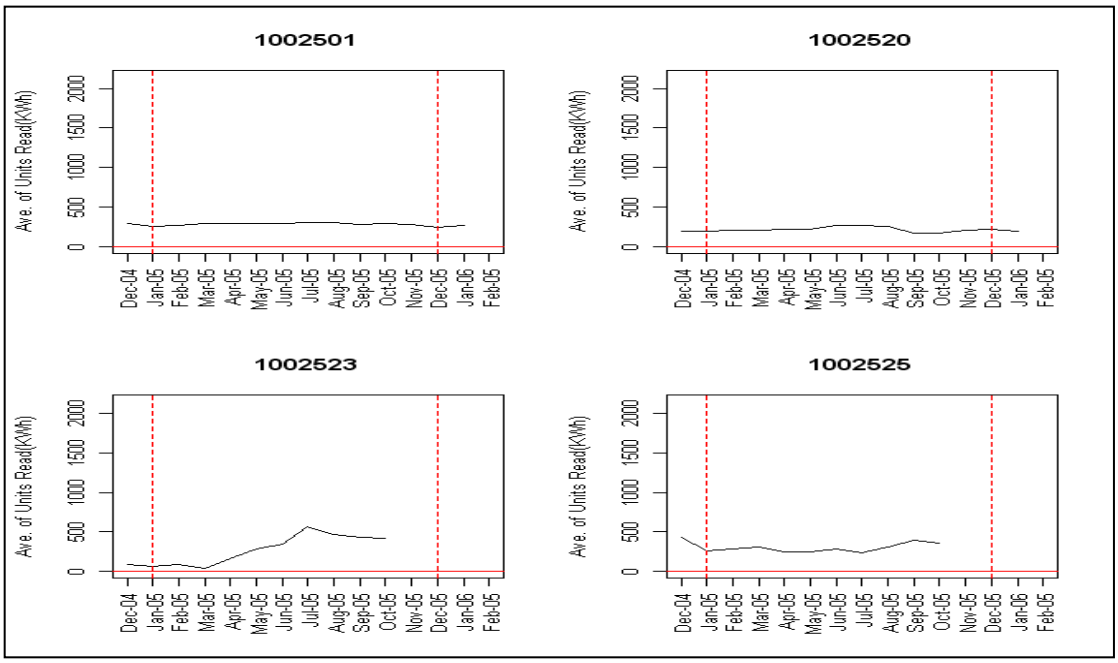


Figure 12: Observed monthly consumption

5.2 INDEPENDENT VARIABLES

5.2.1 Alternative Energy Sources

Table 3: Number of households that use the following alternatives energy sources for cooking and space heating

Alternative source	Task	Do not use	Use
Charcoal	Cooking	176	3
	Space Heating	179	0
Coal	Cooking	146	33
	Space Heating	146	33
Gas	Cooking	174	4
	Space Heating	178	1
Paraffin	Cooking	121	58
	Space Heating	144	35
Wood	Cooking	151	28
	Space Heating	135	44

5.2.2 Appliances

In Table 4: the rows show types of appliances and the column number shows the number of units of that type of appliance. For example, the number 148 in the white grid shows that 148 households had no deep freezers; while the number 3 adjacent to the 148 shows that there were 3 households with 1 deep freezer.

Table 4: Number of households with the following number of units of appliances

Appliances	Number of units per type				
	0	1	2	3	4
Deep Freeze	148	3	1		
Fridge Freezer	51	123	4		1
Geyser		44			
Heater		22			
Hotplate		52			
Iron		133	1		
Kettle		124			
Microwave		55			
3-plate Stove		19			
4-plate Stove		59			
Tumble Drier		12			
Hi-Fi Radio		132	5	2	
TV		116	13	1	
Washing Machine		54			
Other		13	3		

In Table 5: the rows show types of appliances and the column number shows the number of **broken units** of that type of appliance. For example, the number 44 in the white grid shows that out of the 44 households that have a geyser (this can be verified from Table 4), none of the households had broken geysers. The number 21 in second row and first column shows that out of the 22 households that had heaters (as seen from Table 4), 21 had non-broken heaters and 1 household had a broken heater.

Table 5: Number of households with the following number of broken units of appliances

Appliances	Number of broken units per type	
	0	1
Geyser	44	0
Heater	21	1
Hotplate	52	0
Iron	134	0
Kettle	12	4
Microwave	5	3
3-plate Stove	17	2
4-plate Stove	56	3
Tumble Drier	12	0
Washing Machine	54	0
Other	15	0

In Table 6: the rows show types of appliances and the columns shows the frequency at which the type of appliance is used. For example, the number 43 in the white grid shows that out of the 44 households that have a geyser (this can be verified from Table 4), 43 of the households used the geysers on a daily basis and 1 household used the geyser on a weekly basis. The number 7 in the second row and first column shows that out of the 22 households that had heaters (as seen from Table 4), 7 households never use their heaters. Reader can proceed with this kind of interpretation.

Table 6: Number of households with the following frequencies of usage of the types of appliance

Appliances	Frequency of usage of type of appliance			
	Never	Monthly	Weekly	Daily
Geyser	0	0	1	43
Heater	7	8	4	2
Hotplate	6	1	9	36
Iron	1	0	12	13
Kettle	0	0	14	16
Microwave	0	0	12	38
3-plate Stove	0	0	2	15
4-plate Stove	1	1	7	47
Tumble Drier	2	4	6	0
Washing Machine	4	0	49	1
Other	2	7	4	3

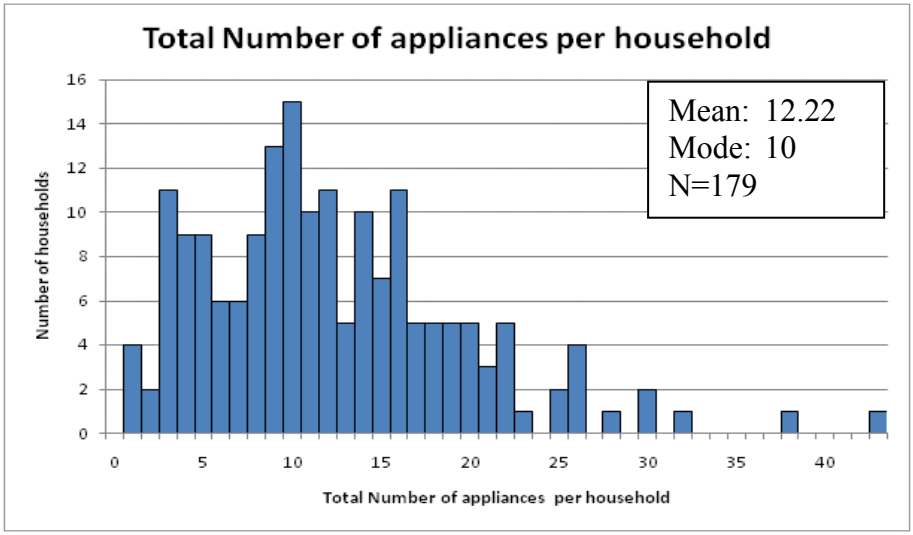


Figure 13: Empirical distribution of total number of appliance per households

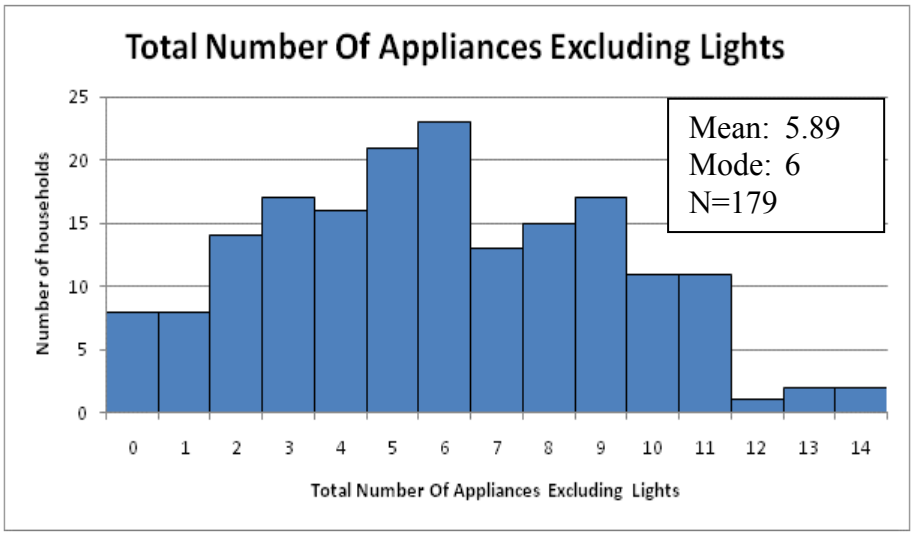


Figure 14: Empirical distribution of total number of appliance excluding lights per households

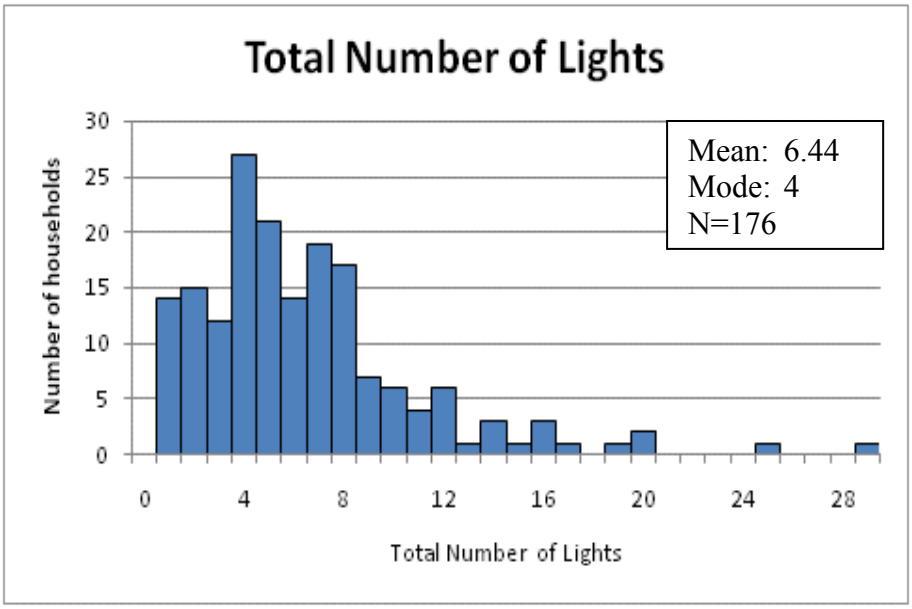


Figure 15: Empirical distribution of total number of lights per households

5.2.3 Cooking Habits

The numbers on white grid indicate total of number of households for which the column number on the row applies. For instance, element[1, 2] indicates that 24 households cook 2 breakfast meals for adults on Saturdays. Adults were defined as the people older than 24 years of age.

Table 7: Total number of households which prepare the shown number and type of meals

Member	Day	Type	1	2	3	4	5	6	7	8
Adults	Saturday	Breakfast	1	24	19	7	1	3	2	
		Dinner	1	25	22	1	2	3	2	
		Lunch	1	33	21	7	1	3	2	
	Sunday	Breakfast	1	23	16	9	1	3	2	
		Dinner	1	2	17	9	1	3	2	
		Lunch	11	33	25	12	2	3	2	
	Week Day	Breakfast	19	18	15	1	1	3		
		Dinner	13	46	26	12	2	3	1	
		Lunch	18	17	17	2	3			
Children	Saturday	Breakfast	17	18	1	6	1			1
		Dinner	18	18	11	6	1			1
		Lunch	2	18	12	7	1			1
	Sunday	Breakfast	17	15	1	5	1			1
		Dinner	16	13	1	5	1			1
		Lunch	2	25	15	6	1			1
	Week Day	Breakfast	18	15	6	6	1			
		Dinner	22	29	15	7	1			1
		Lunch	16	11	6	6	1			

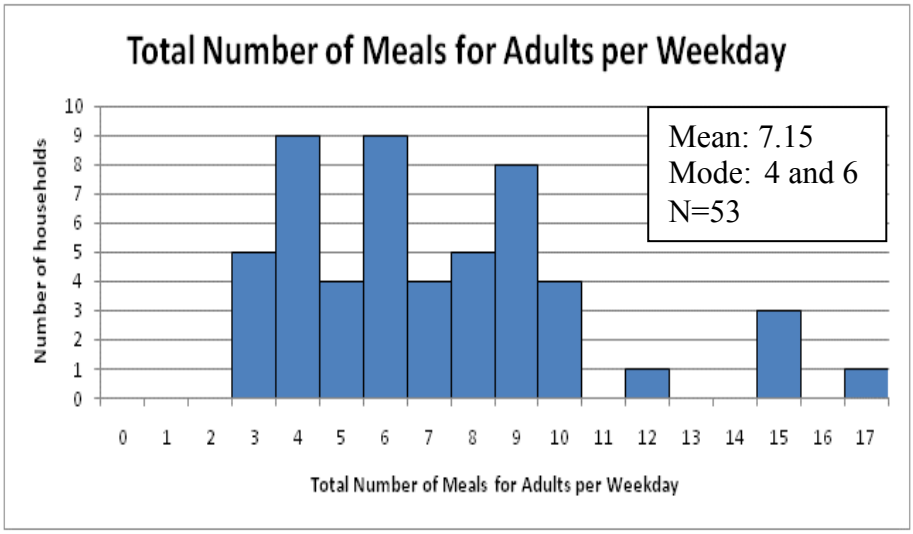


Figure 16: Empirical distribution of total number of adult meals per weekday per households

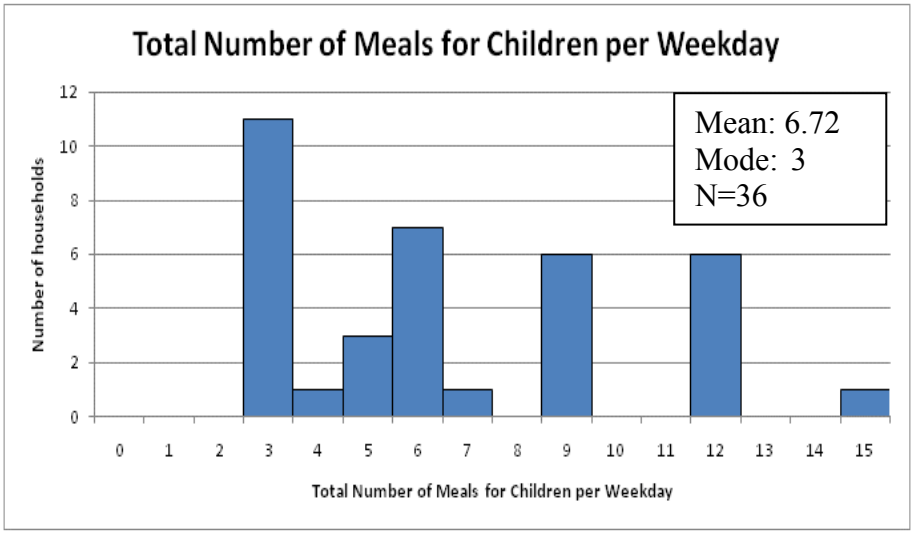


Figure 17: Empirical distribution of total number of child meals per weekday per households

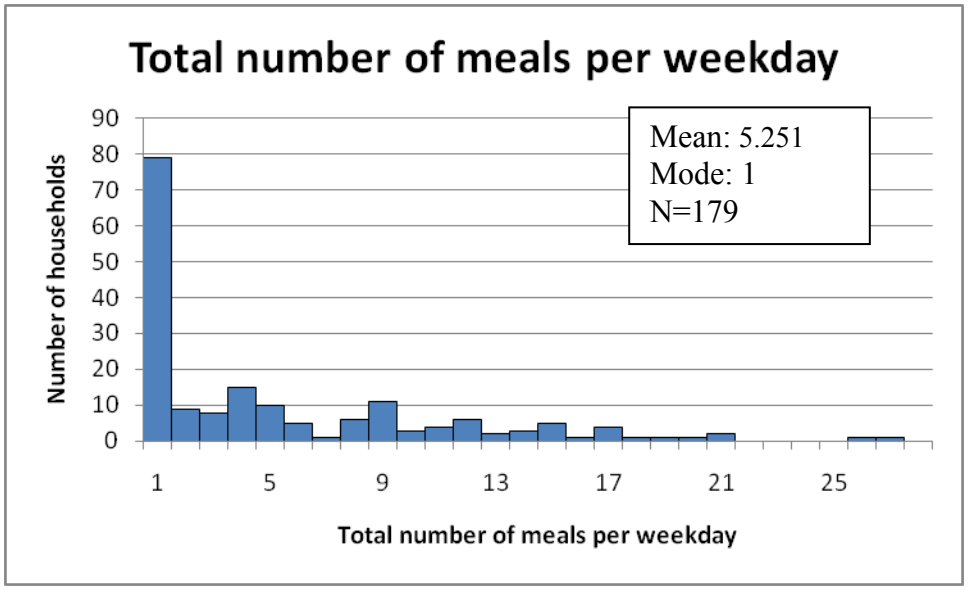


Figure 18: Empirical distribution of total number of meals per weekday per household

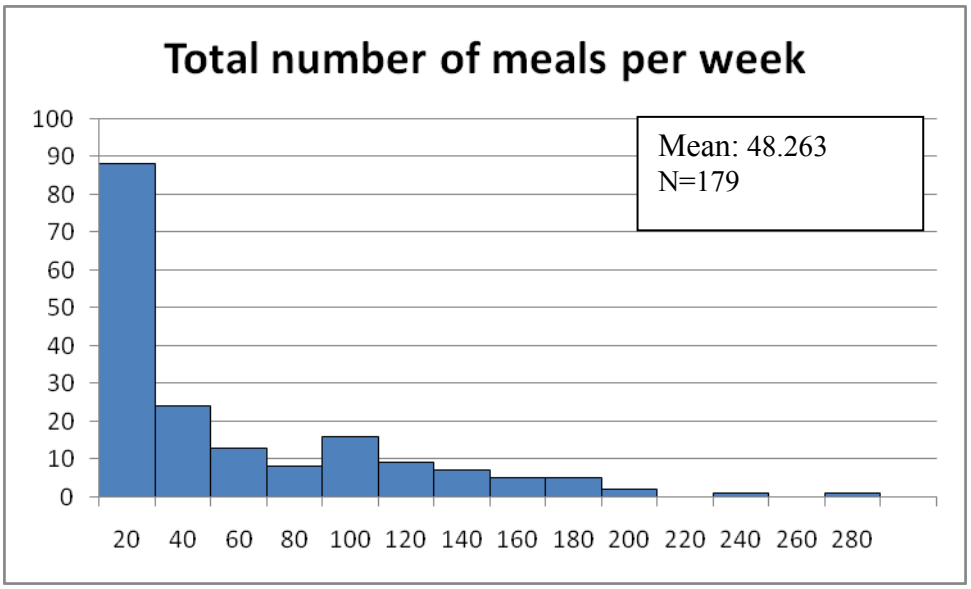


Figure 19: Empirical distribution of total number of meals per week per household

