Analysis of National LR Project

Load and sociodemographic data: 1998

by Schalk Heunis

Table of Contents

1. Introduction

The NRS National Load Research project has been running for four years and a lot of data has been collected from various sites across South Africa. To ensure the relevance of the data, it is important that the gathered data be analysed to ensure that the costs and efforts of the project is concentrated in the right direction. This report is the third annual "review" study undertaken. The work was commissioned by M Dekenah, manager of the NRS LR project.

This report is a follow up on the second review which will be referred to as report 1997 [1]. A number of recommendations were made in report 1997 which was reevaluated in this report (par 2 and 3). A socio demographic review is given in paragraph 4 which shows inaccuracies in the household level predictor sets.

An updated ADMD prediction model is presented in paragraph 5. A number of recommendations were made and conclusions drawn and summary thereof is given in paragraph 6.

Detail analysis of Tafelsig and Orient Hills with an explanation of the methods used is attached as appendix A through C.

The project has been accumulating data for about 4 years, and about 4 GB of residential load data and socio-demographic data has been accumulated to date.

The rate of data accumulation is so high that:

- Project reviewers are not able to fully review the scope of this data collection project based upon the primary results reported.
- Time-to-feedback from any other intensive analysis can be longer than the data collection cycle (i.e. projects are partly out of control)

In these circumstances, the only solution is to specifically commission critical feedback at the end of each calendar year, and formalise the results as a feedback study report

This report is the third annual feedback study which has been undertaken on the data, covering analyses of sociodemographic data, and the connection between the load and sociodemographic data (ie a load model).

Over the past three projects, techniques for evaluation of data quality have advanced very rapidly beyond the normal measures of "lumped statistics". Very powerful tools are being used.

Such feedback fundamental in order in order to manage the project from an expertise-base rather than an experience-base.

2. Recommendations from report 1997

The 1997 statistical review report sets out guidelines in the form of recommendations for further load research. These recommendations can be summarized as follows:

- Populate the holes in the township predictor set, specifically aimed at income and time since electrification.
- Do load research on rural customers (TRI has since taken up this portfolio)
- Investigate geyser ownership and the factors which influence it.

2.1 Holes in the township predictor set

The choice of target townships were made according to these recommendations, although no rural customers were investigated. Figure 2.1 shows the location of the collected datasets in the township descriptor space with income and time since electrification as predictors. The 1998 datasets are displayed as circles and the datasets from previous years as squares.

The data points in the big circles populate to some extent the holes in the township level predictor set. The lower extreme of the data set could however receive some more attention, i.e. poor communities. Rural communities in particular are likely to fall at this lower end of the curve.

Recommendation:

Two objectives emerge from this analysis:

- **Populate the remaining holes in the data set**
- **Target poor communities in urban and rural environments**

2.2 Geyser ownership

The influence of Geyser ownership is pervasive:

Individual loads tend to impact upon the peak when:

- Their profiles are peaky
- They tend to be used at similar times (high correlation)
- They use more energy

Figure 2.2 shows a histogram of geyser ownership as found in the various datasets. The histogram is split in three to show the geyser ownership found in the 1994-1997 datasets, the 1998 datasets and all of the datasets. Tafelsig 1998 and Orient Hills both has partial geyser ownership and may be used to investigate the following:

- factors influencing geyser ownership
- influence of geyser ownership on township demand
- verification of geyser clusters

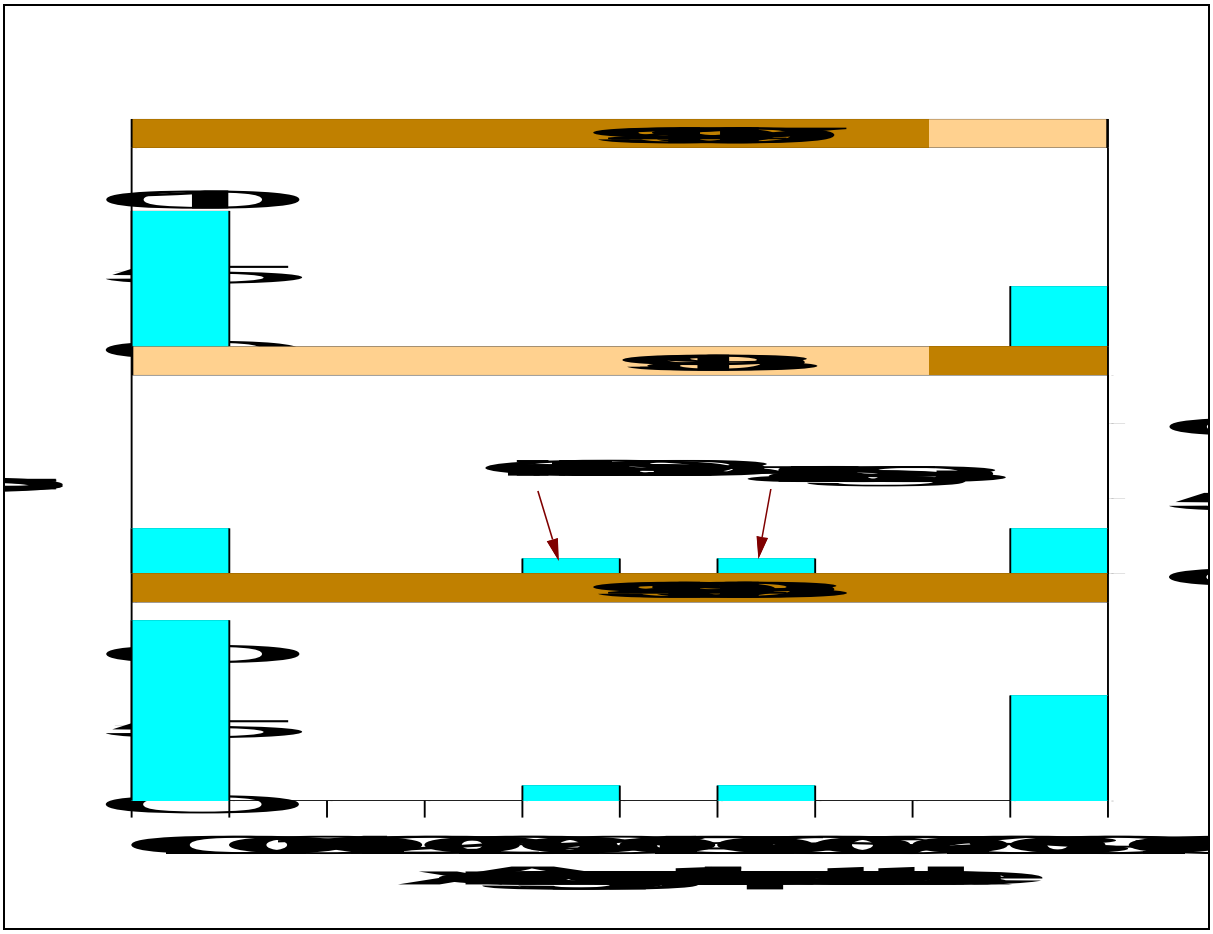

Figure 2. 1 Geyser ownership in datasets for 1994-1997 and 1998

An analysis of Tafelsig 1998 is attached as appendix A. The following results were found in this dataset:

- Geyser ownership has a significant effect on household level demand in this township
- Geyser usage rather than geyser ownership determines the effect of geysers have on the township demand.
- Geyser rating should be taken into account when evaluating the effect of geyser ownership

An analysis of Orient Hills 1998 is attached as appendix B and the following results were found:

- No distinct demand clusters were found in this data set
- Geyser ownership had no significant effect on the household level demand in this data set

The absence of the geyser ownership-demand clusters may be attributed to one or more of the following:

- This data set was recorded from three blocks of flats and a different culture may be present here
- Geyser ownership reported may not agree with actual geyser usage
- Due to the relatively low geyser penetration (40%), the coincidence of geyser usage is not high enough to alter the demand profile. Since the demand profile is used as definition for the household level response, the peak demand could occur at a different time than the peak geyser usage.

- Geyser may be switched off due to the high energy cost or may be out of order
- The presence of load switching relays – this can be ruled out since none were found in Orient hills
- The low geyser penetration (40%)

Recommendation:

It is important to gain an understanding of this phenomena since geyser penetration has a mayor influence on the overall household level response in all the other datasets.

3. Verification of results described in 1997 load survey report using data from 1998 load survey

The 1997 recommendations suggested that as more data became available, the following could be verified:

- Household predictor data set clusters
- Geyser ownership as most significant factor for the household demand

The 1997 data set contained three clusters of household predictors and a summary of the characteristics is given in table 3.1 [1].

Data from 8 more townships are available from the 1998 load survey and were used to further explore the relationships uncovered in the 1997 review report.

