A COMPARISON OF THE PERFORMANCE OF THREE MULTIVARIATE METHODS IN INVESTIGATING THE EFFECTS OF PROVINCE AND POWER USAGE ON THE AMOUNTS OF FIVE POWER MODES IN SOUTH AFRICA.

by

BUSANGA JEROME KANYAMA

Submitted in accordance with the requirements for

the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

UNIVERSITY OF SOUTH AFRICA

Supervisor: PROF. SARMA YADAVALLI

June 2011

DECLARATION

Student number: 36460672

I declare that A COMPARISON OF THE PERFORMANCE OF THREE MULTIVARIATE METHODS IN INVESTIGATING THE EFFECTS OF PROVINCE AND POWER USAGE ON THE AMOUNTS OF THE FIVE POWER MODES IN SOUTH AFRICA is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

_____ _____

SIGNATURE DATE

(Mr B J Kanyama)

DEDICATION

This dissertation is dedicated to my mother Coltide Sakina, my brother Pastor Leonard Kanyama, my wife Apoline Kijana and my children: Heritier Akilimali, Gracia Mwamini, Coltide Sakina, Beatrice Mwaibwa and Apoline Makulata, for giving me the opportunity of furthering my education.

ACKNOWLEDGEMENTS

I would like to thank my colleagues: Prof. Principal Ndlovu, Rajab Ssekuma and Muchengetwa Suwisa, who were kind enough to offer their advice and guidance for this dissertation.

Special thanks to my friend and brothers: Pastor Leonard Kanyama, Innocent Bulenge and Mwinyi Malisawa, for their encouragement and the hands-on spiritual experience they have provided to my family.

Without my supervisor, Prof. Sarma Yadavalli and the editorial skills of Ronel Bouwer this dissertation would not have come to fruition.

I remain extremely thankful to my family: my wife Apoline Kijana, my son Heritier Akilimali and my daughters: Gracia Mwamini, Coltide Sakina, Beatrice Mwaibwa and Apoline Makulata, for their patience and support through this work.

SUMMARY

Researchers perform multivariate techniques MANOVA, discriminant analysis and factor analysis. The most common applications in social science are to identify and test the effects from the analysis. The use of this multivariate technique is uncommon in investigating the effects of power usage and Province in South Africa on the amounts of the five power modes. This dissertation discusses this issue, the methodology and practical problems of the three multivariate techniques. The author examines the applications of each technique in social public research and comparisons are made between the three multivariate techniques.

This dissertation concludes with a discussion of both the concepts of the present multivariate techniques and the results found on the use of the three multivariate techniques in the energy household consumption. The author recommends focusing on the hypotheses of the study or typical questions surrounding of each technique to guide the researcher in choosing the appropriate analysis in the social research, as each technique has some strengths and limitations.

KEY TERMS

Multivariate analysis of variance (MANOVA), discriminant analysis, factor analysis, statistical tests, two-way factor, theory based on MANOVA, canonical correlation, statistical assumptions, partial eta-square, post hoc test, correlation analysis, correlation matrix, factor, component, principal component analysis, exploratory factor analysis and confirmatory factor analysis.

CONTENTS

CHAPTER 1

INTRODUCTION TO STUDY

1.1 INTRODUCTION 2
1.1.1 TEST STATISTIC INTRODUCTION 7
1.1.2 WILKS' LAMBDA INTRODUCTION 7
1.1.3 BARTLETT'S TEST INTRODUCTION 7
1.1.4 TEST STATISTIC OF TUKEY INTRODUCTION 8
1.1.5 BOX'S M TEST INTRODUCTION 9

CHAPTER 2

MULTIVARIATE ANALYSIS OF VARIANCE

2.1 OVERVIEW INTRODUCTION 11

2.2 DEVELOPMENT OF A THEORY BASED ON THE MULTIVARIATE ANALYSIS OF VARIANCE INTRODUCTION INTRODUCTION 12

2.3 STATISTICAL ASSUMPTIONS INTRODUCTION INTRODUCTION 17

2.4 METHODOLOGY OF TWO–WAY USING MANOVA INTRODUCTION 19

2.5 APPLICATION OF MANOVA METHOD USING REAL DATA INTRODUCTION 20

2.5.1 BACKGROUND TO THE REAL DATA OF THE STUDY INTRODUCTION 20

2.5.2 MANOVA TWO– WAY FACTORIAL INTRODUCTION 22

2.6 DATA INTRODUCTION INTRODUCTION 23

2.6.1 DATA SCREENING 25

2.6.2 EXAMINING CORRELATION..... 26

2.6.3 EXAMINING MULTIVARIATE EFFECTS..... 26

2.6.4 FOLLOWING A SIGNIFICANT MULTIVARIATE EFFECT..... 29

2.7 CONCLUSION 32

CHAPTER 3

DISCRIMINANT ANALYSIS

3.1 OVERVIEW 34

3.2 DISCRIMINANT ANALYSIS ASSUMPTIONS AND FUNCTIONALITY 34

3.3 APPLICATION OF DISCRIMINANT ANALYSIS USING REAL DATA..... 36

3.3.1 INTRODUCTION 36

3.3.2 DISCRIMINANTION DATA ANALYSIS..... 39

3.3.2.1 DEPENDENT VARIABLE: PROVINCE 39

3.3.2.2 DEPENDENT VARIABLE: POWER USAGE 45

3.4 CONCLUSION 48

CHAPTER 4

FACTOR ANALYSIS

4.1 INTRODUCTION 49

4.2 METHODOLOGY OF FACTOR ANALYSIS..... 50

4.3 WORKING WITH FACTOR ANALYSIS..... 51

4.3.1 INTRODUCTION 51

4.3.2 PRINCIPAL COMPONENT ANALYSIS SOLUTION..... 52

4.5 CONCLUSION 57

CHAPTER 5

COMPARISON OF DIFFERENT TECHNIQUES

5.1 THEORICAL DISCUSSION. 60

5.2 OVERALL CONCLUSION..... 63

BIBLIOGRAPHY..... 64

CHAPTER 1

INTRODUCTION TO STUDY

1.1 INTRODUCTION

This work analyses the performance of the multivariate techniques using the multivariate analysis of variance (MANOVA), the discriminant analysis and factor analysis. These techniques are most applied in social science to identify and test the effects from the analysis. The use of this multivariate technique is uncommon in investigating the effects of power usage to the nine Province in South Africa on the amounts of the five power modes (electricity, gas, paraffin, solar and other). Indeed, energy household consumption problems are, by definition, multivariate. Two factors were simultaneously used to this multivariate problem as they are known to be a source of power consumption. In a two-way layout, the measurements are recorded at various levels of two factors. Both main effects and combined-effect (interaction) hypotheses will be considered and it is assumed that observations at different combinations of experimental conditions are independent of one another. In this study the author is interested in assessing whether or not the factors involved will have a significant effect on one another. If the two factors do not interact, then the individual effects of the two factors can be investigated separately.

The two-way factor will be examined in the same way as the univariate procedure, but we have to consider two factors at the same time. A factor is an independent variable and we will experiment two independent variables simultaneously. The two independent variables can either both be between groups designs, both repeated measures designs, or mixed designs such as one between groups of independent variables and one repeated measures independent variable.

This section will describe some of the computational details for two-way multivariate. The idea behind “factor” is that there are two variables which affect the dependent variable. Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels.

The treatment groups are formed by making all possible combinations of the two factors. The main effect will involve the independent variables one at a time and the interaction will be ignored in this

procedure. Just the rows or just the columns are used, but not mixed. This part is similar to the two-way analysis of variance. Each of the variances calculated to analyse the main effects are like the between variances.

The interaction effect will be the effect that one factor has with the other factor. In case of interaction the degrees of freedom will be the product of the two degrees of freedom for each factor. The within variations is the sum of squares within each treatment group. For a two-way multivariate all treatment groups must have the same sample size and the degrees of freedom will have one less than the sample size. The total number of treatment groups will be the product of the number of levels for each factor. In this way the within variation is divided by its degrees of freedom. The within group will also be called the error.

A multivariate analysis of variance table will be used to organise data points – indicating the values of a response variable – into groups according to the factor used in each case.

The two sets of experimental conditions have factor “A” and factor “B”, respectively and the analysis of data from experiments will involve two classification variables whose levels are represented by the rows and columns of a two-way layout.

In this study, the objective will be to compare different methods listed below in modelling complex for analysing two-way layout studies:

- Multivariate analysis of variance (MANOVA)
- Discriminant analysis
- Factor analysis

To make appropriate choices among these methods, researchers should understand the statistical models underlying them (Samuel and Thompson, 2006). Each statistical method will present the results using an example of power consumption in the household.

When a number of group comparison strategies need to be chosen, they will include the following:

- To conduct a multivariate analysis of variance (MANOVA) on the multiple dependent variables and then to compute the follow-up analysis of variance (ANOVA) if the MANOVA is significant at the 5% level.
- To conduct a MANOVA on the multiple dependent variables and, if the MANOVA is significant then compute a discriminant analysis to assess linear combination(s) of the dependent variables that differentiate that group.
- To reduce the number of variables, by combining two or more variables into a single component. These factors can be used in further analyses if we need to perform additional analyses using the factors as variables.

Statistical procedures of each model for analysing data will be used when the experiment design include combinations of two-factors, for example factor “A” and factor “B”. The hypotheses of interest for a two-factor experiment will concern the main effects and the combined effect. The strengths and limitations of each approach will be identified by using an application and comparison of method employing real data.

In addition, a statistic will be proposed for testing the combined effect. Two simple models for representing the combined effect of factor “A” and factor “B” – additive and multiplicative – will be considered under the statistical packages that have been used, including SPSS and SAS. The additive model will assume that the combined effect of factor “A” and factor “B” is additive (linear), while the multiplicative model will assume that the combined effect is multiplicative (log-linear).

The study will be conducted on the product of a-levels of factor “A” and b-levels of factor “B”, and “n” independent observations can be detected at each of the (a×b) combinations of levels. The two-way layout will be with one observation per cell for a variety of interaction effects. The experiment procedure will consider the multivariate two-way model in which, in turn, the interaction of the factors will be examined.

The two-way fixed effects model for a vector response will consist of P-component and the k^{th} observation at level “i” of factor “A” and “j” of factor “B” is denoted by X_{ijk} , i = 1, 2, ..., a and j = 1, 2, ..., b and k = 1, 2, ..., n; the two-way fixed-effects for a vector response consisting of P-complements is

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (1)$$

Where $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} \gamma_{ij} = \sum_{j=1}^{b} \gamma_{ij} = 0$

And the vectors are all of order p×1 and ε_{ijk} are assumed to be an $N_p(0, \sum)$ random vector.

μ represents an overall level,

α_i , represents the fixed effect of factor “A”

β_j , represents the fixed effect of factor “B”

γ_{ij}, is the interaction between factor “A” and factor “B”.