5.2.4 Education Profiles

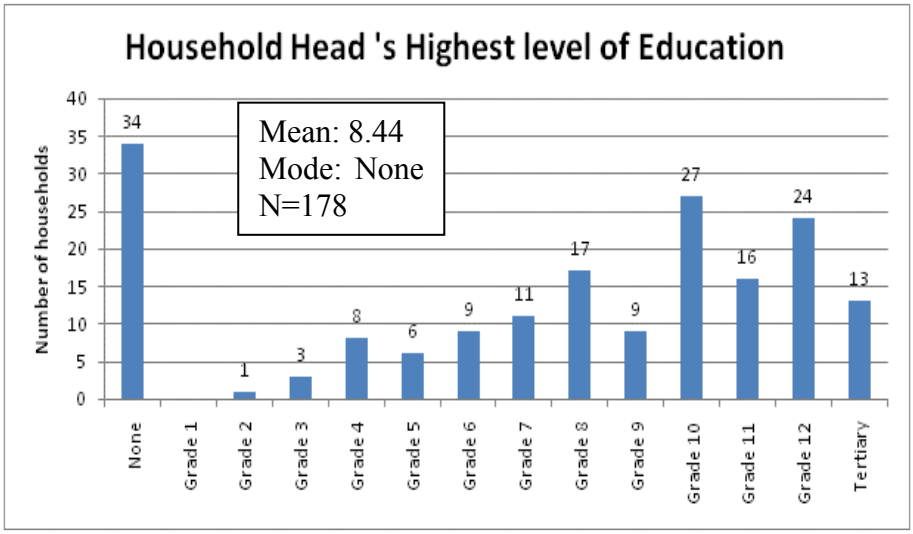


Figure 20: Empirical distribution of highest level of education attained by head per households

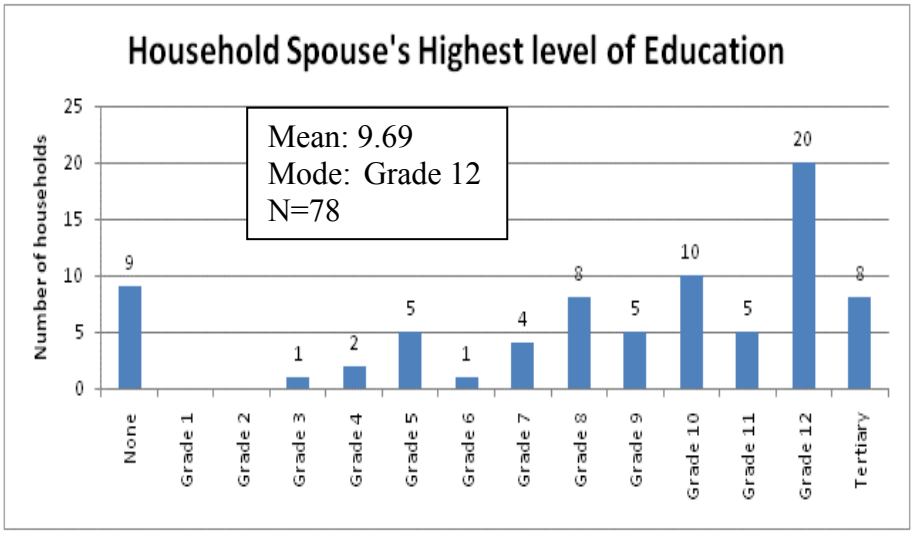


Figure 21: Empirical distribution of highest level of education attained by spouse per households

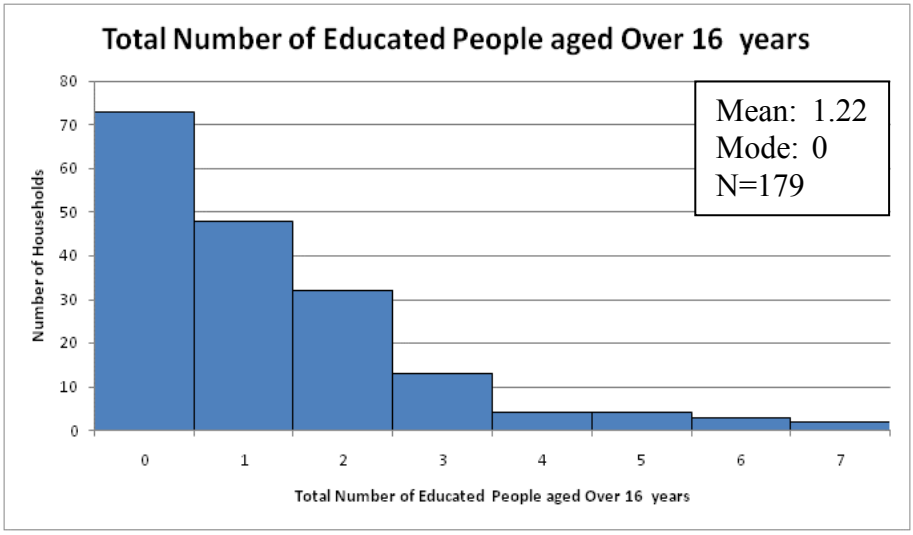


Figure 22: Empirical distribution of total number of educated people aged over 16 years per households

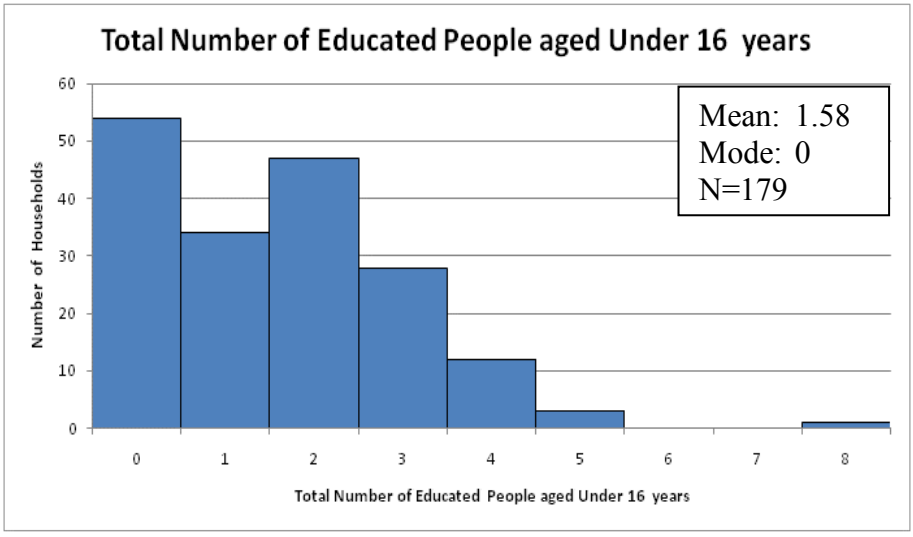


Figure 23: Empirical distribution of total number of educated people aged less than 16 years per households

5.2.5 Employment Profiles.

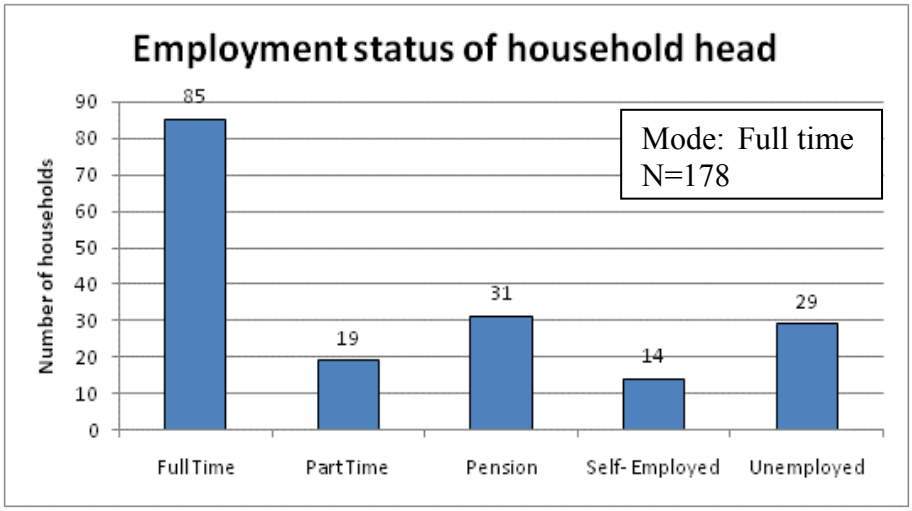


Figure 24: Empirical distribution of employment type of head per households

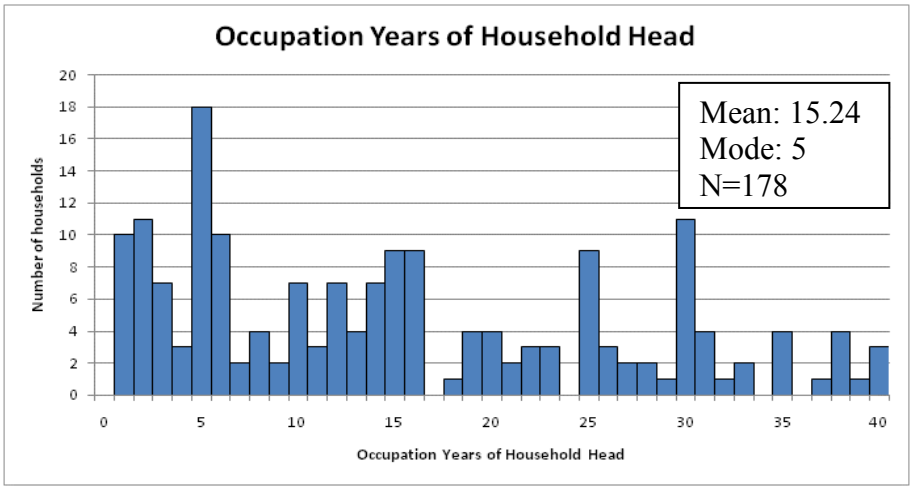


Figure 25: Empirical distribution of occupation years of head per households

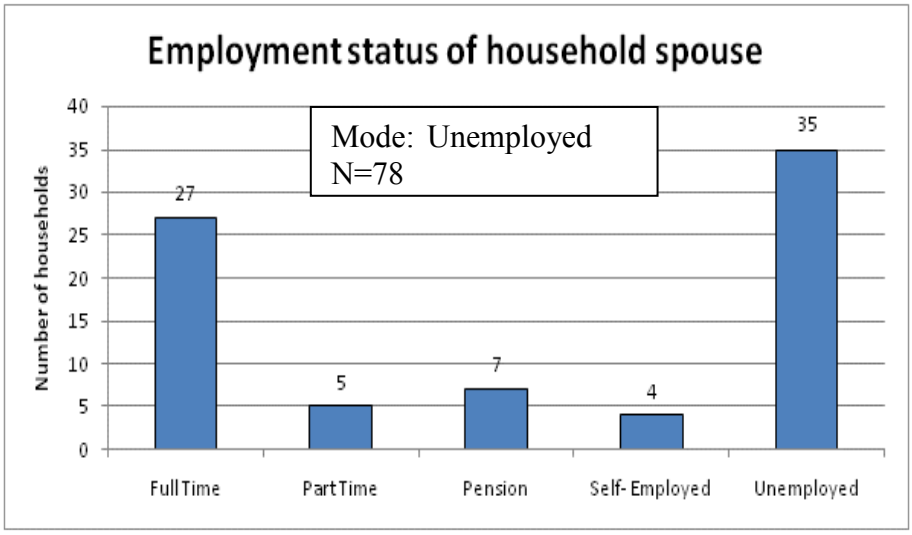


Figure 26: Empirical distribution of employment type of spouse per households

In Table 8, the rows show the employment status while the columns shows the number of people (aged over 16 years) in the house to whom the corresponding employment status applies.

Table 8: Employment status of people aged over 16 in household

	0	1	2	3	4	5	6	7
Full time	158	17	4					
Part time	164	14	1					
Pension	167	12						
Self employed	179							
Unemployed	86	53	24	7	4	3	1	1
TotalNumberOlder16Employed	139	29	9	2				

In all households, members aged 16 and below were unemployed.

5.2.6 Demographic Composition of Households

Table 9: Empirical distribution of total number of females across age groups per household

Females on age range(in years):	0	1	2	3	4	5	6	7	8	9	10	11
16-24	0	39	13	4	0	0	0	0	0	0	0	0
25-34	0	63	5	1	0	0	0	0	0	0	0	0
35-49	0	67	4	0	0	0	0	0	0	0	0	0
>=50	0	54	2	1	0	0	0	0	0	0	0	0
<=16	0	58	21	6	2	1	0	0	0	0	0	0
Total Females	0	14	47	44	41	18	8	4	0	2	0	1

Table 10: Empirical distribution of total number of males across age groups per household

Males on age range(in years):	0	1	2	3	4	5	6	7	8
16-24	138	32	6	3	0	0	0	0	0
25-34	130	46	3	0	0	0	0	0	0
35-49	123	55	1	0	0	0	0	0	0
>=50	155	24	0	0	0	0	0	0	0
<=16	84	58	22	11	3	1	0	0	0
Total Males	20	59	52	27	15	4	1	0	1

Notice that the row "Total Males" in Table 10 is not the sum of the rows above because, for example in the first column; 138 households had no male members in the 16-24 age group, 130 households had no male members in the 25-34 age group, etc , but only 20 households had no male members in total (across age groups).

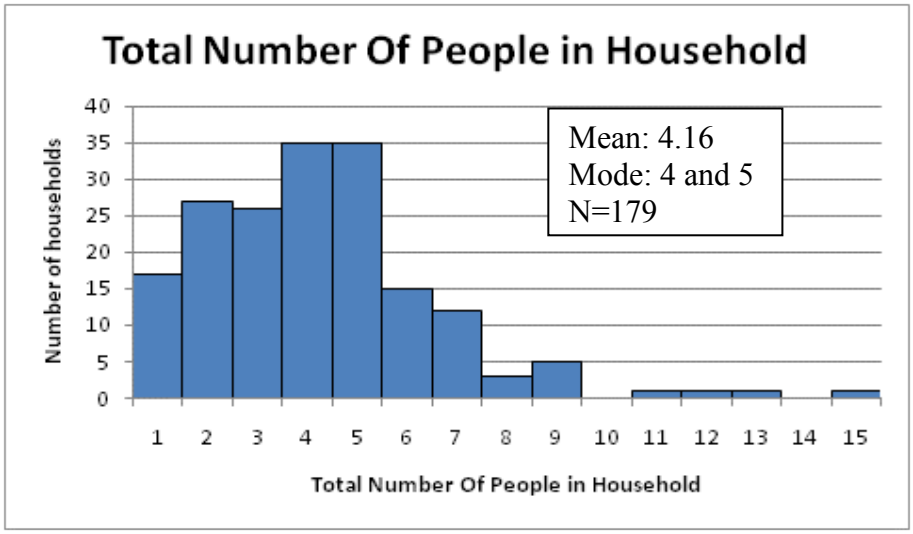


Figure 27: Empirical distribution of total number of people in household

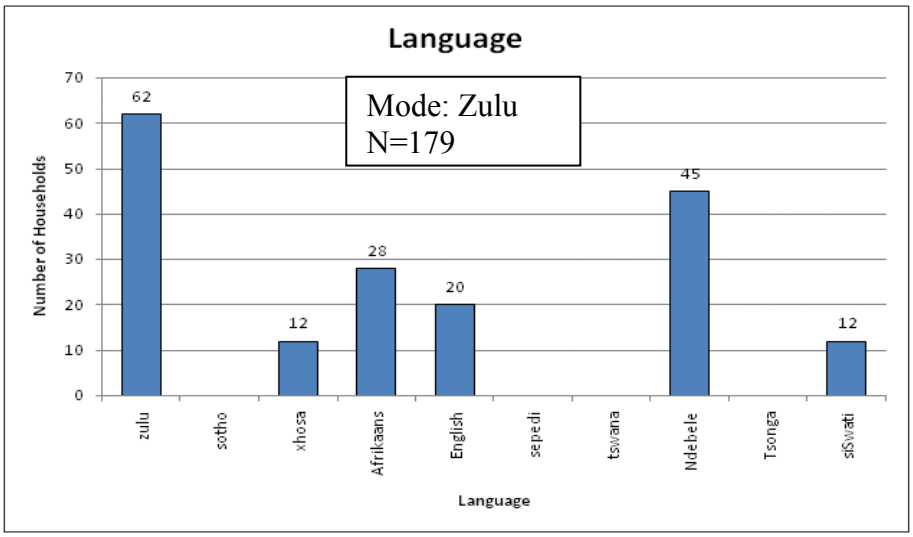


Figure 28: Empirical distribution of home language in household

Table 11: Empirical distribution of head of household

	No	Yes
Is the head of household male?	78	101

5.2.7 Income profiles

Table 12: Empirical distribution of income sources per household

	No	Yes
Is there someone in the house who receives income from an agricultural business?	177	2
Did the interviewee refuse to disclose their income?	176	3
Is there someone in the house who receives income from child grant or a small business?	126	53
Are there any other external sources of	164	15

income?		

