The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

# THE USE OF DATA MINING AS A DECISION MAKING TOOL FOR MUNICIPAL PERFORMANCE MANAGEMENT IN THE WESTERN CAPE

Erica L Rasmussen

A Dissertation

Submitted to

the Science Faculty of the

University of Cape Town

In Partial Fulfillment of the

Requirements for the

Degree of

Master of Science

in

Operations Research in Development

Under the Supervision of

Dr. L Scott and Prof. TJ Stewart

Cape Town, South Africa
November 2007

University of Cape Town

UT 301.3 RASM
833984

# Acknowledgements

To my supervisor, Dr Leanne Scott. Thank you for your unwavering support; advice; encouragement and patience throughout this process.

To my co-supervisor, Prof Theodor Stewart. Thank you for you ever-present experience; advice and open door. Your passion for "the science of better" is genuinely inspiring.

Thank you to the Department of Local Government and Housing of the Western Cape for allowing access to data and relevant personnel. In particular, Pauli Weiderman, thank you for all your organisation and involvement in the process.

To my friends and colleagues from the UCT Statistics Department. Thank you for the productive and supportive environment in which is has been a pleasure to study and work for the past four years. The stress relief of a vuvuzela echoing down the corridors has been invaluable; as have the morning coffee breaks.

To my friends and family. Thank you for your patience and for allowing me the time required to dedicate to this process.

Last, but most certainly not least, to my husband Hendri Cawood. Thank you for everything you have helped out with, a list of which would be too long for this space, but of which I have hopefully recognised and thanked you for as they occurred. Mostly, I would like to thank you for being the constant supportive presence that you are. I truly believe I would not have completed this process without your dedicated love and encouragement.

University of Cape Town

# Table of Contents



University of Cape Town

University of Cape Town

# List of Illustrations



University of Cape Town

University of Cape Town

# List of Tables



University of Cape Town

# Glossary

| | |
|---|---|
| ANOVA | ANalysis Of Variance |
| Centroid | Average point for a cluster of observations |
| Data-defined clusters | Clusters of municipalities defined using the K-Means Clustering technique |
| DM | Data Mining |
| DLGH | Department of Local Government and Housing |
| Factor Structure Coefficients | Used in General Discriminant Analysis, the coefficients indicate the correlation between significant variables and the dimensions calculated by the model |
| Financial type variables | Variables based on a municipality's financial statements and mainly measured in Rands |
| GDA | General Discriminant Analysis |
| KDD | Knowledge Discovery for Databases |
| Opinion-defined clusters | Clusters of municipalities defined through a group exercise with members of the DLGH |
| OR | Operations Research |

University of Cape Town

| | |
|---|---|
| PMS | Performance Management System |
| PVQ | Project Viability Questionnaire |
| Service delivery type variables | Variables based directly on the service delivery of a municipality and mainly measures in counts or percentages |

University of Cape Town

# Abstract

This thesis proposes the use of data mining tools within an operations research process, allowing the integration of ever increasing amounts of data collected worldwide. It is further argued that the wealth of information delivered by DM tools, with their strong visual emphasis, can be used to enhance the transfer of knowledge to stakeholders.

The discipline of operations research could benefit greatly from the methods offered within the field of data mining, used to analyse the ever increasing amounts of data being collected worldwide. However, the subject also offers a wealth of information that could aid in decision making, along with visual representations of this information that might assist in the transferral of knowledge to problem stakeholders.

The advantages offered by data mining are not limited to problem contexts containing high-quality data, but could also assist within the development context where traditionally resources and relevant skills are scarce. The benefits of data mining within this context are illustrated through the use of municipal performance data supplied by the Department of Local Government and Housing in the Western Cape of South Africa. The results of these analyses are presented to the department in order to assess the potential contribution of data mining to decisions surrounding municipal support.

The methods used are able to cluster local municipalities and analyse their differences for profiling. These clusters are then compared to how members of the Department of Local Government and Housing intuitively grouped municipalities in terms of performance. The variables showing significant differences between clusters were financial in nature for those clusters defined by the data itself, but had a more direct service delivery focus when considering the clusters defined by the decision makers.

University of Cape Town

Analysis of the data consistently discovers an anomaly in cluster profiles at a particular period in time. When presented with this phenomenon, the decision makers confirmed the possibility of specific environmental change that may have influenced this. The analysis also indicates the particular significance of a variable relating to a performance feedback mechanism within the municipal support system. The reported values for this variable are consistent with the views held by decision makers concerning municipalities' performance.

Although there was limited scope for interaction with the relevant decision makers, it emerged that those data mining methods offering stronger visual representation yielded better opportunities for stakeholders to develop insight of the problem context. In general, the use of data mining tools, supported by a facilitator, able to interface between tools and the problem owners, appeared beneficial to performance management.

University of Cape Town

# Chapter 1
# Introduction

## Subject

This dissertation proposes the use of data mining as a decision making tool. The role of data mining is not simply limited to being able to better analyse data to be able to make better decisions. Data mining might play the additional role of being able to aid the understanding of relationships presented in data, thus increasing the understanding of the problem context. Along with this, the visual representations of relationships contained in complex data might aid the communication of knowledge from the analyst to decision makers.

## Background

The use of data mining is traditionally associated with business, however there are many non-business sectors that also collect vast amounts of data that require analysis. The particular sector that shall be discussed during the course of this dissertation is that of the Department of Local Government and Housing in the Western Cape of South Africa, who aim to use extensive report data to improve service delivery to the residents of the country.

Data Mining has a potential use not only in terms of being able to analyse vast amounts of data, but also in being able to aid in collecting the 'right' data. This is pertinent to the developmental context which usually implies limited resources and skills. A further use of data mining in the developmental context concerns the communication of knowledge to decision makers, as many of the methods available have a strong visual focus, allowing for better understanding of the problem context and therefore better decisions and increased involvement in the decision making process.

University of Cape Town

## Research Procedure

The procedure that shall be carried out to demonstrate the used of data mining as a decision making tool within a relevant context is as follows:

- Determine which data mining methods would be most useful to the problem context.
- Present and discuss results of these data mining methods to decision makers
- Carry out further analysis based on the results of this feedback

## Procedure Used to Gather Information

Gathering information from decision makers was undertaken through two official meetings with the Department of Local Government and Housing (DLGH) in the Western Cape. The first meeting presented the proposal of this dissertation and subsequently requested the use of relevant data. The data presented in this dissertation is taken from Project Viability Reports published by the DLGH. It was felt that this collection of data gave a good representation of reported municipal performance indicators and was extensive enough to be useful in the proposed analyses. The second meeting with the DLGH presented and discussed the results of the analyses and gathered further information in terms of how members of the DLGH intuitively considered the local municipalities in the Western Cape to have performed during the period corresponding to the reported data.

## Scope and Limitations

The scope of this dissertation is limited to the data mining techniques of Cluster Analysis, General Discriminant Analysis and Classification Trees. The data is limited to the historical Project Viability Reports of local municipalities in the Western Cape of South Africa, for the period June 2002 to June 2005. Feedback from decision makers is limited due to the lack of availability of relevant personnel in the Department of Local Government and Housing.

## Overview of Dissertation

This dissertation begins by discussing the role of data mining as a decision making tool. This is followed by an introduction to the context in which this proposal is applied and the data mining methods chosen. A discussion of the results of the three methods are then presented. The results from the feedback to decision makers and consequent analysis are then presented, finally followed by the conclusions developed from this process.

University of Cape Town

# Chapter 2
# Data Mining as a Decision Tool

It initially appears as if data mining and operations research are two largely independent paradigms that contribute to automated problem solving (Meisel & Mattfeld 2007). However, there has recently been increasing interest in terms of integrating these two disciplines, but with approaches being rooted in either one of the two paradigms, along with being applied in a specific domain (Meisel & Mattfeld 2007). This is leading to a lack of true integration of the two techniques in a way that would be mutually beneficial.

The following chapter discusses the traditional role of the two disciplines, along with their similarities, leading to a proposal of how data mining might be utilised as a decision tool in a variety of problem solving. This expands the current knowledge discovery for databases (KDD) framework to include principles entrenched in operations research, leading to the active involvement of stakeholders in data analysis. Finally, this is discussed specifically in the complex developmental context.

University of Cape Town

## 2.1 Data Mining / Knowledge Discovery for Databases

There are various conflicting definitions of data mining (DM) and knowledge discovery for databases (KDD), along with the role that each play. For the purposes of this discussion the definition of KDD will be as an overall process, including the interpretation and evaluation of data, relying more heavily on human interaction (Firestone 2000). Data mining becomes a step within the KDD process, generating patterns and relationships in data (Firestone 2000); however it is the usefulness of the methods offered by the sphere of data mining that will becomes the focus of this dissertation.

Currently, the use of data mining is typically business or market related (Statsoft Inc 1997), specifically with a focus on prediction. There is simply so much data available that corporations have created data warehouses to store it all (Romeu JL 2001) and require experts, generally statisticians, to be able to analyse it. There is no particular reason as to why DM/KDD be restricted to the

business sector however. There are other sectors that have also begun to collect large quantities of data, without a focus on profit in particular, but rather on being able to gather information from data and utilise it to make better decisions relating to a range of interests.

The sector that shall be discussed in this dissertation pertains directly to government, in particular relating to service delivery of municipalities. Decision makers would benefit greatly from better knowledge concerning the large amount of data reported by municipalities. This would allow them to devise and implement plans to effectively improve service delivery and thereby better the living conditions of residents in the country. This is especially true in the development context where there are less resources and skills available, highlighting the great benefit that this area would gain from a significant addition of knowledge to decisions made.

## 2.2 Operations Research

There are various aspects of Operation Research (OR) that parallel the objectives of DM/KDD. The Encyclopædia Britannica (2007) defines the three essential characteristics of operations research to be systems orientation, the use of interdisciplinary teams and the application of scientific method to the conditions under which the research is conducted.

Operations Research is seen as being explicitly purposeful (Midgley & Reynolds 2002) in terms of focusing on solving the problem at hand. This does not mean to focus primarily on solving the objectives already given by stakeholders, but analysing all aspects of the problem in order to solve an overall objective, perhaps as yet undiscovered, in the best possible way. This obviously involves the ability to define precisely what this overall objective is; however it is almost impossible to know every detail of a situation and therefore Operations Research is known as "the science of better" (Crowder c.2000).

The ultimate objective of operations research is to analyse a current situation, define a possible 'better' situation and create the change required to get there. There are various methods and methodologies available to create this change. INFORMS (2007) includes probability and statistics in the analytical technologies used by OR professions to help measure risk, mine data, test conclusions and make reliable forecasts. This recognises the role that data mining plays in the operations research context, and although there is both evidence of operations research being beneficial to the DM/KDD process (Romeu 2001) as well as the two techniques potentially being used in conjunction with one another as they intersect significantly (Ólafsson 2006), there is little evidence available of DM/KDD being particularly of benefit to the OR process. Even where this

University of Cape Town

proposal is discussed, very recently in terms of the research being undertaken for this dissertation, there is a focus on 'hard' OR methods, rather than the overall OR process (Meisel & Mattfeld 2007).

There is even less evidence of this particular technique being used within the developmental context, most likely due to the data requirements of the methods used. However, of all contexts, it is the developmental one that requires improvement where data mining could be used, with the data that *is* available, to provide more knowledge in a knowledge-poor situation. This would help create systems that are able to make better informed decisions, even if simply to improve the gathering of information in order to analyse it in a more meaningful way.

## 2.3 Combination of disciplines

Many techniques used in operations research are seen as iterative, in terms of improving a situation, re-analysing and then striving to improve it further. Due to their huge dimensions, DM/KDD problems are also frequently iterative. Results from the DM/KDD process are often used as the input to a new iteration or as a final stage until new data becomes available and requires reanalysis (Romeu 2001). When this data becomes available, the problem context changes, requiring another iteration of analysis, which is no different to when an operations research method leads to a changes of situation, requiring reanalysis of the situation to create a 'better' solution to implement.

DM/KDD is also seen as interdisciplinary in terms of analysts, domain experts and other interested parties jointly interpreting analysis results (Romeu 2001). As in operations research, this is one of the major strengths of the process.

A further similarity between the two disciplines is found in the final stage of DM/KDD which concerns the assimilation of the knowledge extracted from the data. There is a belief that in analysing systems data it is possible to obtain information to help resolve a problem in the system, which leads to implementing a relevant course of action (Romeu 2001). This is no different to the problem solving characteristic of operations research. What is vital however, is that this knowledge is not just simply assimilated, but transferred to those stakeholders that can cause the most effective change in situation, rather than simply remaining with the analyst. It is this that indicates that DM/KDD should not be used in isolation, but can benefit by being part of the OR process.

In terms of the developmental context, it is that much more difficult to involve stakeholders in the assimilation of knowledge gained from the DM/KDD process. It is however far more important, as it is these stakeholders that add information in terms of the context of the problem, allowing solutions

University of Cape Town

to be created within the resource constraints given. This is a significant area in which operations research can benefit DM/KDD; however this interaction of disciplines can be mutually beneficial.

The proposal of this dissertation is to integrate data mining as a decision making tool within the operations research process. Not only in terms of data analysis, but in terms of utilising the methods available to be able to fully involve stakeholders in the problem solving process so that they understand and buy into the knowledge gained from the data that they collect. The natural interaction of OR and DM/KDD would concern the use of 'hard' OR techniques; however this is not the only possibilities in terms of combining these two disciplines.

Traditionally, within the paradigm of 'hard' OR, stakeholders are involved in determining the constraints and objectives of a problem; however they are removed from the actual optimisation algorithm. The choice of this algorithm and the results presented are made by the OR practitioner, in the hopes that this is the most useful solution to the problem the stakeholders face. The process is; however acknowledged as iterative in terms of the stakeholder being allowed the opportunity to dispute results and redefine the problem context. The paradigm of 'soft' OR involves the stakeholders of the problem throughout the problem solving process, being far more iterative and discursive, initially being developed in order to be able to solve far more qualitative problems. However, the methods available for use within this paradigm seldom involve mathematics or the use of quantitative data, relying on the human aspect and knowledge of the problem to develop a solution.

This dissertation would seek to bridge the gap between 'hard' and 'soft' OR with the use of data mining. Stakeholders can benefit from the mathematical and statistical algorithms that are able to extract knowledge from data; however it is vital that they remain an integral part of the entire process. It is obvious that there remains a need for a professional to undertake the actual data analysis; however the influence of stakeholders in the process may cause the analyst to change the methods used, dependent on the variable relationships they might seek to explore. Another significant benefit of data mining is that many of the methods offer strong graphical representation of results, with potential for others to be developed, which would allow stakeholders to more easily understand the results gained from the analyses. This in turn allows stakeholders to interpret results more easily within the context of the problem, assisting the analyst to explore new problem areas.

## 2.4 Hypothesis requirements

A significant difference between DM/KDD and OR, is that of the hypothesis. The OR process begins by defining a problem and then seeking a solution; however when large amounts of data are

University of Cape Town

available for analysis, the approach to this analysis alters. Establishing the research hypotheses becomes an intrinsic part of the actual data analysis with the "research hypotheses and relationships between data variables being obtained as a result of (instead of a condition for) these activities" (Romeu 2001)

Data are initially collected with a certain set of objectives in mind; however the achievement of these objectives is not necessarily the only use to which the data can be put, especially when dealing with incredibly large quantities of data. It is possible that there are other, previously unthought of, aspects to the problem that can be addressed by use of the same data.

When attempting to improve a situation it is not vital that a research hypothesis is known prior to the analysis of data. To require the existence of a hypothesis prior to analysis introduces the possibility of restricting the data in such a way that methods fall short in being able to transform it into relevant knowledge (Ólafsson 2006). It is possible that through this analysis itself the problem is better understood and a hypothesis can be formulated.

This process of using the analysis itself to define the problem does not detract from the original objectives that caused the data to be collected, but rather adds objectives and therefore knowledge, to the situation. As the nature of data in general changes, becoming more abundant and complex, the methods and processes by which it is analysed and the results presented, require adaptation. The natural iterative process both of DM/KDD and OR provides the ideal structure to be able to use the same data to both formulate and test a given set of hypotheses. Again, it is vital that stakeholders play a part in this formulation, being given the opportunity to understand the relationships found in the data and integrate this with their expert knowledge of the situation.

University of Cape Town

## 2.5 Communication of Knowledge

Data Mining contains an incredibly important aspect, as with most of type of data analysis, in terms of the visualisation of results. It is that much more difficult within the data mining context however as the data are usually high dimensional and producing useful visualisations is a challenging problem (Ólafsson 2006). This implies however, that many of these types of visualisations have already been developed, with some methods having results that are more easily understood than others.

