

Background Paper Series



Background Paper 2003: 1

**Multivariate Statistical
Techniques**

*Eisenburg
September 2003*

PROVIDE

PROJECT

The Provincial Decision-making Enabling Project

Overview


The Provincial Decision-Making Enabling (PROVIDE) Project aims to facilitate policy design by supplying policymakers with provincial and national level quantitative policy information. The project entails the development of a series of databases (in the format of Social Accounting Matrices) for use in Computable General Equilibrium models.

The National and Provincial Departments of Agriculture are the stakeholders and funders of the PROVIDE Project. The research team is located at Elsenburg in the Western Cape.


PROVIDE Research Team


Project Leader:	Cecilia Punt
Senior Researchers:	Kalie Pauw Esther Mohube
Junior Researchers:	Benedict Gilimani Lillian Rantho Rosemary Leaver
Technical Expert:	Scott McDonald
Associate Researchers:	Lindsay Chant Christine Valente

PROVIDE Contact Details

 Private Bag X1
Elsenburg, 7607
South Africa

 ceciliap@elsenburg.com

 +27-21-8085191

 +27-21-8085210

For the original project proposal and a more detailed description of the project, please visit www.elsenburg.com/provide

Multivariate Statistical Techniques¹

Abstract

This paper serves as an overview of various multivariate statistical techniques that can be used to analyse and describe survey datasets. Such analyses are useful for gaining a better understanding of results and the interpretation thereof. In particular this paper evaluates some of the standard statistical techniques that can be used to analyse consumption patterns of households, both in terms of expenditure on individual goods or baskets of goods. The ultimate aim is to determine whether these techniques – e.g. multivariate tests of significance and cluster analysis techniques – can be successfully employed to assist in forming representative household groups in terms of expenditure patterns. As such this paper is exploratory and does not aim to obtain final results. It is envisaged that more comprehensive research will be done at a later stage of the project. Initial findings suggest that there is merit in using multivariate statistical techniques to form representative household groups, although certain constraints and problems may be encountered.

¹ The main author of this paper is Kalie Pauw, Senior Researcher of the PROVIDE Project.

Table of contents

1.	INTRODUCTION	1
2.	MULTIVARIATE TESTS OF SIGNIFICANCE	1
2.1.	<i>Comparing the mean values for two samples: a single variable case</i>	2
2.2.	<i>Comparing the mean values for two samples: multivariate case</i>	3
2.3.	<i>Comparing univariate and multivariate tests of significance</i>	5
2.4.	<i>Comparing the means of several samples</i>	6
3.	CLUSTER ANALYSIS	6
3.1.	<i>Background</i>	6
3.2.	<i>Multivariate distances</i>	7
3.3.	<i>Clustering techniques</i>	7
4.	CONCLUSIONS	12
5.	BIBLIOGRAPHY	13
6.	APPENDIX: A NOTE ON INCOME AND EXPENDITURE SHARE VARIABLES	14

List of figures

Figure 1:	Expenditure shares – comparing African and Coloured households in South Africa	4
Figure 2:	Comparing cluster sizes of five different clustering experiments	9
Figure 3:	Expenditure patterns – comparing groups 1 to 5	9
Figure 4:	Mean income of household groups 1 to 5	10
Figure 5:	Expenditure and income	11

List of tables

Table 1:	Composition of households within groups	12
Table 2:	Income shares	14
Table 3:	Expenditure shares	15
Table 4:	Top ten expenditure categories including savings and taxes	17
Table 5:	Creating ten broad expenditure groups	18
Table 6:	Breakdown of food	18

1. Introduction

This paper serves as an overview of various multivariate statistical techniques that can be used to analyse and describe survey datasets. The 1995 Income and Expenditure Survey (IES) (SSA, 1997a) and the 1995 October Household Survey (OHS) (SSA, 1997a) is used as a source of information regarding household characteristics, consumption patterns and sources of income. These surveys are particularly useful for three reasons. Firstly, they were one of the first comprehensive national surveys that also included the former TBVC states. Secondly, they were conducted (as far as possible) on the exact same set of households, which makes it possible to merge the two survey datasets (see Hirschowitz and Orkin, 1996). Thirdly, since the IES survey is only performed once every five years, 1995 was the only year in which both surveys were done and for which data is currently available. The combined IES and OHS 1995 survey is useful in the compilation of various Social Accounting Matrices (SAMs) for South Africa.

It is very important to describe and analyse data before it is used in an economic model, as this will ensure that results are interpreted in a sensible way. Of particular importance in micro-macro-type economic models is a fairly detailed description of consumption patterns, sources of income and natural household groupings. These patterns have to be evaluated within the context of income distribution patterns and the demographic characteristics of households. With this in view, this paper will briefly investigate some multivariate techniques (mainly multivariate tests of significance) that may prove to be useful in describing household consumption and income patterns between groups of households. The paper will also review some basic cluster analysis techniques that may be used to find natural household groupings within datasets. It is envisaged that more comprehensive research in this area will be done at a later stage of the project.

2. Multivariate tests of significance

Multivariate tests of significance are mainly concerned with determining whether two or more samples differ significantly on account of a single variable or a number of variables that are contained in the samples. Tests are usually performed on the means of variables (or sets of variables) as well as variations from those means. This paper will mainly be concerned with means difference tests. Several tests exist in the literature, and can be applied to a single variable or a set of variables. This section draws mainly on Manly (1986) and will cover three means difference tests, namely the t-test, Hotelling's test and the likelihood-ratio test.