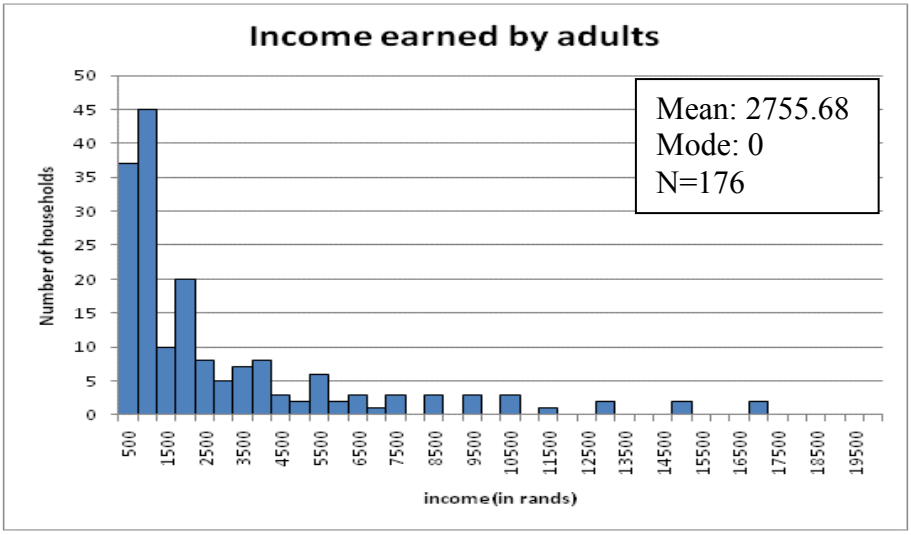


Figure 29: Empirical distribution of level of income earned by adults in household

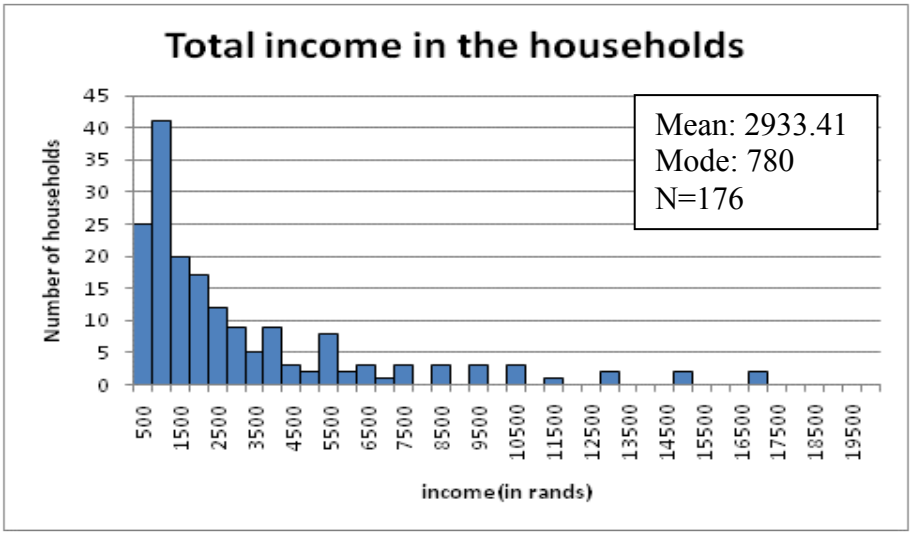


Figure 30: Empirical distribution of level of total income (earned from all) in household

Users of energy models are more interested in using income groups when clustering households (with regards to their usage). Thus, in this analysis, income groups of R500 intervals were used as opposed to actual income levels. Thus income related variables were re-coded into (cardinal) categories as follows:

0=R500, 1=R500-R1000,… and 34=R17000- R17500

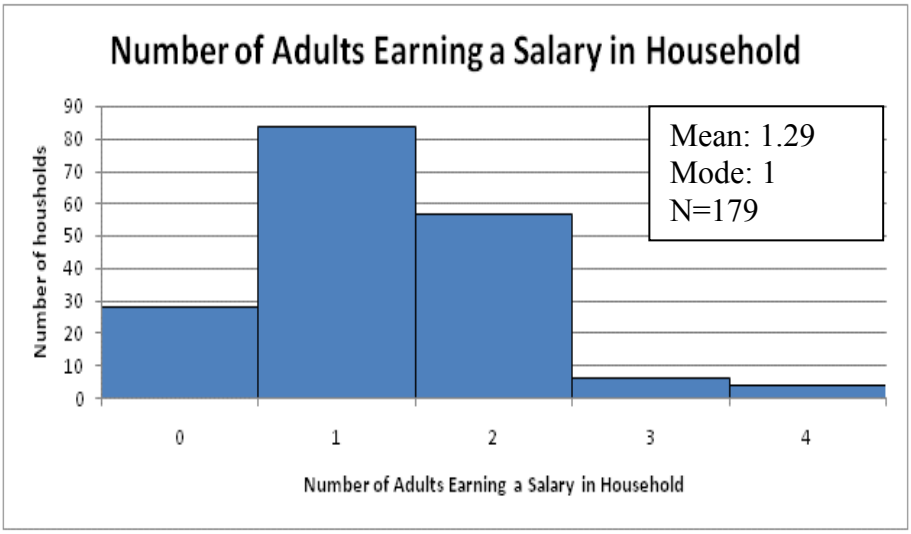


Figure 31: Empirical distribution of number of adults earning a salary in household

5.2.8 Frequency of Different Power Related Problems

Table 13: Frequency of different power related problems across households

Problem	Never	Monthly	Weekly	Daily
How often does electricity go off in the house or general power failure in the entire township?	153	26		
How often do lights get dim at night but the power not tripping completely?	164	5	8	2
How often does the main switch trip or circuit break in the house?	169	6	1	3

5.2.9 Other Variables

Table 14: Empirical distributions of main switch ampere recording

	20 Ampere	60 Ampere
Ampere reading on main switch.	18	161

Table 15: Empirical distribution of other house properties

	No	Yes	Unknown
Own Dwelling?	17	162	0
Is there a business, which uses electricity, run from this home?	177	1	0
Are there other buildings to which electricity is supplied besides the main house?	147	32	0
Does the house have a ceiling?	124	55	0
Does the house have insulation?	176	1	2

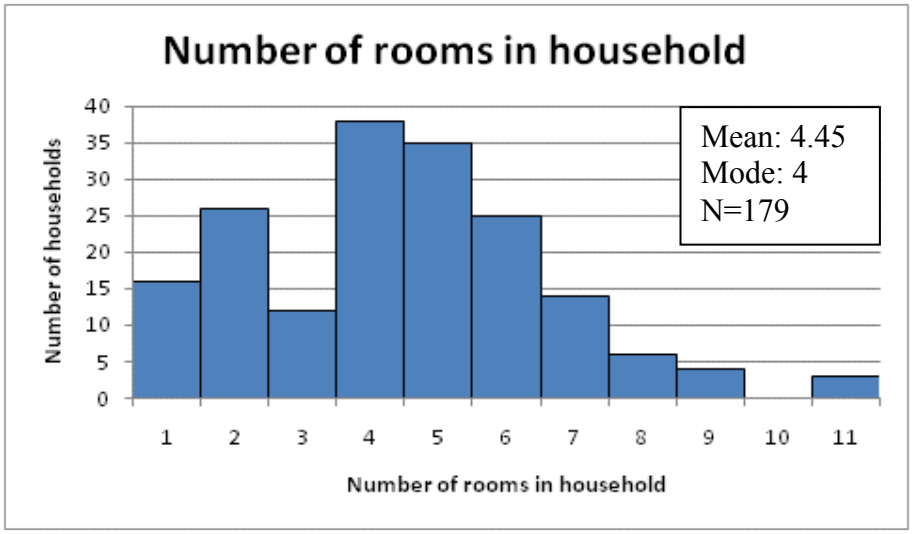


Figure 32: Empirical distribution of number of rooms per household

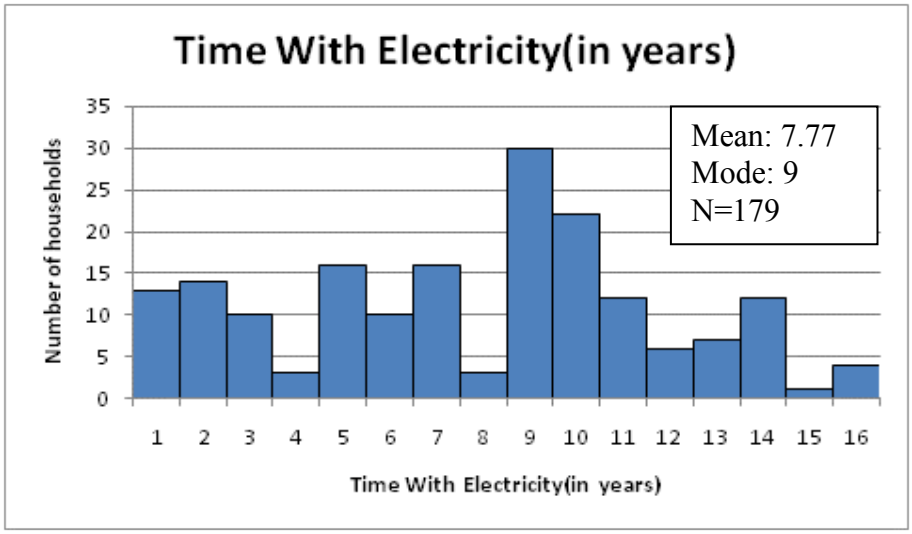


Figure 33: Empirical distribution of time period (in years) over which household has had electricity

5.3 INDEPENDENT VARIABLES VS. RESPONSE VARIABLE

Due to the large number of independent variables included in the study, this exploration section will **not** be done for all independent variables.

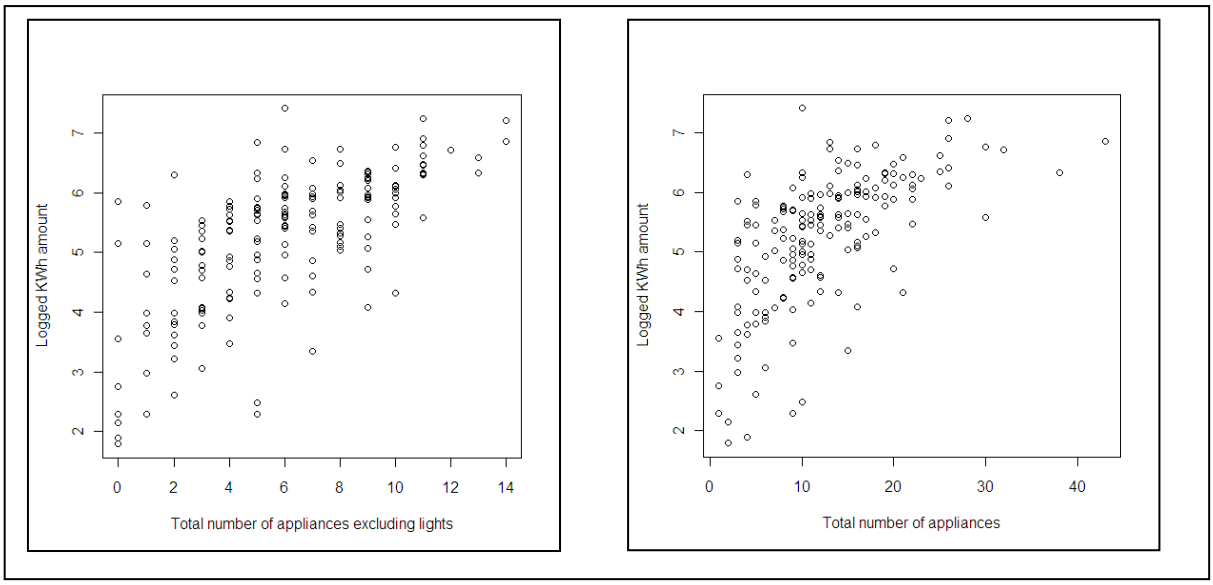


Figure 34: Empirical association between logged KWh amount and number of appliances

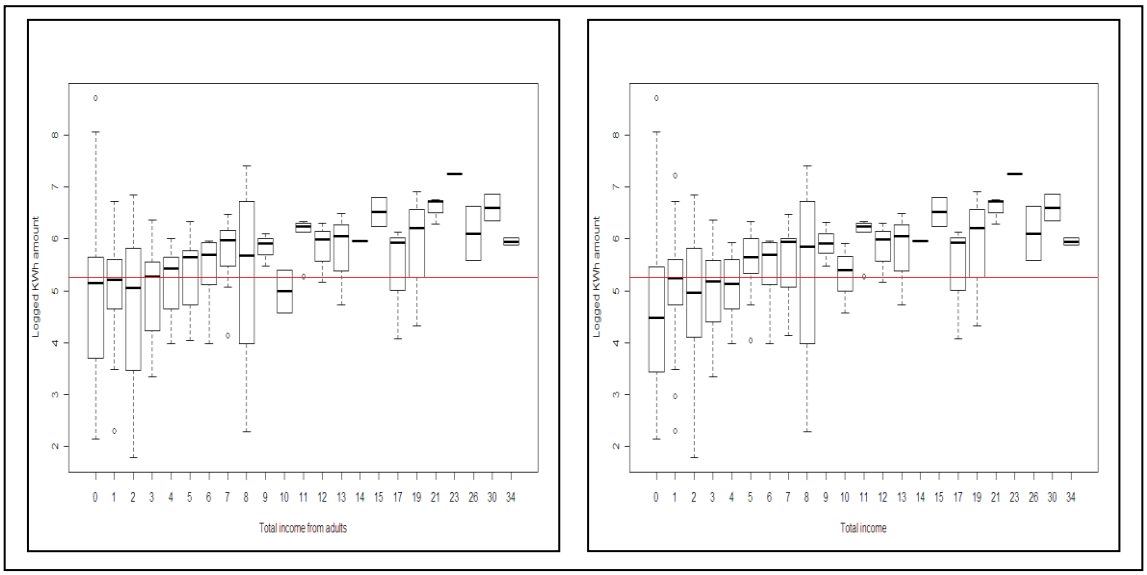


Figure 35: Empirical association between logged KWh amount and income

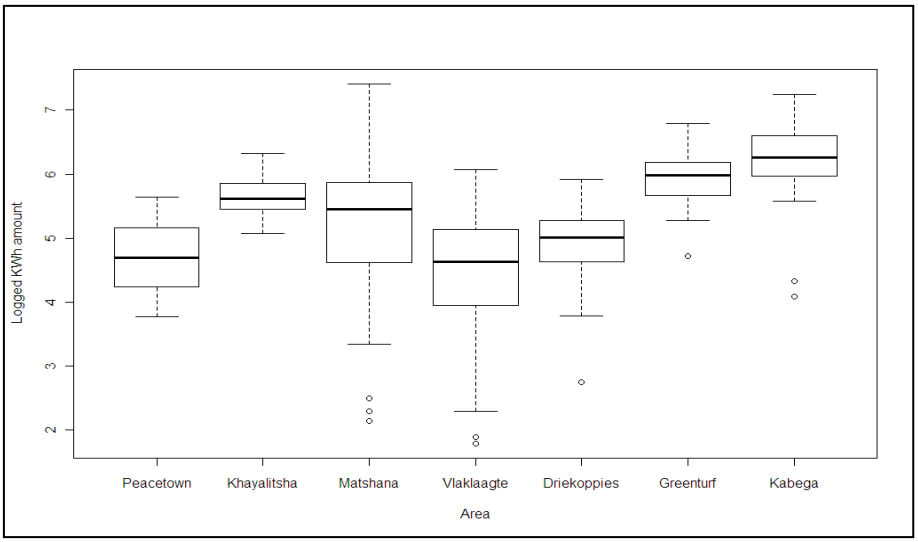


Figure 36: Empirical association between logged KWh amount and area of residence

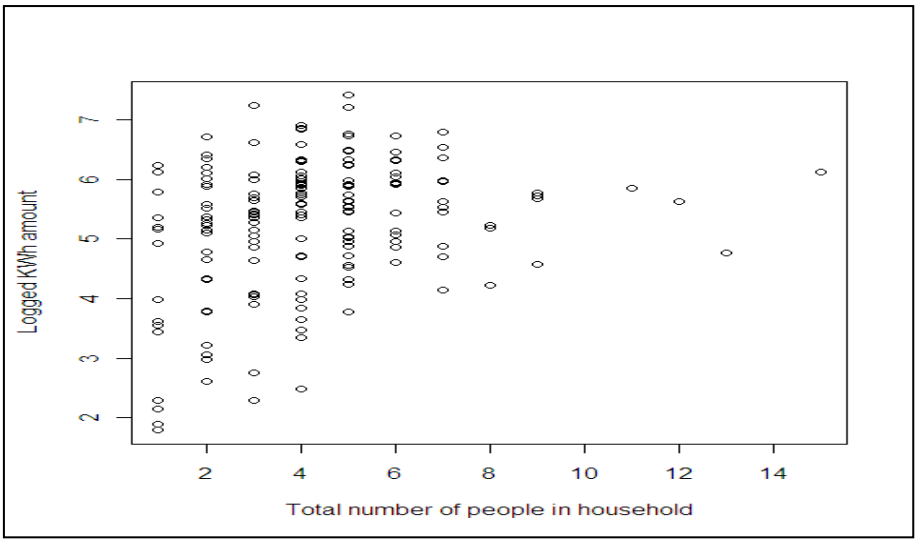


Figure 37: Empirical association between logged KWh amount and total number of people in household

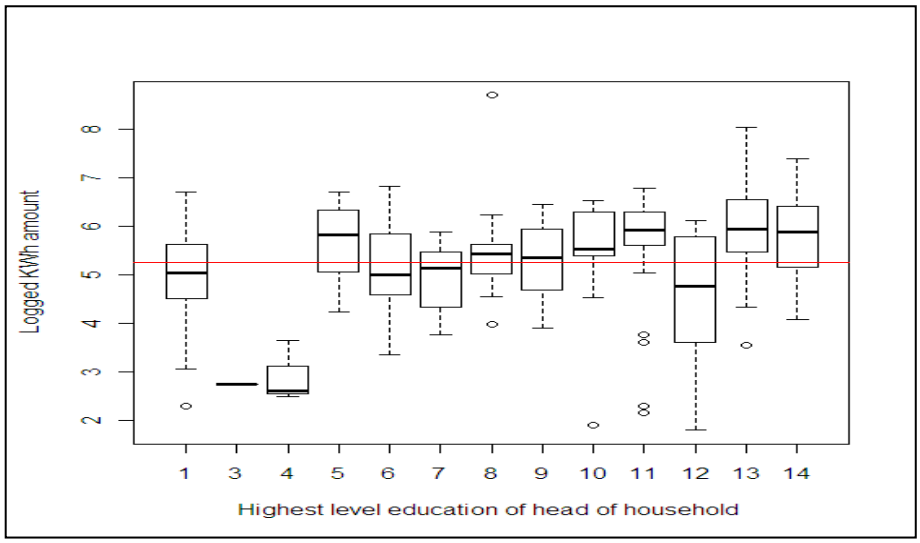


Figure 38: Empirical association between logged KWh amount and the highest level of education by head

6. ANALYSIS

6.1 CART :THE `rpart()` METHOD

6.1.1 Tree building algorithm (with outliers)

For demonstration purposes, Figure 39 was obtained with the minimum number of cases in a node for a split to occur being 2.

Step 1: The maximum tree

Construct the maximum possible tree. The size of the tree was used as an aid to determine the form of the response variable that is most suited for further analysis. What is meant by this is clearer on Figure 39 below.

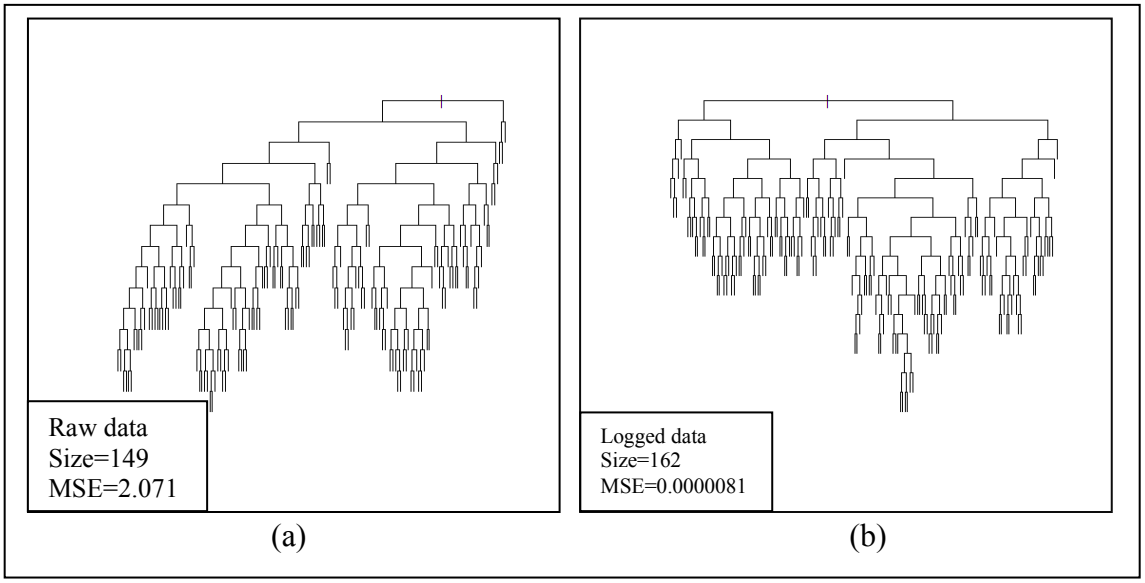


Figure 39: The maximum tree-raw data vs. logged data with outliers (min. number of cases for split = 2)

The tree size for logged data (Figure 39 (b)) is larger and that shows that the algorithm was efficient in detecting associations with logged data. Thus, for the algorithm, we shall take the response as the logged data

Yet it would not be reliable to base splits on two cases, thus we shall consider use the minimum number of cases as 10. As a result, the following maximum tree was obtained

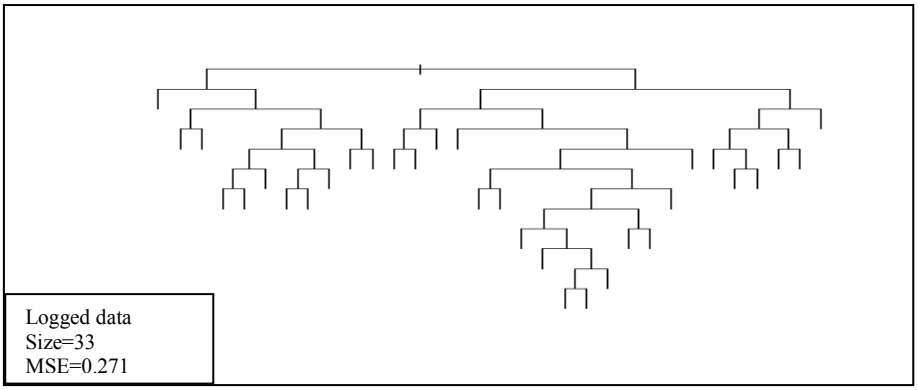


Figure 40: The maximum tree with outliers (min. number of cases for split = 10)

Step 2: Pruning the maximum tree

The maximum tree (with 33 leaves) might be too large to use in practice. As a result, the tree needs to be pruned so as to remove the branches that are least associated with the response variable.