Table 3.1: Characteristics of clusters identified by principle component analysis [1]

Cluster 1	**Cluster 2**	**Cluster 3**
No Geyser ownership / Communal tap	Owns Geyser / Piped water	No Geyser ownership / Communal tap
Higher minimum temperature	Higher minimum temperature	Lower minimum temperature
Lower income	Higher income	Lower income

A demand analysis of the clusters yielded two significant subsets, one with geyser ownership and one without. No significant effect was detected in the demand due to temperature. The new datasets may be used to reevaluate the significance of geyser ownership on demand and the shape of the subsets (demand clusters).

Figure 3.1 shows the results of a tree regression performed on the household level predictors, using the household level response values (see paragraph 7) to steer the regression. The length of the "branches" of the tree indicates what relative proportion of the variance in the response is related to the variance in the household level predictor.

This diagram indicates that most of the household this household level response is well associated with geyser presence, time electrified, and built floor area of the dwellings

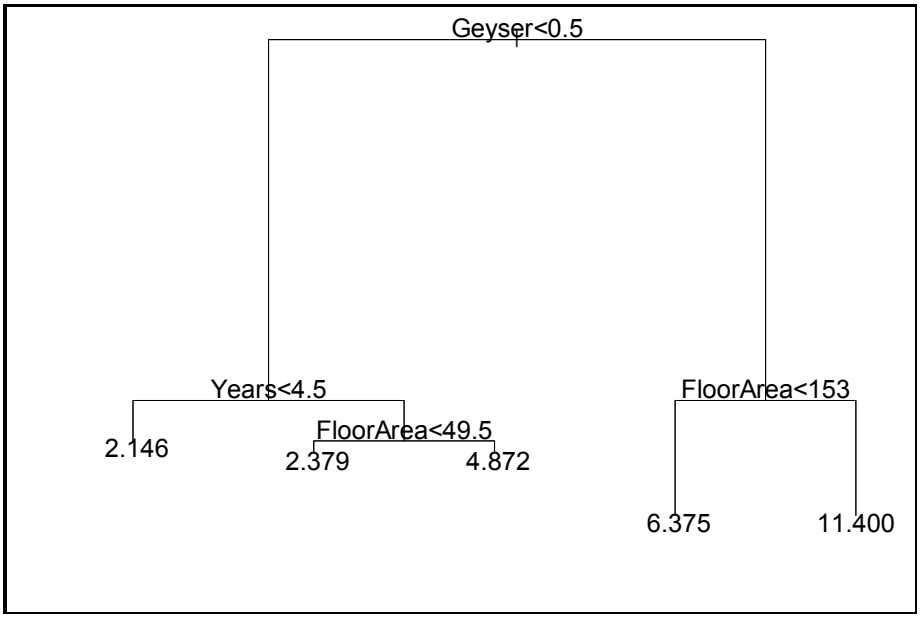

Figure 3. 1 Result of a tree regression showing the predictors which cause the greatest difference in the household level response value (Source: NRS LR Project 1994-1998)

Figure 3.2 shows a histogram of the household level response clusters conditioned on geyser ownership (1994-1998 datasets). Table 3.2 gives a summary of the mean and standard deviation of the two clusters with only the 1994-1997 data, with only the 1998 data and with the complete data set.

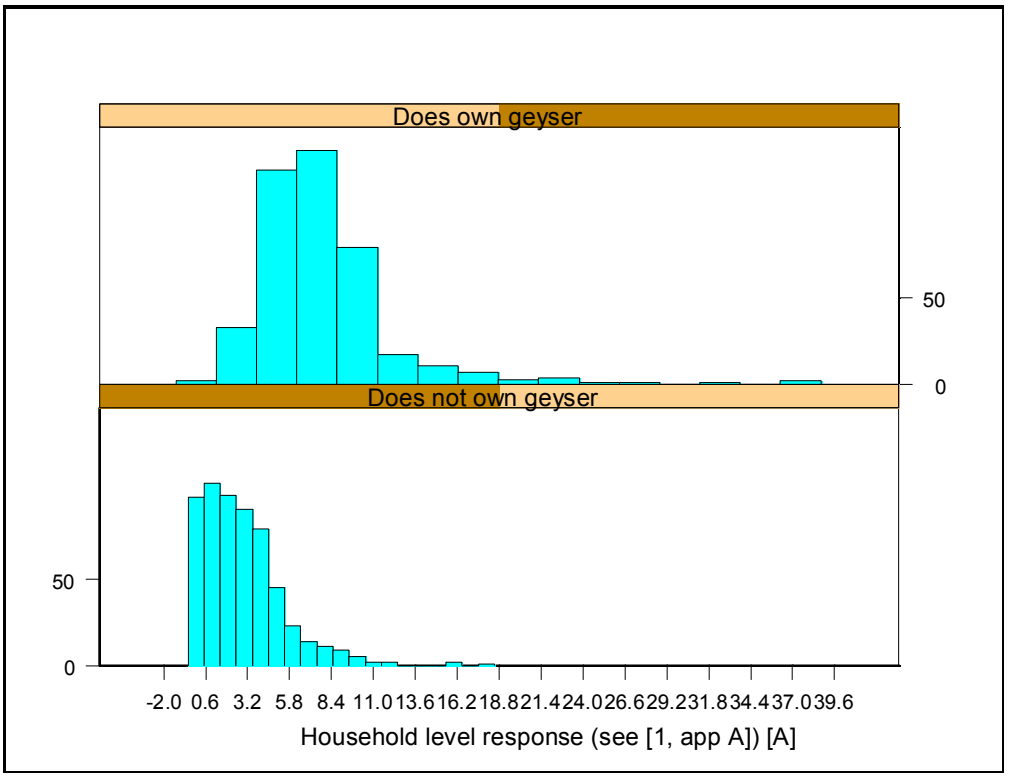

Figure 3. 2 Household level response conditioned on geyser ownership (1994-1998 datasets)

Table 3.2 A summary of the mean and standard deviations of the household level response of the consumers in the two geyser clusters for the different datasets

Data set	Does not own geyser $\mu(\sigma)$	Owns geyser $\mu(\sigma)$
1994-1997	2.8 (2.9)	10.6 (6.2)
1998	4.6 (3.0)	8.6 (5.6)
1994-1998	3.4 (3.0)	9.6 (6.0)

The household level response clusters based on geyser ownership is present in both the older and latest datasets. The difference between the clusters in the 1998 datasets are smaller than for the previous years. This can be attributed to the Tafelsig and Orient Hills datasets as described in Appendix A & B and paragraph 2.2.

Geyser ownership influences the demand of domestic consumers significantly.

Recommendation:

- **The development of a model for the prediction of geyser ownership was recommended in the 1997 review report. The importance of such a model is confirmed with above analysis.**
- **The addition of more data points where partial geyser penetration has occurred or where geyser penetration is currently taking place is necessary for the development of a geyser ownership model. Currently two datasets have partial geyser penetration (see paragraph 2 and appendix A,B).**

3.1 Proposed geyser ownership model

Based on the available data the following logistic regression is proposed for geyser ownership at group level:

$$GP = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad (1)$$

where

GP	Geyser penetration [pu] at group level
x_1	(Piped water supply [pu] at group level) * (Income) * (Time since electrification)
$\beta_{0,1}$	Coefficients of the regression (β_0=-5.7, β_1=0.3 E-3)

Figure 3.3 shows the fitted regression on the geyser penetration in 21 townships. The residual deviance of a non-linear model is the equivalent of the residual sum of squares for a linear model. The residual deviance gives an indication of the size of the remaining deviance in the model and can be compared to the null deviance which is the deviance in the source data. The residual deviance for the proposed geyser penetration model is 2.1% of the null deviance.