2.1. Comparing the mean values for two samples: a single variable case

The t-test (STATA command: *ttest*) compares the mean value of a single variable between two samples. The test is based on the assumption that the variable under consideration follows a normal distribution. It is further assumed that the variance of the variable is the same for the two samples. If, for example, one were interested in determining whether the mean income of a random sample of Indian households in South Africa is significantly different from the mean income of a random sample of Coloured households, the standard approach would be to use the t-test. The t-statistic for this test is calculated as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\left[s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]} \sim t_{n_1+n_2-2}$$

Here \bar{x}_i denotes the mean and n_i the size of the i^{th} sample, for $i = 1$ to 2. The pooled estimate of the standard deviation of the two samples is denoted by s . The t-statistic follows a t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. Consider the following two examples (results appear in Box 1). In example 1 a t-test is used to compare the means of the income from labour as a percentage of total income for Coloured and Indian households in South Africa. Coloured households derive an average of 69.2% of their total household income from labour, while Indian households derive 68.5% from labour. The null hypothesis of equality of the means of these two samples cannot be rejected by this test.

In Example 2 we compare the means of the income share from capital (gross operating surplus) between the same two sets of households. Here the mean for Indian households is significantly larger than that for Coloured households. The null hypothesis of equality of the means can be rejected with almost 100% certainty. The t-test, although very basic, is useful in comparing whether differences between means of variables of two samples are significant. The t-test can also be used to test whether a variable's mean is significantly different from some arbitrary number, e.g. zero.

Box 1: Testing the null hypothesis of equality of means – t-test**Example 1: Two-sample t test with equal variances – income from labour**

```
[Command: ttest cinclab if (race == 2 | race == 3), by(race)]
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Coloured	3766	69.19708	.6512799	39.96758	67.92019	70.47398
Indian	1040	68.51275	1.265807	40.82104	66.02892	70.99658
combined	4806	69.049	.5791608	40.15051	67.91358	70.18442
diff		.6843351	1.406568		-2.073181	3.441852

Degrees of freedom: 4804

Ho: mean(Coloured) - mean(Indian) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = 0.4865	t = 0.4865	t = 0.4865
P < t = 0.6867	P > t = 0.6266	P > t = 0.3133

Example 2: Two-sample t test with equal variances – income from capital

```
[Command: ttest cincgos if (race == 2 | race == 3), by(race)]
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Coloured	3766	2.763308	.2276052	13.96762	2.317067	3.20955
Indian	1040	14.71105	.9862599	31.80593	12.77576	16.64633
combined	4806	5.348753	.2869849	19.89532	4.786131	5.911375
diff		-11.94774	.6753449		-13.27172	-10.62375

Degrees of freedom: 4804

Ho: mean(Coloured) - mean(Indian) = diff = 0

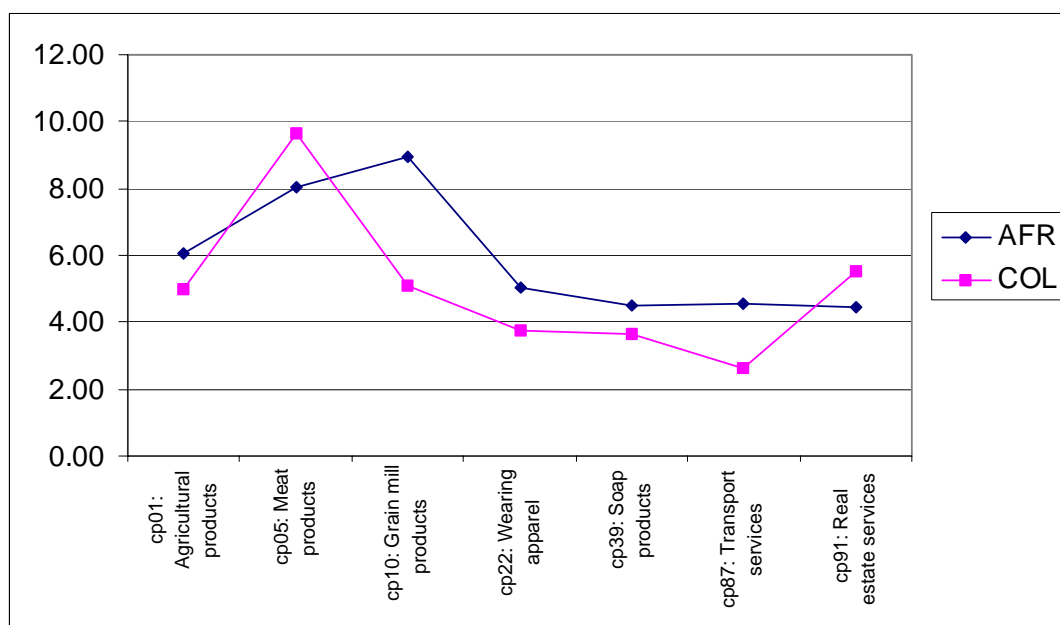
Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = -17.6913	t = -17.6913	t = -17.6913
P < t = 0.0000	P > t = 0.0000	P > t = 1.0000

Source: Authors calculations from IES 1995

2.2. Comparing the mean values for two samples: multivariate case

In some instances the analyst is interested in testing whether the vectors of means of two samples are jointly significantly different from each other. Consider Figure 1, which presents the relative expenditure (cents per R1.00 spent on total consumption) on a list of consumption goods between Coloured and African households. These seven consumption goods are some of the “most popular” consumption goods, selected from a list of 96 expenditure categories (see section 6). On average, households spend a total of 36c out of every R1.00 income (which by definition is equal to the sum of consumption, savings and taxes) on these seven goods. Although these mean (relative) expenditure values graphically appear to be different (Figure 1), it is necessary to perform a statistical test to confirm this.