The `rpart` function uses the cost-complexity method as described under part a) in section 4.2.5, and the mean residual deviance is estimated using a 10-fold cross-validation. Figure 41 shows a sample of the results from the cross-validation process.

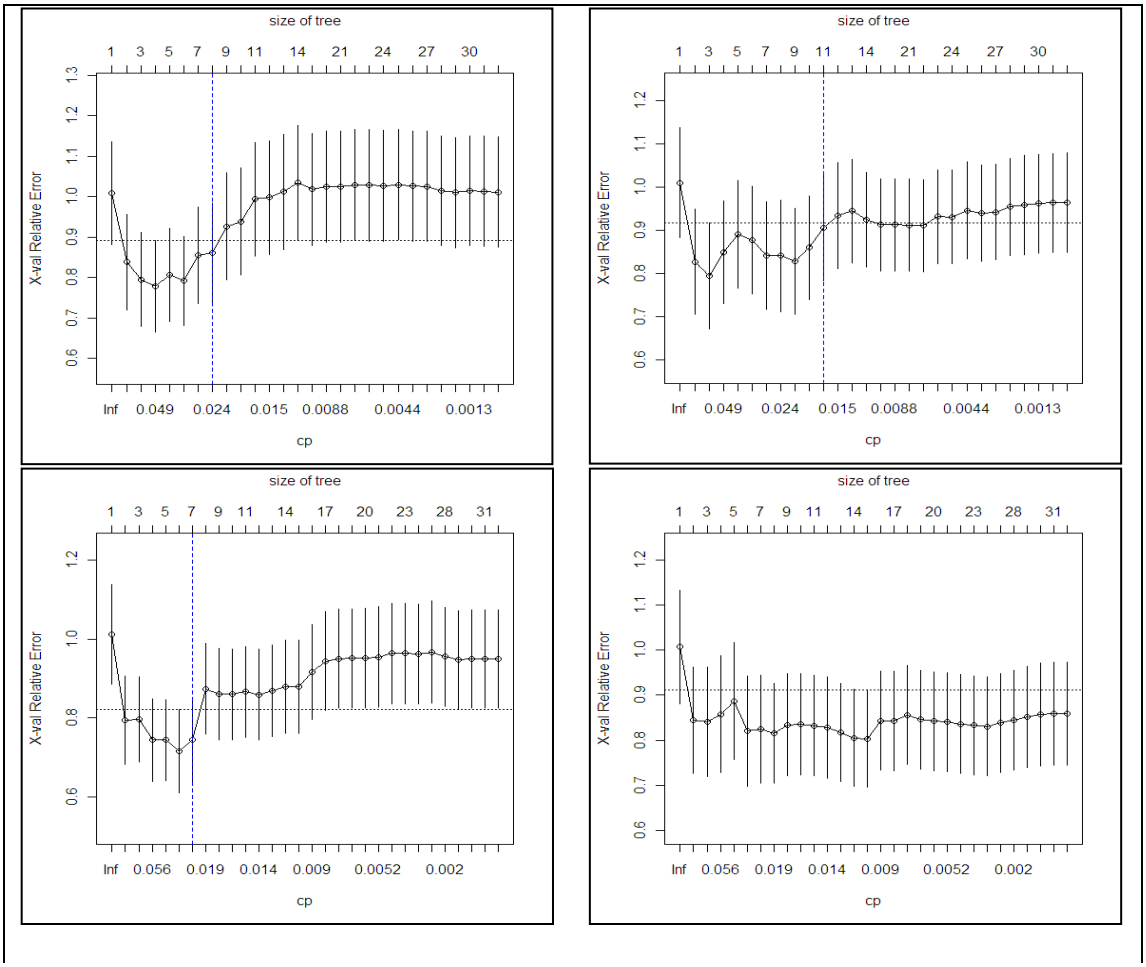


Figure 41: Cross-validation results (with outliers)

On the cross-validation plots, the bottom axis shows the values of the scaled complexity parameter (cp) while the top axis shows the best-sized tree corresponding to each value of the scaled complexity parameter. The scaled parameter is calculated as

$$\text{cp} = \frac{\alpha}{\text{variance of the root node}}$$

The vertical axis shows, for each size and cp-value, the

$$\text{Relative error} = \frac{\text{residual mean deviance of current tree}}{\text{residual mean deviance of tree with root node only}}$$

Cross-validation was done 50 times, and the most frequent "best size" by the 1-SE rule was a tree of size 9.

Step 3: Optimal splits and their surrogates

At each node - unless specified otherwise - if the statement is true for a case, the case goes to the left otherwise to the right. For cases that have missing values for the optimal splits, the surrogate splits on Table 16: Surrogate splits to preliminary tree (with outliers)Table 16.

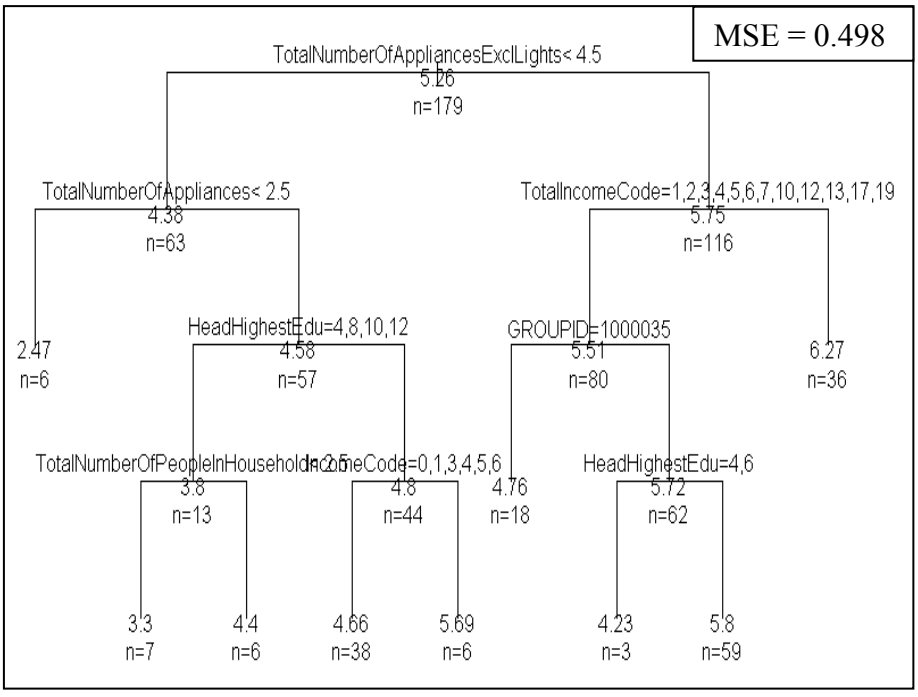


Figure 42: The preliminary optimal tree (with outliers)

Table 16: Surrogate splits to preliminary tree (with outliers)

Optimal split	Surrogate splits	
Total number Of appliances (excluding lights) ≤4?	Total number Of non-broken appliances (excluding lights) Total number of appliances Number of kettles Number of fridge freezers Number of TVs	≤ 4 ≤ 8 ≤1 ≤1 <1
Total number of appliances ≤ 2?	Total number Of non-broken appliances (excluding lights) Total number Of appliances (excluding lights) Wall Material Code = Corrugated iron	=0 =0

Total income code ∈{1,2,3,4,5, 6,7,10,12,13,17,19}	Income code ∈ { 0,8,9,11,14,15,18,19,20,21 } to the right Number of lights ≤11 Total number of appliances ≤20 Total number Of non-broken appliances (excluding lights) ≤10 GROUPID = 1000039 to the right
Head's highest education ∈ {4,8,10, 12}	Income from child grant or small business < 275 to the right GROUPID = 1000033 Gas used for cooking? = yes Total number Of non-broken appliances (excluding lights) ≤1 Total number of appliances (excluding lights) ≤1
GROUPID = 1000035	Home Language = Ndebele Coal used for cooking? = yes Coal used for space heating? = yes Total number of appliances ≤9 Number of lunch meals prepared for children on a weekday ≤2
Total number of people in household ≤ 2	Number of educated people aged under 16 =0 Number of lights ≤3 Number of people aged under 16 =0 Number of females ≤1 Income code ∈ {0, 2}
Income code ∈ {0,1,3,4,5,6}	Total income code ∈ {2, 9} to the right Total number Of non-broken appliances (excluding lights) ≤1 to the right Total number Of appliances (excluding lights) ≤1 to the right Number of fridge freezers =0 Occupation Years < 1.5 to the right
Head's highest education ∈ { 4,6 }	Frequency at which kettle is used = Never Floor Area < 52 Total number Of appliances (excluding lights) ≤10 Frequency at which heater is used = Monthly to the right

6.1.2 Tree building algorithm (without outliers)

The above analysis was repeated without cases with the outlying values 6059.92KWh and 3135.033KWh (as observed from Figure 9 and Figure 10) and the following results were obtained. These values were assessed as outlying by energy experts.

Step 1: Maximum tree

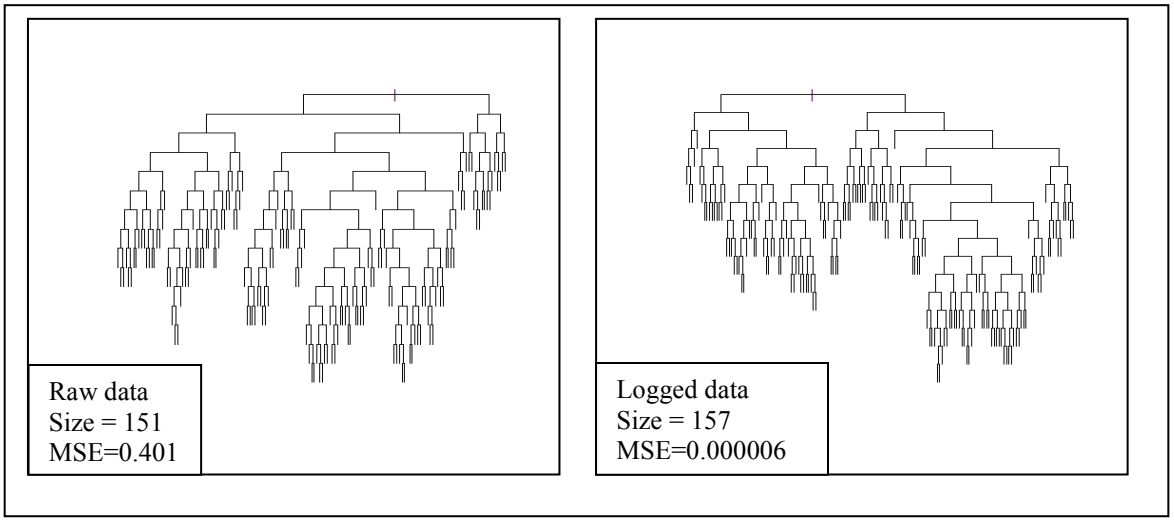


Figure 43: The maximum tree-raw data vs. logged data excluding outliers (min. number of cases for split = 2)

Again we will take the larger size of the tree corresponding to logged data as a hint that the tree was more efficient in detecting associations for between response and the independent variables when the data was less skewed. From here, we increase the

minimum size of a node for which a split is allowed to 10 and the resulting tree is shown in Figure 44.

Figure 45 (below) shows a few plots from the 10-fold cross-validation process. The optimal size for the tree was also found to be 9 but, recall, this is not a *veto* - i.e. one can choose a size that is most preferred.

Figure 46 and Table 17 show the optimal preliminary tree and corresponding surrogate splits (without outliers) respectively.

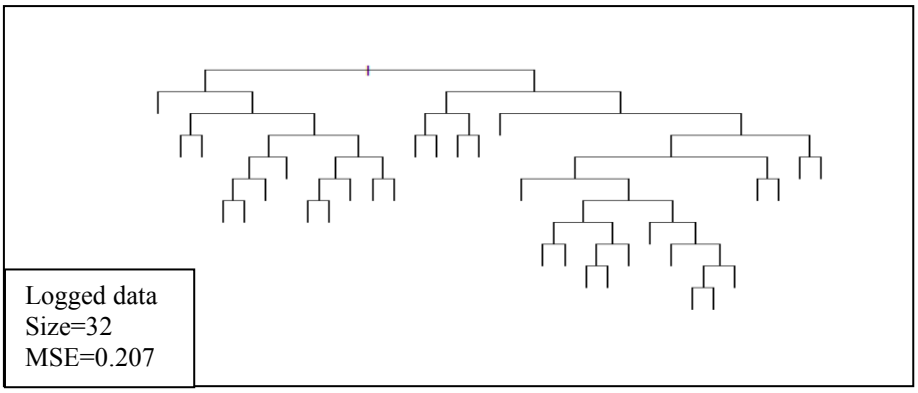


Figure 44: The maximum tree- excluding outliers (min. number of cases for split = 10)

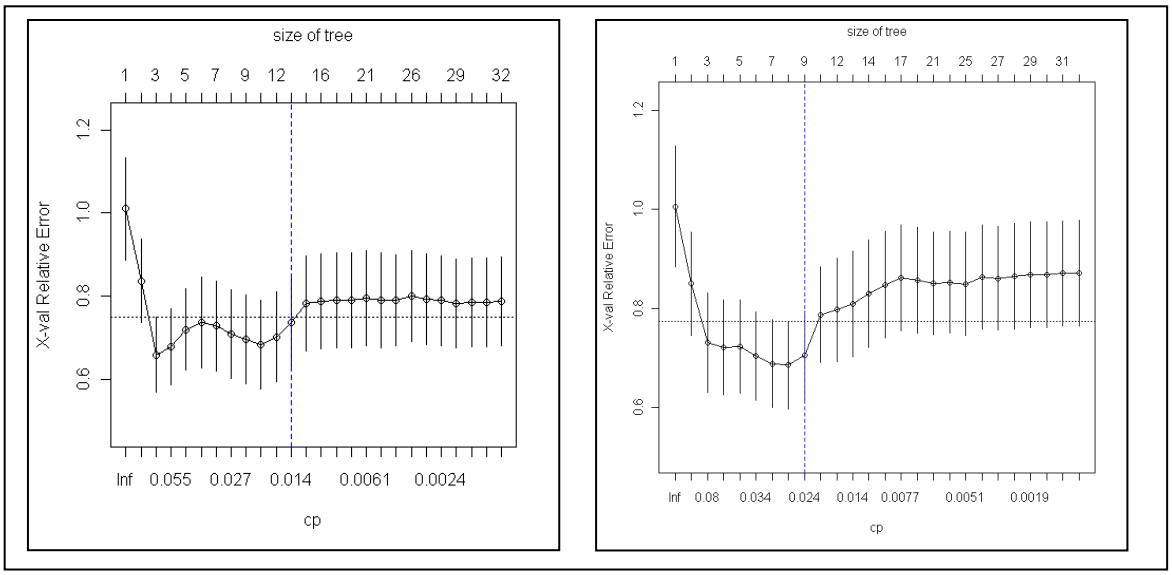


Figure 45: Cross-validation results (excluding outliers)

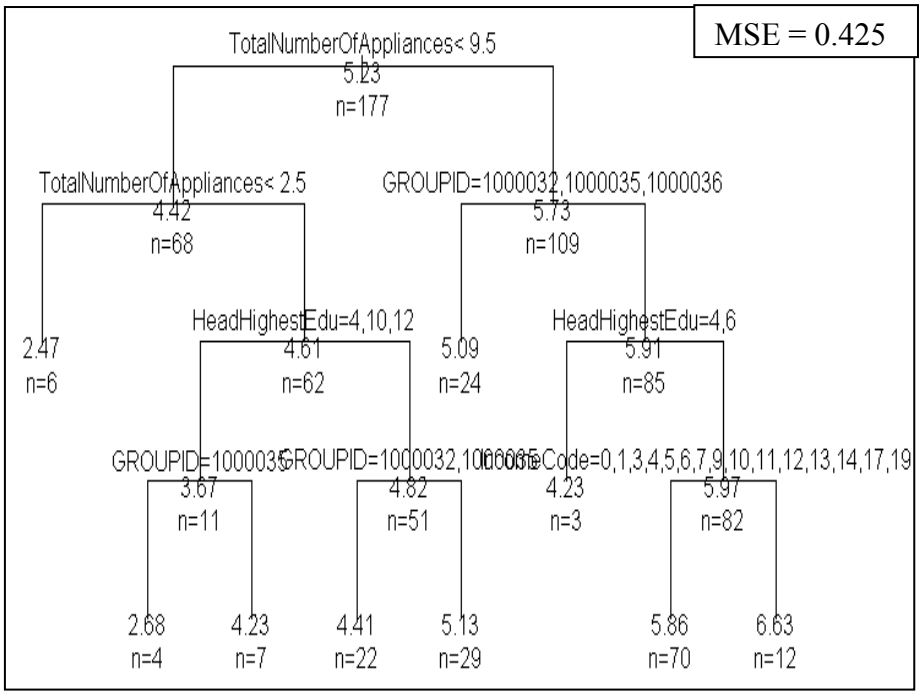


Figure 46: The preliminary optimal tree (excluding outliers)

Table 17: Surrogate splits to preliminary tree (excluding outliers)

Primary split	Surrogate split
Total number of appliances ≤ 9	Number of lights ≤ 4 Total number of appliances excluding lights ≤ 4 Total number of non-broken appliances excluding lights ≤ 4 Rooms < 3.5 Number of Fridge Freezer = 0
Total number of appliances ≤ 2	Total number of non-broken appliances excluding lights = 0 Total number of appliances excluding lights = 0 Wall material = Corrugated iron
GROUPID ∈ {1000032, 1000035, 1000036}	Language ∈ {Ndebele, siSwati} Coal used for cooking? = yes Coal used for heating? = yes Roof material code = Corrugated iron
Head's highest education ∈ {Grade3, Grade 9, Grade 11}	GROUPID = 1000032 Gas used for cooking? = yes Does house have ceiling? = yes
Head's highest education ∈ {Grade3, Grade 6}	Income < R7500 Total number of appliances excluding lights ≤ 10 Total income code < R 9500
GROUPID = 1000035	Hot water source = Coal fire Home language = English Coal used for cooking? = yes Coal used for heating? = yes Total number of people in household < 2.5
GROUPID ∈ {1000032, 1000035}	Home language = Ndebele Coal used for cooking? = yes Coal used for heating? = yes Time with electricity < 8 Paraffin used for heating? = no
Income code ∈ {0, 1, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 17, 19}	Total income < 8000 Total number of appliances < 27 Number of lights < 15.5 Other appliances = {Griller, electric blanket}

6.1.3 The final tree

Figure 42 was based on the data including outliers, thus the results are less credible than the tree that was obtained excluding the outlying observations. As a result, we shall take Figure 46 as the better preliminary tree.