It is important that stakeholders gain the education of how data can work *for* them, not against them. There is a need to convince stakeholders to trust that the results presented are in fact the most useful to their problem situation and the analyst has not simply chosen random results leading to "lies,

damned lies, and statistics" (Twain 1906). The simplest way to create this trust is to allow stakeholders such involvement in the process that they themselves define the results used, within the guidelines set by the analyst. It is important however that the knowledge transfer required in terms of understanding the methods themselves is kept to a minimum and that the focus of the interaction process is on the results themselves.

One of the most effective means of transferring knowledge gained from data analysis to a non-expert is through graphical representation of results, taking advantage of our "innate ability to process visualisations" (Burkhard 2004, p.520). Ideally, the OR practitioner would choose those data mining methods that lend themselves to easily understood graphics. This discourages methods such as neural networks which is considered to be a 'black box' approach and lends itself to methods such as decision trees which clearly illustrates all results in the form of a tree diagram, maintaining the structure of the data and clearly indicating the relationship between variables and observations.

This use of graphical results is especially pertinent within the developmental context, where numerical skill levels may be lacking in terms of being able to generate knowledge from data. Even if an expert analyst was involved in the data analysis process, there would exist the challenge of transferring knowledge to those stakeholders that could implement the changes necessary. By using visual techniques to represent results, even of complex analyses, it is possible to involve all levels of stakeholder, encouraging understanding of the process and commitment to any system developed from the knowledge gathered.

In terms of the political structure of government, stakeholders add understanding of these complexities and the human factor involved in the problem context. In addition to this, service delivery concerns naturally involve people and therefore introduce an extra level of complexity, as many problem settings within a developmental context do. Data collection and analysis appears not to be a priority in terms of decision making, but is rather done in terms of reporting needs. Perhaps if the application of data analysis was made more transparent, stakeholders could understand the need for reporting and how this data helps improve their situation, rather than confounding it.

Finally, it is vital that there exists computing support to back the DM/KDD/OR process. Although there is much current software available that produces graphical output for many data mining procedures, there is a need for this to be developed further. More complex data analysis can be done away from the problem situation, but simple graphical techniques need to be available if analysis is to be done in real time with stakeholders present. Results need to be available almost instantaneously

University of Cape Town

so that various options might be explored.

The 'best' way to represent various relationships between variables may only be discovered as the analysis progresses. Certain relationships would need to exist to be able to challenge the analyst to represent these in an easily understood way, including having stakeholders request the representation of a relationship that the analyst had not previously thought of. However, there is great potential in this area in terms of involving stakeholders in the entire problem solving process, along with benefiting from the mathematical and statistical techniques offered from data mining.

University of Cape Town

# Chapter 3
# Research Context

The idea of using data mining as part of the Operations Research process requires a context in which it might be applied. In the following sections the role of the Department of Local Government and Housing (DLGH) in the Western Cape of South Africa will be discussed, along with the structure of the South African government, in order to better understand the context of the problem presented. Along with this, the available data mining methods are presented with a discussion as to the ones chosen, with the final section presenting further research considerations discovered during the analysis process.

## 3.1 Role of DLGH

The main responsibility of local government is that of service delivery to the community, and that of the DLGH the financing of municipalities in order that they might do this. Within the developing South African context, many municipalities fall short of this goal due to lack of skills and resources. According to a survey commissioned by the *Provincial Treasury* (2007) in 2005, on average more than 10% of households do not have access to basic services, with some racial groups indicating that this increases to up to 30% without access. Obviously, there is a need to improve the situation of service delivery in the Western Cape before even the basic needs of residents are met.

One of the current projects of the DLGH is to create a snapshot of each municipality such that they can be easily classified as failed (red flag), failing (yellow flag) or satisfactory (green flag). It was this particular project which led to the concept of using the available data, from reports that municipalities have submitted, to ascertain whether it might be used to assist in assessing the performance of municipalities in the Western Cape. The information gained from this process would hopefully help in making key decisions around support for municipalities, as well as the data required by the DLGH to be able to undertake these assessments.

A meeting was held with members of the DLGH during May 2006 during which it was discussed that a significant amount of data was available in Project Viability Questionnaire (PVQ) reports.

University of Cape Town

Although this was historical data, it would provide an idea of what could be reported by municipalities and could therefore be used for the analysis proposed.

The usefulness of data mining as a decision tool in government is not limited to this particular situation. It is hoped that by providing a simple illustration of some of the techniques available, it is possible to inform members of government about the greater potential contained within their data. Along with this, it might aid in establishing research collaboration between academic entities and government departments.

## 3.2 Structure of Government

The South African government is separated into three main spheres: national, provincial and local (Burger 2004). The structure of government is laid out as shown in Illustration 3.1.

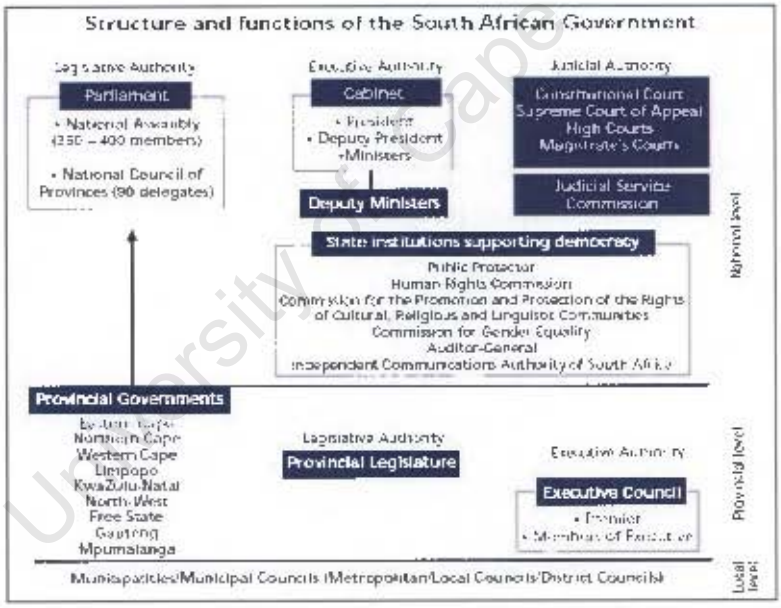



*Illustration 3.1: Structure and functions of the South African government*

For the purposes of this dissertation it is necessary only to discuss the structure and objectives of local government, focusing particularly on the local government of the Western Cape. This is placed at the very bottom of the government structure, but is no less important than any other sector of government, along with being the sector most closely involved with the residents of the country.

University of Cape Town

The objectives of local government as set out in Chapter 7 of the *Constitution of the Republic of South Africa* (1996) are as follows:

a. to provide democratic and accountable government for local communities;
b. to ensure the provision of services to communities in a sustainable manner;
c. to promote social and economic development;
d. to promote a safe and healthy environment; and
e. to encourage the involvement of communities and community organisations in the matters of local government.

These objectives allow local government to not only act in the role of service delivery to communities, but also play a key role in terms of economic development (Burger 2004). It is noted that these "municipal functions are regarded as critical in improving the quality of life of individual and communities" (Ovens 2005, p.5) within the Western Cape and the country as a whole. In order to be able to carry out these responsibilities laid out by the constitution, the *Local Government Municipal Systems Act* (2000) sets out a framework containing "planning, performance management systems, effective use of resources and organisational change in a business context" (Burger 2004, p. 350). Along with this the Act created a system in which municipalities can report on their performance and allows residents to compare this performance with others (Burger 2004).

In terms of the types of municipalities that can be found within the local government structure, there are three categories of municipalities as laid out by the *Municipal Structures Act* (1998). Category A municipalities may only be formed in metropolitan areas and are therefore referred to as metropolitan municipalities. There are six metropolitan municipalities in South Africa, of which Cape Town is the only one located in the Western Cape.

Both category B and C municipalities form non-metropolitan areas, referred to as local and district municipalities respectively (*Municipal Structures Act* 1998). Of these district councils are "primarily responsible for capacity building and district-wide planning" (Burger 2004, p.349). Each local municipality forms a subdivision of a district municipality, sharing the authority with the district municipality into which they fall.

Finally, there exist District Management Areas, which are areas of a district municipality that do not qualify for a local municipality, therefore requiring local government services to be provided directly by the district municipality itself (*Municipal Structures Act* 1998).

University of Cape Town

For the purposes of this dissertation, only the 24 local municipalities of the Western Cape, along with the City of Cape Town will be considered, as outlined in Section 5.1.1.

## 3.3 Data Mining Methods

There are various methods available when using data mining as a tool. These include, but are not limited to (KDnuggets 2002):

- Association Rules
- Cluster Analysis
- Correspondence Analysis
- Discriminant Function Analysis
- Decision/Classification Trees
- Genetic Algorithms
- Logistic Regression
- Neural Networks
- Principal Component Analysis
- Visualisation

University of Cape Town

There is no particular reason to choose one method over another prior to the analysis process, if it is felt that establishing the hypotheses of the problem are in fact part of the process itself. Data mining becomes a means of exploratory data analysis, with each method offering the answer to a different question, or in fact answers to multiple questions.

It is traditionally the decision of the data mining analyst as to which methods may be considered most useful to the client, along with deciding which hypotheses, and subsequent results, might portray the relationships found in the data in the 'best' way possible. The role of the operations research practitioner corresponds with this requirement as it is their responsibility to apply that method which is felt to be most beneficial to the problem at hand. However, it is possible to involve the decision makers in this process, allowing them to advise the analyst as to which methods are of most interest within the decision making context.

It is not necessary to use every method available, but more important perhaps to rather compromise between the availability of data, the expertise of the analyst and the requirements of the client. Examples of inappropriate methods for the situation presented is this dissertation would be Neural

Networks which is classically considered to be a 'black box' approach and therefore would not allow the stakeholders involved to understand the process involved at their current skill level, nor are their specific visualisations involved in this method that aid the transferral of the knowledge gained from the results.

Correspondence Analysis was also considered inappropriate due to the fact that the nature of the data are not conducive to this method. It was attempted briefly to compare clusters of municipalities as defined through the data analysis itself to those defined through a group exercise with the DLGH. (throughout the results of the analyses undertaken, these different types of clusters shall be referred to as *data-defined* clusters and *opinion-defined* clusters). Although the results of this analysis indicated a significant difference to the clustering of municipalities as the data determined, compared to that of how members involved in the problem context would have done so, the details of this were not insightful compared to other techniques and was therefore not considered further.

Another method that was attempted on the data set was Principal Components Analysis. It was hoped that the incredibly large amount of data available could be reduced via this particular method. However, the quality and meaning of the results reflect the underlying structure of the data and the method has no ability to determine the appropriateness of these results other than by using the correlations among the variables (Hair et al. 1998). It is up to the analyst to decide on the appropriateness of results and in this particular case it was felt that, most likely due to the unreliability of the data, there were too few correlations present between variables and the results given were simply not useful enough to be included.

It was felt that the use of some form of Genetic Algorithms might also be useful with such a large data set; however the data was simply too 'messy' and unreliable with multiple, suboptimal solutions found in the few analyses that were undertaken. This method may be able to offer much at a later stage once a more reliable data set can be formed. One of its major strengths compared to some of other methods is being able to draw out non-linear relationships within the data.

The three methods eventually chosen to be used on the data set were Cluster Analysis, General Discriminant Analysis and Classification Trees. A statistical discussion of these particular methods is laid out in Chapter 4.

Cluster Analysis was a natural choice as the DLGH is currently attempting to define municipalities in terms of performance and create three groups from these. This method is able to analyse how the

University of Cape Town

data itself would cluster municipalities, which then leads these results to be compared to how these groupings had been intuitively defined by the 'experts'.

The DLGH need to not only know how the municipalities cluster, but which variables define this clustering so that they might refine their data collection process. Both General Discriminant Analysis and Classification Trees are supervised methods which identify variables that 'best' assign observations into predefined groups. The difference between these two methods are the algorithms used. General Discriminant Analysis could be considered the more powerful of the two; however Classification Trees has a particular visual focus which is easily understood by a non-expert. It was felt that these methods in particular would be useful to the problem at hand and offer visual results that could be understood with little addition training, therefore offering the most appropriate compromise between the needs of the client and the data available.

## 3.4 Further Considerations

Throughout the process of data analysis and interpretation of results, various decisions were required. These were required both in terms of the data available, but also in terms of the analysis process.

### 3.4.1 Data

The data available from the PVQ reports contained three tiers of values: variables; municipalities and time. Analysing two tiers is relatively simple; however the third dimension creates a need to decide how the data should be presented and analysed. The final decision was that each variable should be considered within all quarters of time presented, for each individual municipality. This was due to the fact that the focus of the analyses was on the similarities and differences between municipalities. Along with this, a significant portion of the data was originally missing from the data set and time trends were used to extrapolate values, leading to less reliability in terms of time, but more comparability between municipalities.

A further consideration arose concerning the data when working with the 'size' of municipalities. The data available from the census measured municipalities both in terms of their population and in terms of the number of households within a municipality. It was felt for these particular analyses that household was a better indication of size as the context of the data used was in terms of service delivery, which is measured at household level, not per person. It would however be of interest to complete the same analyses treating municipality size in terms of population and compare the results.

University of Cape Town

It may also be interesting to analyse municipalities in terms of the urban/rural ratio in each municipality. This was not taken into account in the current analysis as the relevant information was not available; however there is a strong possibility that this ratio has a significant influence on municipal performance and should therefore be considered in further research.

The final consideration in terms of the data itself is that it is reported on a quarterly basis. The data set is so large that it may be of interest to attempt to reduce it by transforming the data into annual values. If it is found that the annual values draw out similar relationships to that of the quarterly data, it would indicate that data might not need to be collected on such a regular basis, with smaller reports being done quarterly and an in depth report submitted annually.

### 3.4.2 Analyses

The analyses undertaken are currently done so on clusters as defined by the data itself (data-defined clusters), as well as clusters defined by the subjective intuition of managers of the DLGH (opinion-defined clusters). It may be of interest to use supervised techniques to determine into which district municipality a local municipality might fall. This would identify variables that show clear differences between the districts, hopefully aiding in decisions surrounding the support to certain districts.

University of Cape Town

This was attempted briefly using General Discriminant Analysis and it was found that even after relaxing requirements in the model, it could only classify municipalities with 30% accuracy. This is not too surprising as districts would have been chosen initially to group local municipalities such that there is an equitable spread of capacity. With district municipalities having been formed at the beginning of the period under analysis, it is a result in itself that there are no significant differences between districts. This indicates that the even spread of performance between district municipalities was successful; however if this analysis is rerun at a later stage and significant differences are found, it may be cause for concern.

In terms of the supervised techniques that were chosen, there were certain statistical concessions made in order to best illustrate the techniques with the data available. Ideally, a training and testing subset of data should be used, along with cross validation, to determine the best generalised model given the data. Unfortunately, the structure and reliability of the data did not allow for both this cross validation and the best illustration of the technique as much of the missing data was calculated from, and therefore inherently related to, the existing data. As the results of the analyses are on historical data and are therefore irrelevant in the current context,

the illustration of the techniques available is considered the more important of the conflicting requirements and is therefore chosen as the focus of the analyses.

There is little focus on the time trend effects in the data. This is mainly due to the large amount of missing data which was substituted using time effects and therefore an analysis along these particular lines would be inappropriate for this data set. This does not imply the same for all data available within government and as most data are collected on a regular basis, the effects of time on the data are significant. Any future analysis should account for this, obviously introducing the need for other techniques to be used, both within data mining itself and in terms of the wealth of analysis available to aid decision making in general.

Finally, the understanding of the roles of the variables in the data set was restricted, due to limited interaction with the DLGH. Along with this, the data contained a bias towards being flawed as it was seen within the municipal context as compliance data, along with being historical and therefore less relevant. Although useful in terms of illustrating the techniques available within data analysis, in order to gain further stakeholder participation it would be necessary to do so with more relevant and up-to-date data.

University of Cape Town

# Chapter 4
# Research Methods

The following section discusses the objectives of each of the three methods used for the main analyses. Included is a brief description of how the method works, relevant assumptions and a discussion on the format of the data. Finally there is a brief discussion as to how each method has been applied to the data.

## 4.1 Cluster Analysis

Cluster Analysis is used to define objective groups of observations from a particular selection of variables such that 'similar' observations are found to be of the same class (Manly 1986). This empirically based classification of objects (Hair *et al.* 1998) can not only be used for exploratory purposes, but can also be used to compare a proposed theoretically based classification to that derived from the cluster analysis (Hair *et al.* 1998). If the objective analysis presents unexpected clustering of the data, this may suggest relationships that might require further investigation (Manly 1986).