Figure 1: Expenditure shares – comparing African and Coloured households in South Africa



Source: Authors calculations from IES 1995

A useful statistical test is Hotelling’s T^2 -statistic (STATA command: *hotel*), which is a generalisation of the t-test discussed before. The T^2 -statistic is calculated using two-sample covariance matrices and assumes that the population covariance matrices are the same for the two samples. The transformed statistic follows an F-distribution with p and $(n_1 + n_2 - p - 1)$ degrees of freedom, where p denotes the number of variables under consideration (i.e. 7 in our example):

$$\left[F - \frac{(n_1 + n_2 - p - 1)T^2}{(n_1 + n_2 - 2)p} \right] \sim F_{p, (n_1 + n_2 - p - 1)}$$

The results in Box 2 show that the null hypothesis of equality of the vector of means can be rejected with almost 100% certainty. We can thus conclude that the consumption patterns of a few basic commodities differ significantly between African and Coloured households. This test will form the basis of testing whether expenditure patterns differ significantly between pre-determined household groups.

Box 2: Testing the null hypothesis of equality of means – Hotelling’s test**Example 3: Hotelling’s T-squared test**

```
[Command: hotel cp01 . . . cp91 if race == 1 | race == 2, by(race)]
```

African

Variable	Obs	Mean	Std. Dev.	Min	Max
cp01	19303	5.625334	4.301728	0	51.84049
cp05	19303	7.464734	5.936614	0	75.05963
cp10	19303	8.292812	8.159447	0	78.10651
cp22	19303	4.65693	5.083091	0	53.44328
cp39	19303	4.169181	3.436632	0	63.29114
cp87	19303	4.221662	7.043082	0	89.27464
cp91	19303	4.133249	7.731505	0	75.41795

Coloured

Variable	Obs	Mean	Std. Dev.	Min	Max
cp01	3766	4.752911	3.614746	0	40.53582
cp05	3766	9.177174	6.371441	0	55.98991
cp10	3766	4.871356	5.080822	0	51.91781
cp22	3766	3.575475	3.971164	0	37.08107
cp39	3766	3.480757	2.63273	0	35.07375
cp87	3766	2.497021	5.37922	0	64.17112
cp91	3766	5.28047	7.856653	0	63.58926

2-group Hotelling's T-squared = 1614.5269

F test statistic: $((23069-7-1)/(23069-2)(7)) \times 1614.5269 = 230.5867$

H0: Vectors of means are equal for the two groups

F(7,23061) = 230.5867

Prob > F(7,23061) = 0.0000

Source: Authors calculations from IES 1995

2.3. Comparing univariate and multivariate tests of significance

Manly (1986) warns that it is quite possible that the results of a series of univariate tests differ from that of a multivariate test. A series of univariate (t-tests) may suggest that the null hypothesis of equal means cannot be rejected for some of the variables concerned, while the null hypothesis of equal vectors of means (Hotelling’s test) may be rejected. This can occur because of the “accumulation of the evidence from the individual variables in the overall test” (Manly, 1986: 31). It is also possible that a multivariate test is insignificant while some of the univariate tests are significant, because evidence of difference is “swamped” by evidence of no difference provided by other variables (Manly, 1986: 32). Manly concludes that, despite possible errors a single multivariate test is often a better alternative than a series of univariate tests since it takes into account the correlation between variables.

2.4. Comparing the means of several samples

Both the t-test and Hotelling's test compare the means of a single or multiple variables between two samples. However, sometimes the analyst is interested in comparing several samples and several variables. For this purpose a likelihood-ratio test can be performed. STATA can only perform a likelihood-ratio test after model estimation (e.g. logit or probit estimation), and hence this test is not further considered here.

From the discussion it is clear that Hotelling's T^2 -statistic is a particularly useful tool from the perspective of comparing expenditure patterns between various pre-determined household groups. Tests can be performed on various random samples of households from different races, settlement areas, provinces, or regions. The gender or occupation of the head of the household, educational attainment of household members, size of the household or age distribution (e.g. average age) may also determine expenditure patterns of households. The main limitation remains the fact that only two samples can be compared at any time. This may necessitate the construction of various cross-tabulations to compare the results of various multivariate tests of significance between combinations of different household groupings, either by race, settlement, region, income, or some other classification.

3. **Cluster analysis**

3.1. Background

Whereas means difference tests as a subset of multivariate tests of significance are useful in describing differences between two samples (e.g. household groups), it still requires that the analyst determine how the two samples are selected. Cluster analysis (also sometimes called classification) takes the process a bit further and attempts to solve the following problem (Everitt, 1974: 1):

“Given a sample of N objects or individuals, each of which is measured on each of p variables, devise a classification scheme for grouping the objects into g classes. The number of classes and the characteristics of the classes to be determined.”