The interaction term represents the joint effect of two or more treatments. Interaction terms are created for each combination of treatment variables. Two-way interactions are available, taken two at a time. The number of treatments determines the number of interaction terms possible.

The expected response at the i^{th} level of factor “A” and the j^{th} level of factor “B” is thus

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (2)$$

(Mean response) = Overall average of observation + effects of factor “A” + effects of factor “B” + factor “A” and “B” interaction.

The expected response can be seen as a function of the factor levels with and without interaction, respectively. In the multivariate model for two-way the procedure is the same as in the univariate two-way fixed-effects model with interaction. Thus hypotheses are as follows:

H_0 : No interaction effects ($\gamma_{11} = \gamma_{12} = \gamma_{13} = \ldots = \gamma_{ab} = 0$)

H_1 : At least one $\gamma_{ij} \neq 0$

These hypotheses will specify no factor “A” effects and some factor “A” effects, respectively.

The presence of interaction γ_{ij} will imply that the factor effects (“A” and “B”) are not additive and therefore the interpretation of the results will be complicated (Richard and Dean Johnson 1992: P. 261).

Based on the normal theory assumptions, the most commonly used procedure is the one-degrees of freedom called “Tukey” (1949).

The null hypothesis H_0 is the hypothesis of additive treatment effects for all cells, the most used test in the two-way factors, while the alternative is a general no-additive option.

A test statistic will be conducted while rejecting the hypothesis H_0 for small values of the test statistic (F-values), alternatively P-value < 0.05 significance level. Several test statistics could be used, such as Wilks’ lambda test statistic, Chi-square and Bartlett’s test.

Using Bartlett’s test, the null hypothesis H_0 : no factor “A” effects ($\alpha_1 = \alpha_2 = \alpha_3 = \ldots = \alpha_a = 0$) will be rejected at 5% level if the Bartlett’s value is larger than the critical value, alternatively P-value < 0.05. In a similar manner, factor “B” effects are tested by considering:

H_0 : $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_b = 0$ versusH_1: at least one $\beta_j \neq 0$.

Simultaneous confidence intervals for contrasts in the model parameters can provide insights into the nature of the factor effects. In addition, when interaction effects are negligible, this indicates to concentrate on contrasts in the factor "A" and factor "B" main effects. The Bonferroni approach will apply to the components of the differences α_i- α_m of the factor "A" effects and the components of

β_j- β_q of the factor "B", respectively.

Ordinarily, the test for interaction is carried out before the tests for main factor effects. If interaction effects exist, the factor effects do not have a clear interpretation. From a practical standpoint, it is not advisable to proceed with the additional multivariate tests. Instead, from the responses of P univariate two-way analyses of variance (one for each variable) are often conducted to see if the interaction appears in some responses but not others (Richard A. J. and Dean W.W. 1992: P. 265). Those responses without interaction may be interpreted in terms of additive factor "A" and factor "B" effects.

1.1.1 TEST STATISTICS

1.1.2 Wilks' lambda

The statistic Wilks' lambda is the most common and traditional test in which there are more than two groups formed by the independent variables. It is a multivariate F-test, similar to the F-test in univariate ANOVA. The lower the Wilks's lambda, the greater the differences and the more the given effect contribute to the model. The t-test, Hotelling's T, and F-test are special cases of Wilks's lambda. For large samples, Wilks's lambda can be referred to a Chi-square.

This test will reject the null hypothesis (no interaction effects) at α level when the Wilks' lambda is greater than the critical value (or P-value less than α). For two-way layout the test for interaction is carried out before the tests for main factor effects. If the interaction effects exist, the factor effects do not have a clear interpretation (Richard Johonson & Dean Wichern 1992: P. 264). It is advisable from a practical point of view to proceed with the additional multivariate tests.

1.1.3 Bartlett's test

This is designed to test for equality variances across groups against the alternative that variances are unequal for at least two groups. Equality of variances across samples is called homogeneity of variances. Bartlett's test is sensitive to the departures from normality. It may simply be testing for non-normality. Levene's test is an alternative to the Bartlett test that is less sensitive to departures from normality.

The variances are judged to be unequal if, the Bartlett test statistic is greater than the corresponding critical value (χ^2) at the significance level α. The Chi-square test and analysis of variance are related techniques to Bartlett's test, in which it is available in many general purpose statistical software programs.

1.1.4 Test statistic of Tukey

Tukey's test is similar to the t-test, except that it corrects for experiment-wise error rate, meaning when multiple comparisons are being made, the probability of making a type I error (rejecting the null hypothesis when it is true) increases with t-test. Tukey's test is more suitable for multiple comparisons than doing a number of t-tests.

The test compares the means of every treatment to the means of every other treatment; that is, it applies simultaneously to a set of all pair-wise comparisons μ_i - μ_j and identifies where the difference between two means is greater than the standard error would be expected to allow.

The test statistic Tukey is also used to test the null hypothesis that interaction γ_{ij} = 0 for all "i" and "j" under the assumption that γ_{ij}= δ. α_i. β_j for some constant δ (Cf. Ghosh & Sharma 1963). However, Hartlaub, Dean & Wolfe reported that Tukey's test is not useful for detecting general non-additive effects and is not robust with respect to departures from the model which assumes normal error terms

(Hartlaub, Dean & Wolfe Dec, 1999: P. 864). The Tukey test statistic has reasonably good power when the interaction effects are of the product interaction type γ_{ij}=δ. $\alpha_i . \beta_j$.

There exist different approaches to test for interaction. In the two-way layout Iman (1974) and Conover & Iman (1976) suggested the use of the rank-transform approach, but when there are replications in the cells. However, their methodology is not applicable when there is only a single observation per cell.

Recent studies such as Blair, Sawilowsky & Higgens 1989; Akritas 1990; and Thompson 1991 have shown that the rank transform test for interaction can be confused by the presence of both main effects.

1.1.5 Box's M test

Box's M test is a statistic test which tests the homoscedasticity (equal variation of data) assumption in MANOVA such as that the all covariance is the same for any category.

$H_0 : \sum_1 = \sum_2 = ... = \sum_p$ derived a test statistic based on the likelihood-ratio test. For moderate to small sample sizes, an F approximation is used to compute its significance (Box, 1949).

In chapter 2, we will look at the theory of the multivariate analysis of variance method; assumptions and the methodology will be discussed, including its application using real data.

In chapter 3, the discriminant theory and application will be discussed using real data.

In chapter 4, we will look at factor analysis technique, with its relevant approach theory and methodology being discussed using an example. The statistical packages that will be used for various techniques include SPSS and SAS.

In chapter 5, comparison of different methods will be discussed, including a recommendation that will be made about various techniques used for the relevant data set.

CHAPTER 2

MULTIVARIATE ANALYSIS OF VARIANCE

2.1 OVERVIEW

This chapter provides the procedure used to carry out the research project and the explanations of methods used to arrive at conclusions, including the research design. Research instruments used in the project are also listed, discussed and justified, showing advantages and disadvantages of using these instruments.

When developing a comparative analysis of using MANOVA, discriminant analysis and factor analysis in real data, stages / procedures to be employed will be developed for each statistical method. The researcher will use the following beforehand:

(1) Perform data screening (i.e. assessing missing data, outliers and assumption violations of linearity, normality and homogeneity of variance-covariance matrices to be addressed).
(2) The correlation matrix: This gives the researcher an opportunity to examine the interrelationships of the variables, not only between the dependent variable and independents variables but also between the independent variables themselves.
(3) The regression weighting coefficients and the t-test of each predictor in the equation.
(4) ANOVAs summary table.

Detecting outliers should be done using univariate summary measures and bivariate graphical techniques. It is important that the sample size is large enough for correlations to be estimated reliably. Correlation coefficients tend to be less reliable when estimated from small samples. The required sample size depends on magnitude of population correlation and of factors. As recommended by Hair et al. (1998): The sample size is small but deemed adequate by meeting the minimum cell size of 20.

The two-way factors will be examined on each method, as well as the issues relating to statistical assumption violations.

2.2 DEVELOPMENT OF A THEORY BASED ON THE MULTIVARIATE ANALYSIS OF VARIANCE

Multivariate statistical analysis is concerned with data collected on several dimensions of the same individual. In one-way classification models, the interest is in comparing the treatment effects, which correspond to a single variable. When there are two factors such as factor “A” and factor “B” with different levels, various models can be obtained by combining the two factors.

If factor “A” has different a-levels and b-levels to factor “B”, the experiment consists of “a × b” treatment combination. In such a case, the two factors will cross with each other, and the design will often be referred to as a two-way classification. The two-way situation is sometimes referred to as a factorial or two-way MANOVA, where the effects of two factors are examined simultaneously on the dependent variables. The issues of independence, homogeneity of variance-covariance matrices and normality are comparable to one-way MANOVA.

The one-way MANOVA address the multivariate analysis of a single factor (independent variable). In the univariate statistical domain, we examine one dependent variable at a time. These single-factor analyses can provide very useful information. This is also true in the multivariate domain, where two or more measures that are dependent are assessed by means of a single (categorical) independent variable.

When researchers work with multiple independent variables in a study and if they opt to analyse one independent variable at a time, they will drive up type I error rates. Furthermore, single-factor independent variable assessments (either univariate or multivariate) do not allow researchers to determine how independent variables jointly affect the dependent variables.

MANOVA enables researchers to examine relationships between dependent variables at each level of the independent variable and provides researchers with statistical guidance to reduce a large set of dependent variables to a smaller number of variables.

In addition, MANOVA helps to identify dependent variables that produce the most independent variable separation. The MANOVA is appropriate to testing hypotheses on parameters of the models relevant to each of the variables involved (Norman Johnson and Fred Leone 1964: P. 934).

The methods described in the univariate analysis of variance (ANOVA) can be extended to cases where more than one variate is measured on each individual. A multivariate hypothesis is constructed through a variance-covariance matrix for each line of the univariate analysis of variance table.

Considering the two factors where both are between the groups design, the model is

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad (2.1)$$

Where i =1, 2... a, j = 1, 2, ..., b and k = 1, 2, ... n

For comparing effects of the two factors and their interaction, each observation can be decomposed in the manner that we calculate the ratio of Wilks' lambda, which can be referred to a Chi-square distribution.

Measuring variation between groups

$$X_{ijk} = \bar{X} + (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X}) + (\bar{X}_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}) + (X_{ijk} - \bar{X}_{ij}) \quad (2.2)$$

Where $\bar{X}$: the overall average of the observation vectors

$\bar{X}_{i.}$: The average of the observation vectors at i^{th} level of factor "A"

$\bar{X}_{.j}$: The average of the observation vectors at j^{th} level of factor "B"

$\bar{X}_{ij}$: The average of the observation vectors at the i^{th} level of factor "A" and at j^{th} of factor "B".