A further result of Cluster Analysis is data reduction in terms of being able to profile observations using the general characteristics of cluster membership rather than that of individual cases (Hair *et al.* 1998). Relationships that may not have been obvious using individual cases may also become apparent with the underlying structure of the data represented in these defined clusters and may be analysed using such techniques as Discriminant Analysis (Section 4.2).

### 4.1.1 Method

There are two main methods of clustering: hierarchical and non-hierarchical. Hierarchical methods begin by calculating the distances of each individual observation to all other observations. Groups are then formed by merging individuals that have the most in common with each other, represented by the shortest distance between observations, until all cases are in a single group (Hair *et al.* 1998). This results in a dendogram as a graphical representation of the

University of Cape Town

order and distances of the merging observations.

In non-hierarchical or partitioning methods, cases are allowed to move in and out of clusters at different stages of the analysis (Hair *et al.* 1998). Initially, group centres are chosen and observations assigned to the centre to which they are closest. The centres of each cluster are then recalculated and observations may then be reassigned if they are closer to another cluster centre than the one to which they currently belong (Hair *et al.* 1998). The process is iterative and ends once stability is achieved for the predetermined number of groups; however a range of values for the number of groups is analysed and results compared in order to determine the best value for the final number of groups (Manly 1986).

### 4.1.2 Standardisation

There are many advantages to standardization in Cluster Analysis. The main reason for its use is that most distance measures are sensitive to differing scales or magnitudes between variables (Hair, J., *et al.*,1998).This is solved by creating z-scores (taking each observation and subtracting the mean of the variable and dividing by its standard deviation) resulting in observations that have zero mean and a standard deviation of one. This allows for easier comparison between variables as they are on the same scale, as well as positive values being above the mean and negative values below. The magnitude represents the number of standard deviations the original value lies from the mean (Hair *et al.* 1998).

A further type of standardisation that can be useful in Cluster Analysis is to standardise each observation not to the sample's average, but instead to that observation's average score across variables. In most cases the factor of interest is the relative importance of one variable to another and whether clusters can be found with similar patterns of importance (Hair *et al.* 1998). This type of standardization removes response effects and is similar to a correlational measure in highlighting the pattern across variables, but the proximity of observations is still used to determine the similarity of values (Hair *et al.* 1998).

### 4.1.3 Assumptions

There are two main subjective effects caused by the researcher when undertaking Cluster Analysis. The first concerns the algorithms used in the analysis, as there are many available, but no one accepted method. The different algorithms do not necessarily give the same results for a given set of data (Manly 1986) and therefore the researcher imposes a structure on the data through choice of methodology (Hair *et al.* 1998). The second subjective effect concerns the

University of Cape Town

choice of variables. There is no means of differentiating relevant from irrelevant variables (Hair *et al.* 1998); the technique only derives clusters that are consistent, yet separate, using all variables.

## 4.2 General Discriminant Analysis

General Discriminant Analysis is used when it is of interest to ascertain how well two or more groups of observations might be separated using a particular set of variables for these observations (Manly 1986). Due to the nature of the technique it is possible, as when using regression, to use a stepwise approach where variables are added to the analysis one by one until adding any more variables does not aid in being able to significantly discriminate further between groups (Manly 1986).

The simplest approach to General Discriminant Analysis is to determine linear combinations of the original independent variables ($\mathbf{Z}_i$) such that the distance between centroids of each group are maximised (Manly 1986). The best separation between groups occurs when the mean value between groups, using the variables of interest, differs significantly, but the values within a group are fairly similar.

It is possible to have more than one discriminant function ($\mathbf{Z}_i$), depending on the number of groups in the analysis; however due to the nature of the technique these have ordered explanatory power and generally the first few functions can account for most of the important differences between groups (Manly 1986). This allows for a graphical representation of the relationship between groups by plotting the individual Z-scores obtained through the discriminant functions for each observation and the group centroids (Statsoft Inc 1997).

Observations that were not previously part of the analysis can be assigned to a group, in terms of the distance to the closest group centroid (Manly 1986).

### 4.2.1 Method

Although there are various methods of discrimination available, one of the main approaches makes use of the mahalanobis distance measure. The mean vectors of the observations in each group are regarded as estimates of the true mean vectors for each group (Manly 1986), allowing the mahalanobis distance from each observation to each group centroid to be calculated and observations assigned to the group to which it is closest. This may not necessary be the group that the observation originally came from and the percentage of correct allocations indicates how

University of Cape Town

well the groups have been separated using the variables of interest (Manly 1986).

The discriminant functions ($Z_i$) that are created as a linear combination of the variables under consideration can be chosen in various ways. One of the main ways in which to do this is to choose the coefficients of the linear combination such that the F-ratio testing the difference between means, using ANOVA, is maximised (Manly 1986). This turns out to be an eigenvalue problem and determines that the number of discriminant functions available is the smaller of the number of variables or the number of observations less one. If there is more than one canonical discriminant function available, the first function maximises the F-ratio for a one-way ANOVA and successive functions maximise this same F-ratio subject to the condition that there is no correlation between discriminant functions within groups (Manly 1986) i.e. that the functions are independent from one another.

Interpreting these discriminant functions involves the coefficients of the variables chosen for the discriminant function. The greater the coefficient the greater the contribution of its respective variable to the discrimination between groups (Statsoft Inc 1997); however it does not indicate between which groups this variable might discriminate well (Statsoft Inc 1997). For this purpose, group centroids and factor structure coefficients are used (Garson 2007). These factors structure coefficients indicate the correlation between each variable and its respective discriminant function (Statsoft Inc 1997).

It is also possible to test for significant differences between the groups using various techniques. Mainly these test for overall differences between means of all the groups or more specifically between the means of any pairs of groups (Manly 1986) as is seen in ANOVA.

University of Cape Town

### 4.2.2 Standardisation

Unlike Cluster Analysis there is no need to standardise the data to zero mean and unit variance prior to the analysis. This is due to the fact that the outcome of this particular analysis is "not affected in any important way by the scaling of individual variables" (Manly 1986, p.87). This point can however be reversed and indicates that it is in fact possible to use standardised data in the analysis and the same format of the overall data set may be used for both Cluster and Discriminant Analysis.

### 4.2.3 Assumptions

There are various assumptions underlying discriminant analysis using one-way ANOVA to

determine canonical discriminant functions, mainly affecting significance tests between groups. Firstly, there is the assumption that the within-group covariance matrix is the same for all groups (Manly 1986) and if this assumption is not met then the results of the analysis may be unreliable. This particularly effects the tests of significance which also include the assumption that within groups the data follow multivariate normal distributions (Manly 1986). Further, when using a stepwise approach to Discriminant Analysis it is possible that given large numbers of variables there is a high chance of significance achieved by chance alone (Manly 1986).

However as some of these assumptions are largely relevant to tests of significance it does not mean that a discriminant analysis cannot be used. The technique may determine discrimination on non-normal populations quite well, leaving the ability to easily determine the significance of the results potentially limited (Manly 1986).

## 4.3 Classification Trees

Classification Trees are hierarchical classifiers based on a supervised learning technique (Vos & Evers 2006). This allows one variable to be predicted from a set of measurements made on some class or object (Breiman *et al.* 1984) and depending on the problem the focus of the analysis may either be "to produce an accurate classifier or to uncover the predictive structure of the problem" (Breiman *et al.* 1984, p.6) thereby characterising the conditions that determine the class to which an object belongs (Breiman *et al.* 1984).

The procedure used gives output which is easily interpreted (Vos & Evers 2006), allowing a clear and natural understanding of the predictive structure of the data (Breiman *et al.* 1984). The technique requires much computation, at each stage completing an automatic stepwise variable selection that reduces complexity (Breiman *et al.* 1984); however it handles non-homogeneous relationships well, making powerful use of conditional information (Breiman *et al.* 1984).

### 4.3.1 Method

The general idea of a Classification Tree is nothing more than splitting up the set of observations into smaller and smaller subsets (Vos & Evers 2006). Therefore the three main elements used to construct a tree are the selection of splits; whether to stop splitting at a particular node; and finally how to assign each terminal node to a specific class (Breiman *et al.* 1984).

Initially the whole set of data (root node) is split and then the child nodes are split again and so forth (Vos & Evers 2006). The splits are computed such that the subsequent subsets of data are

University of Cape Town

more 'pure' than the parent subset (Breiman *et al.* 1984) to the point where no more decrease in 'impurity' is possible, or ideally where there is only one class of observation in each terminating node (Vos & Evers 2006).

There are two main decisions when computing a split in the tree. Firstly, which variable to use for the optimal split, as only a single variable is to be used for each split (Breiman *et al.* 1984), and secondly which values of this variable to assign to the left branch and which to the right (Vos & Evers 2006). At each node to be split an algorithm searches through all the available variables one by one and for each finds the best possible split. After this it compares the best split on each variables and selects the best to use in the tree (Breiman *et al.* 1984).

One of the main advantages to using Classification Trees is that there is no loss of information as the terminal nodes together simply represent the entire original data set, split into its various classes. Each terminal node is classified by a class label, which may or may not be made entirely of observations belonging to this class, and there may be more than one terminal node with the same class label (Breiman *et al.* 1984).

The graphical representation of the tree is not the only way to represent the classification of cases. If all the terminal subsets corresponding to a particular class are considered together, a set of rules may be determined that will classify future cases to this class. As all splits in the tree leading to these subsets must be incorporated, it might make sense to represent these splits "as "If . . . then . . . " rules" (Vos & Evers 2006, p.90). For ease of interpretation it is important to ensure that the number of rules is kept quite small and that the rules themselves are quite simple (Vos & Evers 2006).

University of Cape Town

### 4.3.2 Standardisation

As with General Discriminant Analysis there is again no need to standard the data to zero mean and unit variance prior to the analysis. This is due to the fact that every optimal split in the tree is unaffected by all monotone transformations of individual variables (Breiman *et al.* 1984) and the same cases will go left and right in the split no matter which format of data are used (Breiman *et al.* 1984). As with General Discriminant Analysis, this indicates that although there is no need to standardise the data, the standard form may be used and the same format of the overall data set may be used for all three main analysis techniques.

### 4.3.3 Assumptions

As with Cluster Analysis there exist subjective effects caused by the researcher when undertaking to create a Classification Tree. Again there are many algorithms available, but no one accepted method, and the different algorithms do not necessarily give the same results for a given set of data.

However, concerning the choice of variables, as Classification Trees use a stepwise procedure to select variables (Breiman *et al.* 1984), the most appropriate variables will be chosen in the context of the chosen algorithm. This does not imply that the most appropriate variable is chosen in the context of the problem. A variable may not appear in the splits of a Classification Tree and therefore imply it has little association with the classification of observations; however it may still be important with its true effect being masked by other variables in the analysis (Breiman *et al.* 1984).

Although Classification Trees are easy to interpret, they are often outperformed by other methods in terms of accuracy due to their relatively high variance (Breiman *et al.* 1984). This implies that splits should not be over-interpreted in terms of which is more important than another (Breiman *et al.* 1984).

There are two main sources of variation in a Classification Tree. The first concerns the algorithm that is chosen to determine the 'best' split in a tree in term of decreasing the given definition of impurity within the tree. The second is the chosen size of the tree as growing the full tree until the impurity cannot be decreased any further may lead to overfitting (Vos & Evers 2006). Although it is possible to use stopping criteria as the tree is being grown, it is actually more successful to grown the full tree and then prune it back (Vos & Evers 2006).

University of Cape Town

# Chapter 5
# Data Collection

The data set was created from data made available through the Department of Local Government and Housing (DLGH) in the Western Cape. It is contained in quarterly Project Viability Questionnaires (PVQ) and is restricted to the period of June 2002 through to June 2005. Those reports dating after December 2003 were available electronically; however the remaining reports required data capturing from printed reports. The final set of data was the most recent available at the time of constructing this particular collection of data and the June 2002 report was the earliest report available from the department. Prior to this the municipalities in the Western Cape numbered 136 and the questionnaire was developed to coincide with the amalgamation of these into 5 District Municipalities, 24 local municipalities and the City of Cape Town, totalling 30 municipalities. Any data collected before this amalgamation cannot be used due to the significant changes in restructuring the system. There were further changes made to the questionnaire in December 2003; however there being much in common between the two it was possible to extrapolate observations throughout the period to create a complete data set.

University of Cape Town

The initial data set chosen from these reports included all 30 municipalities and 63 variables across all 13 quarters, resulting in 24 570 observations. There were two main types of variables apparent in the data: those that were taken from the financial statements of a municipality and are mainly measured in Rands (henceforth referred to as *financial type* variables) and those that directly represent the service delivery of a municipality measured mainly in terms of counts or percentages (henceforth referred to as *service delivery type* variables).

## 5.1 Missing Data

There was a substantial amount of data missing from the original data set. This is due to various reasons; including some sections not always being compulsory to complete; a learning period as municipalities adapted to the new system; and sometimes data simply not being available. Various techniques were utilised to create a complete data set, keeping in mind the large amount of data and

the time constraints relevant to calculations that need to be used, attempting to rely on methods with computational bias and much computer usage with as little human intervention as possible. This may cause some of the final observations to differ from those captured in the original questionnaires.

### 5.1.1 Removed

There are three main reasons as to why certain observations were removed from the data set. Firstly, all values relating to the District Municipalities are deleted from the data set as a part of their responsibility of these types of municipalities is the management of municipalities within their district and the final definition of this particular role is still in the process of being defined. Along with this, the financial data for the five District Municipalities portrays a significantly different representation of performance than that of the 25 local municipalities (including the City of Cape Town) and therefore cannot be logically comparable to these local municipalities. The City of Cape Town is not considered to be a District Municipality in this case as it stands on its own as a Metropolitan Municipality and can therefore be considered comparable to the other local municipalities.

The second main reason to exclude observations is simply due to too much missing data. The decision is made that when more than 40% of the data for a particular section of the data set is missing the section is removed. This applies to both variables and municipalities. If 40% or more of all data for a particular municipality is missing, the municipality is deleted from the data set. In this case both *Oudtshoorn* and *Laingsburg* are removed, leaving 23 municipalities to be analysed as shown in Table 5.1.

*Table 5.1: Maps of municipalities chosen*

| All Municipalities and Districts | Municipalities Chosen For Analysis |
|---|---|
|  |  |

University of Cape Town

Looking through the variables, the decision is made such that if more than 40% of observations are missing for a variable over all 13 quarters in the period under analysis then the variable is deleted. This is done as it is highly unlikely that the missing data could be substituted with reliable data using the data already existing. Within the variables themselves, if less than 4 out of 13 observations exist for a particular municipality these are deleted, as extrapolating from so few observations is not considered to be reliable in maintaining the relationship between municipalities.

Finally, all variables are considered and any variables not containing significant variation for analysis are deleted. As a consequence all categorical data is removed, as these variables were mainly descriptive or indicated categories that remained constant over the entire period of analysis. However the process is done in order to ensure that the remaining data are useful to the methods used to analyse the data. Some variables have been collected in a similar manner for various types of basic service delivery and the result of this particular type of deletion is that variables are not consistent across types and therefore less comparable. However, the lack of reliable data is considered to be a more important factor than the consistency of reporting.

Once this process is completed, there are 36 variables across 13 quarters of data for 23 municipalities remaining which results in 10 764 potential observations, of which 4469 (41.52%) are still missing. The data set is much reduced from the original but the percentage of missing data has not reduced substantially. The difference lies in the structure that now allows for far more reliable substitution of missing values. At this point the missing observations can be substituted using various techniques, taking into account the size of the data set and the sections of the data set that are missing.

University of Cape Town

### 5.1.2 Interpolation

In the case where a variable had more than 4 out of 13 observations available within a municipality it is possible to interpolate and/or extrapolate the data to substitute the missing values.

If there are data points missing between existing points simple linear interpolation is used. Although this technique does contain the weakness of the assumption that the relationship between points is linear, due to the large number of observations it is considered acceptable. This technique reduces the number of missing data points to 4341 (40.33%).

If however there are data points missing either earlier or later in the period than any of the existing observations for a particular variable a different method of substitution is considered. Initially extrapolation using a simple linear regression of the existing observations was used; however most of the data are very volatile over time and the resulting figures are unrealistic, many of them tending into the negative. These results are not surprising for extrapolation of a time series; however any serious time series analyses is considered impractical in terms of the size of the data set and the amount of missing data.

An attempted solution to this problem is to utilise a moving average which will contain the data within its existing range, keeping the values positive, but not unrealistically large. The outcome of this technique will differ slightly according to how many observations were originally missing for each variable within a municipality; however it will most likely result in more realistic values than most other options, taking into account the sheer amount of data. This technique reduces the number of missing data points to 1391 (12.92%), leaving 107 sections of the data set where all 13 quarters of observations for a variable are missing within a particular municipality.