Cluster analysis therefore attempts to determine natural groupings or clusters of observations (STATA Reference Manual). Manly (1986: 100) states that cluster analysis can assist the analyst in finding “true groups” within datasets, while it may also be useful for data reduction. Although cluster analysis was developed mainly as a technique within the fields of psychiatry or archaeology, it has in recent years become more and more useful as an exploratory data analysis technique with uses in many fields of study (Everitt, 1993 and Gordon, 1999, as cited in the STATA Reference Manual). This has been made possible by the

vast developments in technology and computerised software, which have made the complex calculations that are necessary for cluster analysis easier.

3.2. Multivariate distances

Cluster analysis is based on the notion of multivariate ‘distances’ between two single observations or samples of observations. Distance measures are sometimes more correctly referred to as measures of (dis)similarity (STATA Reference Manual). The most common measure of distance is the *Euclidian distance*. This is also the default measure of distance used in STATA’s *cluster* commands. Suppose there are N observations, each of which has values for p variables X_1, X_2, \dots, X_p . The values of observation i can then be denoted by $x_{i1}, x_{i2}, \dots, x_{ip}$ and similarly for observation j by $x_{j1}, x_{j2}, \dots, x_{jp}$. The distance (d_{ij}) between observation i and j (for $p = 2$) is then given by the following equation, known as the Euclidian distance:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$$

This formula is based on Pythagoras’ theorem of the length of the side of a triangle. The generalised form of the Euclidian distance for p variables is:

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Various other distance measures have been developed, but the Euclidian distance “may serve as a satisfactory measure for many purposes” (Manly, 1986: 44). It is also most frequently used in cluster analysis (Manly, 1986: 106). In a multivariate framework cluster analysis makes use of matrices of distances between observations. Observations are grouped together if they are close to each other. Grouping is thus based on what is sometimes referred to as the “nearest neighbour” principle (Manly, 1986: 101).

3.3. Clustering techniques

Various clustering techniques are described in the literature (see Everitt, 1974 and Manly, 1986 for details). STATA also allows for a variety of clustering commands. Some of the standard methods are briefly discussed here. The discussion is based mainly on the STATA Reference Manual, which contains a more detailed discussion. There are two basic approaches to clustering:

3.3.1. Partition cluster analysis methods

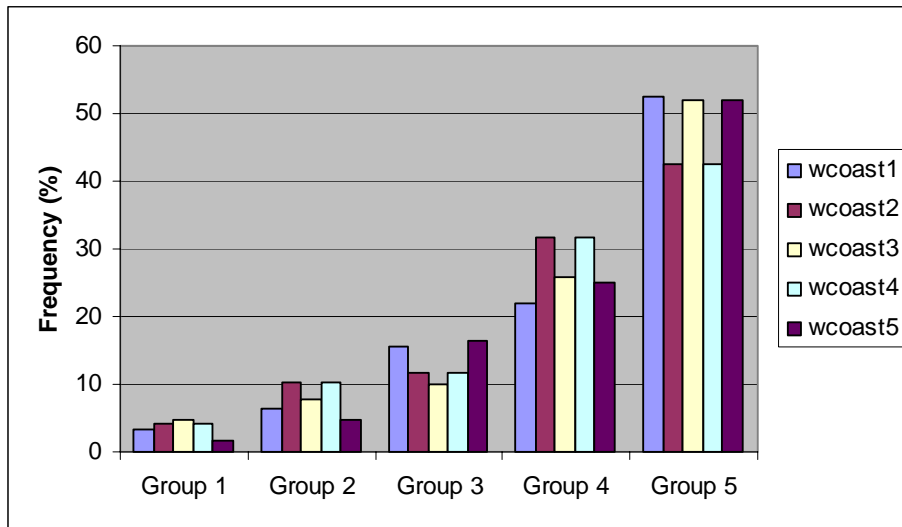
Partition cluster analysis methods divide the observations into a distinct number of non-overlapping groups. STATA can implement two variations of this method, namely *kmeans* and *kmedians*. The *kmeans* approach allows the user to specify the number of clusters (*k*) to be formed. By default the *k* initial group centres are selected randomly, and observations are added to the group with a mean value closest to its own mean. The mean value of the group is recalculated after each round. Observations are now reallocated (if necessary) to a group with a mean value closer to its own mean. The process is repeated until no observations change groups. Since the initial *k* group centres are selected randomly repeated clustering experiments will not necessarily result in the exact same set of groups. However, the fact that this is an iterative process with reallocation of observations after every successive round ensures that the final groups are more or less the same for each clustering experiment. It is further possible to select initial group centres, but this will not affect the final outcome greatly due to recalculation of the group centres after each round.

To see how much final group centres differ if one performs successive clustering experiments, five consecutive clustering experiments were conducted using data on consumption of seven commodities for the West Coast (wcoast) region (Western Cape and Northern Cape provinces). In each of the five clustering experiments five distinct cluster groups were created using the following command in STATA:

```
cluster kmeans cp01 cp05 cp10 cp22 cp39 cp87 cp91, k(5) name(wcoast...)
```