The variation between group means is measured with a weighted sum of squared differences between the sample means and the overall of all the data. Each squared difference is multiplied by the

appropriate group sample size of “a” or “b”. In this sum, this quantity is called sum of squares between groups or SS.

To measure the variation among data points within the groups, the sum of squared deviations between data values and the sample mean in each group are found, after which these quantities are added. This is called the sum of squared within groups.

The total variation in all samples combined is measured by computing the sum of squared deviations between data values and the mean of all data points. This quantity is referred to as the total sum of squares or SSTotal. x_{ijk} Represents the j^{th} observation within the i^{th}group, and $\bar{x}$ is the mean of all observed data values.

Finally, the relationship between SSTotal, SSGroups, and SSError is

SSTotal = SSGroups + SSError

Squaring and summing the deviations (X_{ijk} -$\bar{X}$) gives

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{X})^2 = \sum_{i=1}^{a} a\,n\,(\bar{X}_{i\,.} - \bar{X})^2 + \sum_{j=1}^{b} b\,n\,(\bar{X}_{.\,j} - \bar{X})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b} \mathrm{n}\,(\bar{X}_{ij} - \bar{X}_{i\,.} - \bar{X}_{.\,j} + \bar{X})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{X}_{ij})^2 \quad (2.3)$$

or

$$SS_{corrected} = SS_{factor\,1} + SS_{factor\,2} + SS_{interaction} + SS_{residual}$$

The corresponding degrees of freedom with the sums of squares in the breakup in (3.3) are abn – 1 =

(a - 1) + (b - 1) + (a - 1) (b - 1) + a×b× (n - 1) (2.4)

The MANOVA table for comparing factor and their interaction is follows:

Table 2.1

Source of variation	Matrix of sum of squares and cross-products SSP	Degrees of freedom
Factor 1	$SS_{factor\,1}$ $=\sum_{i=1}^{a} a\, n\, (\bar{X}_{i\,.} - \bar{X})^2$	a - 1
Factor 2	$SS_{factor\,2}$ $=\sum_{j=1}^{b} b\, n\, (\bar{X}_{.\,j} - \bar{X})^2$	b - 1
Interaction	$SS_{interaction}$ $=\sum_{i=1}^{a}\sum_{j=1}^{b} \mathrm{n}\, (\bar{X}_{ij} - \bar{X}_{i\,.} - \bar{X}_{.\,j} + \bar{X})^2$	(a - 1)(b - 1)
Residual (error)	$SS_{residual}$ $=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{X}_{ij})^2$	a×b×(n - 1)
Total corrected	$SSP_{corrected}$ $= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(X_{ijk} - \bar{\mathrm{X}})^2$	a×b×n - 1

In the two-way multivariate model there will be tested for factor “A” and factor “B” main effects as the methods apply quite generally.

In order to construct the likelihood test the following hypotheses for interaction apply:

H_0: No interaction effects ($\gamma_{11} = \gamma_{12} = \gamma_{13} = \ldots = \gamma_{ab} = 0$)

H_1: At least one $\gamma_{ij} \neq 0$

A test statistic will be conducted as the ratio of the determinant $|SSP_{residual}|$ to the determinant $|SSP_{interaction} + SSP_{residual}|$

The Wilks' lambda $\Lambda = \frac{|SSP_{residual}|}{|SSP_{interaction} + SSP_{residual}|}$ (2.5)

The null hypothesis H_0 is rejected for small value of the ratio, at least one $\gamma_{ij} \neq 0$.

For the large samples, Wilks' lambda can be referred to a Chi-square, alternatively in practical purposes we can use Bartlett's test statistic to approximate the Wilks' lambda formula.

Reject H_0 at level "α" if the test statistic is larger than the critical value.

- $\left[ab(n-1) - \frac{P+1-(a-1)(b-1)}{2}\right] \ln\Lambda > \chi^2_{(a-1)(b-1)p(\alpha)}$ (2.6)

  Where, Λ is given by (2.5) and $\chi^2_{(a-1)(b-1)p(\alpha)}$ is the upper $(100\alpha)^{th}$ percentile of a chi-square distribution with $(a-1)(b-1)p$ degrees of freedom.

Chi-square test and Bartlett's test are special (2.6) cases of Wilks' lambda, in which they are available in many general purpose statistical software programs.

Alternatively we can simultaneously use confidence intervals for contrasts in the model parameters to provide insight into the nature of the factor effects. When the interaction effects are negligible, we will focus on factor "A" and factor "B" main effects. This will allow us to examine the interaction of the factors as the two-way interaction does not allow for a possibility of a general interaction term, γ_{ij} as presented above (Table 2.1).

In the two-way MANOVA situation the effects of two factors, "A" and "B", will be examined simultaneously on n-dependent variables. The issues relating to statistical assumptions, violations of

independence, homogeneity of variance-covariance matrices, and normality will be discussed in this section.

There are four commonly used multivariate test statistics: Pillai's trace, Wilks' lambda, Hotelling's trace and Roy's largest root, which can be conducted with SPSS and SAS. The most prominent of these tests in the research literature is Wilks' lambda. To evaluate any of the multivariate test statistics (including Wilks' lambda) SPSS translates the multivariate test value into a multivariate (Rao's) F-statistic, which can be evaluated as much as any other F-value.

A statistically significant multivariate effect will show that the independent variable is associated with differences between the vectors or sets of means. Thus, we can presume that factor effects exist. If effects exist, the next step in this process will be to discover which specific dependent variables are affected. If there are no effects, this will call for the use of separate univariate ANOVAs for each dependent measure with a Bonferroni adjustment to the operational alpha level (0.05 divided by the number of dependent variables) to reduce the possibility of type I error.

When the effects exist, each statistically significant univariate F-statistic can then be further evaluated with a post hoc or multiple comparison tests that assesses every pair wise.

2.3 STATISTICAL ASSUMPTIONS

MANOVA assesses the differences across combinations of dependent variables, as this can construct a linear relationship only between dependent variables. The researcher will examine the data by assessing the following:

- Independence random sampling

  When conducting MANOVA the observations must be independent of one another. The independent variables are categorical in nature and the dependent variables are continuous

variables. MANOVA assumes that homogeneity is present between the variables that are taken for covariates.

- Homogeneity of variance-covariance matrices

  The Box's M test statistic indicates heterogeneity when the test has statistical significance (p-value < 0.01). The null hypothesis is that the variance between groups is equal.

- Multivariate normality

  The standard test for normality is the Kolmogorov-Smirinov statistic. A histogram and normal probability plot distinguish between systematic departures from normality when it shows up as a curve.

- Linearity

  Linearity relationships are assumed between pairs of dependent variables. If nonlinearity relationships are observed, transformations may be in order.

MANOVA is particularly sensitive to outliers or extreme values on the dependent variables. Failure to exclude outliers or transform the data could inflate type I or II error rates. Likewise, missing values in multivariate analysis become more problematic because of the complexity of the dependent variate.

The key concepts and terms:

- Box's M test: In MANOVA, Box's M test is used to know the equality of covariance between the groups. The null hypothesis in MANOVA is that the observed covariance matrices of the dependent variable are equal across groups. Wilks lambda: In MANOVA, Wilks' lambda test is used to know the overall significance of the model, when the overall model is significant, then we can predict the individual significance of the variable. There are other overall significance tests to be used, which include Pillai's trace, Hotelling's trace, Roy's largest root test, etc.

- Levene's test: In MANOVA, Levene's test is used to know whether or not the variance between groups is equal. When the test statistic of Levene is not significant this shows equal variance between groups.
- Partial eta-square: It is analogue to R-square in regression. It shows how much variance is explained by the independent variable.
- Power: Shows the probability of correctly accepting the null hypothesis.
- Post hoc test: In MANOVA, when there is a significant difference between groups, then the post hoc test is performed to know the exact group means, which significantly differ from each other.
- Significance: Similar to ANOVA, probability value is used to make statistical decisions as to whether or not the group means are equal, or if they differ from each other.
- Multivariate F-statistics: F-statistics is simply derived by dividing the means sum of the square for the source variable by the source variable mean error.

- Canonical correlation analysis: A technique by which a linear combination of p-predictors on the one hand and q-dependent variables on the other is determined in such a way that the correlation between these linear combinations in the total sample is as large as possible. It is more elaborated than multiple regression analysis. This means canonical correlation analysis seeks relationships between two sets of variables.

2.4 METHODOLOGY OF A TWO-WAY USING MULTIVARIATE ANALYSIS OF VARIANCE (MANOVA)

The question surrounding multivariate is how the dependent variables differ as a whole across the groups. These differences on individual dependent variables are of less interest than their collective effects. Therefore, the full power of MANOVA is utilised in the situation by assessing both the overall

differences and the differences among the combinations of dependent variables. This question will serve well to detect multivariate differences utilising MANOVA's ability and the univariate two-way fixed-effects model with interaction in which the measurements are as follows:

(1) A two-way factorial between of MANOVA will be conducted after the multiple comparisons for observed testing means have been completed on the dependent variables.

(2) A test of equality of covariance matrices for statistical significance, using Wilks' lambda, Pillai's trace, Hotelling's trace and Roy's largest root to assess the multivariate effects.

The null hypothesis is that the observed covariance matrices of the dependent variables are equal across groups. The significant p-value < 0.05 will indicate that the dependent variable covariance matrices are not equal across the levels of the independent variables.

(3) A Bartlett's test of sphericity for correlation between the dependent variables will be presented. The null hypothesis is that the residual covariance matrix is presented to an identity matrix. The significant test p-value < 0.05 will indicate sufficient correlation between the dependent variables and therefore the two-way MANOVA will be proceeding with the analysis. Failing to satisfy this condition, we can proceed with the univariate ANOVAs for the main effects on each dependent variable with Bonferroni adjusted procedure.

(4) The multivariate tests which correspond to the multivariate main effects and interaction results. If F-value is statistically significant (p-value < 0.05), this will indicate that differences between the groups on the dependent variate exist. The effects of each independent variable and the interaction of the independent variables will also be discussed in this section.

(5) Univariate ANOVAs will be conducted on each dependent measure separately to determine the locus of the statistically significant multivariate main effect of the dependent variable. The null hypothesis is that the factor "A" influences the factor "B". MANOVA will be performed in SPSS and SAS.

2.5 APPLICATION OF MANOVA METHOD USING REAL DATA

2.5.1 Background to the real data of the study

Eskom is a South African electricity public utility, established as the electricity supply commission by the government of South Africa. The utility is the largest producer of electricity in Africa. The company is divided into generation, transmission and distribution divisions and together Eskom generates approximately 95% of electricity used in South Africa.

Due to South African government's privatisation of Eskom in the late 1990s, the South African economy is now adversely affected. Eskom has the potential to contribute significantly to the economic growth of the country by taking advantage of its capacity to produce electricity. Electricity is manufactured at the lowest possible cost in order to keep power bills low, and with the lowest possible impact on the environment. Electricity is the flow of electrical power or charges and it is both a basic part of nature and one of our most widely used forms of energy.