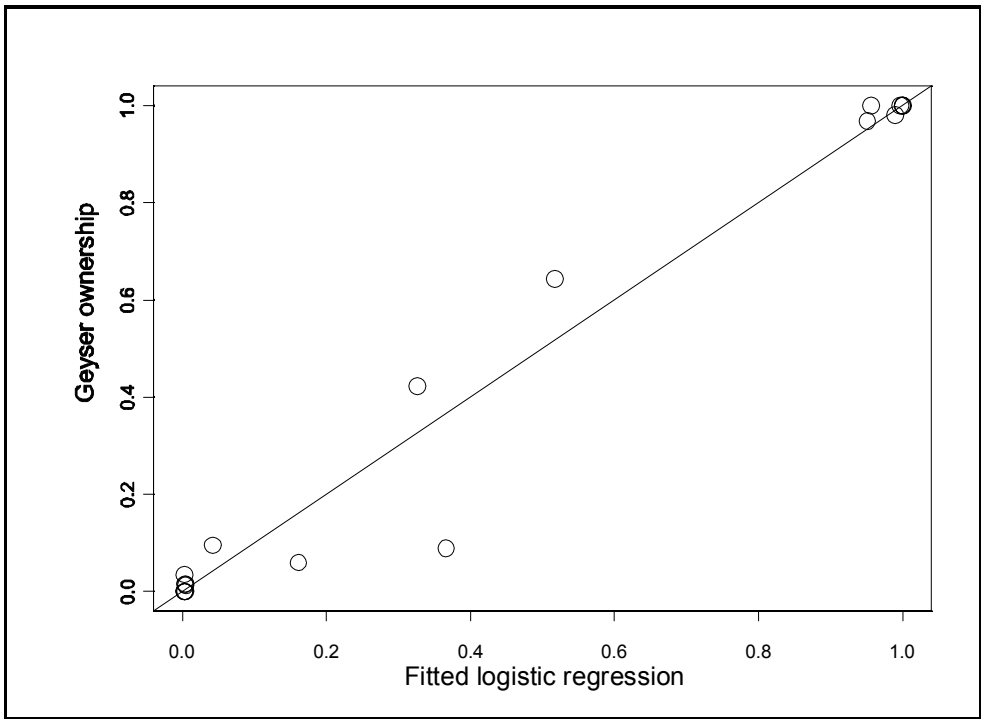

Figure 3. 3 Logistic regression model for geyser ownership based on income, time since electrification and piped water penetration (Source: NRS LR Project 1994-1998, Residual deviance=2.1% of null deviance, N=21)

Recommendation:

This is a proposed model based on the current data set and should be updated / verified as more datasets become available.

4. Review - Socio demographic parameters

The socio demographic parameters contain both missing and inaccurate data. The following two paragraphs summarize and analyze the errors.

4.1 Missing data

A total of 1377 household predictor sets were found with the following missing data per township, the percentages of missing data is for a particular township. A summary of missing data is given in the total row.

Table 4.1: Percentage missing data for four key factors per township (Status at 99/1/7)

Township name	Floor area	Income	Cb size	Time since electrification
Claremont 96	2%	20%	2%	68%
Claremont 97	1%	10%		45%
Claremont 98		10%		2%
Cloetesville 94/5	3%	5%		
Helderberg 97	2%	2%		98%
Helderberg 98		6%	47%	
Kwazakhele 95/6	32%	23%	98%	2%
Lotus Prk 98				
Manyasteng 97			1%	
Manyatseng 96	2%	7%	2%	
Orient Hills 98				
Sweetwaters 96	2%			
Sweetwaters 97			1%	
Tafelsig 98		1%		
Umgaga 98				
Umlazi 98		2%	2%	2%
Walmer 98	2%			
Walmer Est 97				
Total	3%	5%	9%	12%

Total number of complete household level predictor sets: 1 134 (82%)

4.2 Inaccuracy in the predictor set

There are several ways in which to evaluate the predictor set. The system applied depends upon the data use.
Here we are primarily interested in overall system accuracy.

Recall that sociodemographic data is collected in the following steps:

- Field data collection

- Back check to some consumers
- Form check (look for obvious errors)
- Entry of data onto database & check

The accuracy of household level predictors may be evaluated using a year to year relationship analysis. The following paragraph describes the proposed evaluation method and some typical results calculated for Claremont 1997 to 1998.

4.2.1 Method to estimate typical error

The datasets collected from each township can be traced from year to year be using the logger ID and the channel no assigned to it. This assumes that these identifiers did not change from year1 to year2. This assumption was checked by comparing the name and street address of each of the consumers from year1 to year2. Claremont 1997 to 1998 showed consistent name and street address and was used to obtain an indication of repeatability of the sociodemographic parameters.

Recommendation:

It is recommended that a unique identifier is allocated for each physical location measured.

4.2.2 Errors in income, floor area, years electrified

Figure 4.1, 4.2 and 4.3 shows scatter plots of the household income, floor area and time since electrification recorded in Claremont from 1997 to 1998. The standard error is given in each case as an indication of the size of the error for each of the predictors (Table 4.1)

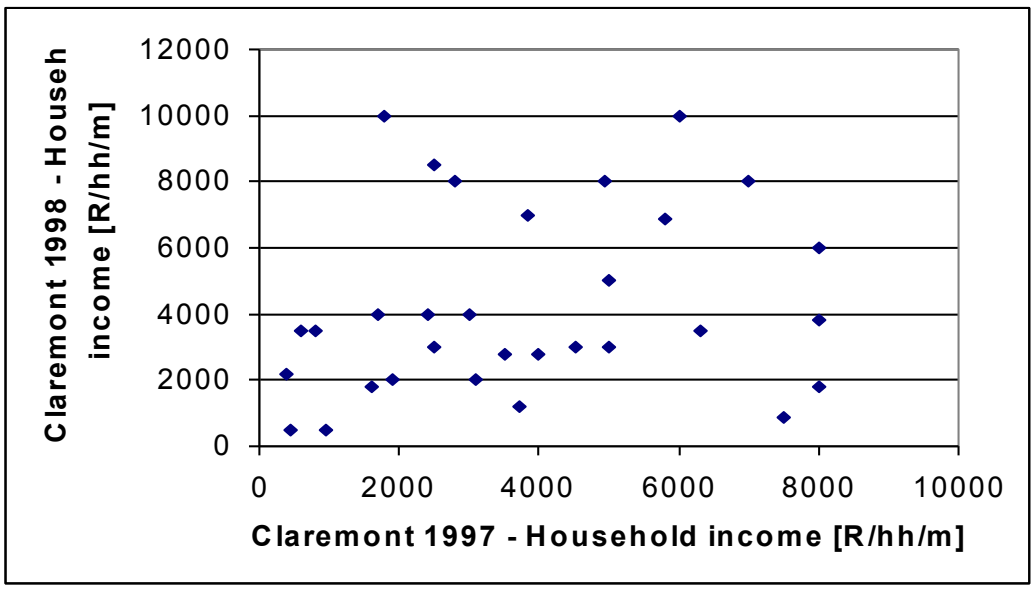

Figure 4. 1 Scatter plot of reported household income for the same households from 1997 to 1998

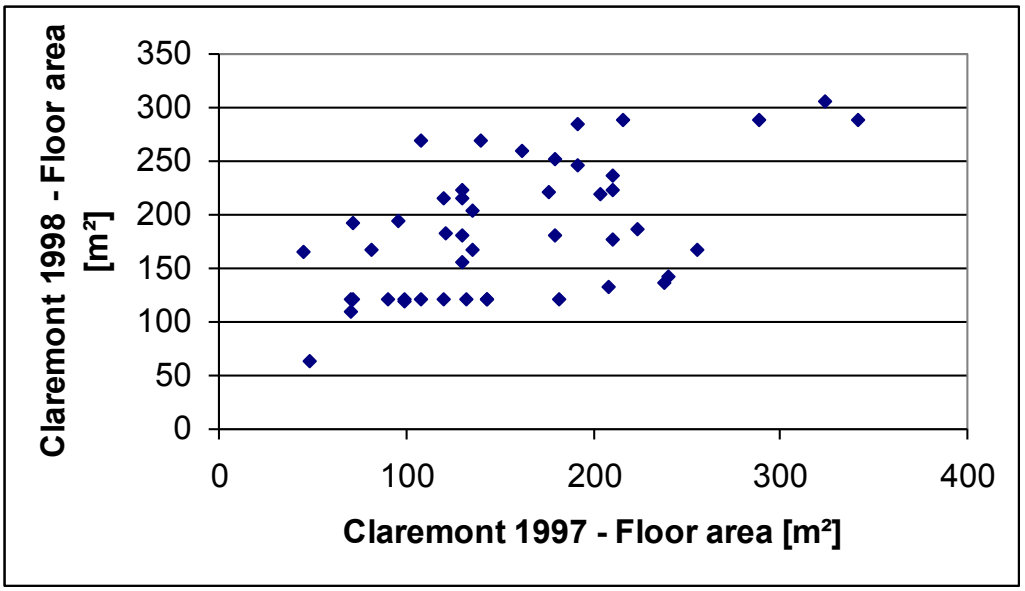

Figure 4. 2 Scatter plot of reported floor area for the same households from 1997 to 1998