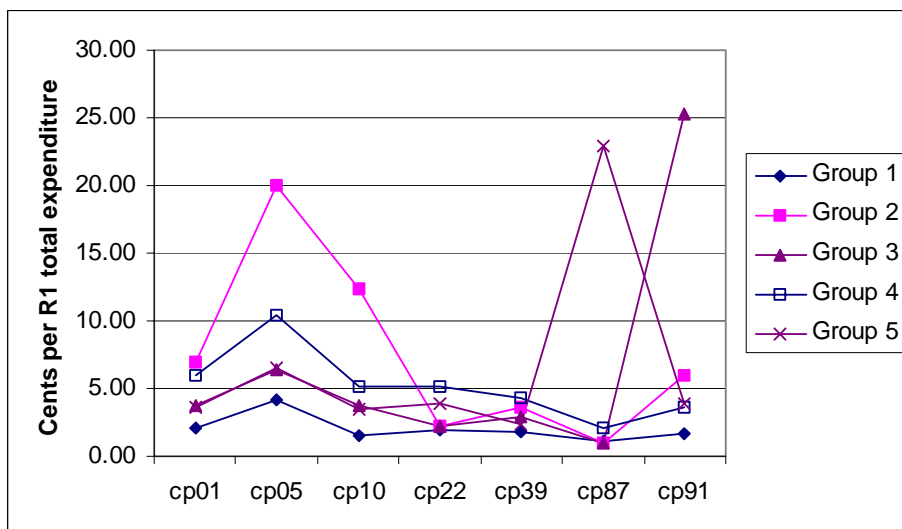
These five experiments were stored as wcoast1 to wcoast5. The number of households contained in each of five groups created in each of the five experiments is shown in Figure 2. The results for each clustering experiment are very similar. The cluster groups of experiment wcoast4 were selected for further exploration because this experiment's group centres were closest to the average group centres of all five experiments.

Figure 2: Comparing cluster sizes of five different clustering experiments



Next expenditure patterns between cluster groups were analysed using cluster groups formed in experiment 4 (wcoast4). All possible paired combinations of groups 1 to 5 within wcoast4 were tested using Hotelling’s T^2 -test. In all instances the null hypothesis of equality of the vector of means was rejected with a great degree of certainty (see Figure 3).

Figure 3: Expenditure patterns – comparing groups 1 to 5



The five groups are also quite different in terms of income levels (Figure 4), suggesting that income is an important determinant of the differences in expenditure patterns. Engel’s Law states that households will spend a smaller proportion of their income on food as income

rises. Looking at Figure 3 it is clear that group 2 spends relatively more money on food products such as agricultural products (cp01), meat products (cp05) and grain products (cp10) than, for example, group 1. Figure 4 below shows that group 2 is the poorest group on average, while group 1 is the richest. This supports Engel’s Law.

Figure 5 confirms that there exists a negative relationship between income and share of expenditure on basic commodities, especially when comparing groups 1 and 2, the richest and poorest groups respectively. The seven commodities included in the analysis can all be seen as basic commodities. In the figure the term ‘relative expenditure’ now refers to the sum of the seven consumption goods under consideration. The three food products, wearing apparel (cp22), soap products (cp39), transport services (cp87) and real estate services/housing (cp91) can all be considered necessities. As income rises, the proportion of expenditure allocated to these essential goods generally decreases, since more money can be freed up to spend on luxury goods. However, the figure shows that there are exceptions, such as expenditure on luxury housing versus expenditure on basic housing. Similarly an expensive vehicle is a luxury good, while public transport is not. Thus, when including some of these other non-food categories of expenditure in the analysis, a positive relationship between income and the proportion spent on these goods may be seen in some cases (e.g. groups 3, 4 and 5). For these cases Figure 3 is a more useful tool of analysis as it provides information on individual goods.