Alternatively, there are other economical methods available to generate electricity from the conversion of other sources of energy, such as coal, nuclear and solar energy, gas and paraffin. These are called primary sources. Energy sources tend to play many roles in our daily lives; including lighting, heating and cooking in our homes.

Using coal in the household to generate electricity is not ideal because no matter how carefully it is burnt as there are gaseous and solid emissions. The gases that are emitted include sulphur dioxide, carbon dioxide and oxides of nitrogen; the first two of which are regarded as having a climate change effect on the environment. Despite of its negative impact on our daily lives, coal is the most economical way to generate electricity in Mpumalanga and the Northern Province.

Natural gas is an environment fuel source, a major feedstock for fertilisers, and a potent greenhouse gas. Before natural gas can be used as a fuel, it must undergo processing to remove all materials other than methane.

The experimental design in this study is motivated through a survey of household power consumption in South Africa during 2007. This can be presented by two factors. The first factor is the power usage with three levels, namely cooking, heating and lighting the abovementioned three levels of power usage are known to be a source of power consumption that is apparent in several measures of the expenses in the household. Household energy consumption is the energy consumed in homes to meet the needs of the residents themselves. The energy consumption of households is often called the residential energy consumed in household dwellings. It is thought that the power consumption might be different in either the rural areas compared to urban areas or from one Province to another. Another factor is the Province with nine levels, namely Gauteng, Eastern Cape, Free State, KwaZulu-Natal, Limpopo, Mpumalanga, North West, Northern Cape and Western Cape.

2.5.2 MANOVA two-way factorial

The focus of the present section will be on using SPSS to analyse a factorial design where the researcher has two categorical independent variables (power usage and Province) and five independent quantitative measures (electricity, gas, paraffin, solar and other) and conceptually related dependent variables.

The hypotheses test of the study

The study aims to indicate the following:

(1) Whether or not there is sufficient correlation between the dependent variables.

(2) Whether or not there are differences between the five power groups on the dependent variate, i.e. whether or not the power mode varies in the population.

(3) Whether or not the difference could produce significant multivariate effect on the power usages, cooking, heating and lighting as well as the factor, Province.

The tests of the hypotheses suggested by (1) to (3) above are presented for two-way layout on the five power modes' consumption with nine independent observation vectors in each combination.

In this section, the main effects and interaction effects of factors on the household energy consumption will be discussed.

With our data screening presentation, issues relating to missing values, outliers, linearity, normality and homogeneity of variance-covariance matrices were addressed.

The purpose is to investigate the main effects between the two factors and the power mode in the expenses of the household.

2.6 DATA

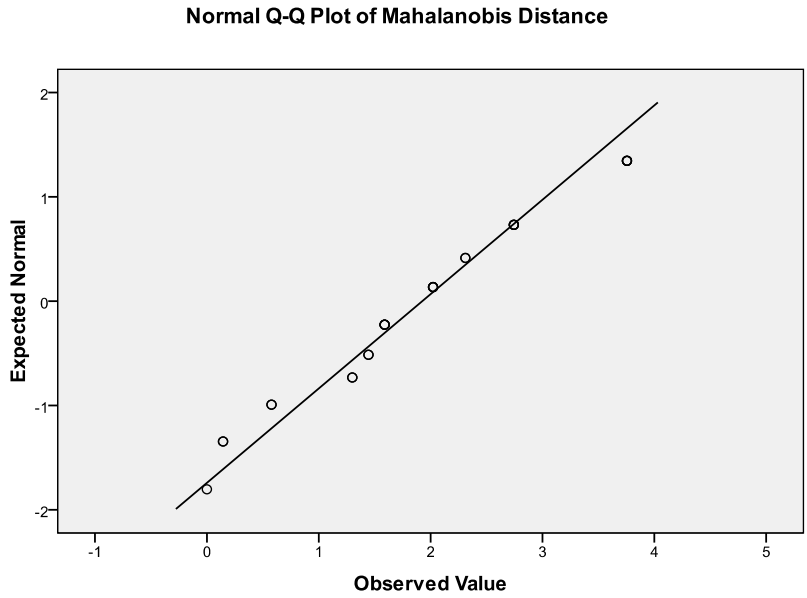
Two-way MANOVA household expenditure data sets will be used, which was obtained from the unit record file of the Stats SA and contains information collected from households in 2007.

		Power mode					
	Province	Electricity	Gas	Paraffin	Solar	Other	Total
Cooking	Eastern Cape	718863	43830	370096	714	453237	1586740
	Free State	604059	17475	136235	319	294780	1052868
	Gauteng	2580888	29670	525102	489	39431	3175580
	KwaZulu-Natal	1362151	53206	295295	754	522723	2234129
	Limpopo	489538	13780	98980	645	602943	1205886
	Mpumalanga	523808	10987	125502	374	660671	1321342
	North West	599909	16169	193282	181	809541	1619082
	Northern Cape	204332	10809	21094	512	236747	473494
	Western Cape	1215436	55691	84514	200	1355841	2711682
	Total	8298984	251617	1850100	4188	4975914	15380803
Lighting	Eastern Cape	1045714	3274	259318	3072	275363	1586741
	Free State	695217	870	27171	635	78977	802870
	Gauteng	2646398	6169	156499	2203	364311	3175580
	KwaZulu-Natal	1596354	4010	57417	11728	564619	2234128
	Limpopo	987422	1177	40042	8304	178989	1215934
	Mpumalanga	772638	1233	26525	1466	138543	940405
	North West	751351	895	34488	504	123885	911123
	Northern Cape	229632	686	6134	1381	26818	264651
	Western Cape	1285543	2448	50987	1111	29086	1369175
	Total	10010269	20762	658581	30404	1780591	12500607
Heating	Eastern Cape	517672	10687	436297	604	965260	1930520
	Free State	438379	14424	167467	474	620744	1241488
	Gauteng	2436348	32342	407373	1093	2877156	5754312
	KwaZulu-Natal	1270768	22751	200753	1710	1495982	2991964
	Limpopo	447069	5652	51279	406	504406	1008812
	Mpumalanga	423462	7725	55286	1755	488228	976456
	North West	536260	9022	126888	595	672765	1345530
	Northern Cape	171753	3338	14935	458	190484	380968
	Western Cape	1094998	21539	170340	1088	1287965	2575930
	Total	7336709	127480	1630618	8183	9102990	18205980
Total		25645962	399859	4139299	42775	14767889	44995784

Usage Power mode

2.6.1 Data screening

Multivariate outliers were assessed by computing a Mahalanobis distance measure for each case with



the SPSS.

Based on this criterion, no multivariate outliers or extreme scores were observed.

Test of Normality

The normality tests (Kolmogorov-Smirnov and Shapiro-Wilks) are statistically not significant, indicating that normality violations are present in the dependent variables. However, the normal Q-Q plots look reasonably normal (i.e. data points are close to the diagonal lines) and hence we judge this data ready for analysis.

2.6.2 Examining the correlation

The Bartlett's test of sphericity, presented also in table 2.1, is statistically significant (approximate Chi-square = 274.446, p-value 0.000 <0.05). This indicates sufficient correlation between the dependent variables to proceed with the analysis.

Table 2.1

Bartlett's Test of Sphericity[a]

Likelihood Ratio	.000
Approx. Chi-Square	274.446
df	14
Sig.	.000

Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.

a. Design: Intercept + Province + Usage

The model design for each dependent variable is Intercept + Province + Usage

2.6.3 Examining the multivariate effects

Four multivariate tests to evaluate any main effects are commonly employed in computerised statistical programs: Pillai's trace, Wilks' lambda, Hotelling's trace, and Roy's largest root. The most prominent of these tests in the research literature is Wilks' lambda. Because Wilks' lambda is an inverse criterion, smaller values provide more evidence of treatment effects (Stevens, 2002). To evaluate any multivariate test statistics (including Wilks' lambda) SPSS translates the multivariate test value into a multivariate F-statistic, which can be evaluated as much as any other F-value.

The multivariate tests' output table appears in table 2.2 (a and b). This table is composed of two parts. The top portion of the table (Intercept) evaluates whether the overall power mode mean differs from zero. Because of its statistical significance, we may conclude that it does differ from zero, indicating that the power mode varies in the population.

Of more importance is the evaluation of the effect of the independent variable (Province and Usage) in the second half of the table.

By examining first the multivariate main effect of Province "a", Wilks' lambda value = 0.001, which is subsequently translated into an F-value of 5.915 and evaluated at hypothesis (between groups) and error (within groups) degrees of freedom of 40 and 55. This F-value is statistically significant (p-value 0.000 < 0.05), indicating differences between the five power groups on the dependent variate. As indicated in the last column of the output, the partial eta-squared value tells us that this main effect accounts for only about 76.6% of the total variance.

Lastly, the multivariate main effect of the power usage produced a Wilks' lambda =0.058, which is subsequently translated into an F-value of 7.533 and evaluated at hypothesis (between groups) and error (within groups) degrees of freedom of 10 and 24. This F-value is statistically significant (p-value 0.000 < 0.05), indicating differences between the five power groups on the dependent variate. As indicated in the last column of the output, the partial eta-squared value tells us that this main effect accounts for only about 75.8% of the total variance.

Table 2.2 (a)

Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.996	588.962[a]	5.000	12.000	.000	.996
	Wilks' Lambda	.004	588.962[a]	5.000	12.000	.000	.996
	Hotelling's Trace	245.401	588.962[a]	5.000	12.000	.000	.996
	Roy's Largest Root	245.401	588.962[a]	5.000	12.000	.000	.996
Province	Pillai's Trace	2.533	2.054	40.000	80.000	.003	.507
	Wilks' Lambda	.001	5.915	40.000	55.101	.000	.766
	Hotelling's Trace	130.127	33.833	40.000	52.000	.000	.963
	Roy's Largest Root	125.687	251.373[b]	8.000	16.000	.000	.992
Usage	Pillai's Trace	1.437	6.633	10.000	26.000	.000	.718
	Wilks' Lambda	.058	7.533[a]	10.000	24.000	.000	.758
	Hotelling's Trace	7.646	8.410	10.000	22.000	.000	.793
	Roy's Largest Root	6.299	16.378[b]	5.000	13.000	.000	.863

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept + Province + Usage

Table 2.2 (b)

Multivariate Tests

	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Pillai's trace	2.533	2.054	40.000	80.000	.003	.507
Wilks' lambda	.001	5.915	40.000	55.101	.000	.766
Hotelling's trace	130.127	33.833	40.000	52.000	.000	.963
Roy's largest root	125.687	251.373[a]	8.000	16.000	.000	.992

Each F tests the multivariate effect of Province. These tests are based on the linearly independent pairwise comparisons among the estimated marginal means.

a. The statistic is an upper bound on F that yields a lower bound on the significance level.