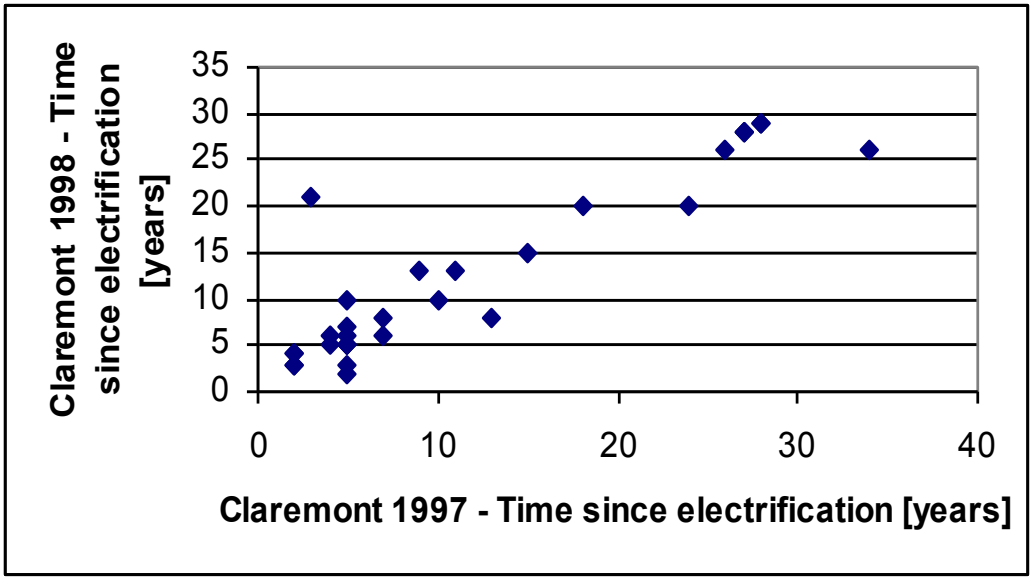

Figure 4. 3 Scatter plot of reported time since electrification for the same households from 1997 to 1998

Table 4.1 Standard error at household level and at group level for income, floor area and time since electrification

	Household income [R/hh/m]	Floor area [m²]	Time since electrification [years]
Standard error at household level	2 188	63	2.5
Standard error at group level	297	8	0.3

4.2.3 Error for other socio demographic data

The repeatability of the socio demographic information at household level for Claremont 1997 and 1998 can be measured in terms of the standard error. The standard error is the error that is obtained if the 1997 dataset is used to predict the 1998 dataset using a linear regression.

The standard error is of the same dimension as the measured statistic, while the coefficient of variation of error is scaled to the mean of the statistic and is therefore dimensionless. Table 4.2 is a list of standard errors and coefficients of variation for different key socio demographic indicators as measured in Claremont 1997-1998.

Table 4.2 Standard error and coefficient of variation at household level for Claremont 1997-1998

	Standard error	Coefficient of variation
Income (R/hh/m)	2 188	0.49
Floor area (m²)	63	0.31
Time since electrification (years)	2.5	0.19
Three plate stove	0.41	2.57
Four plate stove	0.41	0.51
Hotplate	0.18	5.56
Deep freeze	0.41	0.97
Fridge	0.21	0.23

Note that income, floor area and time since electrification is repeated from table 4.1.

Conclusion:

The error in the predictors at household level reduces significantly at group level and suggest that the group level predictor values may be more significant than the household level values, e.g. the average income of the group from which a consumer was sampled is more significant than the reported household income.

To the knowledge of the authors, this is the first time that year-on-year repeatability of sociodemographic data has been assessed for a research project.

It is hoped that further development in this area will allow us to usefully manage between the sources of error.

This typical quality of the sociodemographic data collection *system* repeatability at household level explains in part why household load models have not been extracted by other researchers in the past. Normally market research work is tested against secondary data from the same year, and comparisons are based upon statistical tests between grouped data.

4.2.4 Improved modeling

The analysis in paragraph 4.3.2 indicates that some of the sociodemographic information about the group in which a consumer is sampled might be more significant than the information about the individual consumer. This is confirmed by repeating the tree analysis on the household level response, but instead of using the income per household, the average household income for the group is used. Similar replacements were made for time since electrification and floor area of a dwelling. The tree was pruned with cost complexity pruning to 5 nodes and the result is shown in figure 4.4.

Note that group level income replaces geyser penetration as the most significant predictor. It should however also be noted that geyser penetration can quite successfully be predicted by the group level income and time since electrification. (see paragraph 3.1)

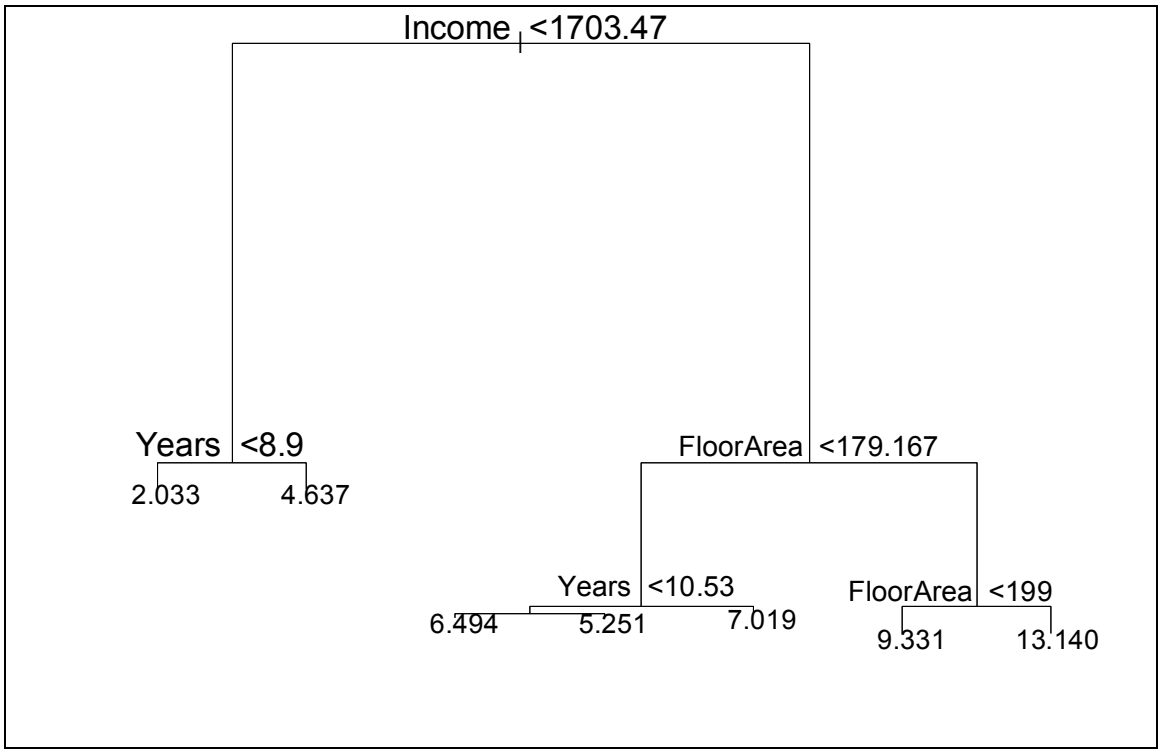

Figure 4.4 Tree analysis of household level response using group level predictors

Conclusion:

Traditional predictors, household income and time since electrification, are prone to large errors at household level but can successfully be used at group level. The ADMD model could be improved by taken this fact into account (see paragraph 5)

If the development of a household load model is important, then the form and implementation of the market research work from which the Sociodemographic information is extracted, needs to be re-evaluated in terms of accuracy and repeatability:

- **does a question measure the intended variable from different consumers= accuracy**
- **does the same question get the same answer at different times from the same consumers= repeatability**

Recommendation:

It is recommended that a control mechanism should be implemented to identify predictor sets with large errors. This mechanism should be implemented as a first phase in the predictor setprocessing

5. Updated ADMD model

The ADMD model can be updated using the new results found as discussed in above paragraphs as follows:

- Paragraph 2 - Geyser size should be taken into account when considering demand
- Paragraph 3 - Proposed geyser ownership model
- Paragraph 4 - Income and time since electrification at *group level* are significant predictors

Figure 5.1 shows the ADMD predictions and measurements using

a. A linear regression with
*ADMD=2.15 (*Geyser ownership*)(*Geyser size [k*W])* + 2 E-4 (*Income*)(*Time since electrification*) + 3.2*
SE=1.44, R^2=0.93

b. A linear regression as in (a) but with geyser ownership predicted using the logistic regression described in paragraph 3.3
SE=1.25, R^2=0.94