Figure 4: Mean income of household groups 1 to 5

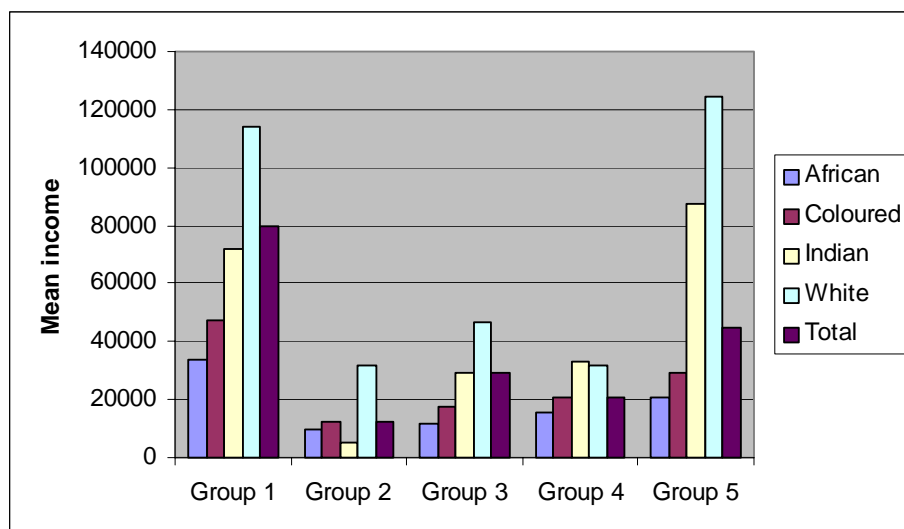


Figure 5: Expenditure and income

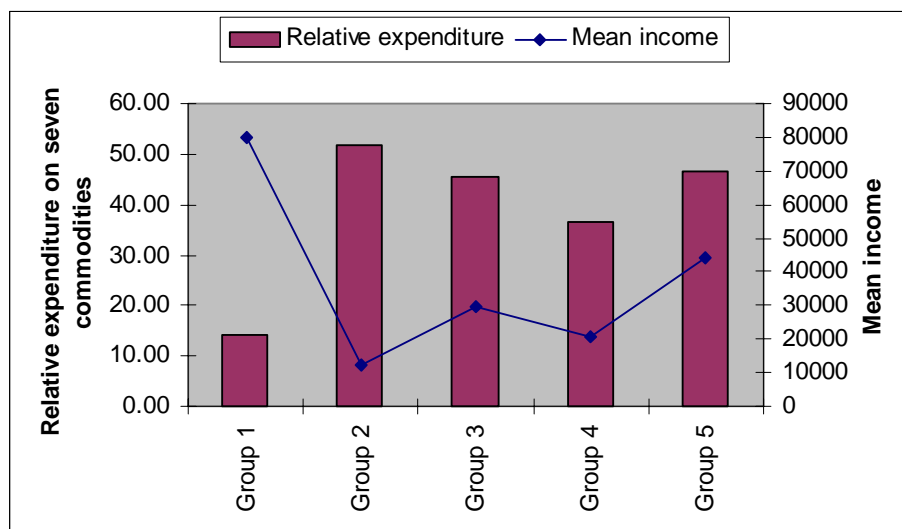


Figure 4 is of further interest as it shows that the richest households of all race groups, except Indians, fall within group 1. This suggests that, at least to some extent, income levels are a stronger determinant of expenditure patterns than race. Rich Indian households find themselves in group 5 – a group characterised by a relatively large expenditure on transport services. However, it must be added that Indian households only make up about 1.33% of the wcoast region’s population, which indicates a need for caution before substantive deductions can be made regarding this population group.

Table 1 provides a summary of the racial breakdown within groups as well as the distribution of different racial groups between cluster groups. Group 1 is made up predominantly of White households (50%) and groups 2 and 4 mainly of Coloured households (59% and 62% respectively). None of the other groups contain a majority of any particular racial group. Given that 78% of households in the Western Cape are either Coloured (48%) or White (30%) it is not surprising that these two groups are fairly well represented within each of the groups, except for group 2, which only has 4% White households. This also happens to be the group with the lowest average household income. Just over 40% of African and Coloured households find themselves in group 4. Most Indian households are in group 1 (56%), while White households are mainly found in group 1, the largest of the five cluster groups.

Table 1: Composition of households within groups

	Group 1	Group 2	Group 3	Group 4	Group 5	Total
<u>Within group composition</u>						
African	11%	37%	15%	26%	38%	20%
Coloured	37%	59%	40%	62%	43%	48%
Indian	2%	0%	3%	1%	1%	1%
White	50%	4%	42%	11%	19%	30%
Total	100%	100%	100%	100%	100%	100%
<u>Within race composition</u>						
African	23%	21%	7%	41%	8%	100%
Coloured	33%	14%	9%	41%	4%	100%
Indian	56%	2%	21%	18%	3%	100%
White	70%	2%	14%	11%	3%	100%
Total	43%	12%	10%	32%	4%	100%

Source: IES 1995

3.3.2. Hierarchical cluster analysis methods

There are two types of hierarchical clustering methods, namely agglomerative or divisive methods. Agglomerative methods begin by considering each observation as a separate group, and forms groups by combining the closest groups. The process is repeated (one 'merge' is performed at a time) until all observations belong to the same group. Divisive methods, on the other hand, start with all observation in one group. The group is split, with observations allowed to move to the nearest group. The process is repeated until all observations are in their own separate groups. Once one of these hierarchical clustering procedures have been completed, it is possible to draw a dendrogram, which is a graphical representation of the groups that have been formed. Hierarchical clustering is very limited as it can only be applied to small datasets due to computing constraints. This option is therefore not considered any further.

4. Conclusions

Multivariate statistical techniques provide a very useful tool for analysing and describing large survey datasets. Such an *apriori* analysis can be useful as it provides insight into the type of data one is dealing with. This can assist in forming representative households for use in a Social Accounting Matrix. Multivariate tests of significance can be useful for describing expenditure patterns of various pre-determined household groups, either for single variables or vectors of variables. As suggested, tests can be performed on various samples of households based on race, location, provinces, or regions. Further analysis may also show that the gender or occupation of the head of the household, educational attainment of household members, size of the household or age distribution may also prove to be a significant determinant of expenditure patterns of households.

Cluster analysis can be used to find natural or true household groups in a dataset. Instead of choosing household groups beforehand – a process that is often biased – the analyst can use cluster analysis to find these groups using a completely numerical technique. Various clustering techniques exist. Hierarchical clustering techniques are problematic since it can only be applied to small datasets. Since national survey datasets are usually large this approach is not very useful. The *kmeans* (or the related *kmedians*) procedure, on the other hand, is a viable clustering option when dealing with large datasets. This paper does not intend to draw conclusions for the cluster groups formed in the wcoast dataset as it merely intended to show how these techniques could be applied.

In conclusion, cluster analysis is certainly useful for describing data and finding true or natural household groups. At this stage it is unlikely that cluster analysis will be used as a first step of disaggregation. For example, it may be necessary to define household groups along racial lines first, depending on the type of policy questions that one wishes to answer with the data. The viability of cluster analysis will also be explored further by applying it to larger datasets.