### 5.1.3 Regression

In the case where all 13 observations for a particular variable are missing within a particular municipality a method of multiple linear regression is used in order to estimate values while maintaining the relationship between municipalities.

The only variables that are complete for all municipalities and across all 13 quarters at this point are *BankBalance*; *CDTotal; DebtOutstanding; Investments* and *WaterIndHouse*. These are used as the independent variables in the multiple linear regression.

If a variable is completely missing for a particular municipality across all 13 quarters, it is possible to estimate those values using the data from the remaining municipalities for that same variable. The relationship between the variable and the independent variables mentioned in the paragraph is determined using the data that is available from the other municipalities and then it is applied to the independent variables for the particular municipality to determine the missing values. This ensures that the relationship between the municipalities is maintained. If more than one municipality has the same variable completely missing across all 13 quarter the analysis is done separately for each of them, using only the data that was originally available from the other municipalities. This process finally completes the substitution of missing data.

University of Cape Town

## 5.2 Meta Data

The following section explains the data to be used in all the main analyses. This includes not only the variables chosen and their definitions, but also the geographical location of the municipalities chosen.

### 5.2.1 Variables

As mentioned above, due to a considerable amount of missing data, both *Laingsburg* and *Oudtshoorn* are removed from the analysis. This leaves the following 23 municipalities:

Matzikama
Cederberg
Bergrivier
Saldanha
Swartland
Witzenberg
Drakenstein
Stellenbosch
Breede Valley
Breede River / Winelands
Theewaterskloof
Overstrand
Cape Agulhas
Swellendam
Kannaland
Hessequa (previously Langeberg)
Mossel Bay
George
Bitou (previously PlettenburgBay)
Knysna
Prince Albert
Beaufort West
City of Cape Town

Also, as discussed, any variable containing more than 40% missing data, or not containing significant variation is removed from the analysis. This leaves the following 36 variables shown in Table 5.2 in alphabetical order, with a brief explanation of the variable is to the right of it.

University of Cape Town

*Table 5.2: Variable Descriptions*

| | |
|---|---|
| BankBalance | Bank Balance |
| CD30-60days | Outstanding debtors 30~60 days for service |
| CD60-90days | Outstanding debtors 60~90 days for service |
| CD90+days | Outstanding debtors 90+ days for service |
| CDCurrent | Current/30 day outstanding debtors for service |
| CDElectricity | Consumer Debtors by nature of service: Electricity |
| CDRates | Consumer Debtors by nature of service: Rates |
| CDRefuseRem | Consumer Debtors by nature of service: Refuse Removal |
| CDSewerage | Consumer Debtors by nature of service: Sewerage/Sanitation |
| CDTotal | Total consumer debtors |
| CDWater | Consumer Debtors by nature of service: Water |
| CrBulkElec | Creditors: Bulk Electricity |
| CrPAYEDed | Creditors: PAYE Deductions |
| CrPension | Creditors: Pension/Retirement deductions |
| CrTotal | Total creditors |
| CrTrade | Creditors: Trade |
| CrVAT | Creditors: VAT |
| DebtOutstanding | Total outstanding debtors |
| ElecDisconnection | Electricity losses due to illegal connections (%) |

University of Cape Town

| | |
|---|---|
| ElecIndHouse | Provision of free basic electricity: Number of Indigent Households |
| ElecIndMonetary | Provision of free basic electricity: Monetary Value |
| ElecReconnection | Number of electricity reconnections over the past three months |
| Investments | Investments |
| LossElecTrans | Electricity transmission losses (%) |
| LossWaterTrans | Water transmission losses (%) |
| Overdraft | Overdraft facility of main bank account |
| PMS%Implemented | Performance Management System (PMS) developed/implemented (%) |
| RRIndHouse | Provision of free basic refuse removal: Number of Indigent Households |
| RRIndMonetary | Provision of free basic refuse removal: Monetary Value |
| SanIndHouse | Provision of free basic sanitation: Number of Indigent Households |
| SanIndMonetary | Provision of free basic sanitation: Monetary Value |
| WaterDisconnection | Water losses due to illegal connections (%) |
| WaterIndHouse | Provision of free basic water: Number of Indigent Households |
| WaterIndMonetary | Provision of free basic water: Monetary Value |
| WaterOthHouse | Provision of free basic water: Number of Other Residential Households |
| WaterReconnection | Number of water reconnections over the past three months |

University of Cape Town

### 5.2.2 Municipalities

It may be of interest to consider the geographical location of certain municipalities, particularly when considering how the data clusters them. The location and boundaries of each municipality in the analysis, along with its relevant district, are shown overleaf:

**West Coast District**

Matzikama

Cederberg

Bergrivier

Saldanha Bay

Swartland

**Boland District**

Witzenberg

Drakenstein

Stellenbosch

Breede Valley

Breede River / Winelands

University of Cape Town

**Overberg District**

Theewaterskloof

Overstrand

Cape Agulhas

Swellendam

**Eden District**

Kannaland

Hessequa

Mossel Bay

George

Bitou

Knysna

University of Cape Town

**Central Karoo District**

Prince Albert

Beaufort West

**City of Cape Town**

City of Cape Town

University of Cape Town

## 5.3 Standardisation

In order to be able to analyse the performance of municipalities throughout the Western Cape, they must be comparable. This can be considered in various ways.

### 5.3.1 Census

One of the most important factors is to standardise the data in terms of the size of each municipality. This size can be measured in two different ways from the data available: population in a municipality or the number of households in each municipality.

These figures are reported every quarter by each municipality, but some reports appear to increase or decrease across quarters at a rate which seems impossible and are therefore to be considered unreliable. Although less accurate, as the figures do not alter throughout time, a far more robust estimate of population size and number of households is taken from the 2001 census data (*Census South Africa* 2001). These figures are shown in Table 5.3 and clearly indicate (note in particular the City of Cape Town) the need for standardisation to be able to more accurately compare municipalities. It was felt that, as the problem context related to service delivery, which is measured per household in the variables available, standardisation by number of households

would be most appropriate, although population could also be considered in future analysis.

*Table 5.3: Municipal population and household data*

| Municipality | Population Size | Number of Households |
|---|---|---|
| Matzikama | 50209 | 14495 |
| Cederberg | 39326 | 11223 |
| Bergrivier | 46325 | 13363 |
| Saldanha | 70439 | 18918 |
| Swartland | 72116 | 18757 |
| Witzenberg | 93567 | 20460 |
| Drakenstein | 194418 | 46265 |
| Stellenbosch | 118708 | 35124 |
| Breede Valley | 146029 | 35094 |
| Breede River / Winelands | 81272 | 21213 |
| Theewaterskloof | 93277 | 24362 |
| Overstrand | 55469 | 19020 |
| Cape Agulhas | 26469 | 7652 |
| Swellendam | 28075 | 7617 |
| Kannaland | 23971 | 6154 |
| Hessequa | 44115 | 12660 |
| Mossel Bay | 71494 | 20258 |
| George | 135408 | 36190 |
| Oudtshoorn | 84691 | 18411 |
| Bitou | 29183 | 9944 |
| Knysna | 51469 | 14973 |
| Laingsburg | 6679 | 1943 |
| Prince Albert | 10513 | 2613 |
| Beaufort West | 37106 | 9106 |
| City of Cape Town | 2892243 | 779239 |

### 5.3.2 Statistical

To ensure that the municipalities are completely comparable, as well as the variables being so, each variables across all 13 quarters for each municipality is statistically standardised. This results in the mean for each variable across all 13 quarters being zero with a standard deviation of one, becoming completely comparable in terms of measurement. A second standardisation is undertaken in terms of standardising each variable across the 13 periods for each municipality, allowing for better comparability between municipalities and variables over time.

Not only is this statistical standardisation necessary for many of the techniques used, but when attempting to predict further classifications within General Discriminant Analysis or Classification Trees it allows for simpler computation.

University of Cape Town

# Chapter 6
# Cluster Analysis

Within the Department of Local Government and Housing (DLGH) the focus is currently on classifying municipalities into three groups. Due to this the following analysis will use K-Means Clustering to cover clustering the municipalities into three groups, but will also consider the analysis using both two and four clusters. This will examine whether the current focus is best or whether it is only necessary to classify into two groups, with less effort, or whether a more detailed analysis need be considered and the most appropriate grouping might actually be four clusters. It is important to note that this analysis creates the clusters from the data alone and there is no human interaction with the algorithm.

For better interpretation and comparability when using Cluster Analysis, all data has not only been standardised across variables so that they might be compared on the same scale, but also across observations to remove response effects within variables and highlight the patterns across variables.

The use of all variables in this particular analysis is due to there being no reason to suspect that any variable should take precedence over another. In fact, it is through treating all variables with equal importance that it is possible to ascertain which ones explain the variation in the data better than others. As about 150 out of the total 468 variables show significant differences between clusters at the 5% level of significance; a 0.1% level of significance will be used to reduce the number of significant variables to around 40. A potential weakness in this technique is that due to the large number of variables available, it is possible that some may be found to be significant by chance alone and this should be taken into account when interpreting results.

## 6.1 Discussion

The results of the 2, 3 and 4 group Cluster Analysis are given in Appendix I, II and III. Using these results it is possible to ascertain which number of clusters might be the most appropriate, as well as determine the profiles of these clusters.

University of Cape Town

There are clearly enough significant variables available in the results to be able to discriminate between 2, 3 or 4 clusters, as in each of the three cases about 150 variables are found to be significant at the 5% level and about 40 at the 0.1% level. A further consideration is the purpose of the clustering in the context of the DLGH. Ideally, as the DLGH is focusing on three clusters of municipalities when dealing with performance, it would be appropriate for the data to show this same profile. However, with the number of municipalities in each cluster under consideration, it appears that a 3 cluster solution is unbalanced with *Drakenstein* and *Bitou* on their own and of the two remaining clusters one is half the size of the other as shown on the geographical map in Illustration 6.1.

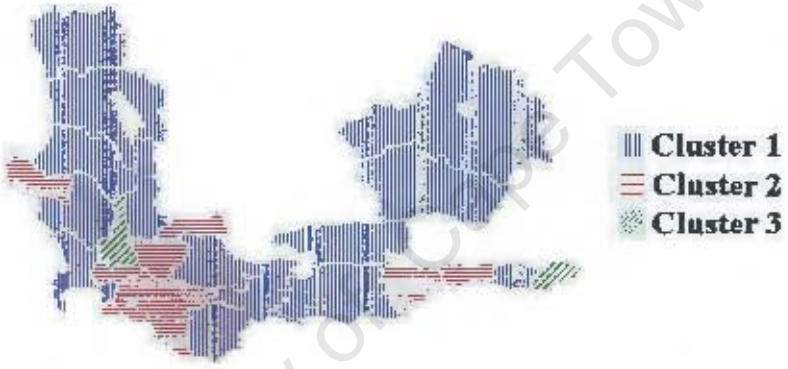



*Illustration 6.1: Geographical three cluster membership*

University of Cape Town

Traditionally in Cluster Analysis it is desirable to result in clusters of similar size with sufficient numbers of observations in each cluster; however this would have to be considered within the problem context. Even in the 2 cluster solution one cluster is almost half the size of the other; however this can still be considered appropriate as there are sufficient numbers in each cluster. If a larger number of clusters is required then the 4 cluster solution is more desirable as although there are still two clusters with only two or three municipalities, overall they are better balanced in terms of numbers. However, within the given problem context this requirement may not be as important as first appears, depending on the requirements of the stakeholders involved. There may be a simple need to distinguish those few municipalities that are outperforming, or underperforming, the others where numbers have little consideration or there might be a need to create tiers of municipal support, in which case more balanced numbers are likely to be desirable.

As this analysis is undertaken in terms of the illustration of the technique, along with one of the main objectives of cluster analysis is that there be as few clusters as possible, along with enough

significant differences between the clusters to be able to gather a clear profile, the two cluster solution will be considered in terms of profiling. The group membership of the two cluster results, along with their distance to the cluster average, are shown in Table 6.1.

*Table 6.1: Municipal cluster membership in two cluster solution*

| Cluster 1 | Distance | Cluster 2 | Distance |
|---|---|---|---|
| Matzikama | 0.6232 | Saldanha | 0.8222 |
| Cederberg | 0.6617 | Drakenstein | 0.9043 |
| Bergrivier | 0.5516 | Stellenbosch | 0.8788 |
| Swartland | 0.7135 | Breede Valley | 0.7853 |
| Witzenberg | 0.7687 | Theewaterskloof | 0.8376 |
| Breede River / Winelands | 0.8262 | Overstrand | 0.7764 |
| Cape Agulhas | 0.7013 | Mossel Bay | 0.8260 |
| Swellendam | 0.5241 | George | 0.8078 |
| Kannaland | 0.5940 | | |
| Langeberg | 0.8815 | | |
| Broo | 0.6897 | | |
| Knysna | 0.8445 | | |
| Prince Albert | 0.4928 | | |
| Beaufort West | 0.8103 | | |
| City of Cape Town | 0.9555 | | |

The variables that show significant differences between the two clusters can be considered in terms of time as shown in Illustration 6.2.

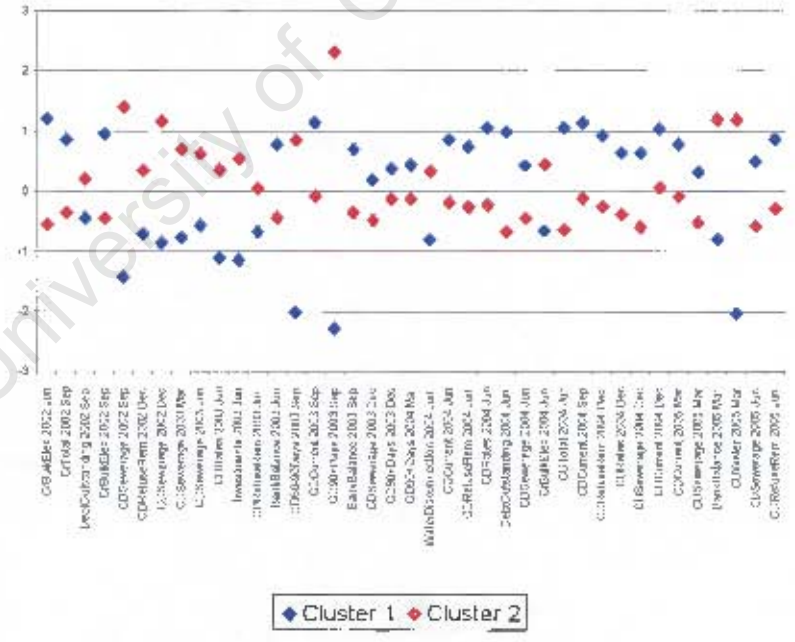



*Illustration 6.2: Significant variables in cluster analysis over time*

Unusual observations are *BankBalance* and *CDWater* for March 2005 where Cluster 1 is much lower than Cluster 2 in a period where for most significant variables the opposite is true. In the earlier periods *BankBalance* June 2003 and *CDCurrent* September 2003 are higher in Cluster 1 than Cluster

University of Cape Town

2, which again contradicts the patterns seen with most of the other variables. There appears to be a clear distinction between earlier and later periods where Cluster 1 initially has lower scores than Cluster 2, but after mid 2003 this changes and Cluster 1 shows higher average scores for the significant variables than Cluster 2.

If the different types of variables are considered, of the service delivery variables only the percentage of water losses due to illegal connections is considered significant. This is only found to be significant for the period June 2004 and Cluster 2 is much higher than Cluster 1. Concerning the financial type variables, an interesting point to note is that medium-to-long-term debtors are higher in Cluster 2 in September 2003 with no such indications before this point, implying that there may have been some specific event to cause this to occur at this particular time. However, after this point Cluster 2 indicates significantly lower current debtors, indicating that whatever occurred in this time period has a significant effect on later periods.

Another factor that might be considered at this point is that investments in Cluster 2 were significantly higher than that of Cluster 1 in June 2003 and may have had some affect in the changes of cluster profiles around September 2003, as it is not picked up as significantly different in any other reported quarter.

It is not clear at this point as to what level of debt relates to good or poor service delivery, but it is clear that this is one of the most significant factors that discriminates between the profiles of the two clusters. The level of this, along with bank balance, changes over time; reversing the profile of the clusters at the beginning of the period to that at the end. The extent to which this change occurs is indicated in Table 6.2.

University of Cape Town

*Table 6.2: Significant variable levels prior/post September 2003*

| | Prior September 2003 | | Post September 2003 | |
|---|---|---|---|---|
| | Cluster 1 | Cluster 2 | Cluster 1 | Cluster 2 |
| BankBalance | High | Low | Low | High |
| CDCurrent | - | - | High | Low |
| CDRates | Low | High | High | Low |
| CDRefuseRem | Low | High | High | Low |
| CDSewerage | Low | High | High | Low |
| CrBulkElec | High | Low | Low | High |

The general profiles of the clusters indicate, as would be expected, that when a municipality has low

debtors and high creditors it is likely to have a high bank balance, and visa versa. The period surrounding June-September 2003 is not the only one of interest however. The quarter that indicates the most number of variables with significant differences between the clusters is in fact June 2004, which is shown in Illustration 6.3.