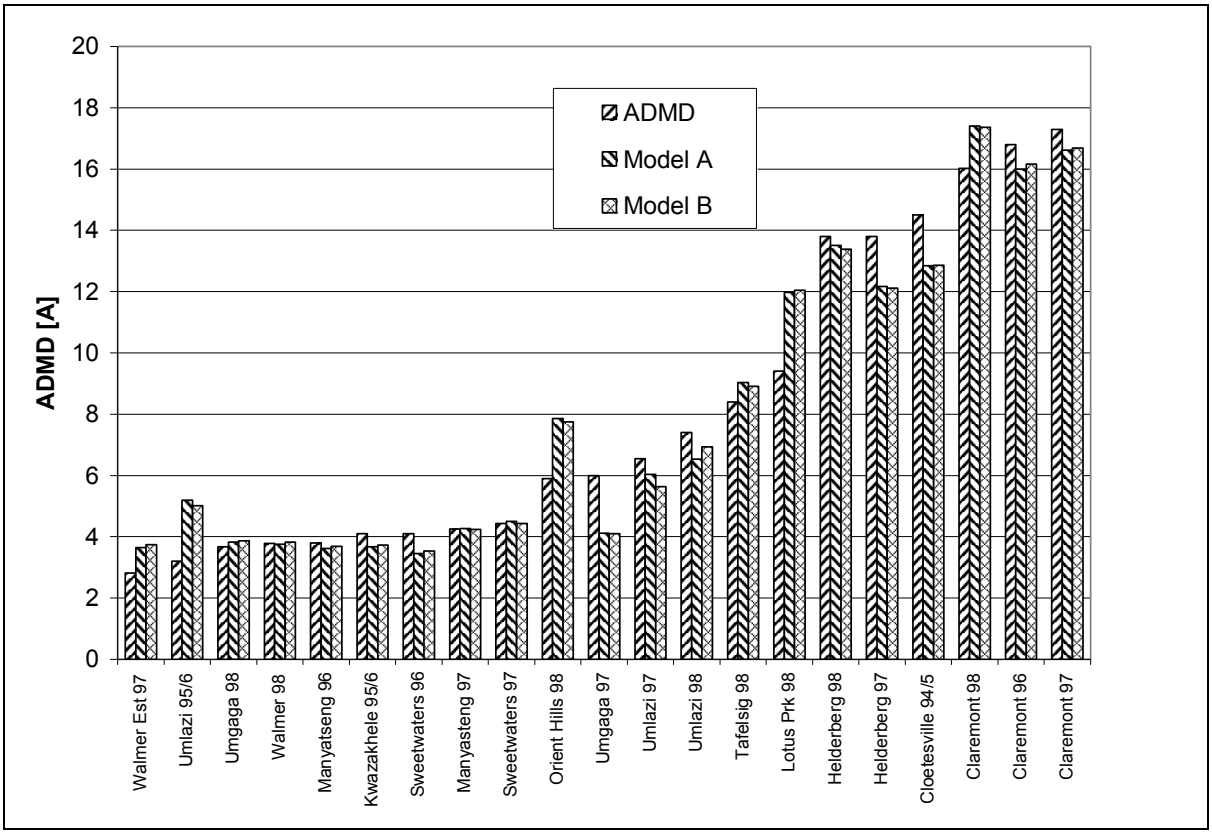

Figure 5. 1 A bar graph showing the measured ADMD and the predicted ADMD with both model A and model B (Source: NRS LR Project 1994-1998)

Using this approach therefore, about 6% of the variance in the model of Admd is still unexplained. This amounts to an overall error of 1.25A, i.e. 280W. The model has the following deficiencies:

- Model does not display saturation at long times after electrification
- It appears to have a useful time range from ? to 11 years, and R/hh/m to R/hh/m
- It does not yet cover poor and rural communities

Conclusions:

- **The ADMD prediction model should be re-evaluated as new residential LR data sets become available.**
- **Buying power or some similar inflation-independent measure of income should be used in the predictions. This should account for the effect of inflation on the model.**
- **A time-since-electrification saturation factor be built into the model to counter unrealistic longer term extrapolations**

6. Recommendations and conclusions

The recommendations and conclusions are printed in **bold** at the end of each section and are repeated here to summarize:

From paragraph 2.1 – "Holes in the predictor set"

Two objectives emerge from this analysis:

- Populate the remaining holes in the data set
- Target poor communities in urban and rural environments

From paragraph 2.2 – "Geyser ownership"

It is important to gain an understanding of this phenomena since geyser penetration has a mayor influence on the overall household level response in all the other datasets.

From paragraph 3 –" Verification of results described in 1997 load survey report using data from 1998 load survey"

- The development of a model for the prediction of geyser ownership was recommended in the 1997 review report. The importance of such a model is confirmed in the analysis.
- The addition of more data points where partial geyser penetration has occurred or where geyser penetration is currently taking place is necessary for the development of a geyser ownership model. Currently two datasets have partial geyser penetration (see paragraph 2 and appendix A,B).

From paragraph 3.1- "Proposed geyser ownership model"

This is a proposed model based on the current data set and should be updated / verified as more datasets become available.

From paragraph 4 – "Review - Socio demographic parameters"

Par 4.2.1 – "Method to estimate typical error"

It is recommended that a unique identifier is allocated for each physical location measured.

Par 4.2.3 – "Error for other socio demographic data"

The error in the predictors at household level reduces significantly at group level and suggest that the group level predictor values may be more significant than the household level values, e.g. the average income of the group from which a consumer was sampled is more significant than the reported household income.

Par 4.2.4 – "Improved modeling"

Traditional predictors, household income and time since electrification, are prone to large errors at household level but can successfully be used at group level. The ADMD model could be improved by taken this fact into account (see paragraph 5)

If the development of a household load model is important, then the form and implementation of the market research work from which the Sociodemographic information is extracted, needs to be re-evaluated in terms of accuracy and repeatability:

- does a question measure the intended variable from different consumers = accuracy
- does the same question get the same answer at different times from the same consumers = repeatability

It is recommended that a control mechanism should be implemented to identify predictor sets with large errors. This mechanism should be implemented as a first phase in the predictor set processing

From paragraph 5 – "Updated ADMD model"

- The ADMD prediction model should be re-evaluated as new residential LR data sets become available.
- Buying power or some similar inflation-independent measure of income should be used in the predictions. This should account for the effect of inflation on the model.
- A time-since-electrification saturation factor be built into the model to counter unrealistic longer term extrapolations

7. Definitions

Household level response: This is a statistic which is calculated per household during the month in which the peak loading of the group in which the household was sampled, occurred. The calculation of the statistic is shown below, but may be seen as a measure of the average energy demand during periods of high system demand.

Group level sociodemographic information: This refers to any sociodemographic information that is calculated as the average of the sociodemographic information of the households in a specific group

Household level sociodemographic information: This refers to the sociodemographic information as captured in questionaires from a sample of households in a specific township.

8. References

[1] Heunis SW, “Analysis of National Load Research Project: Load & socio-demographic data”, Marcus Dekenah Consulting cc, 1998/5/15

Appendix A - Analysis of Tafelsig 1998

The analysis of the 1994-1997 NLR data revealed that geyser penetration has a significant effect on household demand. However none of the townships has partial geyser penetration. Tafelsig 1998 has partial geyser penetration and a study within the boundaries of a township is a good test of this significance.

A tree analysis of the household level response in Tafelsig 1998 (fig A.1) was performed with the same predictor set used on the 1994-1997 data. Figure A1 shows that within the context of this township, geyser ownership has the greatest influence on the individual household demands. This suggests the presence of two household level demand clusters based on geyser ownership.

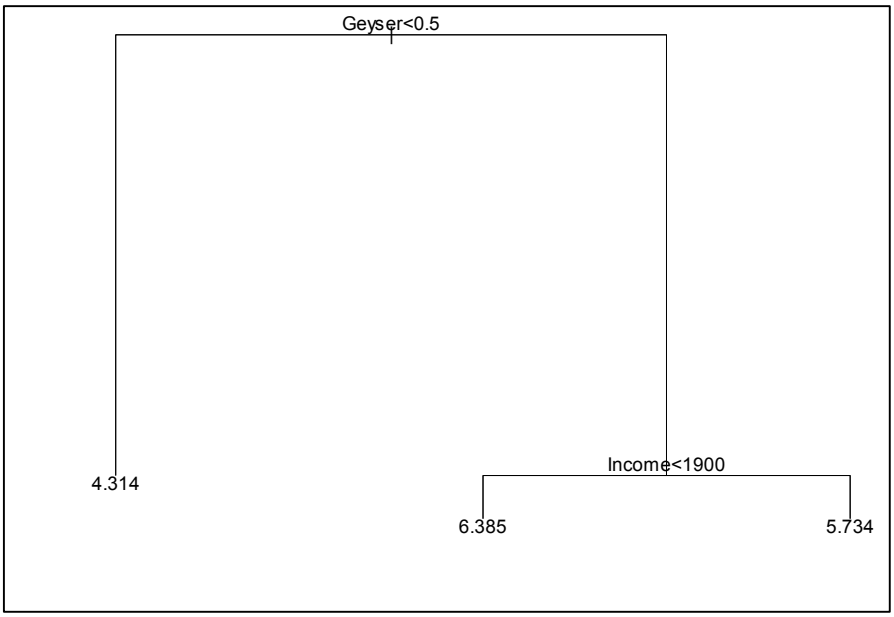

Figure A. 1 Tree analysis of the Tafelsig data set (the number of households per node was set at minimum 30