The trees obtained from the `rpart` algorithm (Figure 42 and Figure 46) are regarded as preliminary trees because sometimes they might needs to be restructured in a way in that is easier to apply in practice.

For instance, in reference to the first child node on the right ("Total income code ∈ {0, 1, 3, 4, 5, 6, 7, 10, 12, 13, 17, 19}"), notice that total income codes 0, 8, 9, 11, 14, 15, 16 and 18 did not appear in the list. At this point of the tree, given the cardinal nature of the values of the income classes, it would be impractical to send cases with level of income that fall into categories 0, 8, 9, 11, 14, 15, 16 and 18 to the right child node (with a higher average consumption level) rather than to the left along with other cases whose level of income falls across the categories {1, 2, 3, 4, 5, 6, 7, 10, 12, 13, 17, 19}.

The reason the income classes 0, 8, 9, 11, 14, 15, 16 and 18 did not occur in the suggested split at this point is either because: among the group that has more than nine appliances, there are too few cases from the respective income groups (0, 8, 9, 11, 14, 15, 16 and 18) such that we can ignore that they were not included in split and conclude the real difference (in consumption of electricity) is between the cases that fall in and below 19th income class (i.e. with total income of R10 000 and below) and those that fall in classes above 19 (i.e. total income larger than R10 000). This is true for income classes like 15 and 18, in which there were absolutely no households that fell in these 2 classes.

Similar reasoning can be used for the highest level of education obtained by head: among the household that between 2 and 9 appliances in total, there were two Grade 3 respondents, three with Grade 9 and six with Grade 11 - i.e. a total of eleven out of 62, thus - with the data at hand - we just regard this as a difference (in electricity consumption) between households whereby the head has attained Grade 11 (at most) and those whereby the head attained at least Grade 11.

Figure 47 shows the final tree, with each parent node showing the optimal question (or rule) at that point in the tree, the average electricity consumption (in KWh) and the number of cases in the node. The leaves only show the average electricity consumption (in KWh) and the number of cases.

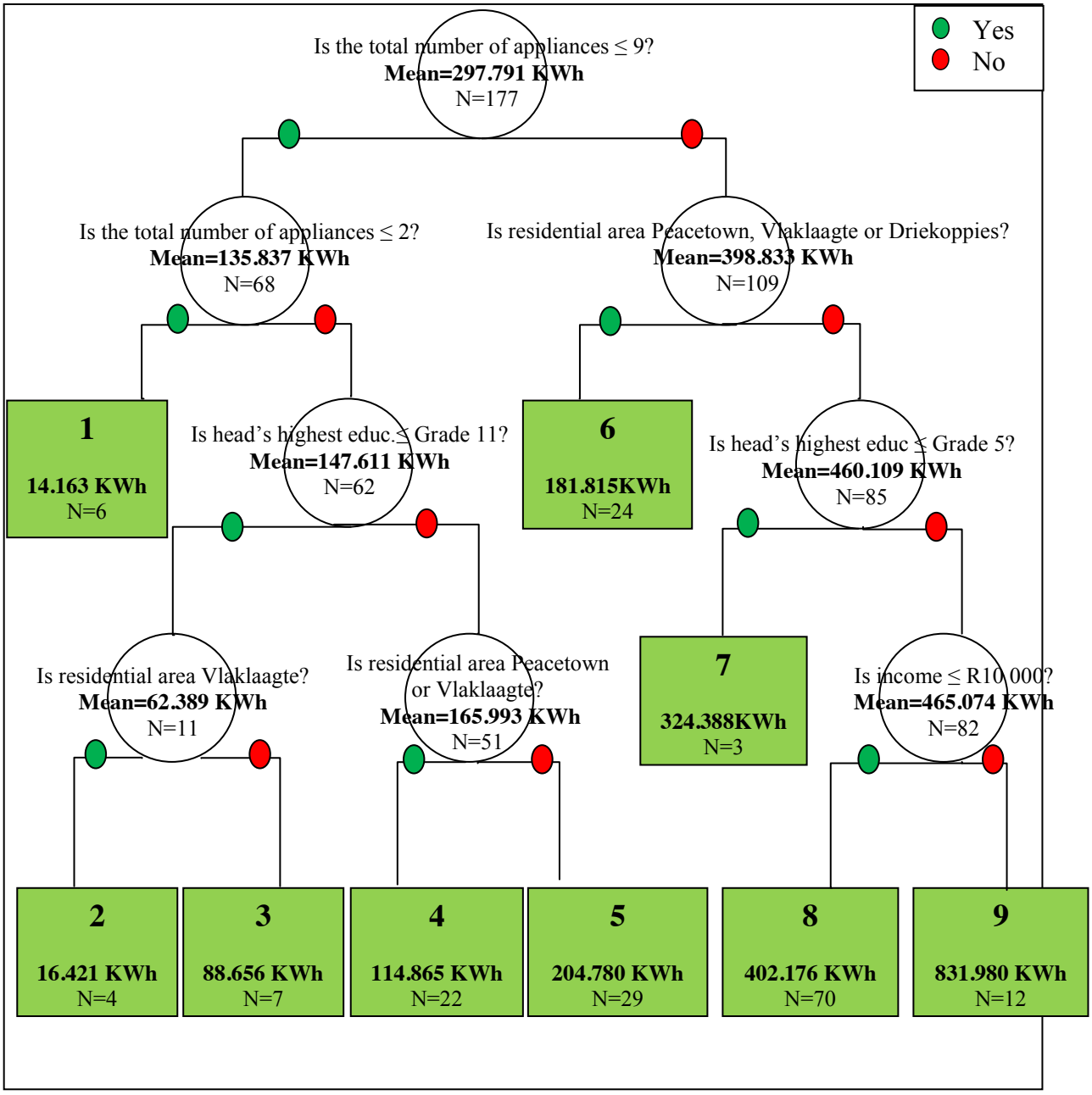
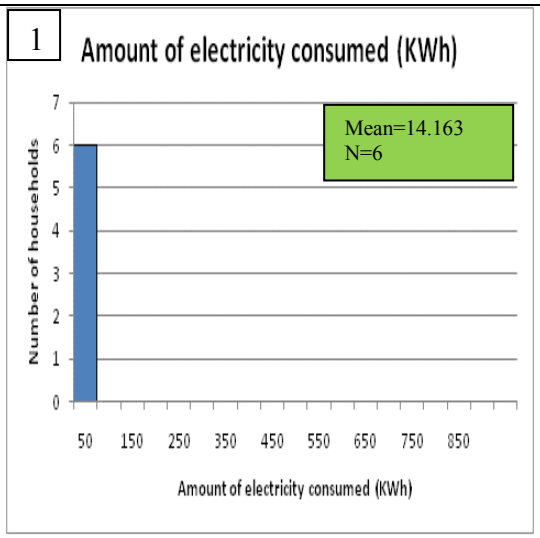


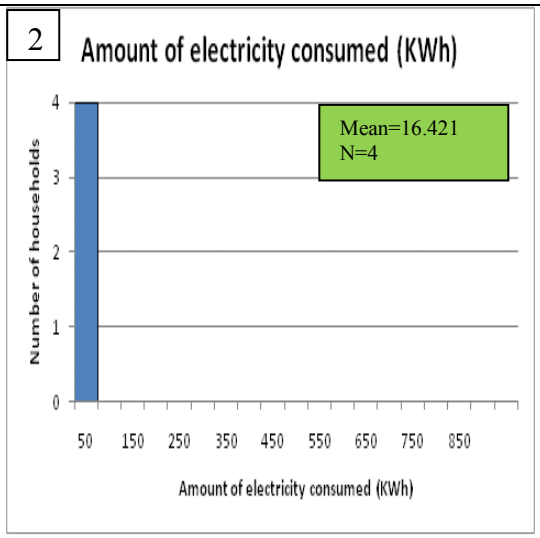
Figure 47: The final tree

leaves in the same numbering. For instance, plot 1 of Figure 48 shows the distribution of amount of electricity used within leaf 1.

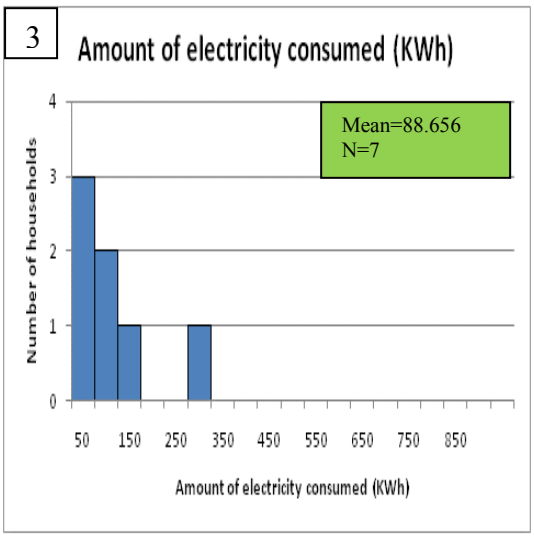
1


Amount of electricity consumed (KWh)
Mean=14.163
N=6
Number of households
0
1
2
3
4
5
6
7
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

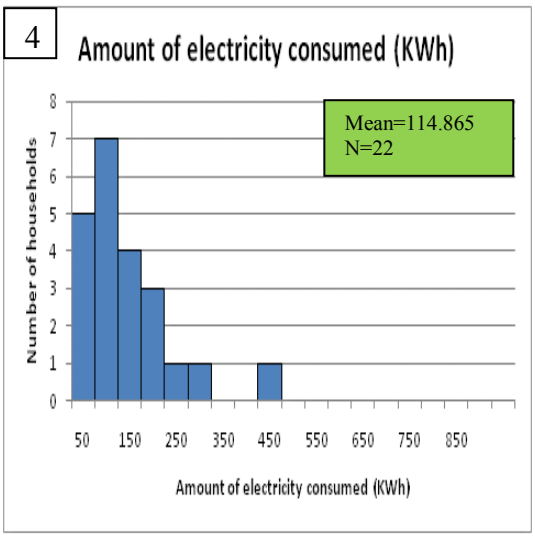
2


Amount of electricity consumed (KWh)
Mean=16.421
N=4
Number of households
0
1
2
3
4
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

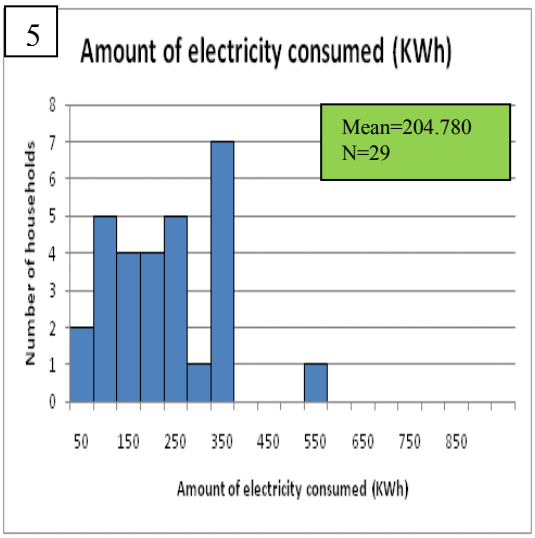
3


Amount of electricity consumed (KWh)
Mean=88.656
N=7
Number of households
0
1
2
3
4
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

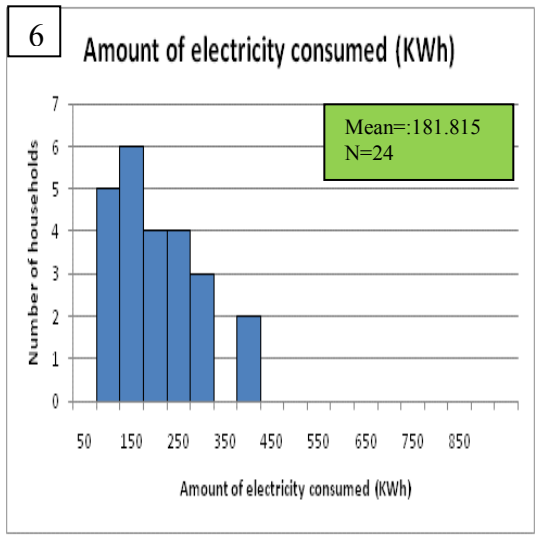
4


Amount of electricity consumed (KWh)
Mean=114.865
N=22
Number of households
0
1
2
3
4
5
6
7
8
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

5


Amount of electricity consumed (KWh)
Mean=204.780
N=29
Number of households
0
1
2
3
4
5
6
7
8
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

6


Amount of electricity consumed (KWh)
Mean=:181.815
N=24
Number of households
0
1
2
3
4
5
6
7
50
150
250
350
450
550
650
750
850
Amount of electricity consumed (KWh)

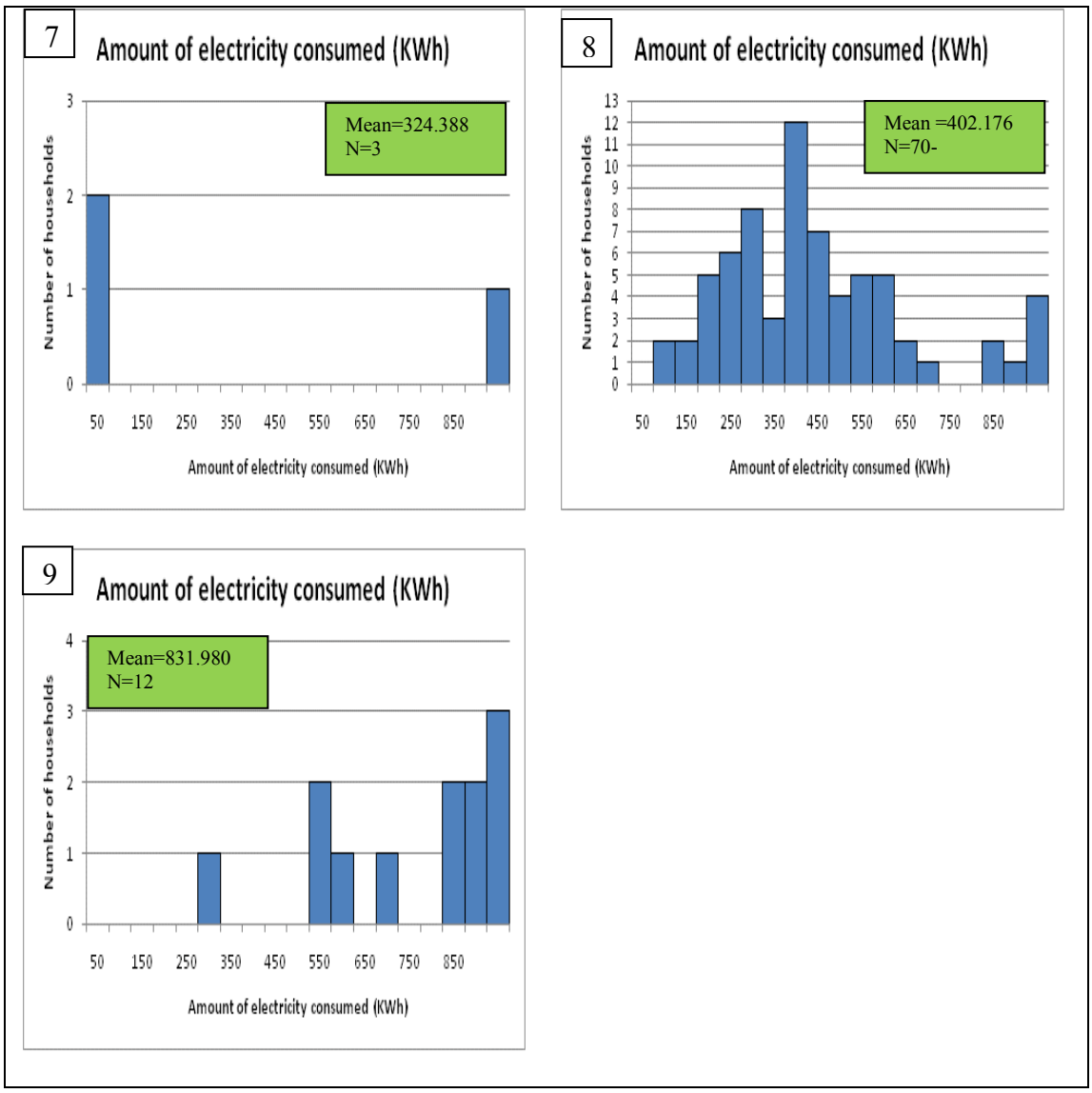


Figure 48: Distribution of electricity consumed per leaf

6.1.4 Interpretations[5]

- Total number of appliances

  The feature that has the major influence on the amount of electricity used in a household is the number of appliances. From this dataset, the difference is marked when number of appliances exceed 9. On average, households that have ≤ 9 appliances used an average of 135.836 KWh while those with ≥ 9 used 398.833 KWh.

  Among the households which have ≤ 9 appliances; there is a marked difference between the households that have at most 2 appliances and those that have more than 2 appliances. The former group had an average consumption of 14.163 KWh while the latter had an average of 147.611 KWh.

[5] Recall that response is the **monthly** consumption of electricity.

- Total monthly income in household

There are two distinguishable income groups. For households which have more than 9 appliances, reside in the areas: Khayelitsha, Matshana, Greenturf or Kabega and the head of the household has obtained a Grade 5 (at least), the association exhibited in **Error! Reference source not found.**(a) was observed.