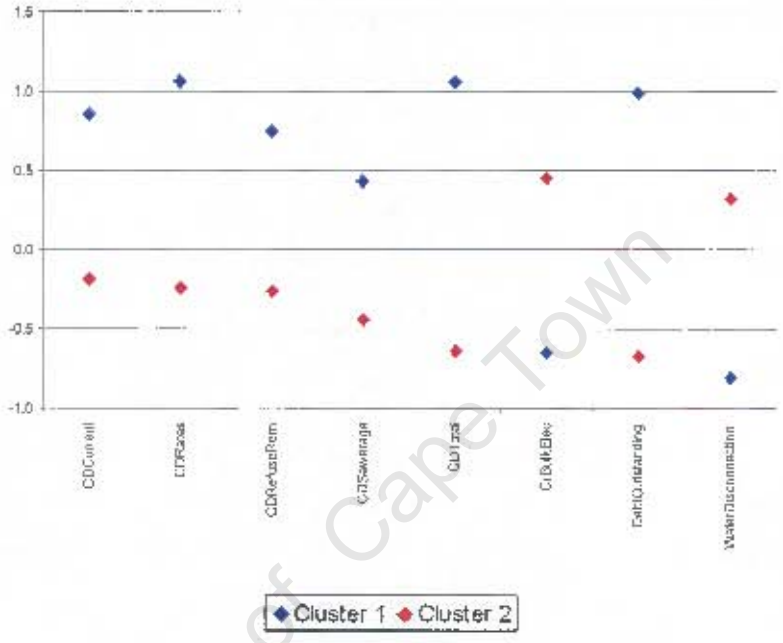



*Illustration 6.3: Significant variables for cluster analysis in June 2004*

For this period in particular various types of debtors in Cluster 2 are significantly lower than that of Cluster 1; however it appears that creditors concerning bulk electricity in particular and the percentage of water losses due to illegal connections is significantly higher for Cluster 2 than Cluster 1 in this quarter. This indicates that simply keeping debtors low as good financial practise does not necessarily indicate good service delivery management.

It is interesting to note that there is no significant difference between the clusters in terms of consumer debtors by nature of service relating specifically to water (*CDWater*), nor in terms of electricity losses due to illegal connections (*ElecDisconnection*) implying that there may not necessarily be a connection between specific financial indicators and the actual service delivery figures they relate to. This is confirmed by the fact that the correlation between *CDWater* and *WaterDisconnection*, along with the correlation between *CrBulkElec* and *ElecDisconnection*, for June 2004 is less than 0.2 indicating the absence of a relationship between these variables.

University of Cape Town

When looking at the relationship between these variables across all periods there are no strong correlations, unless the correlations are considered specifically in terms of the two clusters. There are still no strong correlations evident in Cluster 1 for any of the periods with all correlations except one below 0.3; however in Cluster 2 there are consistently higher correlations between these particular variables. For December 2003 in particular the correlation between *CrBulkElec* and *ElecDisonnection* is 0.87 indicating a very strong positive relationship. This indicates that there may be potential to use financials as indicators of service delivery; however the current system would need to be improved dramatically in terms of reliability before this could be considered a reality. Realistically, there is always the need to consider actual service delivery figures and consider these in the context of the reported financial figures.

University of Cape Town

# Chapter 7
# General Discriminant Analysis

Clustering Analysis has determined how the municipalities naturally cluster into 2, 3 or 4 groups and that there are around 150 variables that have at least one mean different from the others across clusters at the 5% level of significance. However, many of these variables could be explaining the same variation in the data and it would be desirable to reduce the number of variables, also ensuring that the variables chosen are independent from one another and therefore explain different structures of variation in the data.

General Discriminant Analysis is used to determine which variables have the highest discriminatory power between clusters and combining this with a step-wise analysis ensures that no multicollinearity is present in the final model. The parameters initially chosen for the model are such that variables enter the model at the 1% level of significance (p-to-enter = 0.01) and to leave the model they must no longer be significant at the 5% level (p-to-remove = 0.5). These parameters are chosen specifically to limit the number of variables in the model and thereby allow for a model containing better ability to generalise. As these initial clusters have been created using the structures inherent in the data itself, it is likely that General Discriminant Analysis will perform well and identify the same significant variables as that in the Cluster Analysis; however the technique also removes any redundancy in these variables.

It is possible to classify 'new' municipalities into specific clusters using the discriminant functions (Appendix IX) developed from these models; thus it would be more conservative to use the fewest variables possible in the model, as not to encourage overfitting, therefore allowing for the best generalisation and hopefully appropriate classification of further municipalities.

Keeping this in mind, when using General Discriminant Analysis, it is possible to test whether the distance between group centroids is significant. For these particular analyses there are so many variables available, even when restricting the parameters as mentioned above, that it is decided to continue the stepwise analysis

University of Cape Town

only until these significant differences between clusters appear and there is a 100% correct classification rate using the model. This is not necessarily statistically sound, but as the purpose of the analysis here is as an illustration, as well as the groups being determined from the data itself, this particular limitation of the model is considered acceptable.

There is a further consideration that to be able to choose the 'correct' variables for the model more would have to be known about the context of the problem to be able to pick only relevant and generalisable variables to be used. However for the purposes of these analyses and illustrations of technique, the variables of interest are those that are chosen using the data itself.

## 7.1 Discussion

This technique separates the clusters by using a combination of the variables chosen to be included in the model. If more than one combination (dimension) is used the model also ensures that these dimensions are independent from each other, thereby explaining the most variation possible. This allows for a far more powerful separation of clusters than if the variables had just been used individually as in ANOVA. Although all the results from the analyses are calculated (Appendix VI, VII and VIII) for the purposes of the illustration of these concepts the results from the four cluster analysis shall be used.

Although it would be beneficial to name clusters by their defining characteristics, the analysis did not yield clear or obvious results. This was confounded by the level of understanding of which variables might indicate a group of municipalities showing good or bad performance in terms of service delivery, as one of the objectives of the analysis was to aid in this understanding. Therefore the clusters are simply referred to by their numeric labelling given in the results of the analysis.

University of Cape Town

### 7.1.1 Separation Using Variables and Dimensions

There are four variables chosen to be included in the four cluster model in order to be able to separate the clusters of municipalities with 100% accuracy. The most significant of these is *CD60-90Days* 2003 December; however to illustrate the effect of the combination of variables, the first two chosen variables shall be considered and *CDSewerage* 2005 March shall be included at this point. The clusters of municipalities simply plotted against these two variables is shown in Illustration 7.1.

It is clear in this graph that *CD60-90Days* 2003 December clearly separates part of Cluster 3, *Bitou* in fact, from the other municipalities; however it is a poor discriminator when attempting

to separate the other three clusters from each other. Looking at *CDSewerage* 2005 March, it appears that this particular variable separates out the cluster centres quite effectively, but only begins to define the separation between entire clusters. If cut-off points were attempted on this variable it may be possible to just separate Clusters 1 and 2; however municipalities in Clusters 3 and 4 would be clearly misclassified.

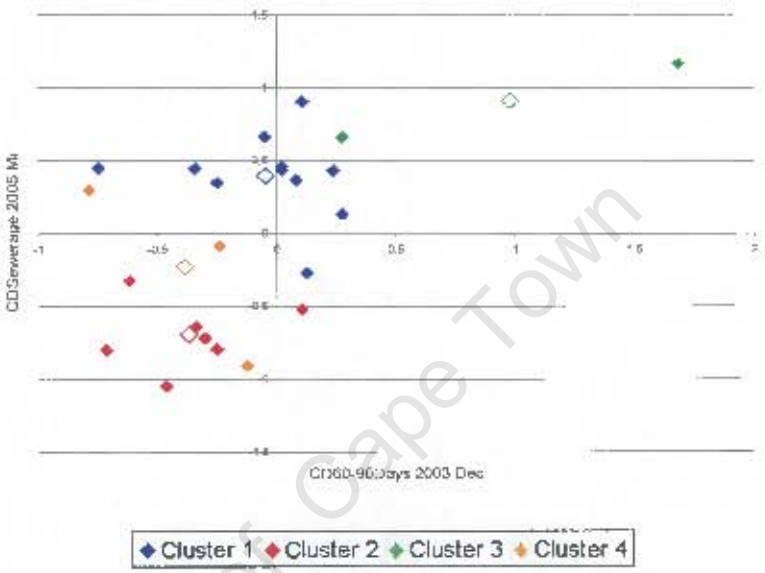



*Illustration 7.1: Municipalities against first two significant variables in four cluster GDA model*

The clusters of municipalities are shown on the first two dimensions of the General Discriminant Analysis model, that includes these same two variables, is shown in Illustration 7.2.

University of Cape Town

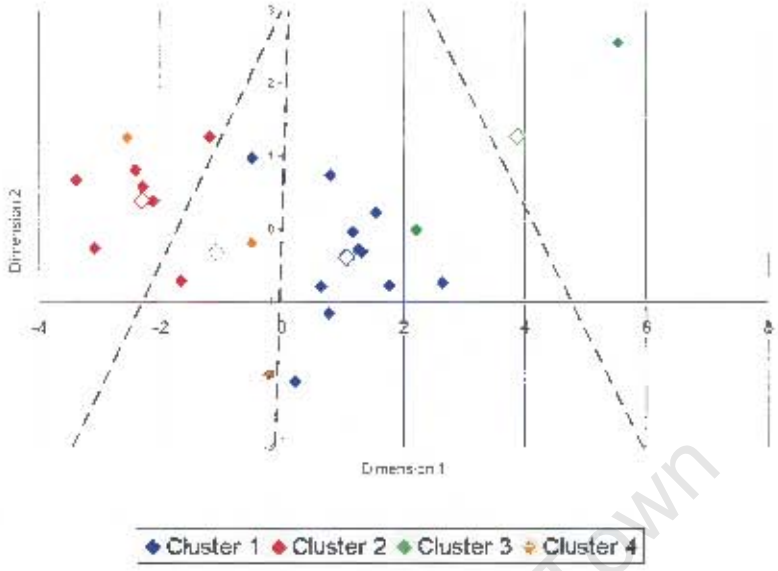



*Illustration 7.2: Municipalities against first two dimensions in four cluster GDA model containing two variables*

*Bitou* is once again clearly separated from every other municipality and the clusters are still not clearly separated. However the separation between all the clusters is significantly more distinct than in the previous graph. This is due to the fact that both dimensions include part of the explanation gained from the two variables, transformed in such as way as to best separate the clusters. Just using cut-off points on Dimension 1 it would be possible to separate Cluster 1, 2 and 3. This would only cause *Breede River / Winelands* to be misclassified as part of Cluster 3 or *Drakenstein* as part of Cluster 1, depending where the cut-off point between these two clusters is made. It can be seen clearly that at this point that if both dimensions are considered, there are four municipalities misclassified by the model, using the dotted cut-off lines as a marker of the midpoint between clusters.

The graph also indicates that *Cederberg* from Cluster 1 and the *City of Cape Town* from Cluster 4 have similar profiles for this model, as they are placed very close together at the bottom of the graph. This shows that there is a strong possibility at this point that either of these two municipalities might be misclassified by the model, as might *Witzenberg* from Cluster 1 which is placed at the top left and a significant distance away from the rest of its cluster. It is interesting to note that Beaufort West from Cluster 4, which was originally part of Cluster 1, is actually placed as the top-left observation belonging to Cluster 4. This is closest to the cluster centre of Cluster 2 and therefore implies that Beaufort West has more in common with Cluster 2; however it should be kept in mind that so far the model only includes two variables and therefore does not

University of Cape Town

necessarily represent a good generalisation of the relationship between municipalities.

Although the first three clusters can be relatively well defined by the model including only these two variables, it is obvious that even if Dimension 2 is considered, to separate out Cluster 4 from Clusters 1 and 2, other variables will be required. The final two variables that the model requires to be able to classify municipalities with 100% accuracy are *ElecIndMonetary* 2004 March and *CDSewerage* 2002 June. The graph of the four clusters of municipalities on the first two dimensions given by the model, together explaining 92.95% of the variation of the model, is shown in Illustration 7.3.

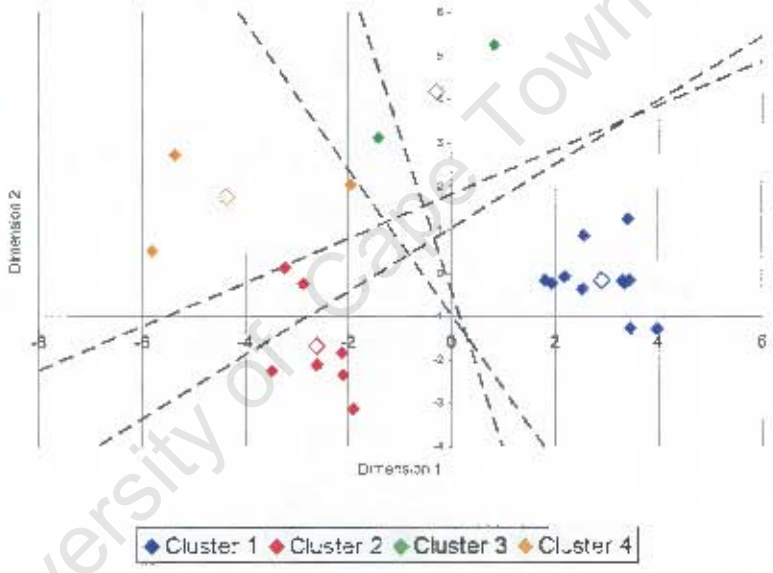



*Illustration 7.3: Municipalities against first two dimensions in four cluster GDA model containing four variables*

University of Cape Town

The four clusters are clearly very well separated by the model, with 100% accuracy being shown by no municipalities lying over their respective dotted cut-off points indicating the midpoint between clusters. Any further variables added to the model at this point would only further enhance this separation, becoming superfluous. Using a combination of the four variables, Dimension 1 clearly shows the separation between Cluster 1, Cluster 3 and Clusters 2 and 4. The model does however require Dimension 2 to be able to fully separate out the clusters as Clusters 2 and 4 overlap on Dimension 1. This does however confirm the fact that Dimension 1 explains 70.14% of the variation of the model, to which Dimension 2 adds 22.18%. The final dimension only adds 7.08% explanation and therefore can be considered negligible.

Concerning Cluster 4, both the *City of Cape Town* and *Beaufort West*, which were very close to Cluster 1 and 2 respectively in the model with only the first two variables included, are now clearly separated out to the left. *Knysna* remains between Clusters 1 and 2 on Dimension 1, but is now separated out using Dimension 2.

The only other notable feature on the graph is that *Mossel Bay* and *George* from Cluster 2 are placed at the top of their cluster, very close to Cluster 4 and therefore are likely to have much in common with the municipalities from Cluster 4, although not enough to have joined them in a separate cluster.

Now that it is clear that the models are capable of separating the clusters using combinations of significant variables, it is possible to analyse these variables in terms of attempting to ascertain whether there is any similarity between the models or any time effect evident. The actual profiles and defining characteristics of the clusters for all three models is given in Appendix VI, VII and VIII.

### 7.1.2 Time Effects

The above discussion concerned only the variables from the four cluster model; however to analyse the effect of the variables on the cluster profiles in terms of time, the significant variables from all three General Discriminant Analysis models will be considered. It is interesting to note that the variables that relate to financial data in particular all concern a municipality's consumer debtors; however there is no similar distinguishing feature in those variables that relate to service delivery.

Illustration 7.4 shows the cluster averages for all the significant variables, using four clusters. This particular choice is made as Cluster 4 is made up completely of municipalities originating from Cluster 1 and Cluster 3 consists of the two municipalities that had the least in common with Clusters 1 and 2 originally.