5. Bibliography

- Manly, BFJ (1986). *Multivariate Statistical Methods. A primer*. Chapman and Hall: London.
- Everitt, B (1974). *Cluster Analysis*. Heinemann Educational Books: London.
- Hirschowitz, R and Orkin, M (1996). *Living in South Africa. Selected findings of the 1995 October Household Survey*. CSS: Pretoria.
- STATA Version 7 Reference Manuals (Various Volumes). STATA Press: Texas.
- SSA (1997a). *Income and Expenditure Survey, 1995*. Pretoria: Statistics South Africa.
- SSA (1997b). *October Household Survey, 1995*. Pretoria: Statistics South Africa.

6. Appendix: A note on income and expenditure share variables

Various income and expenditure share variables were created during the initial data extraction phase of this Project (see *Technical Paper* 2003: 1). Total household expenditure (*extot*) was defined as the sum of the expenditure and savings variables (*p01 – p96* and *hhtotals*, *hhindtax*, *hhinctax*, *hhsav* and *hhother*). Each of these variables was divided by *extot* and multiplied by 100 to create an expenditure share variable. This variable shows the amount spent on each type of expenditure or savings category for every R1.00 spent by a given household. Similarly, total income (*inctot*), defined as the sum of *inclab*, *incgos*, *inctrans*, *inccorp*, *incgov* and *incother*, was used to create income share variables. Each of these share variable starts with *c** (there is no particular reason for this naming convention).

The rationale behind using these relative expenditure or income variables is simple. It effectively controls for variations in the level of expenditure, therefore income patterns between rich and poor can easily be compared. Although theorems such as Engel's Law states that the proportion spent on food will decrease as income rises, this approach may show that income is not always a determinant of a certain expenditure pattern. For example, race may prove to be a stronger determinant of expenditure patterns than income – and this will only be possible to show when one uses relative expenditure shares.

Consider Table 2 below. This shows the relative income shares from various sources. Income from labour is clearly the largest source of income – 58.93c out of every R1.00 earned is received from labour services. No changes were necessary, and all further analyses of the income side are done using these variables.

Table 2: Income shares

INCOME SHARES	
Cinclab	58.93
Cincgos	5.66
Cintran	7.91
Cinccorp	5.73
Cincgov	16.17
Cincothe	5.61
	100.00

Table 3 lists all the expenditure share variables. The variable *chother* was netted out and this amount was added proportionally to the other variables. Only expenditure share variables exceeding R0.01 out of every R1.00 spent was included.

Table 3: Expenditure shares

Description	Variable name	Average expenditure share	Weighted average share
Maintenance, gifts and lobola	chhtotals	0.98	1.05
"Indirect" taxes	chindtax	1.37	1.46
Direct taxes (income)	chinctax	7.94	8.46
Savings	chhsav	4.17	4.44
Other (this category will be netted out)	chhother	6.17	0.00
Agricultural products	cp01	4.78	5.10
Coal and lignite products	cp02	0.12	0.13
Gold and uranium ore products	cp03	0.00	0.00
Other mining products	cp04	0.00	0.00
Meat products	cp05	7.20	7.68
Fish products	cp06	0.88	0.94
Fruit and vegetables products	cp07	0.86	0.92
Oils and fats products	cp08	1.08	1.15
Dairy products	cp09	2.29	2.44
Grain mill products	cp10	6.39	6.81
Animal feeds	cp11	0.05	0.05
Bakery products	cp12	2.49	2.65
Sugar products	cp13	2.28	2.43
Confectionary products	cp14	0.17	0.18
Other food products	cp15	3.28	3.50
Beverages and tobacco products	cp16	3.34	3.56
Textile products	cp17	0.07	0.08
Made-up textile products	cp18	0.75	0.80
Carpets	cp19	0.09	0.09
Other textile products	cp20	0.01	0.01
Knitting mill products	cp21	0.47	0.51
Wearing apparel	cp22	3.94	4.20
Leather products	cp23	0.00	0.00
Handbags	cp24	0.07	0.07
Footwear	cp25	1.60	1.71
Wood products	cp26	0.02	0.02
Paper products	cp27	0.00	0.00
Containers of paper	cp28	0.00	0.00
Other paper products	cp29	0.74	0.79
Published and printed products	cp30	0.36	0.39
Recorded media products	cp31	0.01	0.02
Petroleum products	cp32	1.83	1.96
Basic chemical products	cp33	0.00	0.00
Fertilizers	cp34	0.01	0.01
Primary plastic products	cp35	0.00	0.00
Pesticides	cp36	0.01	0.01
Paints	cp37	0.00	0.00
Pharmaceutical products	cp38	0.21	0.22
Soap products	cp39	3.66	3.90
Other chemical products	cp40	0.27	0.29
Rubber tyres	cp41	0.10	0.11
Other rubber products	cp42	0.00	0.00
Plastic products	cp43	0.04	0.05
Glass products	cp44	0.02	0.02