2.6.4 Following a significant multivariate effect

A statistically significant multivariate effect informs us that the independent variable is associated with differences between the vectors or sets of means. This calls us in turn to presume that treatment effects exist and the next question is to discover which specific dependent variables are affected. This uses separate univariate ANOVAs for each dependent variable with a Bonferroni adjustment to operational alpha level (0.05 is divided by the number of the dependent variable) to reduce the possibility of type I error.

Because both multivariate main effects in Table 2.2 were found to be statistically significant, we can proceed with a separate assessment of each dependent variable for each main effect. This process is begun with an inspection univariate ANOVAs for both main effects on each dependent variable with a Bonferroni adjusted alpha level of 0.05/2 = 0.025.

These univariate F-tests can be found in the test of between-subject's effects table also in Table 2.3. Of particular interest is the upper-middle portion of the table, labelled province and power usage that depict separate univariate ANOVAs for each dependent variable on each main effect.

The F-values obtained from these analyses are identical to running separate univariate ANOVAs for each dependent measure. This output summarises standard ANOVA output (i.e. sum of squares, degrees of freedom, mean squares, F-values and significance level) which can be also viewed in Table 2.3.

Table 2.3

Source	Dependent Variable	Types III sum of squares	df	Mean Square	F	P-value	Partial Eta square
Corrected Model	Electricity	1.217E13	10	1.217E12	141.271	0.000	0.989
	Gas	4.8888E9	10	4.888E8	5.116	0.002	0.762
	Paraffin	4.673E11	10	4.673E10	11.572	0.000	0.879
	Solar	9.337E7	10	9337081.193	1.925	0.117	0.546
	Other	5.192E12	10	5.192E11	1.852	0.131	0.537

Table 2.4 shows a split decision for Province main effects as we find that Gas, Solar and Other were statistically not significant (p-value > 0.05), whereas Electricity and Paraffin were statistically significant, as p-value < 0.05.

Hence, the differences between the five power groups on Electricity and Paraffin produced the statistically significant multivariate main effect of Province.

Conversely, for the power usage main effect we find that the dependent variables (Electricity, Gas, Paraffin, Solar and Other) were statistically significant, P-value < 0.05. With the three levels of the power usage, we know that cooking, lighting and heating differ significantly by subject.

Table 2.4

Source	Dependent Variable	Types III sum of squares	df	Mean Square	F	P-value	Partial Eta square
Corrected Model	Electricity	2.436E13	1	2.436E13	2827.744	0.000	0.994
	Gas	5.922E9	1	5.922E9	61.988	0.000	0.795
	Paraffin	6.346E11	1	6.346E11	157.135	0.000	0.908
	Solar	6.777E7	1	6.777E7	13.970	0.002	0.466
	Other	9.316E12	1	9.316E12	33.237	0.000	0.675
Province	Electricity	1.176E13	8	1.470E12	170.676	0.000	0.988
	Gas	1.921E9	8	2.401E8	2.514	0.055	0.557
	Paraffin	3.780E11	8	4.725E10	11.699	0.000	0.854
	Solar	4.904E7	8	6129631.481	1.264	0.327	0.387
	Other	2.197E12	8	2.746E11	0.980	0.485	0.324
Usage	Electricity	4.075E11	2	2.037E11	23.652	0.000	0.747
	Gas	2.966E9	2	1.483E9	15.526	0.000	0.660
	Paraffin	8.936E10	2	4.468E10	11.064	0.001	0.580
	Solar	4.433E7	2	2.217E7	4.570	0.027	0.364
	Other	2.995E12	2	1.497E12	5.343	0.017	0.400

Because the main effect of Province was statistically significant for the Electricity and Paraffin dependent measure, we can examine exactly where the significant difference lies with a post hoc examination of the five treatment means. These estimated marginal means are displayed next in Table 2.5, where it is noted that electricity scored the highest of the power usage group with a mean of 949850.444.

Table 2.5

1. Grand Mean

Dependent Variable	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Electricity	949850.444	17862.208	911984.254	987716.634
Gas	14809.593	1880.995	10822.061	18797.124
Paraffin	153307.370	12230.017	127380.893	179233.848
Solar	1584.259	423.865	685.705	2482.813
Other	587388.704	101885.755	371400.552	803376.855

2.7 CONCLUSION

A two-way between-subjects multivariate analysis of variance (MANOVA) was conducted on the five dependent variables: electricity, gas, paraffin, solar and other. The independent variables were power usage, having cooking, lighting and heating levels and Province, having Gauteng, Mpumalanga, Free State, Limpopo, North West, Western Cape, KwaZulu-Natal and Eastern Cape levels.

Extreme scores, outliers, or statistically assumption violations were not noted in the present data. Using Wilks' lambda, the dependent variables were significantly affected by the main effects of Province, Wilks' lambda = 0.001, F-value = 5.915, P-value =0.000 < 0.05 and the partial Eta-square ($= R^2$) = 76.6% and the power usage, Wilks' lambda = 0.058, F-value = 7.533, P-value = 0.000 < 0.05 and the partial Eta-square ($= R^2$) = 75.8%.

CHAPTER 3

DISCRIMINANT ANALYSIS

3.1 OVERVIEW

In this section discriminant analysis will be discussed in detail in a closed parallel manner with the multivariate analysis of variance (MANOVA) procedure for two-factor designs with equal cell frequencies.

The discriminant function analysis design predicts membership in one or more than two groups. The predictors are continuous variables (Duarte Silva & Stam, 1995). The goal of discriminant analysis is to define discriminant functions. The model in discriminant function analysis is linear, thus the independent and dependent variables have constant relationship to each other. The weights are calculated to derive a discriminant score for each case and the mean discriminant scores for the groups are called the centroid.

The sample size needs to be taken into account with discriminant analysis. This statistical technique permits the groups to be of different sample sizes but the sample size of the smallest group should exceed the number of independent variables.

3.2 DISCRIMINANT ANALYSIS ASSUMPTIONS AND FUNCTIONALITY

Comparable to MANOVA, the dependent variable is a categorical variable. The assumptions for discriminant analysis are as follows:

- Linearity
- Normality
- Independence of predictors
- Homoscedasticity
- Absence of multicollinearity
- The influence of outliers

Discriminant analysis is highly sensitive to outliers as this problem should be resolved prior to the analysis. With discriminant analysis we assume that the continuous independent variables are normally distributed, violation of this assumption will suggest opting for logistic regression. A discriminant function analysis is similar to logistic regression in the way that we use to develop a weighted linear composite to predict membership in two or more groups (Gooley & Lohnes, 1971).

Discriminant function analysis can be used for two purposes such as:

(1) Prediction, referred to as predictive discriminant analysis and

(2) Explanation, referred to as descriptive discriminant analysis (Huberty, 1994)

Lawrence, Glenn & Guarino (2006) indicated that descriptive discriminant analysis is often used as a follow-up analysis to a significant multivariate analysis of variance (MANOVA) to determine the structure of the linear combination of the dependent variables. This shows that discriminant analysis is computationally identical to MANOVA.

Descriptive discriminant analysis has a focus on revealing major differences among the groups (Stevens, 2002, P. 285).

The discriminant function will weight n-independent variables such that two or more dependent variable groups will be differentiated. One way to evaluate the solution that will be examined is based on how accurately the independent variables were classified into groups.

The equation for discriminant score is as follows:

D = a + b_1X_1 +b_2X_2 + b_3X_3 + ……. + b_nX_n

Where

D represents the predicted score in the dependent variables.

a, is the intercept

b_i (i = 1, 2, 3… n) are the coefficients associated with the independent variables.

X_i (I = 1, 2, 3… n) are the independent variables in the equation.

In this section, the researcher needs to determine how the energy in the household is consumed, based on both Province and power usage (cooking, heating and lighting). The researcher will endeavour to know the following:

- ✓ Can the Province be a factor that could be used to classify the expenses of power consumption in the household?
- ✓ Can the power usage be used as factor to classify the expenses of the power consumed in the household?
- ✓ What are the chances of making mistakes when using these factors?

In the example that will be used, mistakes occur whenever a power mode-source (electricity, gas, paraffin, solar and other) will be classified into the wrong category expense in the household. Thus, an error will occur when, for example, household expenses for cooking is predicted to be caused by lighting or heating. Alternatively, an expense consumed in a particular Province is allocated to another. It is also noted that these two kinds of errors are probably not equally serious. Discriminant analysis is a multivariate technique that can be used to control the power consumption and classify expenses of the households into the appropriate factor (Province or power usage).

3.3 APPLICATION OF DISCRIMINANT ANALYSIS USING REAL DATA

3.3.1 Introduction

The analysis will consider five continuous independent variables: electricity, gas, paraffin and other, with two categorical variables: Province with nine levels (Gauteng, Limpopo, KwaZulu-Natal, Northern Cape, Western Cape, Free State, North West, Mpumalanga and Eastern Cape) and power usage with three levels (cooking, heating and lighting). This data is analysed with SPSS software.

The structure matrix reports the discriminant loadings of the variables in the discriminant analysis. The standardised discriminant analysis coefficients are used to assess each independent variable with its contribution to the discriminant function.

The researcher may consider eliminating variables that do not significantly contribute to prediction. The relative importance will be assessed by standardised canonical discriminant function coefficients.

The hypotheses for discriminant function are:

H_0 : The means of the two groups on the discriminant function are equal

H_1 : The means of the two groups are not equal

Several methods are available to test if the discriminant model is statistically significant. For the purpose of this dissertation Wilks' lambda is presented. This test varies from 0 to 1 and will tell us about the variance of the categorical groups' variable that is not explained by the discriminant analysis. Alternatively the F-test of Wilks' lambda will indicate which variables are statistically significant. The independent variable that will fail to contribute a significant amount of prediction could be considered for deletion from the model (Lawrence Meyers, Glenn Guarino, P. 262).

The square of the canonical correlation is similar to the coefficient of determination in a multiple regression analysis.

In the classification table, the rows are the observed categories of the independent variable and columns are the predicted categories of the independent variables. The percentage of cases on the diagonal will be the percentage of the correct classifications.

3.3.2 Discrimination Data Analysis

3.3.2.1 Dependent variance: Province

The findings indicate that all observations are treated equally. The test of equality of the groups' means illustrated significant differences in means of the predictors between the nine groups. The F-tests in table 3.1a are all significant for electricity (p-value 0.000 < 0.05) and paraffin (p- value 0.01 < 0.05), indicating that the energy consumed in the household differ in terms of Province on these independent variables. However, there is no significant difference between the nine Provinces on gas (p-value 0.494 > 0.05), solar (p-value 0.534 > 0.05) and other (p-value 0.718 > 0.05). Alternatively the smaller the Wilks's lambda, the more important that independent variable is to the discriminant function (Lawrence, Glenn Gamst & Guarino 2006, P. 270).