University of Cape Town

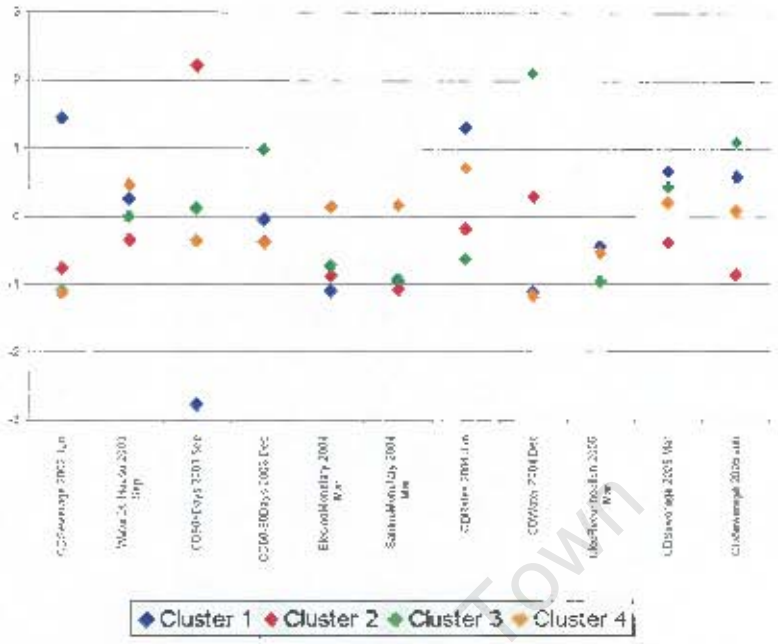



*Illustration 7.4: Four cluster averages across all significant GDA variables*

The most noticeable gap between clusters is for *CD90−Days* 2003 September, where Clusters 3 and 4 show moderate values, but Clusters 1 and 2 show the most extreme values of the entire graph. Other noticeable variables are *CDSewerage* 2002 June which clearly separates Cluster 1 from the other three clusters, and *CDWater* 2004 December which separates Cluster 3 from Cluster 2 from Clusters 1 and 4. It is not that surprising that this variable does not separate Clusters 1 and 4 as it was found significant for the three cluster model and the municipalities from Cluster 4 all come from Cluster 1, giving the model no reason to treat them differently. It is however surprising that the four cluster model chose *CDSewerage* 2002 June as most significant in being able to separate the four clusters as it appears at face value that there are various other variables that would be able to do this more clearly.

Variables that separate the four clusters relatively evenly are *WaterOthHouse* 2003 September, *CDRates* 2004 June and *CDSewerage* 2005 March and June. All of these, except *CDSewerage* 2005 March chosen in the four cluster model, were chosen as significant in the three cluster model. This is a little surprising as these same variables are able to separate Cluster 4 quite clearly from the other three clusters without simply treating it as part of Cluster 1, therefore indicating that they may contribute to the differences between Cluster 4 and the remaining municipalities in Cluster 1.

University of Cape Town

### 7.1.3 Variable Changes Over Time

It is interesting that across all the models only *CDSewerage* is seen more than once. It appears once in the three cluster model and twice in the four cluster model. This indicates that the information this particular variable offers changes significantly over time, as it appears in the model as significant for two distinct periods in time. It may therefore be interesting to consider this variable for the four clusters over the entire range of quarters as shown in Illustration 7.5.

The variation in the graph is largely due to actual reported values. The only municipalities that had all their data missing for this variable originally were *Matzikama* and *Cape Agulhas*, which both form part of Cluster 1 and had these values extrapolated from their relationship to the other municipalities.

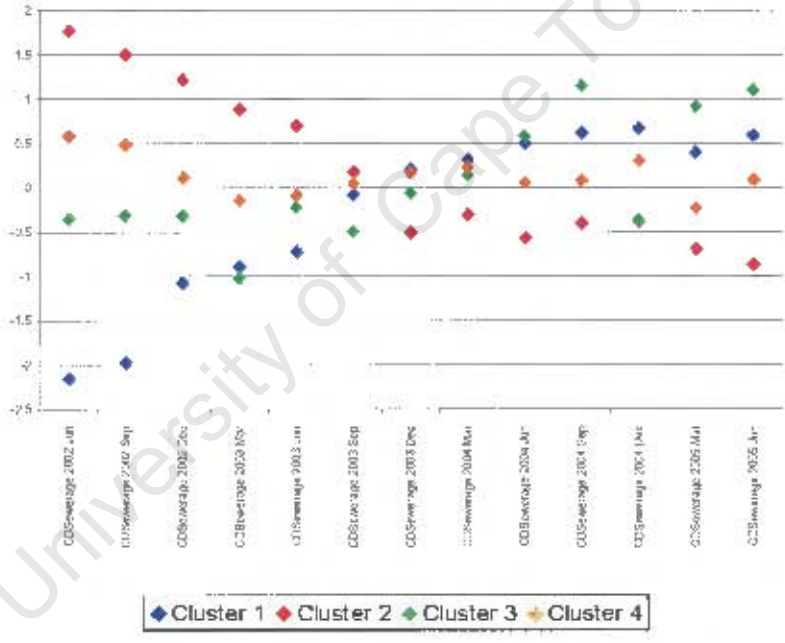



*Illustration 7.5: CDSewerage across all quarters for four clusters*

Periods of interest concern the dips made by Cluster 3 in March 2003 and December 2004, as it is only Cluster 3 that behaves in this fashion. The general trend of values shows that Cluster 2 begins with the highest values, but at the end of the period shows the lowest. A similar reversal of profiles is shown by Clusters 1 and 3, with Cluster 4 showing relatively moderate values across the entire period. This reversal of profiles across clusters appears to occur around September 2003, as has been shown for other variables in the previous chapter, again questioning what occurs during this period to cause this significant change.

This phenomenon was presented to members of the Department of Local Government and

University of Cape Town

Housing (DLGH) at the feedback meeting. Various hypotheses were discussed as to what environmental changes occurred during this particular period to cause this change in values, obviously being more difficult to ascertain due to working within historical context. Although these scenarios would need to be followed up, it is interesting to note that there is a strong possibility that these changes are not simply due to the data itself, but have been caused by external factors and are found to have a significant influence during the data analysis.

## 7.2 Application of Model

The following section shows how General Discriminant Analysis might be applied in terms of attempting to assign a 'new' municipality into one of the predetermined clusters. Obviously, this 'new' municipality is only one that has not been included in the analysis that builds the model. If there exists a municipality that did not contain enough data to be included in the analyses, but does contain data from the variables picked as significant in the analyses, it can be assigned to a cluster using the data from these particular variables without needing any of the data which it has not been able to report. This would allow for the possibility that when a significant portion of data are missing in a municipality, making it near impossible to identify its overall performance, it could be classified as belonging to a particular cluster and therefore most likely have a similar performance profile to the other municipalities in that cluster.

Although it would be ideal to be able to classify another municipality into its respective cluster using the discriminant functions, both *Oudtshoorn* or *Laingsburg*, that were originally removed from the analysis, lack full data for the relevant variables.

*Oudtshoorn* does however have most of the data available and at least one observation for every variable under consideration, if not in the correct time period, for all models. As the two cluster model has the fewest variables needed it will be used for the purposes of illustration and *Oudtshoorn* will be considered using the following standardised values calculated from the original data:

| | | | |
|---|---|---|---|
| CD90+Days | 2003 | September | 0.3378 |
| ElecReconnection | 2004 | March | -0.5262 |
| SandIndMonetary | 2005 | March | -1.2727 |

The discriminant function for the two cluster model is:

$$Z = 0.6865 - 1.1271x_{CD90+Days} - 0.7577x_{ElecReconnection} + 2.2186x_{SandIndMonetary}$$



University of Cape Town

Therefore the discriminant value for *Oudtshoorn* using the above values is:

$$Z_{\text{Oudtshoorn}} = -2.1192$$

The graph of when this particular value is plotted against the other observations from the two cluster general discriminant analysis, along with the two group centroids, is in Illustration 7.6.

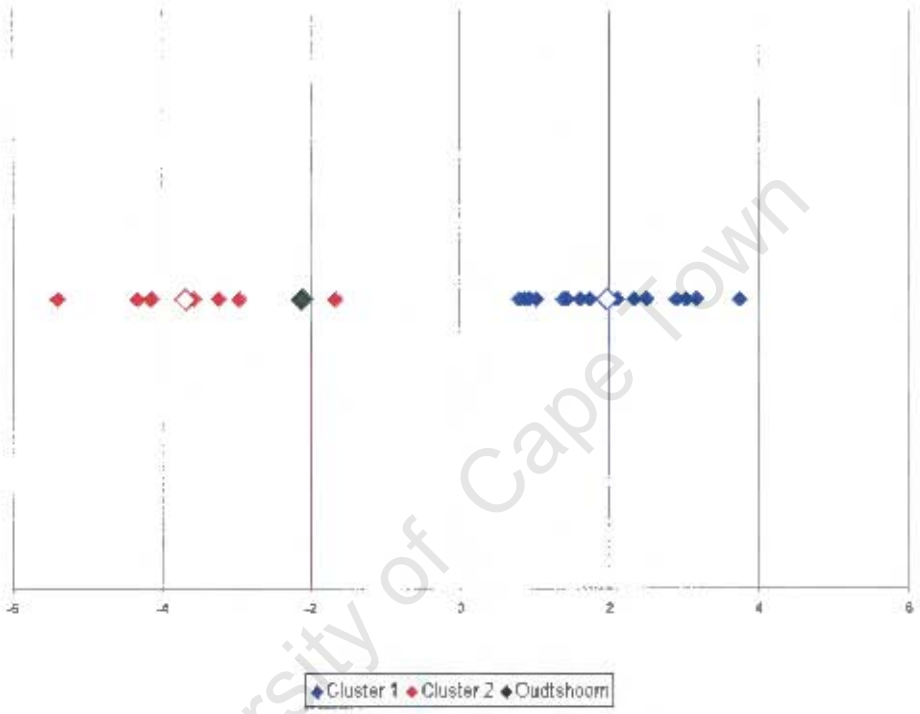



*Illustration 7.6: Oudtshoorn discriminant value on two cluster GDA model*

It is clear from the graph above that according to the two cluster discriminant model and the values that have been used to represent *Oudtshoorn*, the model would classify this municipality as part of Cluster 2. If considered purely on the distance between *Oudtshoorn* and the next closest municipality using these particular variables, it appears that *Oudtshoorn* has the most in common with *Saldanha*, both of which are the 'closest' municipalities of Cluster 2 to Cluster 1 indicating that they have the most in common with it, but not enough to actually form part of Cluster 1.

Further discussion concerning the 'true' classification of *Oudtshoorn* can be found in Chapter 9; however it is interesting to note that it is possible to gain some insight into how this municipality might be grouped with the other municipalities, even when missing significant portions of reported data.

University of Cape Town

# Chapter 8
# Classification Trees

Although the variables that successfully discriminate between clusters have now been identified, they may not all be necessary to be able to successfully classify municipalities into specific clusters. Along with this, although the variables may have been identified, there is no simple cut-off point for each variable that helps to determine into which cluster a municipality will fall. The following analysis uses Classification Trees to determine not only the best variables that are strictly necessary to be able to classify municipalities into clusters when using all observations, but also what levels of each variable lead to this classification.

Classification Trees can be represented graphically, or by using a set of decision rules, both of which allow other 'new' municipalities to be classified into one of the predefined groups. As with the consideration of overfitting in General Discriminant Analysis, it is noted that there are few cases to be classified and many more variables available for the choice of splitting. However; as there are at most 4 classification classes and the tree will therefore never be particularly large, a tree with only observations from one class in each terminal node can be considered, leading to a 100% correct classification rate. Although this particular technique has been grossly simplified and may not be particularly useful generalising to new data, it is meant as an illustration. It is felt that the lack of reliability of the data justifies the use of the entire tree, especially as it has been grown to classify clusters that were determined from the structure inherent in the data itself. This decision implies that there will be no attempt made to prune the tree to its 'ideal' size.

In spite of these limitations, an attempt will be made to classify *Oudtshoorn* in each analysis as this municipality was not originally used due to lack of data, but has reported data available for all 13 periods for the specific variables identified in the trees.

## 8.1 Discussion

The statistical results and discussion of all three Classification Tree analyses are given in Appendix

University of Cape Town

X, XI and XII. To illustrate the ease to which the decision rules are represented graphically, the tree classifying all municipalities into two clusters is shown in Illustration 8.1.

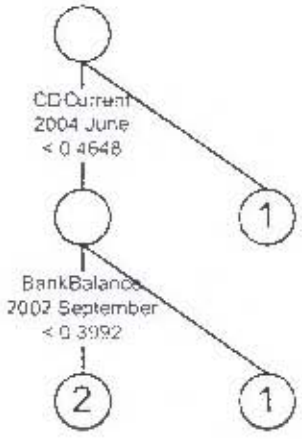



*Illustration 8.1: Classification tree for 2 clusters*

It may not be clear as to why *CDCurrent* 2004 June is chosen as the variable which 'best' splits the municipalities into the two clusters. Briefly considering this particular variable for all four quarters of 2004 (Appendix XIII) it can be seen that although the two clusters clearly overlap in all cases, it is June 2004 which shows the best split of the two clusters as seen in Illustration 8.2.

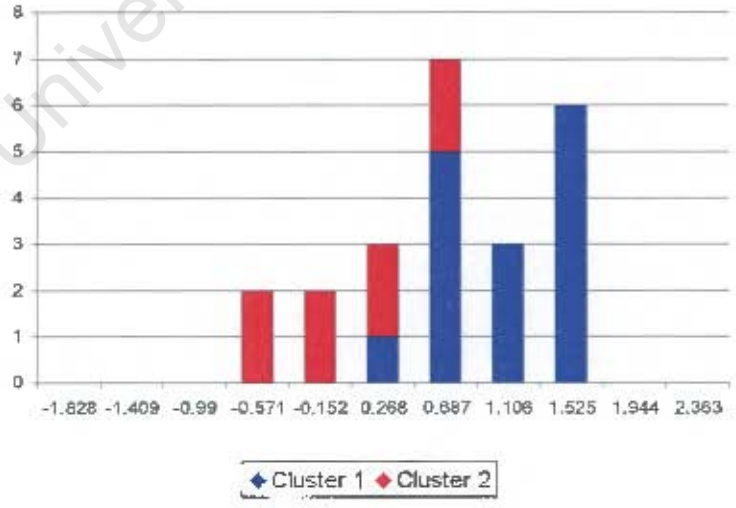



*Illustration 8.2: Municipal values of CDCurrent 2004 June for 2 clusters*

The split is however still not completely clear and in fact *Bitou* initially splits with Cluster 2 in the

University of Cape Town

tree and requires a further split for all the municipalities to be able to be classified correctly. The tree itself is deceiving in this regard as the structure implies that the number of municipalities in each terminal node is evenly spread. A clear graphical illustration of exactly how each of the municipalities have been classified is in Illustration 8.3.

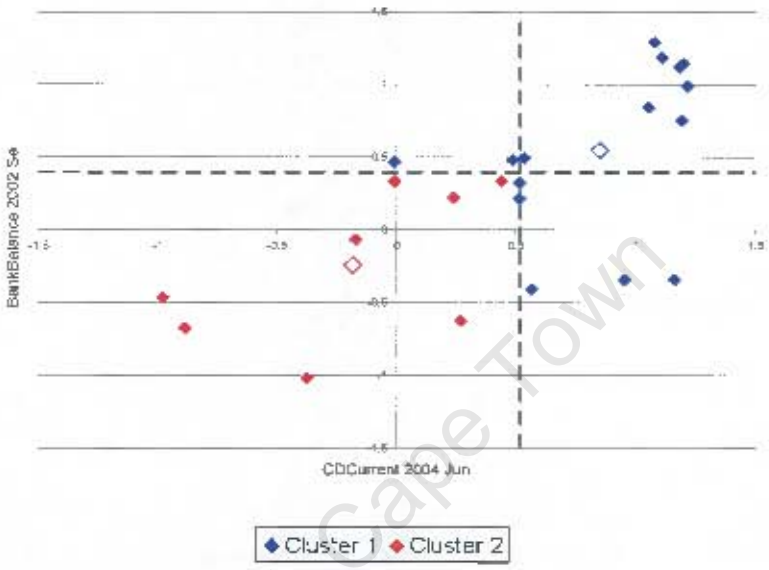



*Illustration 8.3: Municipalities plotted on significant variables for 2 clusters*

The initial split on *CDCurrent* 2004 June is shown on the x-axis where the cut-off point almost perfectly separates the two clusters, leaving only *Bitou* from Cluster 1 with Cluster 2. *BankBalance* 2003 March is then chosen by the model to split *Bitou* from Cluster 2 as shown with the cut-off point on the y-axis.

It is much more clearly shown from this graph how unevenly the spread of municipalities is in the tree. In fact, for this particular tree it would be more useful as a generalisation if only the first split was used and less than 100% accuracy of classification was accepted.

When considering the tree for three clusters (Appendix XI), it is once again *CDCurrent* 2004 June which is chosen for the initial split separating Cluster 1 from the other two clusters. This graph showing the values from all municipalities for this particular variable (Appendix XIII) again clearly shows this split.