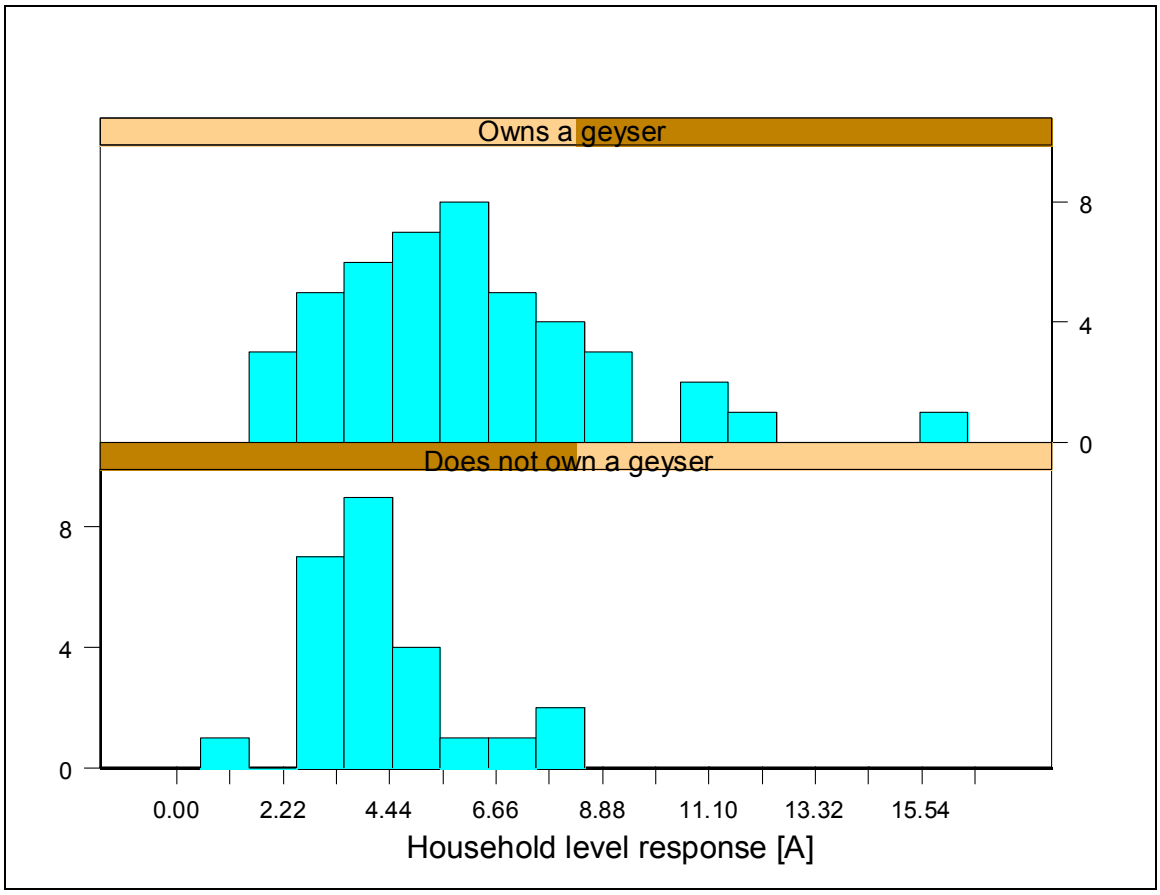

Figure A. 2 A geyser ownership conditioned histogram of the household level response [1,Appendix A]

Figure A.2 shows a geyser ownership conditioned histogram of the individual household level response ([1], Appendix A). These two histograms represent the geyser clusters. The distance (in mean demand) between the clusters is 1.8 A compared to the distance between the clusters in the 1994-1998 datasets of 6 A.

The difference can be ascribed to the following:

- Geysers with a smaller rating (relative to the 1994-1997 communities) are generally found in Tafelsig 1998.(see appendix C). A difference of 3A is estimated using the technique described in appendix C.
- Some customers reported hot water cylinders but no evidence of geyser usage could be found when the individual consumers were analysed.
- The load data for Tafelsig 1998 was recorded for 1 month only and may not be representative of the peak demand period.

Figure A.3 shows the clusters conditioned on geyser usage instead of geyser ownership. The distance between the clusters based on geyser usage is 2.5 A.

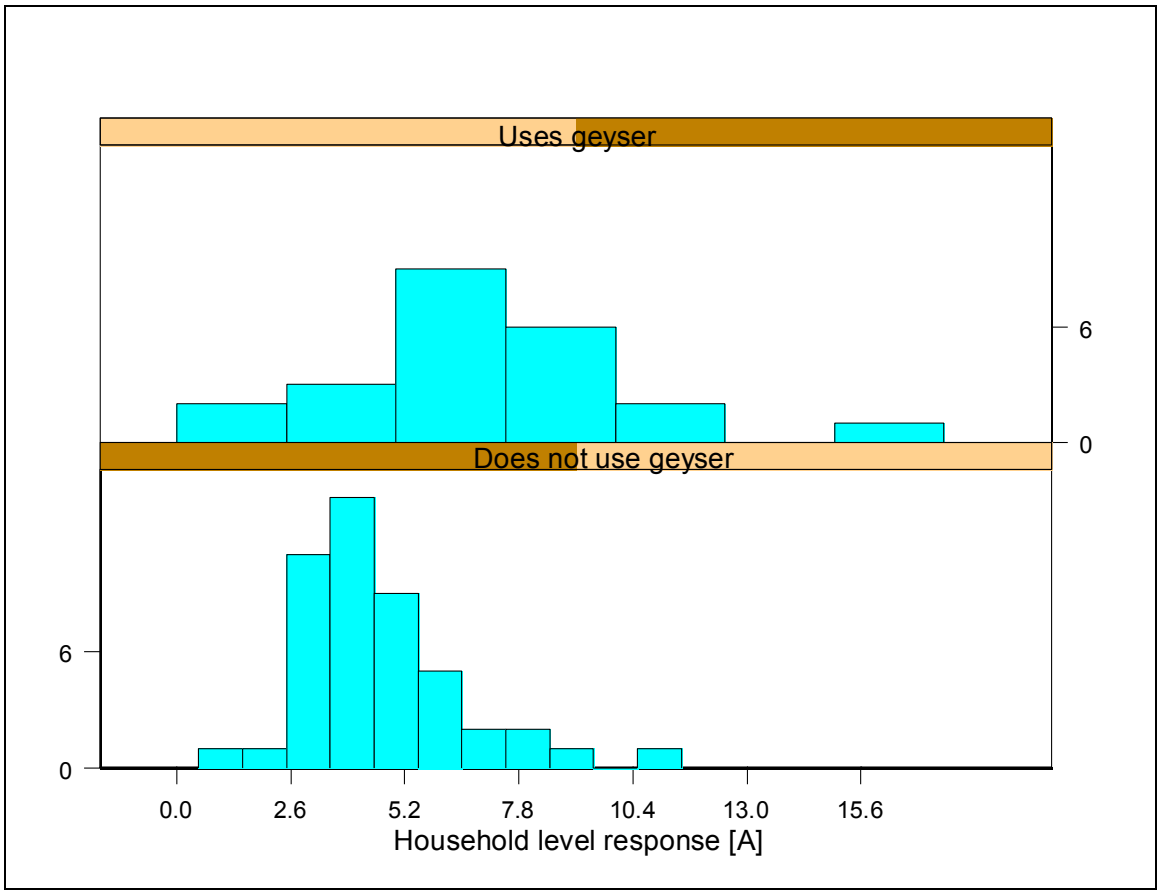

Figure A. 3 Household level response conditioned on geyser usage (see appendix B)

Appendix B - Analysis of Orient Hills 98

Figure B1 shows a histogram of the household level response for the consumers in Orient Hills, conditioned on geyser ownership. It is clear that no distinct household demand/geyser ownership clusters are present in this data set.