Ceramicware	cp45	0.00	0.00
Ceramic products	cp46	0.00	0.00
Cement	cp47	0.00	0.00
Other non-metallic products	cp48	0.17	0.18
Iron and steel products	cp49	0.00	0.00
Non-ferrous metals	cp50	0.00	0.00
Structural metal products	cp51	0.00	0.00
Treated metal products	cp52	0.00	0.00
General hardware products	cp53	0.03	0.04
Other fabricated metal products	cp54	0.03	0.04
Engines	cp55	0.00	0.00
Pumps	cp56	0.00	0.00
Gears	cp57	0.00	0.00
Lifting equipment	cp58	0.00	0.00
General machinery	cp59	0.00	0.00
Agricultural machinery	cp60	0.02	0.02
Machine-tools	cp61	0.00	0.00
Mining machinery	cp62	0.00	0.00
Food machinery	cp63	0.00	0.00
Other special machinery	cp64	0.03	0.03
Household appliances	cp65	0.75	0.80
Office machinery	cp66	0.06	0.07
Electric motors	cp67	0.00	0.00
Electricity apparatus	cp68	0.00	0.00
Wire and cable products	cp69	0.00	0.00
Accumulators	cp70	0.03	0.03
Lighting equipment	cp71	0.03	0.03
Other electrical products	cp72	0.00	0.00
Radio and television products	cp73	0.38	0.40
Optical instruments	cp74	0.12	0.13
Motor vehicles	cp75	0.96	1.02
Motor vehicles parts	cp76	0.07	0.07
Other transport products	cp77	0.03	0.04
Furniture	cp78	2.14	2.28
Jewellery	cp79	0.08	0.08
Other manufacturing	cp80	0.78	0.84
Electricity	cp81	3.35	3.57
Water	cp82	1.32	1.40
Buildings	cp83	0.00	0.00
Other constructions	cp84	0.19	0.20
Trade services	cp85	0.23	0.25
Accommodation	cp86	0.86	0.92
Transport services	cp87	3.38	3.61
Communications	cp88	1.56	1.66
FSIM	cp89	0.00	0.00
Insurance services	cp90	1.86	1.99
Real estate services	cp91	4.53	4.83
Other business services	cp92	0.14	0.15
General Government services	cp93	1.17	1.25
Health and social work	cp94	2.34	2.49
Other services / activities	cp95	2.44	2.60
Household domestic services	cp96	0.75	0.80
		100.0	100.0

A simple sort was performed to see on which goods households spend most of their money. The ‘top ten’ expenditure categories (including savings and taxes) appear in Table 4 below. Households spend a total of 52.59c out of every R1.00 spent on these ‘goods’. It is important to notice that some of the various food expenditure categories are quite prominent among the ‘top ten’ expenditure categories.

Table 4: Top ten expenditure categories including savings and taxes

Expenditure categories	Code	Amount spent per R1.00 consumption expenditure (in cents)
Direct taxes (income)	chinctax	8.46
Savings	chhsav	4.44
Agricultural products	cp01	5.10
Meat products	cp05	7.68
Grain mill products	cp10	6.81
Wearing apparel	cp22	4.20
Soap products	cp39	3.90
Electricity	cp81	3.57
Transport services	cp87	3.61
Real estate services	cp91	4.83
TOTAL		52.59

The ten groups above were used as a guideline to construct ten broad expenditure groups. Other variables were added to make these groups more representative of total household expenditure. As expected, food makes up a relatively large share of total expenditure. Note that only expenditure shares exceeding R0.01 out of every R1.00 (see Table 3) were included in these groups, hence the reason why expenditures do not add up to R1.00. Various multivariate statistical analyses performed on the datasets will mainly be based on these ten categories, sometimes excluding direct taxes (group 1) and savings (group 2).²

² This no doubt impacts on the results obtained as taxes and savings often make up a large share of total expenditure. It is therefore important to remember that some of the multivariate experiments performed are merely illustrative.

Table 5: Creating ten broad expenditure groups

Group	Description	Amount	Variables included
Group 1	Direct taxes	8.55	Chinctax
Group 2	Savings	4.32	Chhsav
Group 3	Food and agric products	32.03	cp01, cp05, cp08, cp09, cp10, cp12, cp13, cp15
Group 4	Clothing	5.86	cp22, cp25
Group 5	Soap and household products	3.92	cp39
Group 6	Electricity, water and household fuel	6.95	cp32, cp81, cp82
Group 7	Housing	4.91	cp91
Group 8	Services	9.83	cp88, cp90, cp93, cp94, cp95
Group 9	Transport (incl. purchase of vehicles)	4.63	cp75, cp87
Group 10	Other	8.28	chhtotal, chindtax, cp16, cp78
TOTAL		89.28	

The food category may also be of interest. Table 6 below shows the relative expenditure shares on various food categories. The last column ('weighted') lists the relative expenditure per R1.00 spent on food. Note that the total of the 'amount' column differs from the group 3 expenditure above because of the exclusion of expenditures less than R0.01 before (marked with an asterisk).

Table 6: Breakdown of food

Variable	Description	Amount	Weighted
cp01	Agricultural products	5.16	15.14
cp05	Meat products	7.69	22.54
cp06*	Fish products	0.94	2.76
cp07*	Fruit and vegetables products	0.91	2.67
cp08	Oils and fats products	1.15	3.38
cp09	Dairy products	2.44	7.17
cp10	Grain mill products	6.94	20.34
cp11*	Animal feeds	0.05	0.14
cp12	Bakery products	2.66	7.80
cp13	Sugar products	2.48	7.26
cp14*	Confectionary products	0.18	0.53
cp15	Other food products	3.50	10.26
		34.11	100.00

Background Papers in this Series

Number	Title	Date
BP2003: 1	Multivariate Statistical Techniques	September 2003
BP2003: 2	Household Expenditure Patterns in South Africa – 1995	September 2003
BP2003: 3	Demographics of South African Households – 1995	September 2003
BP2003: 4	Social Accounting Matrices	September 2003
BP2003: 5	Functional forms used in CGE models: Modelling production and commodity flows	September 2003

Other PROVIDE Publications

Technical Paper Series
Working Papers
Research Reports