However, when splitting Cluster 2 and 3 from each other at the next step, the model again chooses



University of Cape Town

*BankBalance* at this level, but from the period 2003 March rather than 2002 September. Illustration 8.4 clearly shows how this variable would be useless in being able to split Cluster 1 from the other two clusters; however if Cluster 1 had already been separated from them, then *BankBalance* 2003 March easily separates Cluster 2 and 3.

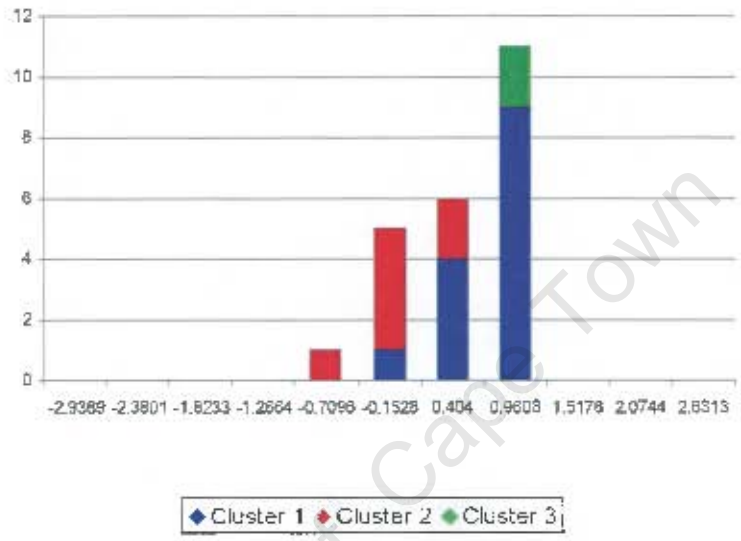



*Illustration 8.4: Municipal values of BankBalance 2003 March for 3 clusters*

It is clear that the reason this particular variable was not chosen in the two cluster tree is that *Bitou* is joined by *Drakenstein* in Cluster 3 and therefore the two cluster model would then have split *Drakenstein* along with *Bitou* into a terminal node and it would therefore be misclassified as belonging to Cluster 1. The tree would require a further split to completely separate the clusters.

Finally, when considering the four cluster tree (Appendix XII) it is *CDRates* 2004 June that is chosen for the initial split, unlike the other two trees that both used *CDCurrent* 2004 for this first split. It is interesting to note however that both of these variables are from the same period. As all of the municipalities in Cluster 4 originate from Cluster 1 it is clear that this change in variable choice is necessary. The difference between the split of the four clusters using *CDCurrent* (Appendix XIII) and that of using *CDRates* is significant, with *CDRates* showing a far better split of clusters as seen in Illustration 8.5.

University of Cape Town

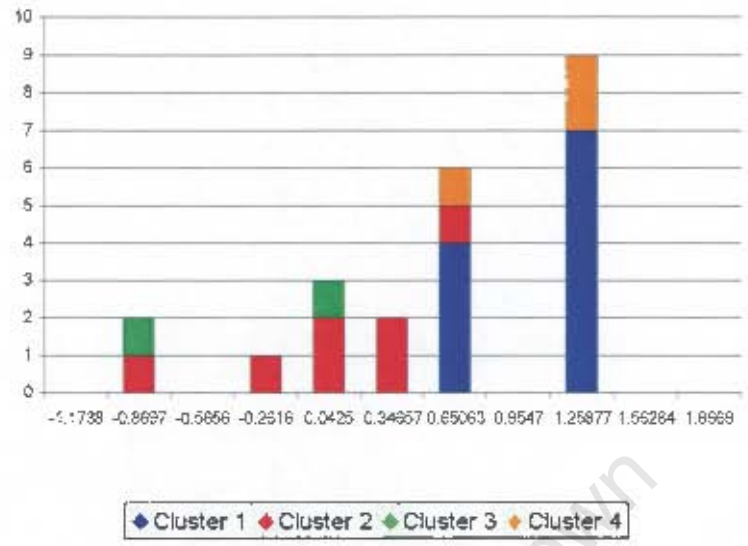



*Illustration 8.5: Municipal values of CDRates 2004 June for 4 clusters*

This slight change indicates that when defining the clusters, there has been a particular focus on consumer debtors for the period June 2004. It is only when separating *Knysa, Beaufort West* and the *City of Cape Town* from the other municipalities that the focus shifts slightly from current consumer debtors to consumer debtors relating specifically to rates.

As seen with the two cluster tree, in the four cluster tree there is one municipality that splits incorrectly initially and therefore requires a further split to be classified into the correct cluster. In the two cluster tree *Bitou* splits with Cluster 2 which was not surprising; however in the four cluster analysis *Knysna* initially splits with Cluster 1 instead of with the other 3 clusters. This implies that although *Knysna* is part of Cluster 4, and the three municipalities in Cluster 4 all originate from Cluster 1, *Knysna* has the most in common with the rest of Cluster 1 when compared to the other two municipalities in Cluster 4.

To clearly separate Cluster 4 from the other three clusters, *BankBalance* is chosen by the model. In separating *Knysna* of Cluster 4 from Cluster 1 the period of June 2003 is used; however when separating the rest of Cluster 4 from Cluster 3, the period June 2002 is considered. This indicates that *BankBalance* is an important variable when defining the clusters; however is it useful across more than one period. It is interesting to note that for this particular tree, every split uses variables from June, further indicating some importance from this quarter in each year.

University of Cape Town

A final note is that not one of the service delivery type variables are chosen by any of the three models to split the clusters. Although only a few of this type of variable were identified in creating the data-driven clusters initially, the fact that none are identified as significant in this particular analysis indicates that the patterns inherent in their data is less clear than for the financial type variables.

## 8.2 Application of Model

Although there are no 'new' municipalities to classify using the classification trees, both *Laingsburg* and *Oudtshoorn* were removed from the analysis when cleaning the data as they were considered to have too much missing data to be analysed reliably. Unfortunately, *Laingsburg* continues not to have the data requirements needed to be able to classify it in this analysis: however *Oudtshoorn* has the required reported data across all 13 periods for each of the variables that have been found significant in the Classification Trees. The relevant data, once standardised by the number of households in the municipality, as well as statistically, is shown in Table 8.1.

| | | | |
|---|---|---|---|
| BankBalance | 2002 | June | -1.49246 |
| BankBalance | 2002 | September | -1.49246 |
| BankBalance | 2003 | March | 1.30249 |
| BankBalance | 2003 | June | 1.31136 |
| CDCurrent | 2004 | June | -1.07724 |
| CDRates | 2004 | June | -1.28437 |
| CDRefuseRem | 2002 | June | 0.54622 |

*Table 8.1: Relevant standardised values for Oudtshoorn*

It is possible to simply run *Oudtshoorn* through the decision rules as given in each analysis (Appendix X, XI and XII), but perhaps it is more interesting to consider graphically, giving a better idea of how *Oudtshoorn* in placed compared to the other municipalities. The placement of *Oudtshoorn* for the two and three cluster trees, with the relevant variables and cut-off points is shown in Illustrations 8.6 and 8.7.

University of Cape Town

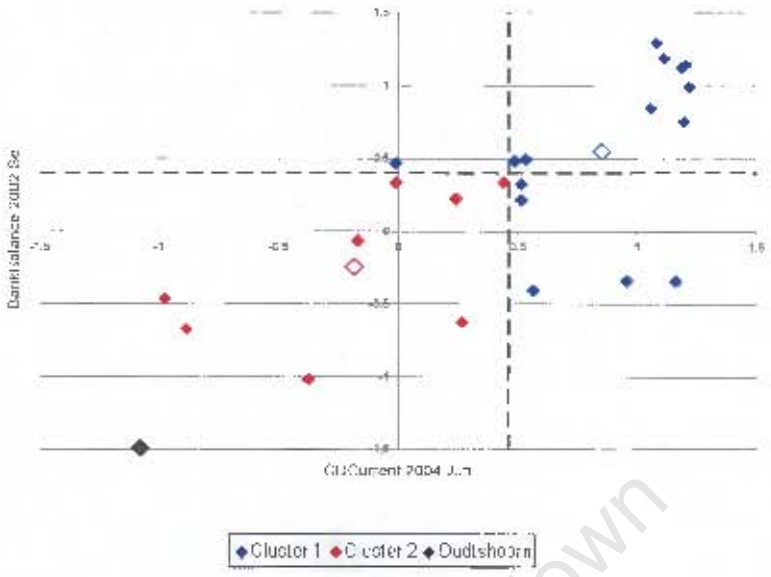



*Illustration 8.6: Oudtshoorn plotted on significant variables for 2 clusters*

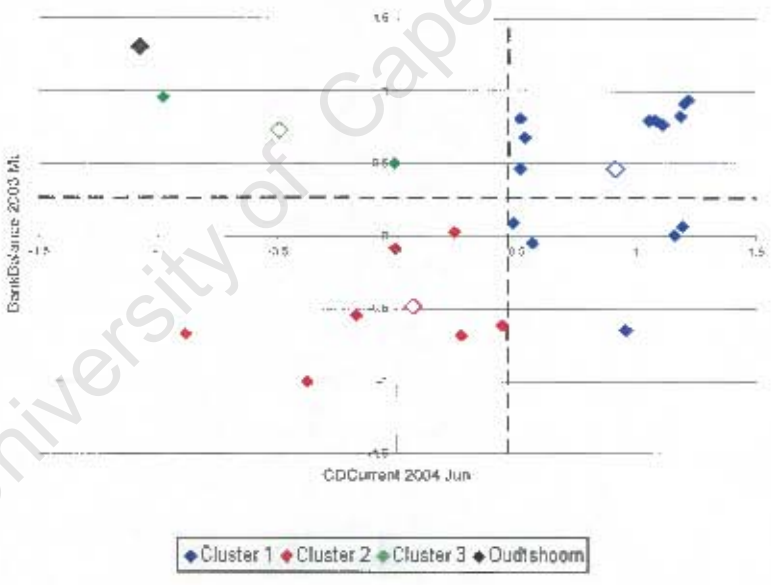



*Illustration 8.7: Oudtshoorn plotted on significant variables for 3 clusters*

The two graphs above show the clustering and decision rules of the 2 cluster and 3 cluster analysis. It is clear in both cases, where *CDCurrent* 2004 June is the most significant splitting variable, that *Oudtshoorn* is definitely not classified as belonging to Cluster 1. In the 2 cluster analysis, it is then classified as belonging to Cluster 2, but once *Bitou* and *Drakenstein* split off in the 3 cluster analysis, *Oudtshoorn* is classified as having more in common with these two municipalities and is classified as Cluster 3.

University of Cape Town

The graphical classification for the 4 cluster tree (Appendix XII) is shown in Illustration 8.8

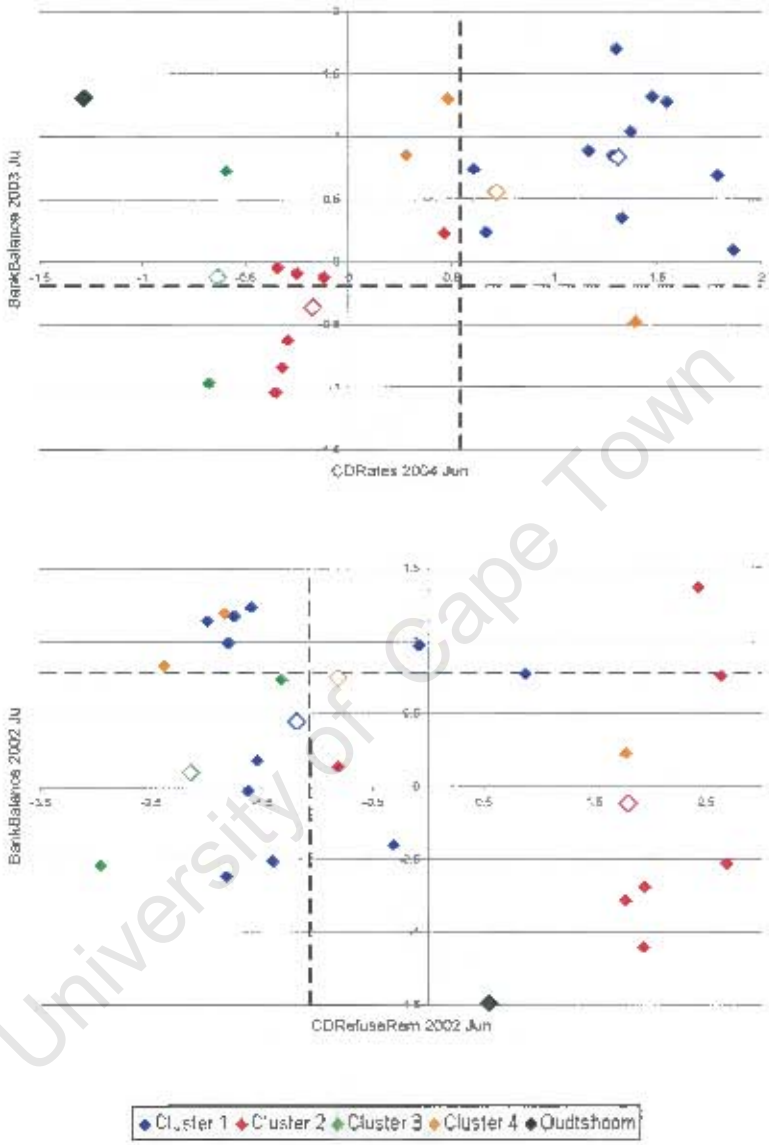



*Illustration 8.8: Oudtshoorn plotted on significant variables for 4 clusters*

In the first graph it is clear that *Oudtshoorn* does not split down the main right branch of the tree with Cluster 1 and therefore it is necessary to consider the second graph from the main left branch. Considering one level further, *CDRefuseRem* 2002 June classifies *Oudtshoorn* as belonging to Cluster 2. This indicates that there is some discrepancy overall as to whether *Oudtshoorn* has more in common with *Bitou* and *Drakenstein* or the rest of the municipalities that make up Cluster 2 in all of the analyses. As there are more municipalities in Cluster 2 and *Oudtshoorn* is classified as belonging

University of Cape Town

to this cluster twice out of the three analyses, this is likely to be the most appropriate classification when no other information is available. However, if possible, more analyses should be done to better determine its profile, although it appears it is reasonable to conclude that *Oudtshoorn* does not belong in Cluster 1 for any of the analyses. Again, further discussion as to the 'true' classification of *Oudtshoorn* can be found in Chapter 9.

University of Cape Town

# Chapter 9
# Interfacing with the DLGH

Although the previous chapters have been a constructive illustration of potential techniques that can be used to help the Department of Local Government and Housing (DLGH) gain more knowledge from the large amounts of data that is collected on a regular basis, as yet there has been no discussion on feedback from the DLGH itself. The following chapter discusses the feedback session with the DLGH and the analysis, using the techniques that have been illustrated previously, on further data gained from a group exercise with some of the participants.

University of Cape Town

## 9.1 DLGH Meeting

A meeting was held with the DLGH of the Western Cape, at their offices in Cape Town, during the afternoon of 16 May 2007.

The two main purposes of this meeting were to present the results of the analyses that had been performed on the data obtained from their PVQ reports, as well as allowing members to partake in a group exercise in order to gain a better understanding of the subjective opinions as to the performance of municipalities during the period under analysis.

Ten members of the DLGH attended the feedback meeting (Appendix XV), of which only five partook in the group exercise (Appendix XV), indicating that the opinions expressed within this chapter are based on a small sample size and may not be representative of the DLGH as a whole.

### 9.1.1 Feedback Presentation

The first part of the meeting involved a presentation of some of the results from the analyses that had been completed so far (Appendix XIV). Firstly, the data used was explained in terms of its origin and what transformations had occurred to create a complete data set that could be used for the analyses chosen. It was stressed at this point that the quality of the data was very poor and the results may not be realistic, but that this was also a point of interest in terms of the feedback that

the members partaking in the meeting might be able to give. Along with this, the members of the meeting were reminded that the data under analysis was for the period June 2002 until June 2005 and it was important that discussions and feedback was made with this constraint in mind.

Each analysis technique was presented with a brief explanation as to what it might be used for within the local government context, along with an example of results taken from the relevant analyses. As the majority of people in the meeting had never seen the particular techniques used and more complexity would detract from the usefulness of these techniques, the examples given in the presentation were chosen specifically from the two cluster results. These results were explained in terms of the data and discussed by the group within the local government and municipal context.