The researcher may suggest in a future analysis to consider eliminating gas, solar and other from the model. It should be noted that the discriminant analysis is robust to the violation of homogeneity of variance assumption, provided the data do not contain extreme outliers.

Table 3.1a

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Electricity	.044	48.531	8	18	.000
Gas	.701	.962	8	18	.494
Paraffin	.289	5.523	8	18	.001
Solar	.713	.905	8	18	.534
Other	.773	.661	8	18	.718

Table 3.2a presents the Eigen values. The larger the eigenvalue, the more of the variance of the nine group dependent variables is explained by the discriminant function. There are five discriminant functions in this study, listed in descending order of importance, as shown in column 1. The third

column lists the percentage of variance explained and the last column shows the canonical correlation. The function “1” to “3” exceeds the criterion of 0.05 for a strong relationship. The squaring of these values provides the coefficients of determination which represents the percentage of variance explained in the dependent variance.

Table 3.2a

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	74.607[a]	96.4	96.4	.993
2	1.822[a]	2.4	98.7	.804
3	.863[a]	1.1	99.8	.681
4	.116[a]	.1	100.0	.322
5	.018[a]	.0	100.0	.134

a. First 5 canonical discriminant functions were used in the analysis.

Table 3.3a serves a purpose distinct from the Wilks’ lambda in the normal ANOVAs table. In this table the Wilks’ lambda tests the significance of the eigenvalue for each discriminant function. There are five functions in this table, but it is only one function, “1 through 5” which is significant (P-value 0.000 < 0.05).

Table 3.3a

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 5	.002	116.140	40	.000
2 through 5	.167	33.955	28	.202
3 through 5	.473	14.241	18	.713
4 through 5	.880	2.424	10	.992
5	.982	.344	4	.987

Table 3.4a presents standardised Canonical Discriminant Function Coefficients, indicating the relative importance of the independent variable in predicting Province. From this table the best independent variable in predicting the dependent variable is noted. Function 1: electricity (1.680), function 2: Paraffin (1.218), function 3: solar (1.134), function 4: Gas (0.746) and function 5: other (1.021).

Table 3.4a

Standardized Canonical Discriminant Function Coefficients

	Function				
	1	2	3	4	5
Electricity	1.680	-.252	-.098	-.057	-.081
Gas	-.374	-.421	.967	.746	-.374
Paraffin	.668	1.218	-.117	-.009	.035
Solar	-.818	.556	1.134	-.376	.009
Other	.702	-.279	.041	-.078	1.021

Table 3.5a presents the simple correlation of each variable with the discriminant function(s). Function 1 is correlated to electricity (0.535), Function 2 is correlated to paraffin (0.740), Function 3 is correlated to solar (0.605), Function 4 is correlated to the following: gas (0.937), solar (- 0.791), paraffin (0.636) and electricity (0.549), Function 5 is correlated to other (0.944).

Table 3.5a

Structure Matrix

	Function				
	1	2	3	4	5
Paraffin	.137	.740*	-.054	.636	.159
Gas	.055	.009	.345	.937*	.022
Solar	.012	.048	.605	-.791*	-.079
Electricity	.535	-.251	.330	-.549*	-.491
Other	.057	-.043	.149	.284	.944*

Table 3.6a displays the classification function coefficients

Classification Function Coefficients

	Province								
	Eastern Cape	Free State	Gauteng	Kwazulu Natal	Limpopo	Mpumalanga	North West	Northern Cape	Western Cape
Electricity	8.740E-5	6.044E-5	.000	.000	5.659E-5	5.720E-5	6.844E-5	1.775E-5	.000
Gas	.000	.000	.000	.000	-8.781E-5	.000	.000	-7.446E-6	.000
Paraffin	.000	4.893E-5	.000	.000	4.423E-5	4.215E-5	5.520E-5	1.092E-5	7.067E-5
Solar	-.002	-.002	-.009	-.003	-.001	-.002	-.002	.000	-.004
Other	9.482E-6	6.850E-6	3.013E-5	1.528E-5	6.951E-6	7.009E-6	8.324E-6	2.163E-6	1.449E-5
(Constant)	-53.808	-22.446	-381.250	-100.613	-21.095	-20.187	-28.044	-4.059	-80.864

Fisher's linear discriminant functions

In table 3.7a the classification results demonstrate how well the discriminant functions were able to classify the cases for each group of the dependent variable. The discriminant function correctly classified 77.8% of all the cases. The prediction of power consumption in the household for Province (Gauteng, Limpopo, Mpumalanga, Northern Cape, Eastern Cape, Free State, North West, Western

Cape and KwaZulu-Natal) was with the overall classification rate of 77.8%. There was a greater success rate for the Provinces: Eastern Cape, Gauteng, KwaZulu-Natal, Northern Cape and Western Cape, who were 100% correctly classified, than for the rest of the Provinces, who were incorrectly classified.

Table 3.7a

Classification Results

			Predicted Group Membership									
		Province	Eastern Cape	Free State	Gauteng	Kwazulu Natal	Limpopo	Mpumalanga	North West	Northern Cape	Western Cape	Total
Original	Count	Eastern Cape	3	0	0	0	0	0	0	0	0	3
		Free State	0	2	0	0	0	0	1	0	0	3
		Gauteng	0	0	3	0	0	0	0	0	0	3
		Kwazulu Natal	0	0	0	3	0	0	0	0	0	3
		Limpopo	0	0	0	0	1	2	0	0	0	3
		Mpumalanga	0	1	0	0	0	1	1	0	0	3
		North West	0	1	0	0	0	0	2	0	0	3
		Northern Cape	0	0	0	0	0	0	0	3	0	3
		Western Cape	0	0	0	0	0	0	0	0	3	3
	%	Eastern Cape	100.0	.0	.0	.0	.0	.0	.0	.0	.0	100.0
		Free State	.0	66.7	.0	.0	.0	.0	33.3	.0	.0	100.0
		Gauteng	.0	.0	100.0	.0	.0	.0	.0	.0	.0	100.0
		Kwazulu Natal	.0	.0	.0	100.0	.0	.0	.0	.0	.0	100.0
		Limpopo	.0	.0	.0	.0	33.3	66.7	.0	.0	.0	100.0
		Mpumalanga	.0	33.3	.0	.0	.0	33.3	33.3	.0	.0	100.0
		North West	.0	33.3	.0	.0	.0	.0	66.7	.0	.0	100.0
		Northern Cape	.0	.0	.0	.0	.0	.0	.0	100.0	.0	100.0
		Western Cape	.0	.0	.0	.0	.0	.0	.0	.0	100.0	100.0

a. 77.8% of original grouped cases correctly classified.

3.3.2.2 Dependent variable: Power usage

The test of equality of the groups' means reflected any significant differences in means of the predictors between the three groups. The F-tests in table 3.1b are significant for gas (p-value 0.001 < 0.05), solar (p-value 0.027 < 0.05) and other (p-value 0.012 < 0.05), indicating that the energy consumed in the household differ in terms of power usage (cooking, heating and lighting) on these independent variables. However, there is no significant difference between the three groups on electricity (p-value 0.668 > 0.05) and paraffin (p-value 0.110 > 0.05).

Table 3.1b

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Electricity	.967	.411	2	24	.668
Gas	.538	10.319	2	24	.001
Paraffin	.832	2.423	2	24	.110
Solar	.741	4.201	2	24	.027
Other	.691	5.379	2	24	.012

Table 3.2b presents the Eigenvalues. There are two discriminant functions, listed in descending order of importance as shown into column 1, function 1 (1.994) and function 2 (0.808). The third column lists the percentage of variance explained and the last column shows the canonical correlation. The two functions exceed the criterion of 0.05 for a strong relationship. By squaring these values the coefficients of determination which represent the percentage of variance explained in the dependent variance are provided.

Table 3.2b

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.994[a]	71.2	71.2	.816
2	.808[a]	28.8	100.0	.669

a. First 2 canonical discriminant functions were used in the analysis.

Table 3.3b below tests the significance of the eigenvalue for each discriminant function. In this table there are two significant functions (P-value < 0.05).

Table 3.3b

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 2	.185	37.154	10	.000
2	.553	13.031	4	.011

Table 3.4b shows standardised Canonical Discriminant function Coefficients, indicating the relative importance of the independent variable in predicting power usage. From this table the best independent variables in predicting the dependent variable are noted as function 1: electricity (-0.974), gas (0.754) and paraffin (0.516) and function 2: gas (-0.927) and other (1.143).

Table 3.4b

Standardised Canonical Discriminant Function Coefficients

	Function	
	1	2
Electricity	-.974	-.364
Gas	.754	-.927
Paraffin	.516	.287
Solar	-.361	-.106
Other	.346	1.143

Table 3.5b presents the simple correlation of each variable with the discriminant function(s). Gas is correlated to function 1 (0.608) and other is correlated with function 2 (0.522).

Table 3.5b

Structure Matrix

	Function	
	1	2
Gas	.608*	-.390
Solar	-.419*	-.024
Paraffin	.318*	.003
Electricity	-.112*	-.106
Other	.338	.522*

In Table 3.6b the discriminant function correctly classified 88.9% of all the cases. The prediction of power consumption in the household into power usage (cooking, heating and lighting) has the overall classification rate of 88.9%. There was a greater success rate –100% correctly classified – for power usage: heating and lighting, while 66.7% of cooking was correctly classified and 33.7 were incorrectly classified.

Table 3.6b

Classification Results

		Power Usage	Predicted Group Membership			Total
			Cooking	Lighting	Heating	
Original	Count	Cooking	6	0	3	9
		Lighting	0	9	0	9
		Heating	0	0	9	9
	%	Cooking	66.7	.0	33.3	100.0
		Lighting	.0	100.0	.0	100.0
		Heating	.0	.0	100.0	100.0

88.9% of original grouped cases correctly classified.

3.4 CONCLUSION

A simultaneously discriminant analysis of both Province and power usage was constructed to determine whether the five independent variables: electricity, gas, solar paraffin and other could predict the consumption of energy in the household. The overall Wilks' lambda was significant for Province: function1 with $\Lambda = 0.02$ $\chi^2 = 116.14$, P-value < 0.05 indicating that the overall independent variables differentiated between the nine groups of Province (Gauteng, KwaZulu-Natal, Limpopo, Free State, Western Cape, Northern Cape, Mpumalanga, Eastern Cape and North West). Similarly the overall Wilks' lambda was significant for power usage: (a) function 1 with $\Lambda = 0.185$ $\chi^2 = 37.154$ and (b) function 2 with $\Lambda = 0.553$ $\chi^2 = 13.031$, P-value < 0.05 indicating that the overall independent variables differentiated between the three groups of power usage (cooking, heating and lighting).

CHAPTER 4

FACTOR ANALYSIS

4.1 INTRODUCTION

The development of factor analysis can be seen as a solution to statistical techniques that may be applied to a group of variables in which none has been specified as a dependent variable or an independent variable. In recent decades factor analysis seems to have found a rightful place as a family of methods which is useful for certain limited purposes.