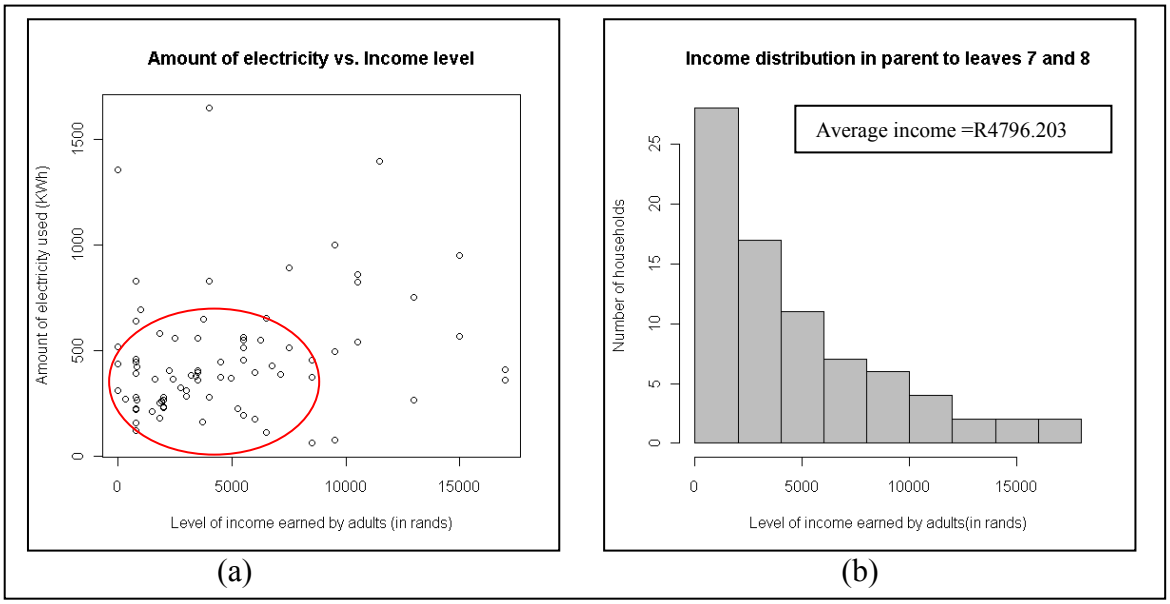


Figure 49: Income vs. amount of electricity used in parent node leaves 7 & 8

The income levels in these groups are quite diverse but most (88%) of these households earn below R10 000. Although the data is sparse for income levels beyond R10 000, it can be observed that for all these household consumption is relatively higher.

On average these households - which have more than 9 appliances, reside in the areas: Khayelitsha, Matshana, Greenturf or Kabega and the head of the household has obtained Grade 5 or beyond and income (earned by adults) ≤ R10 000 consumed an average of 402.176 KWh of electricity, which is notably lower than the average amount electricity consumed by households that have a similar profile except that income level is higher than R10 000 (831.980 KWh).

The income effect is partly through the number of appliances. Households with more income have a higher level of purchasing power and thus can buy and use more appliances.

In reference to Figure 50, across all households in the study it can be observed that there is somewhat a positive association between level of income earned by the adults in the household and the total number of appliances, this association is even more striking when income exceeds R5000. Below an income level of R5000, the total number of appliances does not vary much as income varies. This could be due to the idea that there are several basic appliances (such as lights, stove and fridge) that most households will have

and there are appliances that are not necessities (such as tumble driers and washing machines). As a result, the households with higher levels of income are more likely to purchase these non-basic appliances along with the basic appliances.

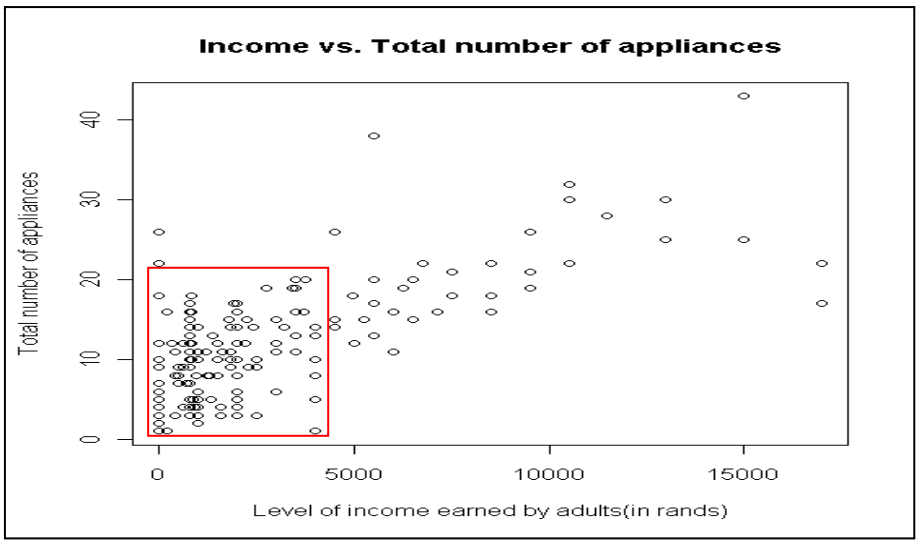
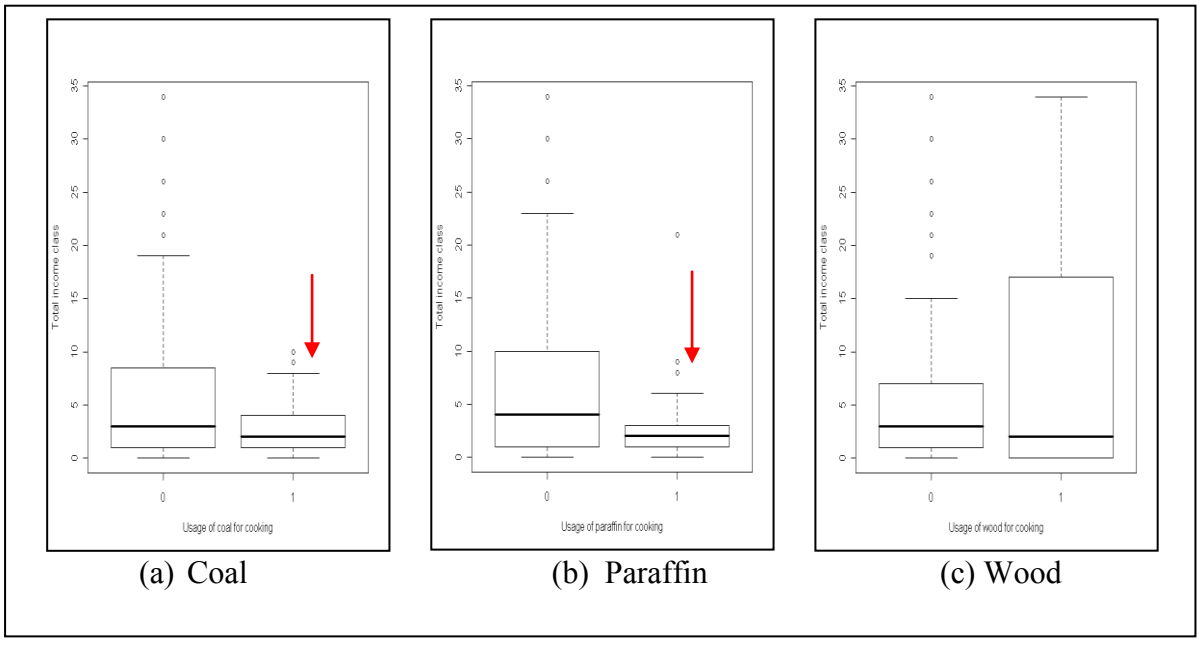


Figure 50: Relationship between total income and the number of appliances

The income effect could also be due the *perceived* cost of electricity relative to alternative energy sources among income groups. From Table 3, the most used alternatives for cooking were coal, paraffin and wood, thus we shall only consider these.



(a) Coal (b) Paraffin (c) Wood

Figure 51: Total income vs. Usage of alternative sources for cooking

Except for wood-users, households that use coal or paraffin for cooking are mostly low income households (as shown by the arrows in Figure 51).

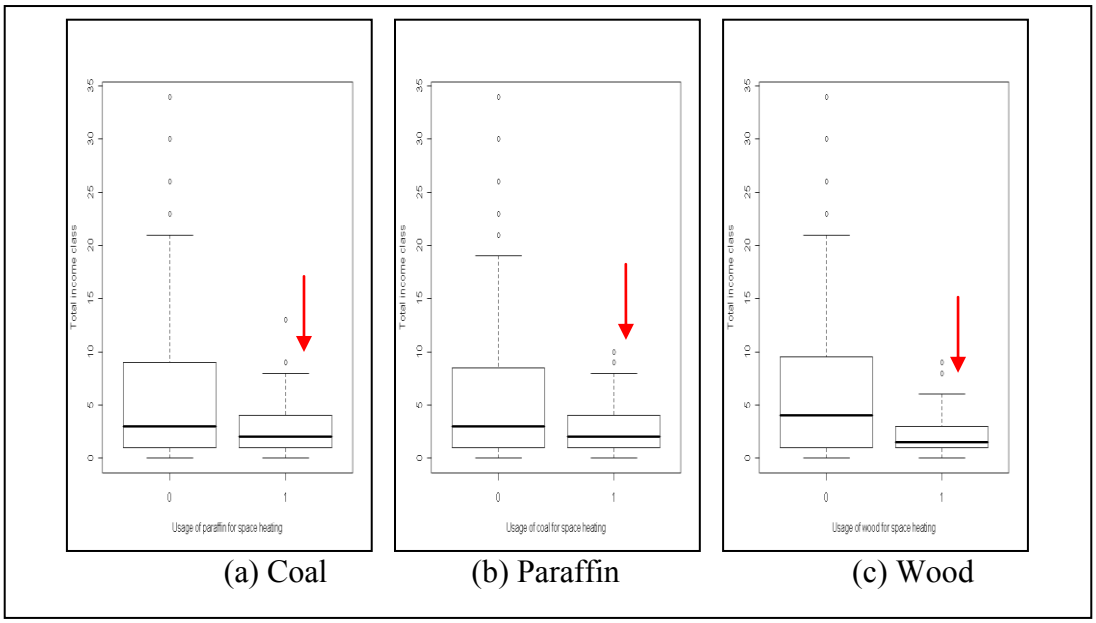


Figure 52: Total income vs. Usage of alternative sources for space heating

Households that use coal, wood or paraffin for space heating are mostly low income households. Figure 51 and Figure 52 could be evidence of the relative low cost of alternative energy sources to electricity, as perceived by low-income households.

- Highest level of education obtained by head of household

Among the households which have number of appliances ranging from 3 to 9, the households with head's highest level of education being lower than Grade 11 used significantly less electricity (about 62.389 KWh) than those with head's highest level of education being higher than Grade 11 (165.993 KWh). This could largely be due to the association between education and income.

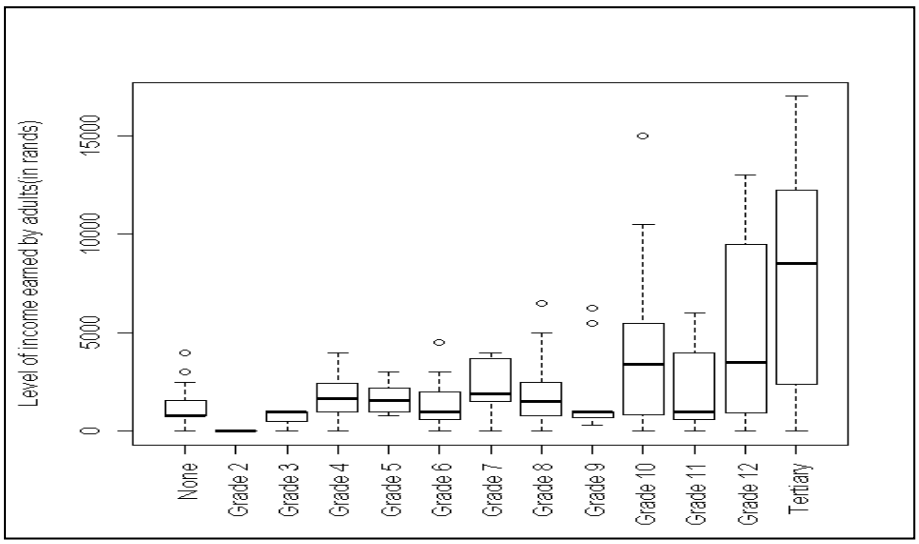


Figure 53: Total income vs. highest level of education obtained by head.

- Residential area

  Among the households that have the total number of appliances ranging between 3 and 9, with the head of the household having attained a Grade 11 (at most); households that reside in Vlaklaagte (which were 4 in total in the parent node to leaf 2 and 3) used considerably less electricity (16.421 KWh on average) compared to similar households in Matshana (4 households), Driekoppies (2 households) and Khayelitsha (1 household) with an average of 88.656 KWh. See Figure 54 (a).

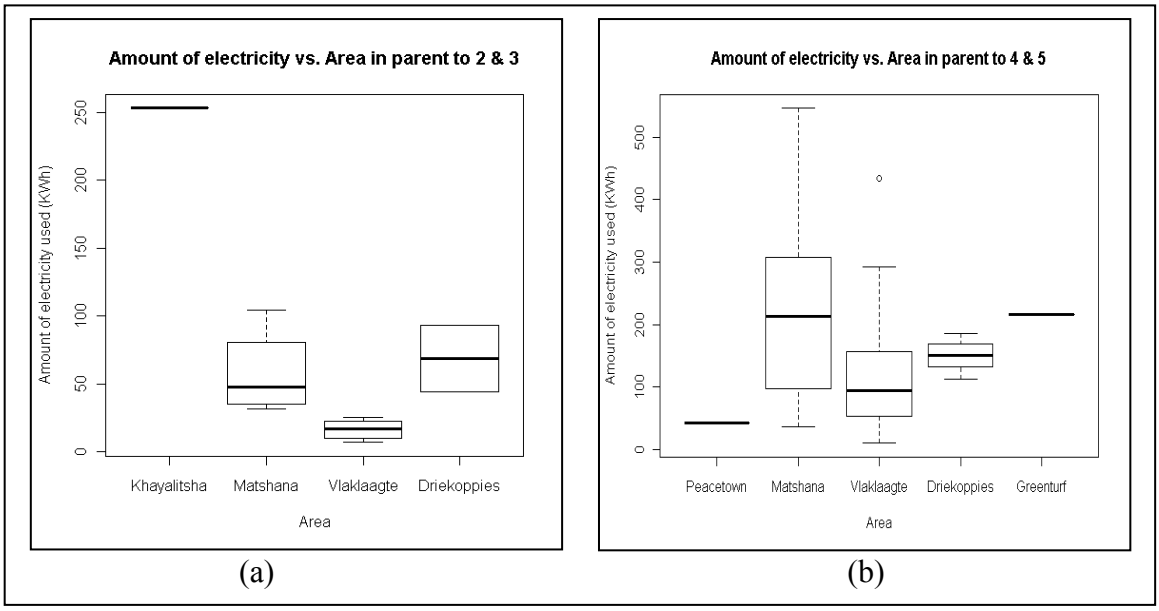


Figure 54: Amount of electricity vs. Residential area

In reference to Figure 54 (b); among the households that have the total number of appliances ranging between 3 and 9, with the head of the household having attained a Grade 11 (at least), households from Peacetown (1 household) and Vlaklaagte (21 households) used less electricity on average (114.865 KWh) compared to similar households from Matshana (25 households), Driekoppies (3 households) and Greenturf (1 household) with an average of 204.780 KWh

From Figure 55 and the optimal tree, it can be observed that the households which have more than 9 appliances and reside in Peacetown (2 households), Vlaklaagte (17 households) or Driekoppies (5 households) used less electricity on average (181.815KWh) compared to households that also had more than 9 appliances but reside in Khayelitsha, Matshana Greenturf or Kabega (460.109 KWh).

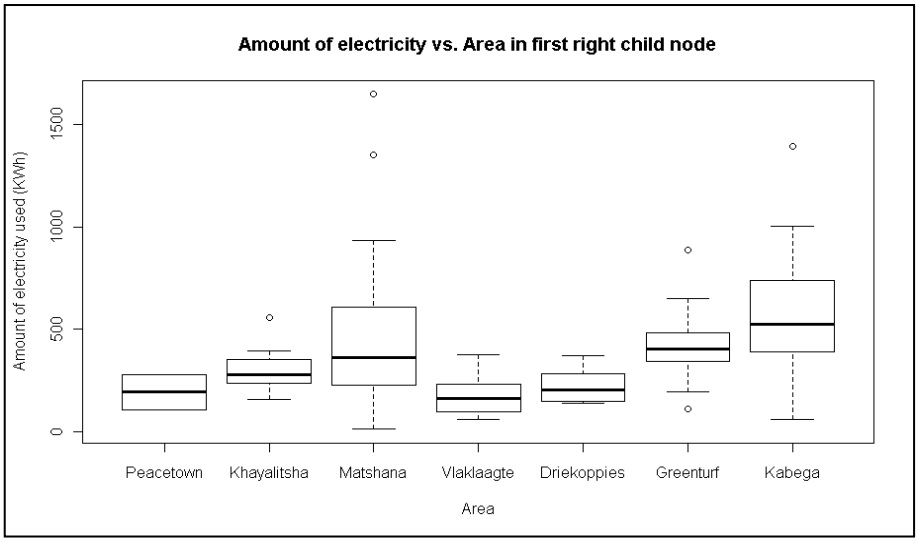


Figure 55: Amount of electricity vs. Residential Area (2)

Overall Peacetown, Vlaklaagte and Driekoppies seem to be are associated with low electricity consumption. , Figure 56 shows that most households from these areas have low levels of income and subsequently lower appliance ownership. Greenturf and Kabega are relatively affluent areas and subsequently high appliance ownership. Thus households from Greenturf and Kabega are likely to use more electricity.

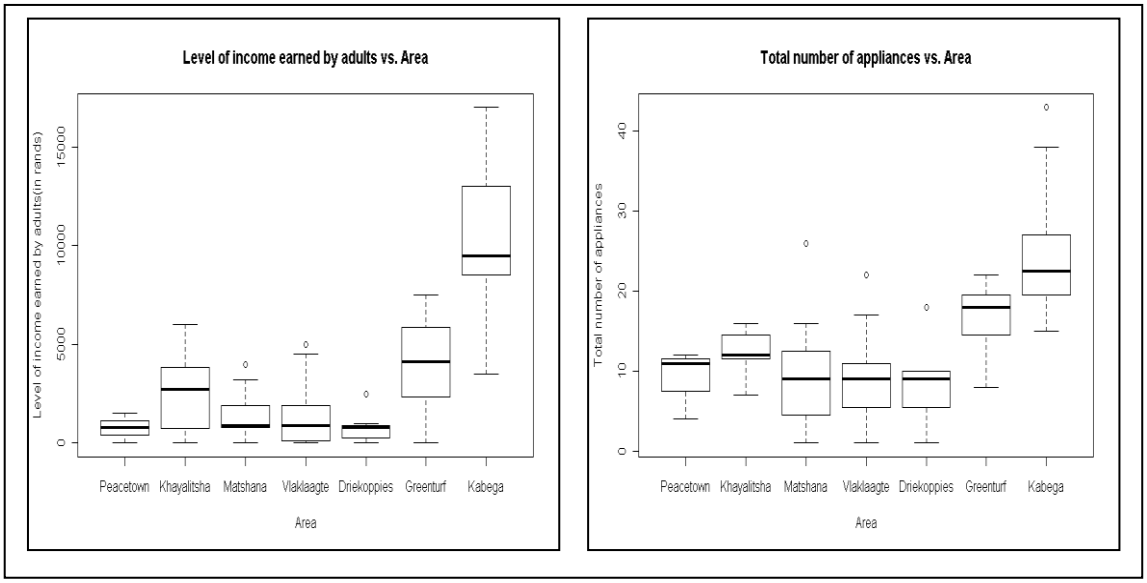


Figure 56: Income vs. Residential area

6.2 GENERALISED LINEAR MODELS

6.2.1 Model building

In reference to section 4.1, when fitting a generalised linear model, the underlying distribution of the outcome of interest has to be identified.