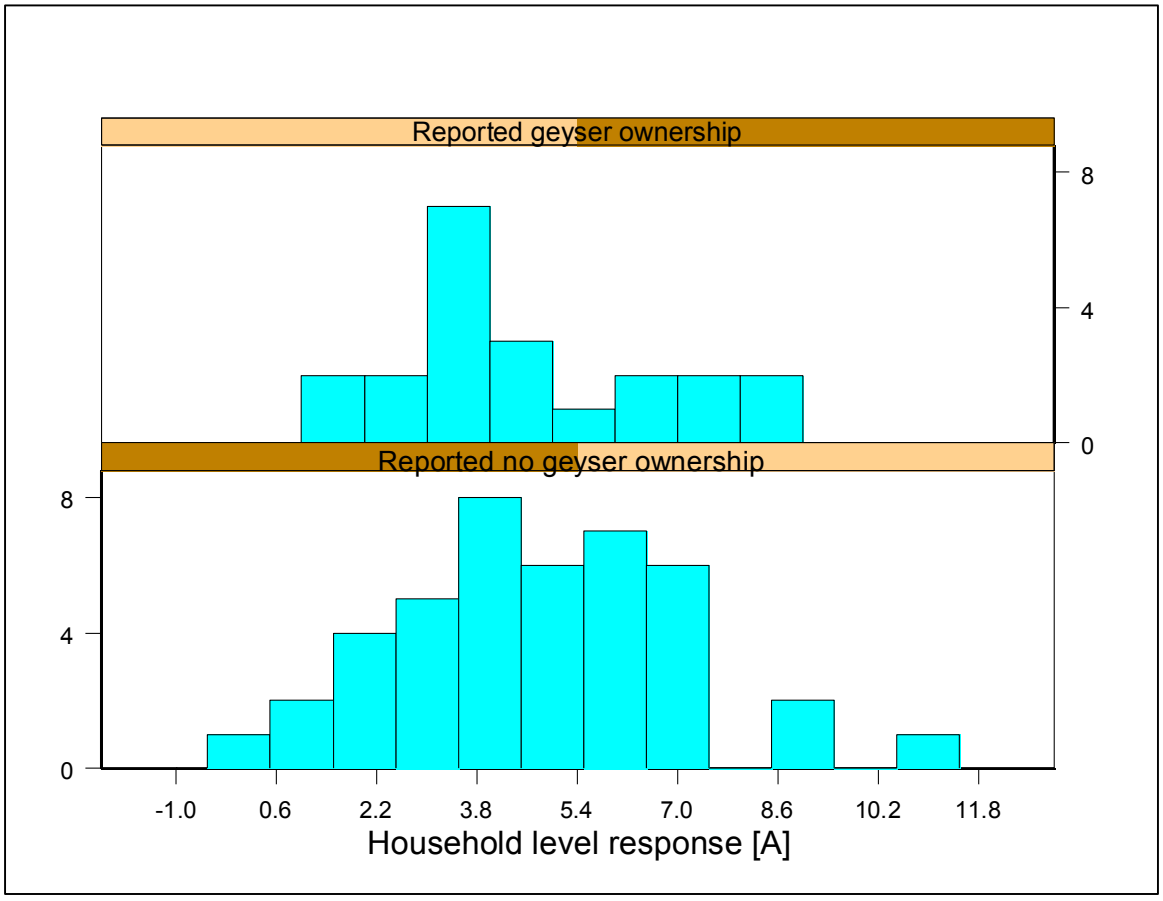

Figure B. 1 A histogram of the household level response conditioned on geyser ownership for the customers in the Orient Hills data set

Further investigation is required to determine what sets Orient Hills apart from all the other datasets.

Appendix C - Method used to estimate geyser usage and geyser size

The geyser usage in the datasets were estimated by analysing the early morning maximum demand of each of the consumers. The analysis is based on the following assumptions

- All the geysers is rate 2~3 kW.
- Geyser lose heat and periodically have to switch on to counter the heat loss
- Very few other appliances are used during the period 0:00 to 4:00
- By taking the maximum change in 5 minute demand for two weeks, the effect of alliasing can be eliminated.

The estimate of the geyser rating is clearly not very accurate but serves as an indication of the typical size found in the township. To compare reported geyser ownership and geyser usage, the maximum 5 minute change in demand during the period 0:00 to 4:00 is analysed. Figure C1 shows a histogram of maximum change [A] conditioned on reported geyser ownership.

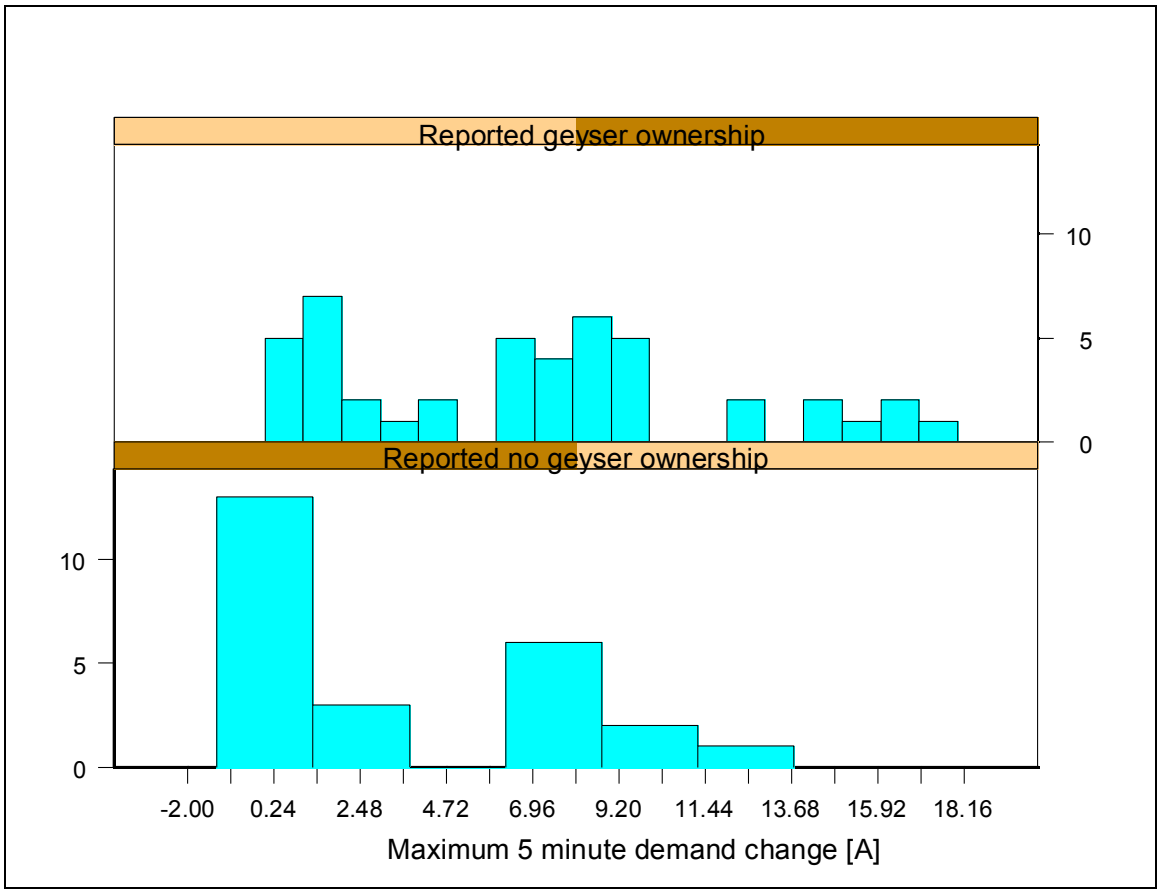

Figure C. 1 Maximum 5 minute demand change during early mornings for Tafelsig 1998

A number of the consumers who reported geyser ownership do not show geyser switching during the early hours of the morning. This could be due to the following

- The geysers are not switched on during the period in question

- The geysers are not working properly
- One of the assumptions of the analysis is incorrect
- The geyser ownership reported is incorrect