Many statistical methods are used to study the relation between independent and dependent variables. Factor analysis differs from other multivariate techniques in terms of intercorrelations among variables in a single set. The goal in factor analysis is to discover something about the nature of the independent variables, to summarise patterns of intercorrelations among variables and to test theory about underling structure. This in turn, will reduce a large number of variables to a smaller number of factors. These problems are common in psychology and natural science and are the reason for factor analysis being so popular in these fields. Therefore factor analysis presents one such possible methodology which can be used to great effects in the human science.

Factor analysis, like most multivariate techniques, examines the pattern of correlations between the observed variables. The variables that are highly correlated, either positively or negatively, are likely to be influenced by the same factors, while those that are relatively uncorrelated are likely to be influenced by different factors (Jamie Decoster 1998, P. 1).

The hypotheses in factor analysis are as follows:

(1) How many different factors are needed to explain the pattern of the relationship among these variables?
(2) What is the nature of these factors?
(3) How well do the hypothesised factors explain the observed data?

(4) How much unique variance does each observed variable include?

Mathematically, several linear combinations of observed variables, called components or factors, are produced by combining scores on observed variables, some of which are correlated, however imperfectly, with each of the factors.

Using the factor scores

Several things can be done with factor analysis results, but the most common is to use factor scores based on the factor structure. These factor scores can then be used in analyses just like any other variable.

Because the results of a factor analysis can be strongly influenced by the presence of error in original data, Hair, et al. (1992) recommend using factor scores if the scales used to collect the original data are "well-constructed, valid and reliable" instruments.

4.2 METHODOLOGY OF FACTOR ANALYSIS

In factor analysis, the researcher is interested in discovering which variables in a data set form coherent subgroups that are relatively independent of one another. There are two approaches to factor analysis: exploratory factor analysis and confirmatory factor analysis. In exploratory analysis, the nature of variables influencing a set of responses can be discovered. In confirmatory factor analysis, a specified set of variables can be tested in a predicted way. In factor analysis the data is in the form of correlations.

The following steps will be performed in factor analysis:

(1) Data collection: The variables are measured on the same experiment units.

(2) We obtain the correlation matrix between each of the variables.

(3) Extraction of initial factor solution: Submit the correlations into a computer program to extract factors.

(4) Rotation of the factors: By rotating the factor we find a factor solution that is equal to that obtained in the initial extraction, but have the simplest interpretation.

(5) Interpretation of the factors: Each of the variables will be related to each of the factors. The factor loading produced by the rotation can be interpreted as standardised regression coefficients.

The more factors one permits, the better the fit and the greater the percent of variance in the data explained by the factor solution. The selection of the number of factors is probably critical. Eigenvalues represent variance; therefore any factor with an eigenvalue less than 1 is not as important. The number of factors with eigenvalues greater than 1 is an estimate of the maximum number of factors.

4.3 WORKING WITH FACTOR ANALYSIS

4.3.1 Introduction

Kaiser-Meyer-Olkin (KMO) and Bartlett's test of sphericity produce the KMO measure of the sampling adequacy of how the correlations are for factor analysis.

Kaiser (1970 and 1974) indicated that a value of 0.70 or above is considered adequate while Bartlett's test provides a test of the following hypotheses:

H0: the variables are not correlated, versus

H1: the variables are correlated

Reject Ho if p-value < 0.05 level of significance to proceed with the factor analysis.

The Scree plot will provide the information for the researcher to determine the number of factors or components.

4.3.2 Principal component analysis solution

The assessment of the results of the KMO and Bartlett's test of sphericity in table 4.1 does not look good. This means that the KMO value is less than the heuristic of 0.70; indicating that the correlations matrix is inadequate for factor analysis and principal component analysis. Likewise, a significant Bartlett's test enables us to reject the null hypothesis Ho of the lack of sufficient correlation between the variables. This result gives us confidence to proceed with the analysis.

Table 4.1

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.478
Bartlett's Test of Sphericity	Approx. Chi-Square	57.004
	df	21
	Sig.	.000

The communalities below currently indicate the degree to which each variable is participating or contributing to the component solution. No variable appears to be particularly low for removal from the analysis and the analysis is therefore continued.

Communalities

	Initial	Extraction
Electricity	1.000	.605
Gas	1.000	.801
Paraffin	1.000	.802
Solar	1.000	.678
Other	1.000	.819
Province	1.000	.574
Power Usage	1.000	.882

Extraction Method: Principal Component Analysis.

Table 4.2 below indicates that three factors accounted for about 74 % of the total variance. In practice, a robust solution should account for at least 50% of the variance (Tabachnick & Fidell, 2001b).

Table 4.2

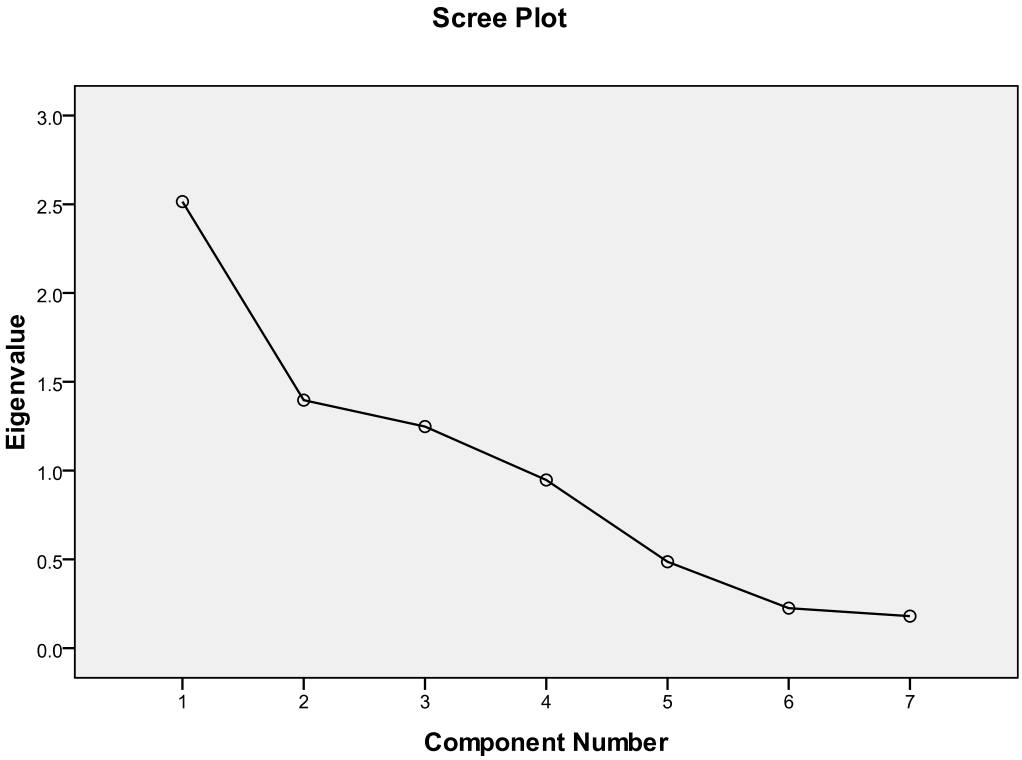
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.515	35.923	35.923	2.515	35.923	35.923	2.157	30.819	30.819
2	1.397	19.956	55.879	1.397	19.956	55.879	1.696	24.224	55.044
3	1.248	17.829	73.707	1.248	17.829	73.707	1.306	18.664	73.707
4	.947	13.534	87.241						
5	.487	6.956	94.197						
6	.226	3.225	97.422						
7	.180	2.578	100.000						

Extraction Method: Principal Component Analysis.

Figure 4.3 shows the Scree plot for the initial solution. This graphical method helps the researcher to determine how many components or factors should be included in the solution. The curve turn is at component 3, indicating a transition point between components with high and low Eigenvalues. This confirms the previous observation, derived from the total variance explained table 4.2, that three components best describe our principal components solution.

Figure 4.3



The unrotated component matrix shows the values in the array that have correlations between the variables and the components. The rotate component matrix presented in table 4.4 displays variables ordered by correlations within each of the component as follows:

Component 1: had paraffin (0.787), electricity (0.771) and Province (- 0.737)

Component 2: had gas (0.784) and solar (- 0.762)

Component 3: had power usage (0.891) and other (0.694)

The first three items were found to correlate to the first component. These items seem to be related to the power household consumption; thus, this component is referred to as power source consumed inversely to the Province. The next two items are all related to household power consumption that is

gaseous and solid emissions. The last two items appear to be related to the primary economical source of energy from our daily lives at home.

The goal of data reduction has been achieved by reducing an array of seven power household consumption factors into three uncorrelated principal components. These new composite variables can now be used as dependent variables in statistical analysis.

Table 4.3

Component Matrix

	Component		
	1	2	3
Paraffin	.891	.034	-.078
Gas	.756	-.476	.048
Electricity	.671	.345	-.186
Province	-.522	-.385	.391
Solar	-.144	.719	-.375
Power Usage	-.131	.607	.704
Other	.622	.128	.645

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Table 4.4

Rotated Component Matrix

	Component		
	1	2	3
Paraffin	**.787**	.422	.064
Electricity	**.771**	.027	.101
Province	**-.737**	.158	.073
Gas	.420	**.784**	-.103
Solar	.312	**-.762**	.025
Power Usage	-.119	-.271	**.891**
Other	.336	.474	**.694**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Component Transformation Matrix

Component	1	2	3
1	.836	.534	.127
2	.409	-.760	.505
3	-.367	.370	.854

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

4.5 CONCLUSION

An exploratory factor analysis, using a principal component extraction method and a Varimax rotation of seven items was conducted. Prior to running the analysis with SPSS, the data was screened by examining descriptive statistics on each item, correlation matrix and possible univariate and multivariate

assumption violation. The Bartlett's test of sphericity was significant (P-value < 0.05), indicating sufficient correlation between the variables to proceed with analysis.

Using the Kaiser-Guttman criterion of Eigenvalues greater than 1.0, a three-factor solution accounted about 74% of the total variance. The present three-factor model was deemed the best solution because of its conceptual clarity and case of interpretability.

CHAPTER 5

COMPARISON OF DIFFERENT TECHNIQUES

5.1 THEORICAL DISCUSSION

This section points out the differences and similarities between the two-way multivariate MANOVA, discriminant analysis and factor analysis. In the previous chapters, separate sections of sets of data were analysed and in turn the results obtained from the three techniques will be compared. The three methods are multivariate approaches with the intention of the multiple response outcomes. The data involves more than one variable and all the techniques focus on terms such as correlation, linear combinations, factors and functions. These techniques look at the phenomena in multiple levels of analysis.