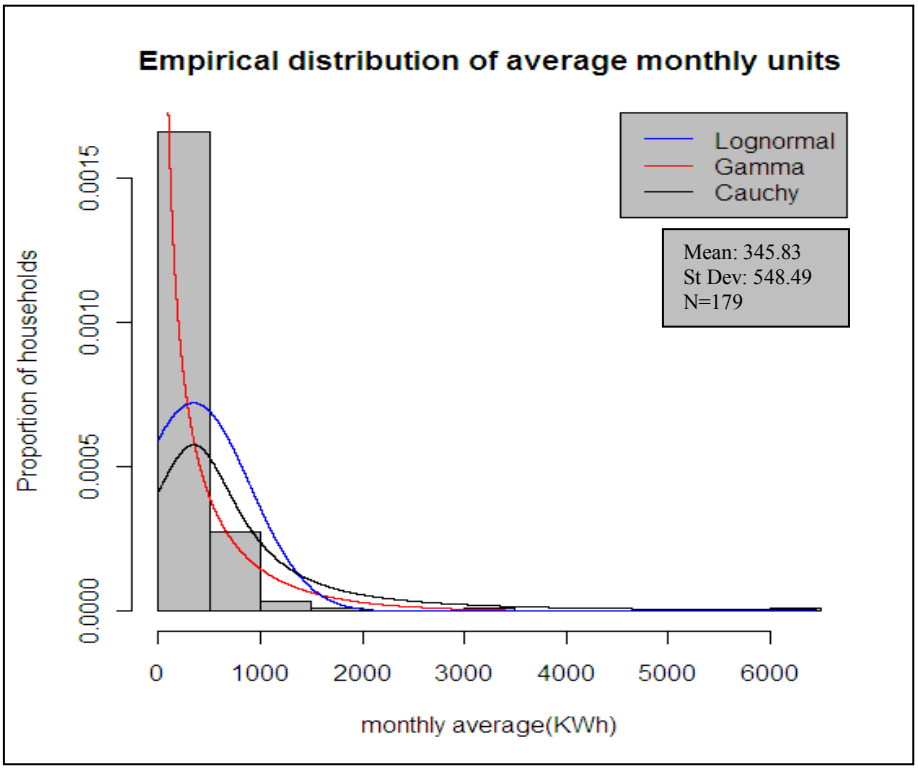


Figure 57: Empirical distribution of average units read

The empirical distribution of the average units read has a tail to the right. The superimposed Gamma distribution fits the empirical distribution better than the lognormal distribution.

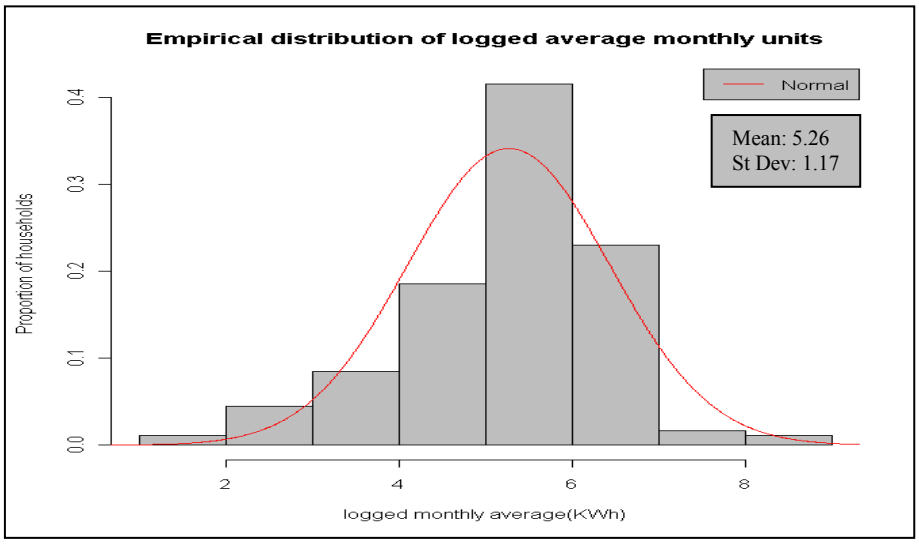


Figure 58: Empirical distribution of logged values

The lognormal distribution does not quite fit the response values because the logged values, even though almost symmetric (has a bit of a tail to the right), has "fat tails" and is slightly more peaked that its normal distribution counterpart.

```
        Shapiro-Wilk normality test
data:  logged
W = 0.9568, p-value = 3.233e-05
```

Figure 59: Result from the Shapiro-Wilk test for the logged average units read

The results of the Shapiro-Wilk test validate that the logged distribution is not quite a normal one (p-value < 0.0001).

Assuming the underlying distribution to be the Gamma, the density function is:

$$Y_i \sim \text{Gamma}(\alpha, \lambda) \;\Rightarrow\; f(y_i|\alpha,\lambda) = \frac{\lambda^{\alpha} y^{\alpha-1} \exp(-\lambda y)}{\Gamma(\alpha)} \qquad i = 1,2,3,\ldots 179$$

- The "iterative weighted least squares" method was used in estimating the GLM coefficients.

- Although the gamma appeared to fit the data better than the lognormal distribution, both the `inverse` and `log` link were used, but the iteratively reweighted least squares algorithm used did not converge. This often indicates that the assumed distribution is not quite correct, i.e. the Gamma does not fit the data well. Thus, we will resort to the Lognormal distribution, with density function

$$Y_i \sim \text{Lognormal}(\mu, \sigma^2) \;\Rightarrow \text{Z}_i = \log(Y_i) \sim \text{N}(\mu, \sigma^2)$$

$$f(z_i|\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp \frac{(z_i-\mu)^2}{2\sigma^2} \qquad i = 1,2,3,\ldots 179$$

"Undirected" stepwise variable selection was used with the AIC criteria by function `step` and `glm` (in `R`). This procedure could not take place with missing values, as a result, the variables that had many missing entries had to be dropped. In general, when encountering a missing value for an observation, that specific observation would be deleted but in this data set there so many missing values such that if the corresponding observations were deleted then there would not be any data left. From a total of 131 independent variables, only 96 remained.

Figure 60 shows the results of the optimal model and Figure 61 shows the diagnostic plots for the optimal model.

```
Coefficients:
                                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)                      3.940459   0.263847  14.935  < 2e-16
TotalNumberOfAppliancesExclLights  0.167675   0.028743   5.834  2.98e-08
TotalNumberOfPeopleInHousehold     0.099592   0.030112   3.307  0.00117
AltFuelCoalCook1                -0.739188   0.169513  -4.361  2.33e-05
PQTrips2                        -0.219804   0.354028  -0.621  0.53558
PQTrips3                         0.068356   0.816818   0.084  0.93341
PQTrips4                         1.635618   0.491277   3.329  0.00108
FloorArea                        0.005054   0.001947   2.596  0.01031
ApplianceFridgeFreezerNumber     0.239013   0.146292   1.634  0.10429
TimeWithElectricity             -0.024826   0.017262  -1.438  0.15236
SupplyOfOutBuildings1            0.261173   0.164435   1.588  0.11422
CookChildrenSundayDinner         0.228927   0.092437   2.477  0.01432
CookChildrenWeekDayDinner       -0.142738   0.087863  -1.625  0.10625
PQLightsDim2                    -0.852855   0.377805  -2.257  0.02535
PQLightsDim3                     0.339452   0.312475   1.086  0.27899
PQLightsDim4                     0.366156   0.576482   0.635  0.52625
OwnDwelling1                    -0.363635   0.226651  -1.604  0.11063

(Dispersion parameter for gaussian family taken to be 0.6347717)

    Null deviance: 238.10  on 174  degrees of freedom
Residual deviance: 100.29  on 158  degrees of freedom
AIC: 435.21
```

Figure 60: GLM Results

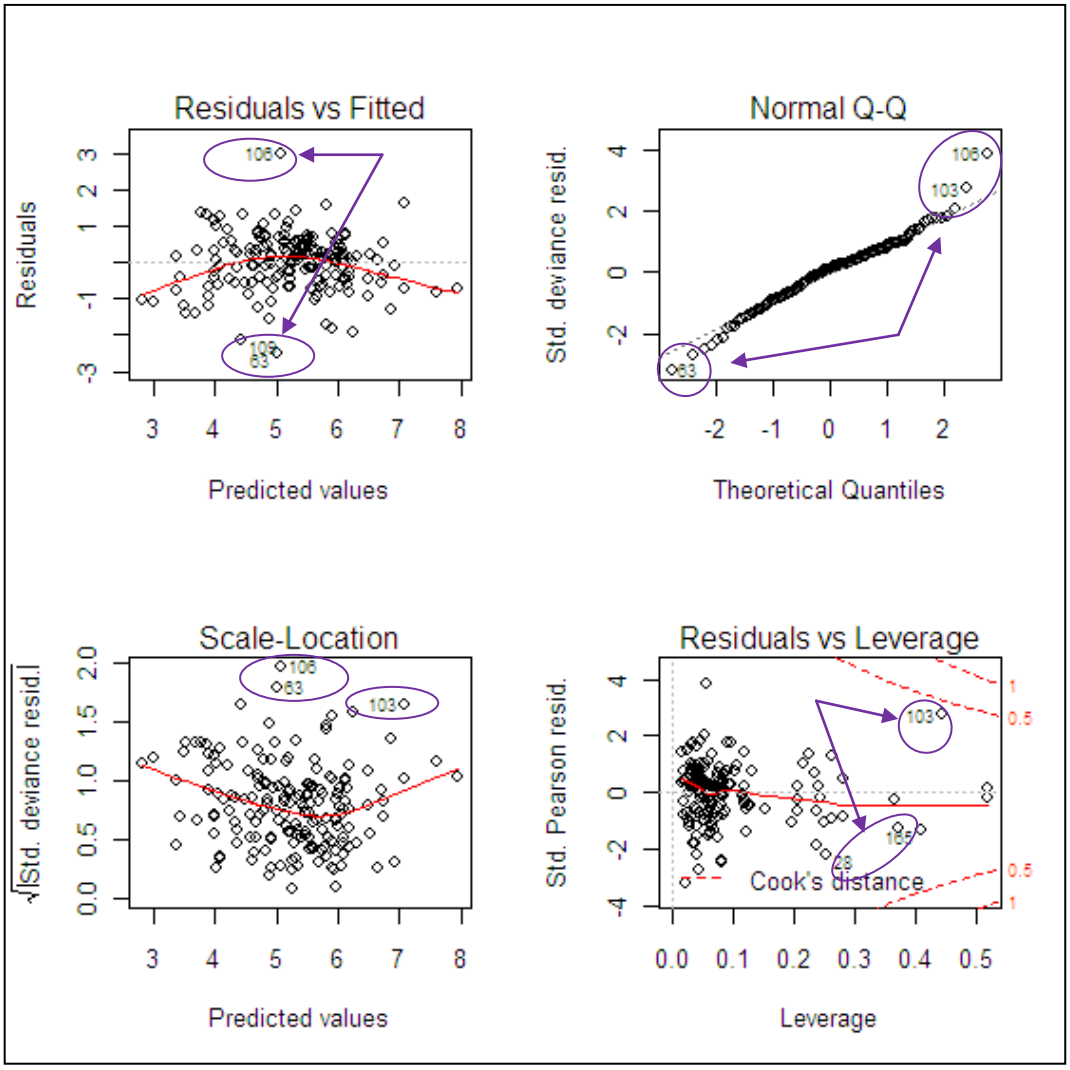


Figure 61: Diagnostic plots

Normal Q-Q and Residual vs. Fitted plots:

Observations 103 and 106 were the outliers noted at outset which had consumption values 6059.920KWh and 3135.033KWh respectively, while observation 63 had a consumption value of 12.085KWh which is extremely low. These values are also distinguished on the Scale-Location plot, showing that they each had high variance which violates the assumption of residuals with equal variance underlying GLM.

Residual vs. Leverage:

Observation 28,103 and 165 was the influential observations, with 103 being the more outstanding observation. , but none of these values were excluded from further analysis.

6.2.2 Analysis without upper-outliers

The stepwise selection was repeated without the upper outliers (6059.920KWh and 3135.033KWh). Again, when assuming that the response variable had an underlying Gamma distribution, the algorithm did not converge with both the `log` and `inverse` link functions. As a result, the lognormal distribution was assumed.

Figure 62 and Figure 63 show the results from the optimal GLM and the corresponding diagnostic plots.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.892560   0.226617  17.177  < 2e-16 ***
TotalNumberOfAppliancesExclLights  0.184550   0.031194   5.916 1.94e-08 ***
LanguageID3                      0.401943   0.257808   1.559 0.120953
LanguageID4                     -0.129089   0.247300  -0.522 0.602398
LanguageID5                     -0.395761   0.278634  -1.420 0.157449
LanguageID8                     -0.524720   0.228315  -2.298 0.022843 *
LanguageID13                    -0.425890   0.271417  -1.569 0.118592
TotalNumberOfPeopleInHousehold   0.100019   0.026033   3.842 0.000176 ***
FloorArea                        0.005957   0.001820   3.273 0.001303 **
TimeWithElectricity             -0.040148   0.018384  -2.184 0.030424 *
ApplianceOtherNumber            -0.282576   0.187630  -1.506 0.134032
SupplyOfOutBuildings1            0.241197   0.155897   1.547 0.123802
AltFuelCoalCook1                -0.372785   0.244696  -1.523 0.129618

(Dispersion parameter for gaussian family taken to be 0.5485095)

Null deviance: 218.075  on 172  degrees of freedom
Residual deviance:  87.762  on 160  degrees of freedom
AIC: 401.54