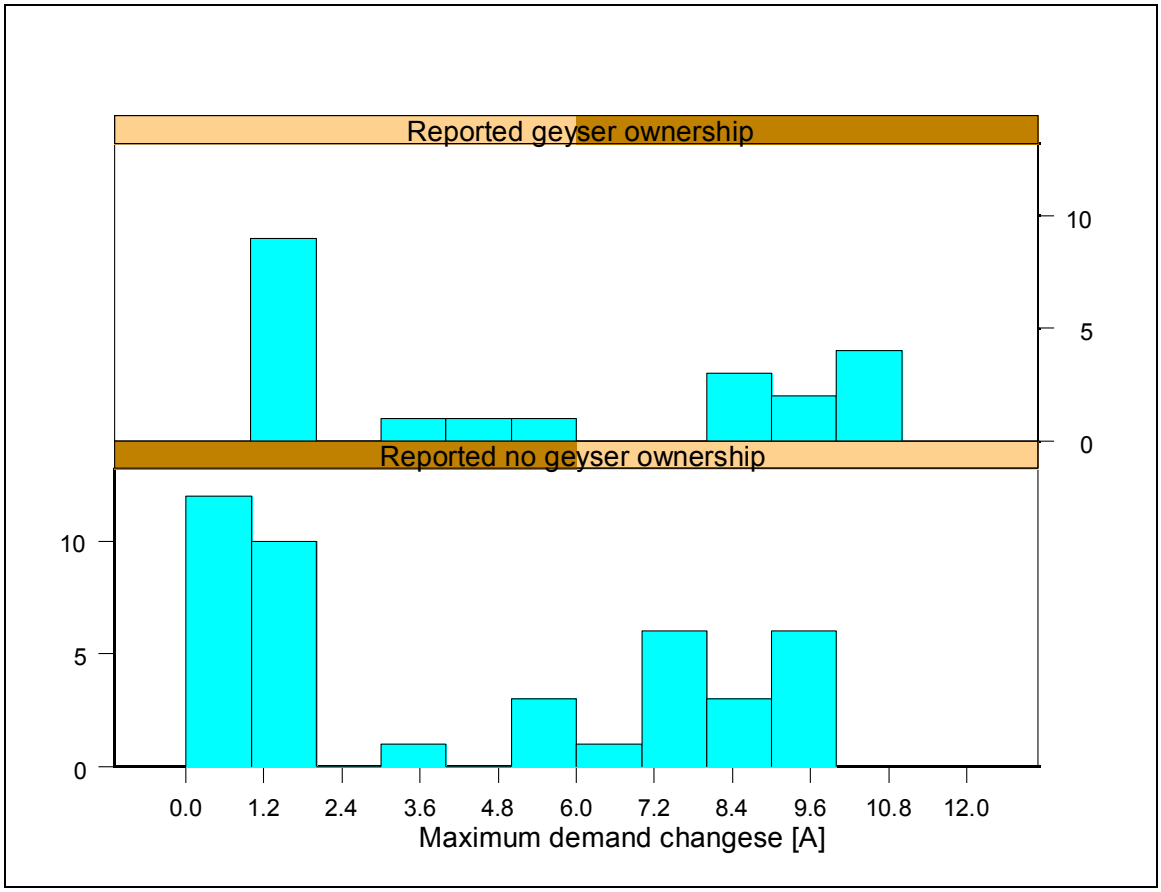

Figure C. 2 Maximum 5 minute demand change during early morning for Orient Hills 1998

A similar picture can be seen for Orient Hills 1998 (figure C2), however the graph may suggest that some of the customers that reported geyser absence may in fact have a geyser.

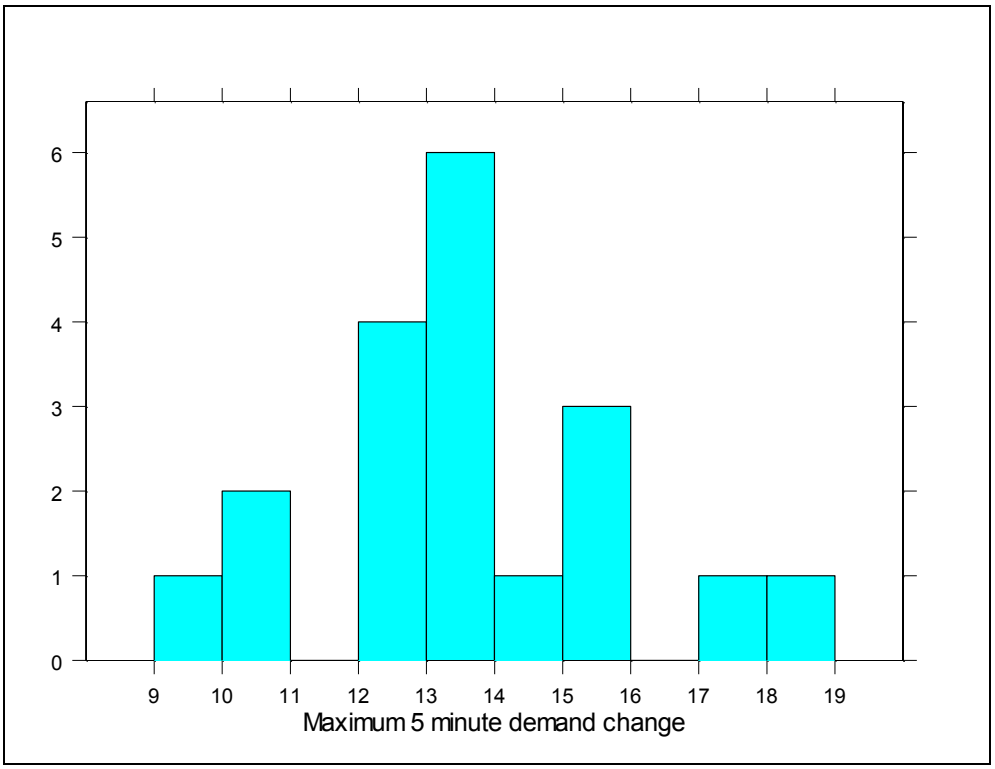

Figure C. 3 Maximum 5 minute demand change during early morning for Claremont 1998

Figure C3 shows the same analysis performed on the customers in Claremont 1998. Claremont reported 100% geyser ownership.

The typical geyser rating is estimated by analysing the distribution of demand changes. Typical geyser rating was assumed to be between 2~3kW and all the demand changes outside this range was discarded. The mean of the remaining demand changes is taken as the typical geyser rating.

This method is not extremely accurate and does not allow for multiple geyser ratings, it does however provide an indication at group level of the expected geyser rating of the group. The typical geyser rating of Tafelsig, Claremont and Orient Hills is as follows:

- Tafelsig 2 kW - 8.6 A
- Claremont 2.8 kW - 12.2 A
- Orient Hills 2.2 kW - 9.2 A

Appendix D - Predictor space set and coding applied

The predictor space set used included the following variables:

Quantitative variables
- Floor area
- Household income level
- Minimum temperature
- Distance to nearest city
- Time since electrification
- Number of rooms in the dwelling
- The size of the circuit breaker feeding the dwelling

Binary variables of appliance ownership (at least one)
- Presence of a deep freeze
- Presence of a fridge
- Presence of a geyser
- Presence of a heater
- Presence of a hotplate
- Presence of an iron
- Presence of a kettle
- Presence of lights
- Presence of an appliance used to produce music
- Presence of a stove with three plates
- Presence of a stove with four plates
- Presence of a TV
- Presence of a washer

Binary variables (other)
- Presence of small business
- Presence of a ceiling
- Presence of insulation
- Ownership of the dwelling
- Presence of a male household head
- The use of coal for heating
- The use of coal for cooking
- The use of paraffin for heating
- The use of paraffin for cooking
- The use of gas for heating
- The use of gas for cooking
- The use of wood for heating
- The use of wood for cooking
- The use of charcoal for heating
- The use of charcoal for cooking
- The presence of outbuildings being supplied from the dwelling

Appendix E - Data analysis

E.1 Data transfer method and hardware used

The consumer load data was transferred using CD Rom (14 townships) and removable hard disk (7 townships). A four gigabyte hard disk was acquired for the data analysis exercise. The hard disk was installed to serve as a removable disk and is intended for large volume data transfer between remote workstations.

The hardware used was a 233 MHz Pentium, with 64 Meg RAM and a 1 gigabyte base hard disk. The total hard disk space available therefore amounts to 5 gigabyte.

E.2 Analysis platform

The data analysis was performed on above mentioned workstation using Windows 95 as operating system. Two software packages were used for the analysis:

1. MS Office 97

Two of the office applications, namely MS Excel 97 and MS Access 97 were used for the initial manipulation of the data. MS Access 97 is a relational database and is capable of importing/linking the load research databases, which are in paradox 4.0 format.

2. S Plus 4.5

S Plus 4.5 is a statistical/data exploration application which has extremely powerful data exploration/visualization capabilities. The application was used for the statistical analysis including the principle component analysis, clustering, tree regressions and multiple regression analysis.

E.3 Analysis process

The analysis process could be summarized in the following steps:

- Condition the raw source data in MS Access

This included setting up the relationships between the data tables, writing source code to extract the household level demand response from the database, extracting the data and linking it to the socio demographic consumer predictor set.

Once the source code was debugged, the extraction process took approximately three hours.

- Transfer the source data to S Plus 4.5 using ODBC

This process is almost instantaneous thanks to the 32 bit ODBC server Microsoft distributes of MS office

- Data analysis with S-plus

This activity contained a series of iterative processes, which included experimenting with different clustering algorithms, scaling of the data set, dealing with missing data, tree regressions, multiple regressions and principle component analysis.

S-plus gives the researcher the opportunity to experiment at a high level with the data. This is done using a simple menu system. It further allows the user to explore at a lower level. This is done by typing high level commands at a command prompt.

The user can explore at an even lower level by writing procedures and function in an object orientated environment.

An extensive range of plots are available with conditioning on each plot type. All the figures in this report was prepared with S-Plus.

The processing time of each data analysis process in S-plus was almost instantaneous which allows the researcher to easily and time efficiently experiment with different data analysis techniques.