All three techniques analyse a complex array of variables, providing greater assurance to get conclusions with less error and more validity. The three methods are a linear combination of variables indicating whether the independent variables or dependent variables form a linear combination of variables to interpret the data. The three techniques provide an idea about multivariate approaches. In MANOVA the linear combination of dependent variables maximises the distinction between groups. Discriminant analysis looks at the effect of several independent variables that are combined to form one or more linear composites and factor analysis is an alternative of the linear discriminant analysis, as it is used to determine how many factors are needed to explain the set of variables.

The primary concern of multivariate techniques is first to predict outcomes based on prior information, such as being able to accurately predict group membership of a given number of variables.

Second, to answer to a question such as; “Which variables are the most important in the prediction of some outcome?"

The three techniques are available to the researcher and the hypotheses of the study or kind of questions of each technique to guide the researcher in choosing the appropriate analysis.

In terms of degrees of the relationship between variables, MANOVA uses the multiple correlations "R" to indicate the relationship of a set of variables to other dependent variables. This includes the coefficient of determination r^2 . Discriminant analysis uses Canonical correlation "R" to indicate the relationship between sets of variables and factor analysis uses the correlation coefficient "R".

The test statistics measure preferred is Wilks' lambda for MANOVA and discriminant analysis. This can be translated into chi-square and F-value when the cell sizes are equal, there are no assumption violations and sample sizes are adequate. There is not a significant test in factor analysis that will test a hypothesis surrounding the number of factors, because factor analysis yields frequently more factors than can be satisfactorily interpreted.

All the three multivariate techniques are very sensitive to the effect of outliers as they have impact on the type I error.

MANOVA is appropriate in situations where the correlations between dependent variables are moderate, while it is not suitable for very high or very low correlation in the dependent variables.

In MANOVA, the discriminant functions are not always easy to interpret because they were designed for separate groups, not to make conceptual sense. Factor analysis appears to be very complex as factor analysts reach different conclusions, contradicting each other only if they all claim absolute theories, not heuristics (Richard B. Darlington, P.3). This means that more than one interpretation can be made of the same data set. Software is very user friendly in all multivariate techniques presented to the researchers.

In a MANOVA model, each equation represents the conditional mean of a dependent variable as a function of explanatory variables, while each equation represents a causal link rather than a more empirical association.

Rotations in factor analysis are different underlying processes but all rotations are equally valid using outcomes. It is impossible to identify the proper rotation using factor analysis alone. Factor analysis can only be as good as the data allows, as it relies on the self-reports on valid and reliable measures.

Naming of the factors can be difficult, since multiple variables can be highly correlated with no apparent reason and if sets of observed variables are highly similar to each other but distinct from other items, factor analysis will assign a factor to them. This makes it difficult to know what the factors actually represent.

In addition to the practical problems of multivariate techniques discussed, researchers continue to criticise the use of categorical data in case of computing the correlation matrix, while factor analysis requires considering all variables.

The most common correlation matrices are:

(1) The Pearson correlation: Can be used when a continuous variable is correlated with another continuous variable.
(2) The Polychoric correlation matrix: Can be used when a categorical variable is correlated with another categorical variable.
(3) The Polyserial correlation matrix: Can be used when a categorical variable is correlated with a continuous variable.

When there are more than two categorical variables, the numerical computation involved in producing this matrix becomes considerable (Dunn, Everitt and Pickles 1993, P. 171).

5.2 OVERALL CONCLUSION

This section gives a discussion on MANOVA, discriminant analysis and factor analysis. Throughout this dissertation, the concepts and stages / procedures of different techniques have been discussed.

The multivariate main effects was statistically significant p-value 0.000 < 0.05, the dependent variables were significantly affected by the main effects of Province and power usage p-value 0.000 < 0.05. The main effect for Province and power usage accounted for 76.6% and 75.8% respectively of the total variance. The dependent variable for each main effect indicated differences between the five power groups on electricity and paraffin produced the statistically significant multivariate effect of Province. Conversely, for the power usage the main effect we find each dependent variable was statistically significant, p-value < 0.05.

The findings of the discriminant analysis indicated that the energy consumed in household differ in terms of first, Province, the test statistic was significant for electricity and paraffin (p-value < 0.05). This suggests in a future analysis to consider eliminating gas, solar and other from the model. Second, power usage, the test statistic was significant for gas, solar and other (p-value < 0.05) of these independent variables.

The factor analysis indicated three factors accounted for about 74% of the total variance. The component 1 had observed paraffin, electricity and Province, the component 2 had observed gas and solar and the component 3 had observed power usage and other.

A correct model can be proposed to each of the three techniques with confidence, because we developed a model based on theory without omitting important variables from the model. Advantages and disadvantages of the present multivariate techniques have been discussed; including criticism regarding the application of each technique. The author suggests that the choice of the technique to use should be based on the question under investigation.

BIBLIOGRAPHY

ANDRE I. KHURI (1996). Minimal sufficient statistics for a general class of mixed models. Department of statistics, University of Florida. Gainesville USA, Statistics & Probability letters.

BARBARA G. TABACHNICK & LINDA S. FIDELL (1983). Using Multivariate Statistics, California State University, Northridge. Harper & Row, publishers, New York.

BRADLEY A. HARTLAUB, ANGELA M. DEAN AND DOUGLAS A. WOLFE. Rank – based test procedures for interaction in the two –way lay out with one observation per cell.

BECHCHLOFER R.E., SANTNER T. J. AND GOLDSMAN D.M. (1995). Design and Analysis of experiments for Statistical selection, screening and multiple comparisons.

CARLOS TADA DOS SANTOS DIAS, WOIJTEK JANUSZ KRZANOWSKI. (2006). Choosing components in the additive main effect and multiplicative interaction (AMMI) Models.

CRAIG K. ENDERS (2003). Performing Multivariate group comparisons group comparisons following a statistically significant Manova, Journal article excerpt.

CHEN J.J. & KODELL R.L. (1987). Analyses of two – way Chronic Studies. National Centre for Toxicological Research, U.S. food and Drug administration, USA Biometrics 43, P. 499 - 509

DUNN, G. EVERITT, B., and PICKLES, A. (1993). Modelling covariances and latent variables using EQS. London: Chapman and Hall.

DALLAS E. JOHNSON (1998). Applied Multivariate Methods for data analysis, Kansa State University. P. 255 – 278

DAVID G. GARSON (2009) Multivariate GLM & Manova: Statnotes, from North Carolina state University. Articles.

DAVID P. NICHOLS (1997). Manova to GLM: Basics of parameterization. Retrieved February 27, 2010 http://www.ats.ucla.edu/stat/spss/library/manglm.htm . P. 1 - 6

DAVID P. NICHOLS (1993). Interpreting MANOVA parameter estimates. Retrieved February 27, 2010. http://support.spss.com/ProductsExt/SPSS/Documentation/Statistics/articles/.

ENERGY INFORMATION ADMINISTRATION: Residential Energy Consumption survey, home energy uses and costs (2005). Retrieved November19, 2009 http://www.eiea.doe.gov/emeu/recs/

FACTOR ANALYSIS – PSYCHOLOGY WIKI. Retrieved December 10, 2010, from http:// Psychology.wiki.com/wiki/Factor_ analysis

GHOSH D.K. SAURASHTRA (1989). Construction of two – way group divisible designs. P. 33 – 334

GUPTA S.C. (1984). Iterative analysis of two and three way designs. P. 95 -102

IAN PRICE Two-way ANOVAs, two – way analysis of variance (2000) University of new England Retrieved. http://www.une.edu.au/WebStat/unit_materials/c7_anova/twoway_anova.htm.

JIANHUA Z. HUANG, HIPENG SHEN AND ANDREAS BUJA. The analysis of two- way Functional data: using two- way Regularized Singular value Decompositions. Department of Statistics, Texas A & M University. Articles.

JOHNSON R. A. AND WICHEM D. W. Applied multivariate Statistical Analysis

JAME S. (2005) Applied Multivariate Analysis using Bayesian and Frequentist methods of inference.

JAMES STEVENS (1992). Applied Multivariate Statistics for the social science. University of Cincinnati, second edition.

JAMES W. GRICE & MICHIKO IWASAKI (2007). A truly multivariate approach to MANOVA. Applied multivariate research, volume 12, P. 199 – 226. Oklahoma State University and University of Washington School of medicine.

JAMIE DECOESTER (1998). Overview of factor analysis. Department of Psychology, University of Alabama. Journal article.

JEFFERS J.N.R. Forestry commission (1996). Two case studies in the application of principal component analysis. Journal article.

LAWRENCE S. MEYER, GLENN GAMST AND A. J. GUARINO (2006). Applied Multivariate Research. Design and interpretation, pp. 259 – 277, 365 – 537.

HAIR A.T.B (1998). Multivariate data Analysis.

HAIR J.F. BARBIN AND ANDERSON R.E. (2010). Multivariate Data Analysis a global perspective 7th edition.

HAIR, J.F. ET AL. (1992) Multivariate data analysis, New York Macmillan, 3rd edition.

MAURICE. M. TATSUOKA and PAUL R. LOHNES (1998). Multivariate Analysis: Techniques for Educational and Psychological Research, University of Illinois at Urbana – Champaign, second edition.

MORRISON DONALD. F. (1978). Multivariate Statistical Methods, University of Pennsylvania, second edition

MULTIVARIATE STATISTICS: FACTOR ANALYSIS. Retrieved December 10, 2010, from http// www.socialresearchmethods.net/tutorial/Flynn/factor.htm.

MULTIVARIATE GLM, MANOVA, AND MANCOVA. Retrieved February 27, 2010. http://faculty.chass.ncsu.edu/garson/PA765/manov.htm. Journal article.

NADAM A BURNEY (1995). Socioeconomic Development and Electricity Consumption. Energy economics, vol. 17, No3, P. 185 – 195

PRAMOND MOHANLAL (1997). Structural Equation Modelling

RUSSEL ECOB (1987) Structural equation modelling by example, applications in educational sociological and behavioural research.

RICHAED A. JOHNSON AND DEAN W. WICHEM (1992). Applied Multivariate Statistical Analysis.

RAVINDRA KHATTREE & DAYANAND N. NAIK (1999). Applied Multivariate Statistics with SAS software, second edition.

RICHARD B. DARLINGTON. Factor analysis. Retrieved May 18, 2010. http://www.psych.cornell.edu/Darlington/factor analysis.htm. Journal article.

SAMUEL B. G. AND MARILYN S. THOMPSON (2600). Structural Equation Modelling for conducting tests of differences in multiple Means. Psychosomatic Medicine, pp. 706 – 716.

STATSOFT ELECTRONIC STATISTICS TEXTBOOK CREATORS OF STATISTICA data analysis software and services, Retrieved December 10, 2010. http: //www.stasoft.com/textbook/Principal-component-factor-analysis.

TIANHUA Z. HUANG, HAIPENG SHEN ANDREAS BUYA. The analysis of two-way functional data using two –way regularized singular value decompositions.