Number of Fisher Scoring iterations: 2
```

Figure 62: GLM results (without outliers)

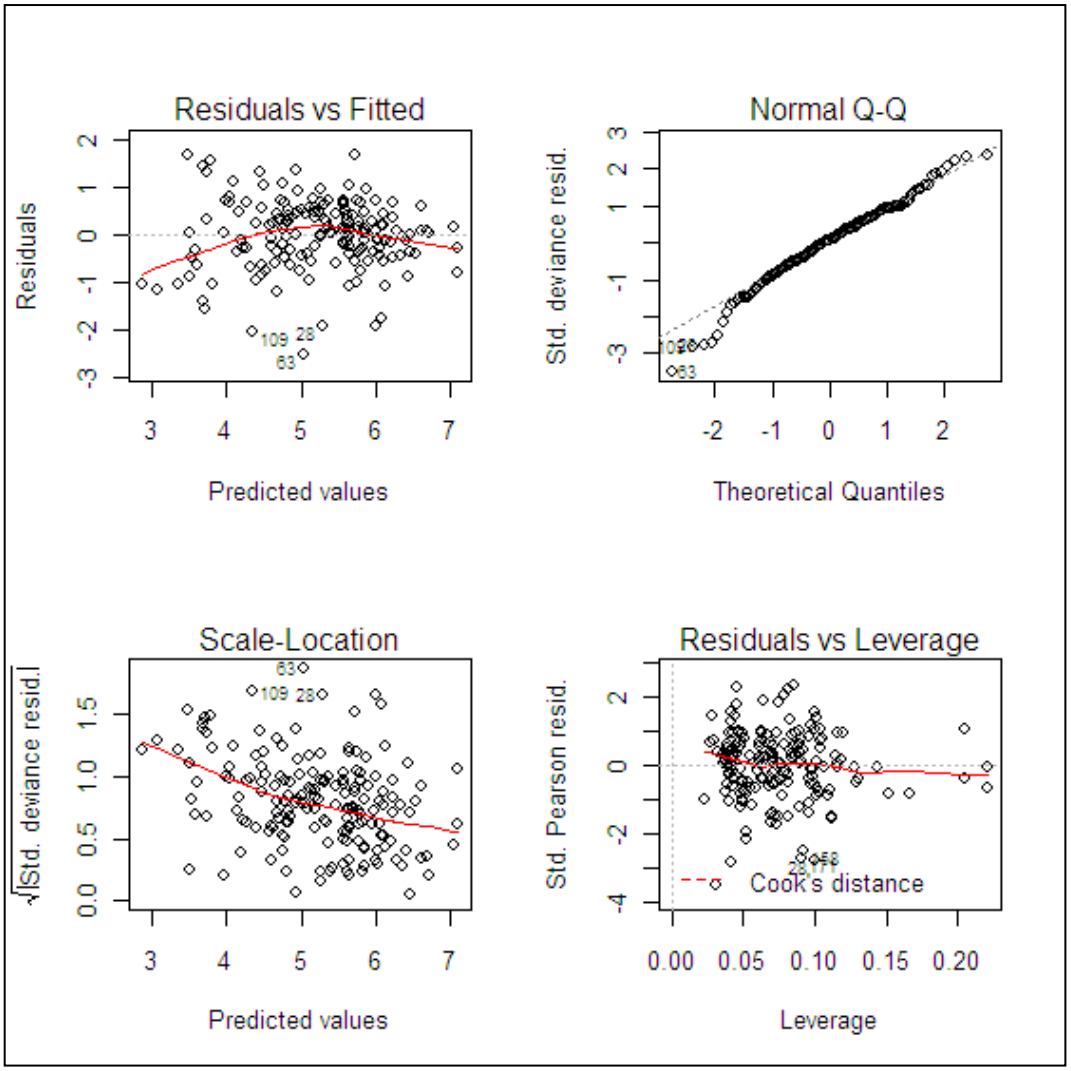


Figure 63: Diagnostic plots (2)

6.2.3 Interpretations

- Total number of appliances excluding lights

  On average, as number of appliances (excluding lights) increase by 1, the logged monthly KWh amount increases by 0.185. That is, the monthly KWh amount will increase by a factor of $e^{0.185} = 1.203$ or, equivalently, will be 20.3% higher from the baseline KWh consumption amount.

  This effect is significant with p-value < 0.00001

- Home language

  i. In the analysis, the baseline home language was set to be Zulu. Households that use Xhosa as a home language appeared to use more electricity, i.e. an increase **by** a factor of $e^{0.401} = 1.495$ (49.5% higher KWh amount) than households with Zulu as a home language.

This *Zulu-Xhosa* difference is somewhat (but not strikingly) significant with p-value = 0.12095.

ii. Households that had Afrikaans as a home language appeared to use slightly less electricity than households with Zulu as a home language. Typically these *Afrikaans households* used an amount of electricity that was $e^{-0.129} = 0.879 = 87.9\%$ of the typical amount used by households that had Zulu as home language. That is 12.1% lower amount of electricity.

This *Zulu-Afrikaans* difference appeared insignificant with p-value = 0.602398.

iii. Households that had English as a home language appeared to use quite less electricity than households with Zulu as a home language. Among these *English households*, electricity consumption was only $e^{-0.396} = 0.673 = 67.3\%$ of the typical electricity used by amount than households with Zulu as a home language. That is, these *English households* used 32.7% less electricity compared to the *Zulu households*.

This *Zulu-English* difference is also somewhat (but not strikingly) significant with p-value = 0.157449.

iv. Households that use Ndebele as a home language also seemed to use notably less electricity compared to the households with Zulu as a home language. On average, these *Ndebele households* used electricity amount only equal to $e^{-0.525} = 0.592 = 59.2\%$ of the amount of electricity used by *Zulu households*. That is, 40.8% lower amount of electricity.

This *Zulu-Ndebele* difference is very significant with p-value = 0.022843.

v. Lastly, households that had siSwati as a home language appeared to use quite less electricity as well. These *siSwati households* used an average of $e^{-0.426} = 0.653 = 65.3\%$ of the amount of electricity used by *Zulu households*. That is, 34.7% lower amount of electricity

This difference is somewhat significant with p-value = 0.118592.

The differences in electricity usage depicted across the language groups is a probably an indication of cultural differences (such as cooking habits, typical number of members per family and attitudes to technology) between these language groups. These differences in electricity usage might also be due to the idea that people who reside in the same area tend to have the same home language and - thus - the observed energy differences might be an indication (or a result) of the socio-economic status of the households across the different residential areas.

- Total number of people in household

  As number of people in household increase by 1, the logged monthly KWh amount increases by 0.100. That is, keeping all the other features constant, the monthly KWh amount of electricity will typically increase **by** a factor of 1.105. That is, a 10.5% increase per person.

  This effect is remarkably significant with p-value = 0.000176.

  This effect could be due to households with more appliances having a higher potential of consumption. As a result, the total number of people in a household becomes a notable driver of electricity usage in households that have more appliances.

- Floor area

  For each squared-metre increase in floor area of a house, the amount of electricity used typically increases **by** a factor of $e^{0.006} = 1.006$. That is, a 0.6% increase for each extra squared-metre.

  This effect is also very significant with p-value = 0.001303.

- Time with electricity

  For a 1-year increase in the length of time for which a household has had electricity, the amount of electricity used typically decreases **to** a factor of $e^{-0.040} = 0.961 = 96.1\%$. That is, 3.9% lower for each extra year.

  This effect is quite significant with p-value = 0.030424.

- Number of (additional) appliances

  For each extra appliance (apart from those not listed in Table 1 Table 4), the amount of electricity used typically decreased **to** a factor of $e^{-0.283} = 0.754$. That is, 24.616% lower. This result is quite counter intuitive but in reference to the data, these additional appliance where mostly digital video display (DVD) systems and VCRs, i.e. appliances with low levels of electricity *intake*.

  This effect could be significant with p-value = 0.134032.

- Supply to external buildings

  Households that had supplied electricity to other (external) buildings typically used amount electricity that is higher **by** a factor of $e^{0.241} = 1.273$; that is 27.3% more than households that did not supply electricity to other (external) buildings.

  This effect might be significant with p-value = 0.123802.

- Usage of coal for cooking

Households that used coal for cooking typically used amount electricity that is $e^{-0.373} = 0.689 = 68.9\%$ of the amount of electricity used by households that did not use coal for cooking. That is, 31.1% lower electricity than households that did not use coal for cooking.
This effect could be significant with p-value = 0.129618.

Due to the large number of variables involved, interaction between the variables will not be considered otherwise they should be considered.

6.3 COMPARISON OF RESULTS: CART vs. GLM

The most influential (to the amount of electricity consumed) households as identified by CART were:

- total number of appliances,
- highest level of education obtained by head of household,
- residential area, and
- income earned by adults.

However, GLM analysis identified the following features as most influential:

- total number of appliances excluding lights
- home language
- total number of people in household
- floor area
- time with electricity
- number of (additional) appliances
- supplying to external buildings
- usage of coal for cooking

SIMILARITIES

Both GLM and CART agree on the number of appliances being the principal feature associated with the amount of electricity used. Although GLM identified the "total number of appliances excluding lights" rather than "total number of appliances" as in CART, from Table 17, it can be observed that these features are closely related as the former was identified as a surrogate for the later.

The residential area of the households (GROUPID) appeared in CART but not in GLM. However, GLM identified total number of people in household, home language, time length for which the household has had electricity and usage of coal for cooking as drivers of amount of electricity while CART identified these features as surrogates for residential area. See Table 17.

DIFFERENCES

The CART algorithm identified the highest level of education obtained by the head of the household and the level of income earned by adults in the household as drivers of the amount of electricity.

GLM, on the other hand, identified the floor area of the house as a significant driver of amount of electricity used. Whether or not the household supplies to external buildings was also identified as being influential but this feature was not a persuasively significant driver.

7. CONCLUSIONS

7.1. COMPETENCE OF METHODS: CART vs. GLM

At the start of the analysis, GLM require distribution checks to be performed and assumptions to be made about the distribution underlying the response variable. GLM further restrict that the distribution assumed must be from the exponential family. On the other hand, CART does not require any distribution assumptions to be made regarding the response but it seemed to be more efficient in investigating associations when the distribution of the response variable was less skew.

During analysis, the CART require several subjective choices to be made; fro instance, the minimum of cases that should be in a node for splitting to be allowed. In GLM, however, there is one key choice to be made and that is the underlying probabilistic models for the response; after this choice, the algorithm is entirely self-conducted. This idea was also evident when the GLM algorithm detected the observations which had outlying electricity consumption values while in CART, the choice to exclude these observations was as per guidance of the field expert.

During analysis, variable selection method for GLM requires that there be no missing values for in the data set. The dataset used in this study has many missing values due to the structure of the questionnaire used in gathering information. In this regard, variables that had many missing values had to be removed and - thus - useful information was lost. The CART algorithm does not require a complete dataset; it uses whatever information available and builds surrogate for ease of usage of the tree in future.

To investigate interactions between the features in GLM, without guidance of the energy experts, all the possible interactions will have to be considered. Considering all possible interactions is quite impractical (and consumes time) in a data set that has many independent variables like the dataset used in this study. On the other hand, CART has an innate ability to detect interactions between independent variables. For instance, from the optimal tree, it can be observed that the level of income (earned by adults) only become influential to amount of electricity used provided that a household has more than 9 appliances, resides in Greenturf, Kabega, Matshana and Khayelitsha and - lastly - the head of the household had obtained a Grade 5 at least. This hierarchical structure of splits shows that there is an interaction between income and education, residential area and number of appliances.

GLM explain the relationship between the response variable and independent variable in a linear form. The linearity assumption might not always be valid. CART - however - only indicates the critical values (of the independent variables) at which a marked difference in the values of the response variable is observed in the emerging subgroups from splitting at the critical values. In cases where there is a linear relationship between an independent variable and

the response; GLM would be more useful because they summarise the linear relationship by estimates of the β coefficients.

When assessing reliability of the results, the probabilistic model assumed in formulating a GLM become very helpful because then we can place confidence intervals and p-values on the results obtained. In CART, these statistical checks cannot be obtained and - thus - we cannot communicate how sure we are about the reliability of the results obtained.

7.2. PRINCIPAL HOUSEHOLD FEATURES

From both analyses, it is evident that the most influential features are the:

- Total number of appliances
- Total number of people in household
- Time length for which household has had electricity
- Usage of coal for cooking
- Home language
- Residential area

These identified features are interlinked.

Both CART and GLM identified the number of appliances as the major driver of the amount of electricity consumed. Households with more appliances are regarded as having a higher potential of consumption. As a result, the total number of people in a household becomes a notable driver of electricity usage in households that have more appliances.

The accumulation of appliances is likely to increase as the time for which the household has electricity increases (assuming that income levels allow further purchasing), but this accumulation is likely to halt (or occur less frequently) after sometime. As a result, electricity usage will typically increase rapidly in the early years of electrification of the household and then slow down after sometime. This rapid increase is - however - largely bounded by level of purchasing power (thus socio-economic status) of a household.

Households that have easy access to alternative sources of energy like coal and wood (such that using electricity is perceived as costly) are likely to use less electricity than households that do not have easy access to these alternatives. The usage of alternative energy sources can also be linked to socio-economic status of households. Low income households have a low purchasing power, as a result cannot *afford* to buy electric appliances that will allow them to actually use electricity. Subsequently, these households will resort to using means such as coal fire or wood fire for several end-uses.

The home language of a household could be an indication of cultural differences (such as cooking habits, typical number of members per family and attitudes to technology) between these language groups. Often, households that reside in the same area use the same home language and - thus - the observed energy differences might be an

indication (or a result) of the socio-economic status of the households across the different residential areas

Residential area is a more complex feature as it generally encapsulates the home language (thus cultural differences), access of (and usage) of alternative energy sources and the socio-economic status (thus purchasing power), and the latter (in turn) affects the extent of appliance ownership usage.

With the above reasoning, we conclude that the CART method was more competent in identifying the household features that differentiate electricity consumption.

LIST OF REFERENCES

1. **Ogunlade Davidson, Harald Winkler,Andrew Kenny, Gisela Prahad, Jabavu Nkomo, Debbie Sparks, Mark Howells, Thomas Alfstad.(2006)** *Energy policies for sustainable development in South Africa*. Cape Town : Energy Research Centre, University of Cape Town.

2. Enerdata. *Enerdata*. [Online] Enerdata. www.enerdata.net.

3. **Mr. Jeff Subramoney, Mr. Johan van Wyk, Ms. Mmabakwena Dithupe, Mr. Allen Molapo,Ms Nombulelo Mahlangu,Ms Ruth Morumudi.(2010)** *DIGEST OF SOUTH AFRICAN ENERGY STATISTICS 2009*. PRETORIA : ENERGY DEPARTMENT, SOUTH AFRICA. 978-1-920448-25-7.

4. **Stephanie C Lemon, Jason Roy, Melissa A. Clark, Peter D. Friedmann, William Rakowski.(2003)** .*Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression*. Worcester : SC Lemon Ph.D., Division of Preventative and Behavioural Medicine, University of Massachusetts Medical School, Worcester. NA 01655.

5. **Leo Breiman, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone(1984).** CLASSIFICATION AND REGRESSION TREES. *CLASSIFICATION AND REGRESSION TREES*. CALIFORNIA : WADSWORTH INTERNATIONAL GROUP.

6. **Timofeev, Roman(2004).** *Classification and Regression Trees(CART) Theory and Applications*. Berlin : Timofeev, Roman.

7. **W.N. Venables, B.D. RIPLEY(1999).** *Statistics and Computing, Modern Applied Statistics with S-PLUS*. New York : Springer-Verlag New York Inc. ISBN 0-387-98825-4.

8. *ALTERNATIVE ENERGY*. **AFRICA, ENERGY RESEARCH CENTRE & SUSTAINABLE ENERGY SOCIETY SOUTHERN(2010).** 1, PARKLANDS : TRINITY MEDIA GROUP, Vol. 1.

9. Nagpaul, PS. Guide to Advanced Data Analysis using IDAMS Software, United Nations Cultural and Scientific Organisation: New Delhi: India., (2010) http://www.unesco.org/webworld.idams.advguide/Chapt10.htm

10. Dr. Jia Li, STAT 557 (Data Mining),Pennsylvania State University, Department of Statistics,(2009). http://www.stat.psu.edu/online/development/stat557/10_trees/03_trees_impurity.html

11. StatSoft, Inc. (2010). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/classification-and-regression-trees/

APPENDIX I

Description: Site synopses of residential areas.

Group ID	Sample Size	Year and Location
1000032	3	2005 Peacetown
1000033	12	2005 Khayelitsha
1000034	59	2005 Matshana
1000035	45	2005 Vlaklaagte
1000036	12	2005 Driekoppies
1000038	24	2005 Greenturf
1000039	24	2005 Kabega

(☐ Location synopsis is incomplete)

i. Matshana, Kwazulu Natal

- Years monitored: 2002, 2003
- This is a very old electrification site, near Empangeni (inland from Richards Bay) on the North Coast of Kwazulu Natal. Terrain is hilly. This site represents the future of other electrification projects because the consumers are poor *and* they have been electrified a very long time. (2002, 2003)
- Most of the dwellings have a 60 A supply. (2002, 2003)
- Houses are mostly modern style, having tin roof & block or mud/clay/daub walls. The built area of the houses is in the region of $60m^2$. (2002, 2003)
- Most consumers (77%) get water from a communal tap and certain consumers (21%) get water from a tap in their yard (2002). Some consumers (16%) get water from a communal tap, but most consumers (75%) get water from a tap in their yard. (2003)
- Average gross income is about R1400/household/month (2002), R1200/household/month (2003) in a poorly serviced area with gravel roads (2002, 2003).
- There is approximately 26% (2002), 33% (2003) hotplate ownership and 50% (2002), 40% (2003) ownership of fridge or fridge-freezer.
- The reported time with electricity is about 7 years (2002), 5 years (2003).

ii. Greenturf, Phillipi, W. Cape

- Years monitored: 2001, 2002, 2003
- This is a new housing development in the urban Phillipi area of Cape Town (2001, 2002, 2003).
- Dwellings are low-cost single or double-storey (semi-detached), in a well-serviced region with tarred roads, street-lights and sewerage. (2001, 2002, 2003)
- All houses are built with block-type walls and tiled roofs (2001, 2002, 2003). The built area of the houses is in the region of 70m^2 (2001), 62m^2 (2002), 60m^2 (2003)
- Average gross income is in the region of R 2100/ household/month, (2001), R2300/ household/month (2002), R2700/ household/month (2003)
- All houses have piped water inside, but only about 65% (2001, 2002), 76% (2003) have hot water storage heaters.
- There is approximately 25% (2001), 36% (2002), 25% (2003) hotplate ownership and 94% (2001, 2002), 93% (2003) ownership of fridge or fridge-freezer.
- The reported time with electricity at this site is about 2 years (2001, 2002), 4 years (2003).

iii. Kabega, Port Elizabeth, E. Cape

- Years monitored: 2001, 2002, 2003 (no surveys collected this year)
- This is a relatively new urban housing area at Port Elizabeth. (2001, 2002)
- Dwellings are single-storey in a well-serviced region with tarred roads, street-lights and sewerage. (2001, 2002)
- All dwellings have tiled roofs with either brick or block plastered walls. (2001, 2002)
- The built area of the houses is in the region of 70m2. (2001, 2002)
- Average income is in the region of R 8300/ household/month (2001), R 8600/ household/month (2002).
- There is almost no hotplate ownership (2001) and full penetration of all the major appliances (2001, 2002). All houses have hot water heaters (2001, 2002).
- The reported time with electricity at this site is about 4 years (2001, 2002).

Reference:

iv. Peace Town, KwaZulu Natal

Individual monthly income (2001)

Persons	2001
None	7857
R1 - 400	552
R401 - 800	999
R801 - 1600	297
R1601 - 3200	138
R3201 - 6400	36
R6401 - 12800	6
R12801 - 25600	0
R25601 - 51200	0
R51201 - 102400	48
R102401 - 204800	0
Over R204801	0

Household size

Households	1996	2001	% change
1	135	264	95.56
2	168	207	23.21
3	164	195	18.90
4	190	228	20.00
5	174	210	20.69
6	177	183	3.39
7	165	153	-7.27
8	98	123	25.51
9	166	87	-47.59

10 and over	122	198	62.30

Source of energy for lighting

Households	1996	2001	% change
Electricity	813	1188	46.13
Gas	6	24	300.00
Paraffin	34	15	-55.88
Candles	721	606	-15.95
Solar	-	0	-
Other	0	9	-

Annual household income

Households	2001
None	603
R1 - 4800	147
R4801 - 9600	459
R9601 - 19200	339
R19201 - 38400	201
R38401 - 76800	63
R76801 - 153600	18
R153601 - 307200	0
R307201 - 614400	0
R614401 - 1228800	9
R1228801 - 2457600	9
Over R2457600	6

v. Khayelitsha, W. Cape

Source of energy for lighting

Households	**1996**	**2001**	**% change**
Electricity	6300	8001	27.00
Gas	9	12	33.33
Paraffin	221	285	28.96
Candles	34	60	76.47
Solar	-	3	-
Other	0	6	-

Annual household income

Households	**2001**
None	2184
R1 - 4800	438
R4801 - 9600	1266
R9601 - 19200	2265
R19201 - 38400	1563

R38401 - 76800	516
R76801 - 153600	108
R153601 - 307200	12
R307201 - 614400	12
R614401 - 1228800	0
R1228801 - 2457600	6
Over R2457600	0

Source of energy for lighting

Households	1996	2001	% change
Electricity	1508	6324	319.36
Gas	9	30	233.33
Paraffin	3987	2025	-49.21
Candles	105	165	57.14
Solar	-	9	-
Other	0	12	-

Annual household income 2001

Households	2001
None	2250
R1 - 4800	801
R4801 - 9600	1239
R9601 - 19200	2364
R19201 - 38400	1353
R38401 - 76800	444
R76801 - 153600	75
R153601 - 307200	18
R307201 - 614400	9
R614401 - 1228800	6
R1228801 - 2457600	9
Over R2457600	3

vi. Vlaklaagte, Mpumalanga

Source of energy for lighting

Households	1996	2001	% change
Electricity	2201	2493	13.27
Gas	2	18	800.00
Paraffin	8	9	12.50
Candles	146	66	-54.79
Solar	-	3	-
Other	0	6	-

Annual household income

Households	2001
None	216
R1 - 4800	423
R4801 - 9600	618
R9601 - 19200	606
R19201 - 38400	375
R38401 - 76800	210
R76801 - 153600	99
R153601 - 307200	30
R307201 - 614400	9
R614401 - 1228800	0
R1228801 - 2457600	9
Over R2457600	0

vii. Driekoppies, Mpumalanga

Source of energy for lighting

Households	1996	2001	% change
Electricity	968	1992	105.79
Gas	9	6	-33.33
Paraffin	651	243	-62.67
Candles	798	906	13.53
Solar	-	9	-
Other	0	12	-

Annual household income 2001

Households	2001
None	1002
R1 - 4800	357
R4801 - 9600	609
R9601 - 19200	420
R19201 - 38400	297
R38401 - 76800	276
R76801 - 153600	138
R153601 - 307200	45
R307201 - 614400	9
R614401 - 1228800	3
R1228801 - 2457600	0
Over R2457600	0

APPENDIX II

Description: South African Energy Supply data.

i. Electricity generation by fuel

	1998	1999	2000	2001	2002	2003	2004	2005	2006
Coal	187758	186859	193419	183541	190019	202464	212406	214533	220991
Nuclear	13601	12837	13010	10719	11991	12663	13365	11293	10026
Hydro	1595	726	1343	2061	2357	3509	4452	1166	5845
Pumped storage	2420	2590	2591	1587	1738	3006	3822	3032	4102
Imports	2375	6673	4719	9200	9496	8194	9818	11079	10624
Exports	4532	4266	4007	6996	7242	10263	13254	13422	13589

ii. Sectoral consumption of electricity

	Industry	Transport	Agriculture	Commerce	Residential
1998	101,867	4,639	5,627	13,974	30,163
1999	99,673	4,429	5,755	17,709	29,511
2000	99,703	5,411	3,954	17,164	28,680
2001	106,469	5,562	4,175	18,301	34,623
2002	115,785	6,246	4,644	18,227	30,418
2003	109,589	5,565	5,142	21,071	34,074
2004	134,384	6,302	6,158	24,990	36,231
2005	113,028	5,545	5,520	27,103	36,970
2006	116,631	3,480	5,841	28,833	39,671