The presentation slide showing the results of the Cluster Analysis using the data objectively, both in terms of listing the members of each cluster and the geographical representation of these clusters, for the two cluster analysis is shown in Illustration 9.1

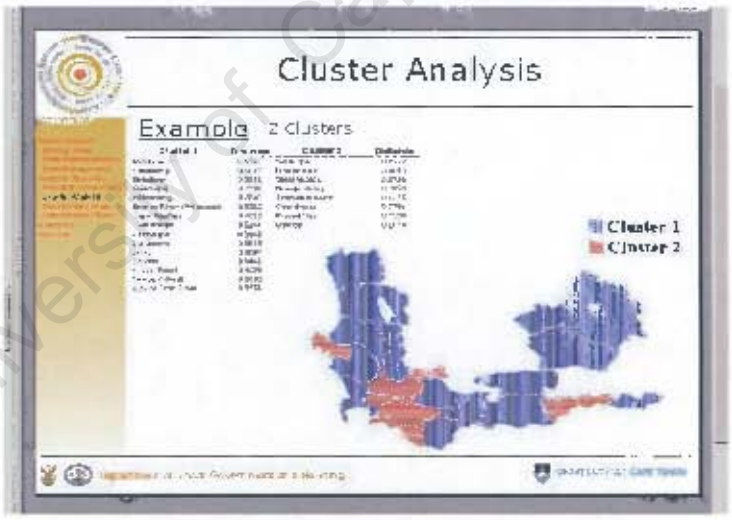



*Illustration 9.1: Presentation slide of Cluster Analysis results*

The reaction from the members of the meeting concerning these particular results were that although the results were based purely structures inherent in the data itself, Cluster 2 generally indicated municipalities that were performing well over the period and Cluster 1 those that didn't show good performance. The only exception to this appeared to be *Theewaterskloof* where a few members felt that it did not belong with the other municipalities in Cluster 2. In contrast to this, an opinion was expressed that during this period in particular, *Theewaterskloof* was actually

University of Cape Town

outperforming expectations and therefore if analysed on relative circumstances, might actually belong with those municipalities that generally performed well.

During the general discussion following the presentation, it was mentioned that of all available techniques, the most useful was probably Cluster Analysis. This is not surprising as Cluster Analysis seems the easiest technique to understand, along with allowing for the clearest illustrations as to the results of the technique.

It was also mentioned that there was a need for time trend analysis in order to be able to predict in advance what might occur within a municipality. Due to this comment the analysis of results contains more focus on interpreting results within the context of time; however not with any focus on being able to predict future results. This is due to the nature of the data, being far too large and unreliable.

Throughout the presentation it was stressed that these particular techniques may be available in various software packages and therefore easily utilised by an unqualified person. However, it is vital that a qualified practitioner is involved in the entire process and is able to present the results of the analyses such that everyone involved understands how the data has been manipulated, along with the limitations of the techniques used. Along with this it was discussed how the results cannot simply be presented, but require discussion with domain experts to be able to be interpreted within the given context.

There was a general view that more analysis of the data would be useful, along with the understanding that for this to occur skilled personnel would need to be included in the entire process. A further point was made that in terms of historical data collection there has been a tendency for the entire system to be overhauled on a semi-regular basis and perhaps it would be more successful to first focus on allowing smaller sections to update as needed, focusing on the most important requirements, without a complete change. Once stable these could then be integrated into an interdependent system. This would also allow for data to be more consistent and help in time trend analysis, along with involving all sections in the interpretation and understanding of data analysis.

### 9.1.2 Group Exercise

Due to time and complexity constraints, the group exercise covered only the scenario of three clusters of municipalities. This particular choice is justified by the fact that the department itself

University of Cape Town

has chosen to analyse municipalities in terms of having good, average or poor performance.

Participating members were given cards showing the names of the 25 local municipalities in the Western Cape and asked to assign them to the groups 'good', 'average' or 'bad' in terms of how they, subjectively, thought these municipalities had performed during the period June 2002 to June 2005.

*Illustration 9.2: Group exercise participants ranking municipalities*

Although it was not expressly requested, there was minimal discussion as to the placement of municipalities during the exercise.

As mentioned, there were fairly significant time constraints during the exercise as shown by the fact that only five members of the original ten that attended the meeting could stay to participate in the exercise. Unfortunately it was not possible to complete the exercise earlier as it was important that everyone involved understood what was needed as explained during the presentation.

It would have been ideal to be able to consult the participants as to why various choices were made; however again due to time, it was not possible to consolidate the results at that point and discuss them. The grouping of municipalities was done at a later stage using rankings and it would have been interesting to compare to how the participating members would have eventually chosen to group municipalities within the context of discussion and further interaction.

University of Cape Town

*Illustration 9.3: Group exercise participants ranking municipalities*

The results showed a fairly even split between municipal placements in the 'good' and 'bad' categories with 35 and 38 placements respectively. There were 52 placements made in the 'average' category which either indicates indecisiveness on the part of the group exercise participants, or that these municipalities are neither noticeably 'good' or 'bad' in terms of performance.


*Illustration 9.4: Poster of municipalities ranked 'average'*

University of Cape Town

Obviously with five members participating in the exercise it is unlikely that they all agree as to where a particular municipality should be placed. To be able to use this information in terms of treating a municipality as belonging to a particular cluster and using this in an analysis, a consensus must be found. The final assignment of municipalities was made by ranking 'good', 'average' and 'bad' as 1, 2 and 3 and comparing the rounded mean, median and mode of the 5 opinions. Surprisingly, there were very few occasions where these three measures of location disagreed.

*Beaufort West*, *Bergrivier*, *Bitou* and *Drakenstein* were each placed at least once in each category; however *Bergrivier* had the majority of members assessing the municipality as 'bad' and *Bitou* had the majority of members assessing the municipality as 'good' and therefore these clusters were chosen. This indicates that *Beaufort West* and *Drakenstein* were the most contested clusters, with their final allocation being made as 'average'.

All other municipalities had a clear bias towards either good/average or average/bad and as the three measures of location for each did not disagree they were assigned to that respective cluster. Interestingly, *Kannaland* was undisputed as being a 'bad' performing municipality and *Overstrand* was undisputed as being a 'good' performing municipality.

The final allocations to clusters is shown in Table 9.1, in the order in which it is felt that the consensus is strongest that a municipality belongs to that particular cluster. This is further indicated by the variability of consensus (measured simply by taking the statistical standard deviation of the five opinions given for each municipality) given in brackets beside each municipality below:

*Table 9.1: Assignment of municipalities to clusters by opinion*

| Good | | Average | | Bad | |
|---|---|---|---|---|---|
| Overstrand | (0.00) | Knysna | (0.40) | Kannaland | (0.00) |
| Breede Valley | (0.40) | Mossel Bay | (0.40) | Cederberg | (0.40) |
| Stellenbosch | (0.40) | Langeberg | (0.40) | Cape Agulhas | (0.49) |
| Swartland | (0.40) | Saldanha | (0.40) | Witzenberg | (0.49) |
| Breede River / Winelands | (0.48) | Theewaterskloof | (0.40) | Bergrivier | (0.80) |
| City of Cape Town | (0.49) | Matzikama | (0.49) | | |
| George | (0.49) | Prince Albert | (0.49) | | |
| Bitou | (0.80) | Swellendam | (0.49) | | |
| | | Beaufort West | (0.75) | | |
| | | Drakenstein | (0.89) | | |

To gather an idea as to how this placement differs to that of the data-defined clusters, the original assignment to clusters is indicated in terms of colour. The data-defined Cluster 1 is shown in blue, Cluster 2 in red and Cluster 3 in green. A simple alphabetical list of cluster allocation, both



University of Cape Town

for data-defined and opinion-defined clusters, along with colours indicating the other classification is shown in Appendix XVI.

It is clear that the data-defined and opinion-defined clusters differ significantly. There was no reason to assume that Cluster 1 of the data-defined clusters would be the best performing cluster of municipalities and so forth; however it is interesting that the data-defined clusters are spread throughout the opinion-defined clusters. It is noteworthy however that the 'bad' performance cluster is made up entirely of municipalities from Cluster 1, clearly indicating that Clusters 2 and 3 have a good/average opinion bias even though the performance of Cluster 1 is unsure.

There was also the opinion given during the meeting that except for *Theewaterskloof* the data-defined Cluster 2 was made up of municipalities that had been performing well for the period. This follows through in the table above where the majority of Cluster 2 forms part of the 'good' cluster; however it does raise the query as to why *Mossel Bay* and *Saldanha* are only ranked 'average' if previously they were seen as part of the section of municipalities that was performing well.

It is interesting to note that both *Bitou* and *Drakenstein* from the data-defined Cluster 3 appear at the bottom of the lists for 'good' and 'average' showing that they are the most contested as belonging to those clusters. It was these two clusters that split from Cluster 1 and Cluster 2 respectively in the two cluster analysis to form Cluster 3 in the three cluster analysis. Once again they indicate that they have the least in common with the other municipalities in the cluster; however this time due to an intuitive feel as to how to municipalities group, not to the data itself.

Neither *Oudtshoorn* nor *Laingsburg* were included in the original analysis due to lack of data; however they were included as part of the group exercise. Unfortunately, knowing how people involved in the system might assess these two municipalities is still not enough information to be able to include them in the analyses. However, what this extra information does add is the knowledge of to which cluster these municipalities *would* be assigned. This allows for the comparison of this, to how the models developed from the General Discriminant Analysis and Classification Trees would have assigned these municipalities given enough data to do so. For both municipalities, the opinion gathered from the group exercise is that they are 'bad' performing municipalities for the period June 2002 to June 2005.

To aid the readability of interpretations in further analyses, Cluster 1 will be considered to be the cluster that indicated 'good' performance, Cluster 2 to be 'average' performance and Cluster 3 'bad' performance.

University of Cape Town

## 9.2 Analysis

The first analysis originally carried out on the data, without any opinion taken from the municipal context, was a Cluster Analysis. This unsupervised method determined how the data naturally groups the municipalities into 2, 3 or 4 clusters. Through the group exercise with the participating members of the DLGH, three clusters have been defined and analyses may now be carried out on these clusters in terms of defining the cluster profiles and using General Discriminant Analysis and Classification Trees to assess which variables, and what levels of these variables, now become important in defining these given groupings. These results can then be compared to what was seen when the data-defined clusters were used.

### 9.2.1 Cluster Analysis

When attempting to profile clusters, it is simplest to do in terms of the difference between the three clusters. Illustration 9.5 shows all variables that indicated a significant difference between the cluster averages at a 1% level of significance. These are not shown in order of significance, but rather are ordered by time to ascertain whether there might be any noticeable patterns throughout the period.

University of Cape Town

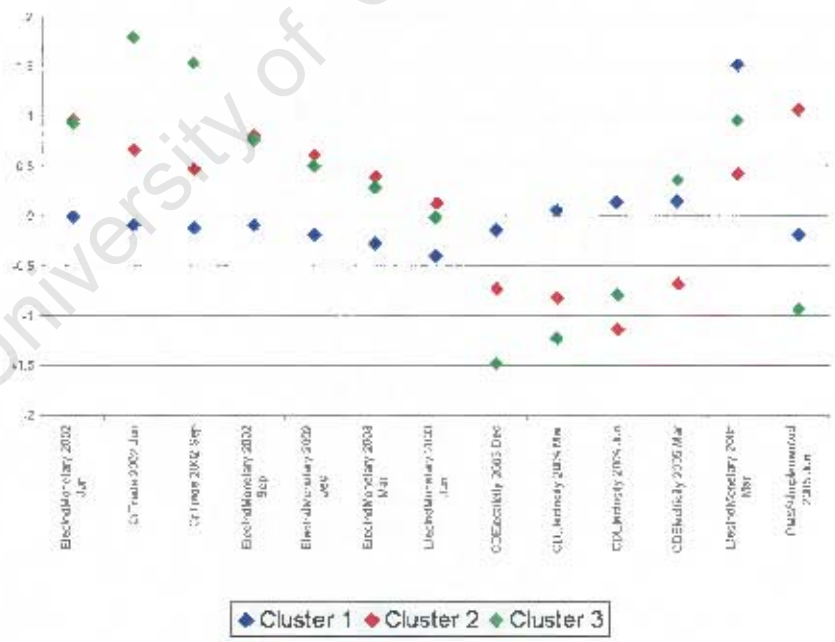



*Illustration 9.5: Opinion-defined cluster averages across significant variables for Cluster Analysis*

It is interesting to note that except for *ElecIndMonetary* 2005 March, Cluster 1 shows relatively moderate values all wavering about the average. Along with this, Clusters 2 and 3 appear to have

very similar values for most of the variables, differing noticeably for *CDElectricity* 2005 March and *PMS%Implemented* 2005 June.

Also interesting is that Cluster 2 and 3 are above Cluster 1 for the earlier periods, but from about June 2003 tend below Cluster 1. The only exception to this is at the end of the overall period, in 2005, where the relationship between the clusters vary.

The relationships between the clusters is what would have been expected looking at those formed only by the data itself, where the profile of Cluster 1 versus Cluster 2 swapped to the exact opposite profiles over time, changing around the period June-September 2003. A similar swap occurs in the opinion-defined clusters, but it is not a complete opposite between Cluster 1 and Clusters 2 and 3, but rather that Cluster 1 remains rather moderate and Clusters 2 and 3 vary widely as time progresses. The pattern prior to June-September 2003 and after this period does however change significantly, indicating that it is not only the data-defined clusters that indicates that this period may have shown some significant changes to cause variation in the reported values from the municipalities.

Electricity plays an important role as *ElecIndMonetary* appears among the 13 significant variables 6 times, with *CDElectricity* appearing 4 times. It may be interesting to consider the relationship between the three clusters for these two variables in particular across all quarters. The graph showing the values for the three clusters of *ElecIndMonetary* across all quarters is shown in Illustration 9.6.

This graph appears rather unusual as there is a large jump in all values at December 2004, which immediately queries the reliability of the data. There is a need to understand if any of the values were originally missing and if this may cause this unusual result or not. Only *Overstrand* of Cluster 1 and *Cape Agulhas* of Cluster 3 originally had every value missing for this variable, with these values being extrapolated from their relationship to the other municipalities. However, for all municipalities any value prior December 2003 was originally missing as it was not available in the printed reports, but only in the electronic data, explaining the smooth curves prior to this point.

University of Cape Town

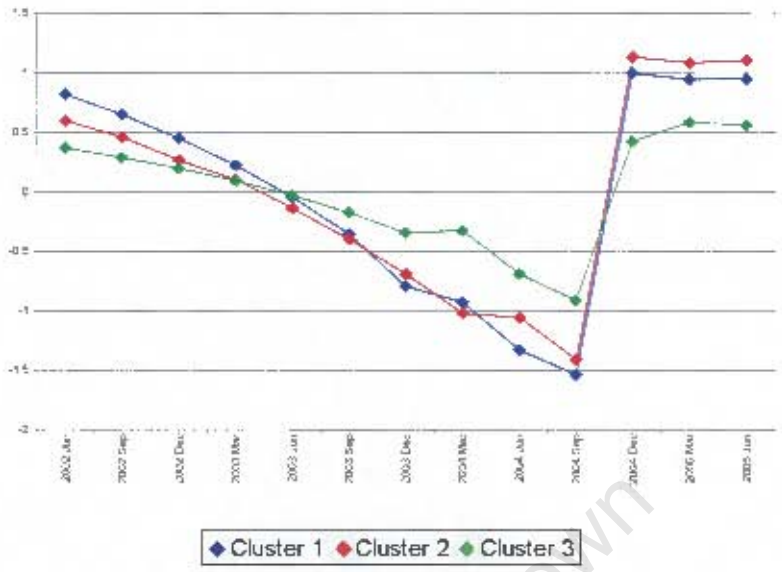



*Illustration 9.6: ElecIndMonetary for Opinion-defined cluster averages across all quarters*

There are various reasons to consider this variable unreliable; however it is interesting to note that the large jump in values at December 2004 is not due to manipulation of the original data, but is rather part of the values actually reported by municipalities. This jump in values is consistent across all municipalities, raising questions as to what may have occurred in this period to cause the reported values to change so drastically. However, the relationship between the clusters is still maintained and this changes over time even if everything prior to December 2003 is ignored.

It might be considered that having high values for the monetary value of the provision of free basic electricity indicates good municipal performance as this might show more electricity provided than if this value was low. If this is true then it is not surprising to see Cluster 3 with some of the lowest values compared to the other two clusters in the later quarters as it is considered to be a cluster of 'poor' performing municipalities. It does not however explain why Cluster 1, containing 'good' performers, has lower values than that of Cluster 2's 'average' performers.

It is uncertain as to whether these reported values are a true reflection of the situation at the time, or whether changes within the system caused the numbers to be reported differently, thereby explaining the distinct jump. Either way, it is certain that this particular variable should be treated with some concern during interpretation, although the relationship between the clusters appears to be maintained.

University of Cape Town

When considering *CDElectricity* a different pattern is seen in the data. The graph showing the average values for the three clusters of this particular variable across all quarters is shown in Illustration 9.7.

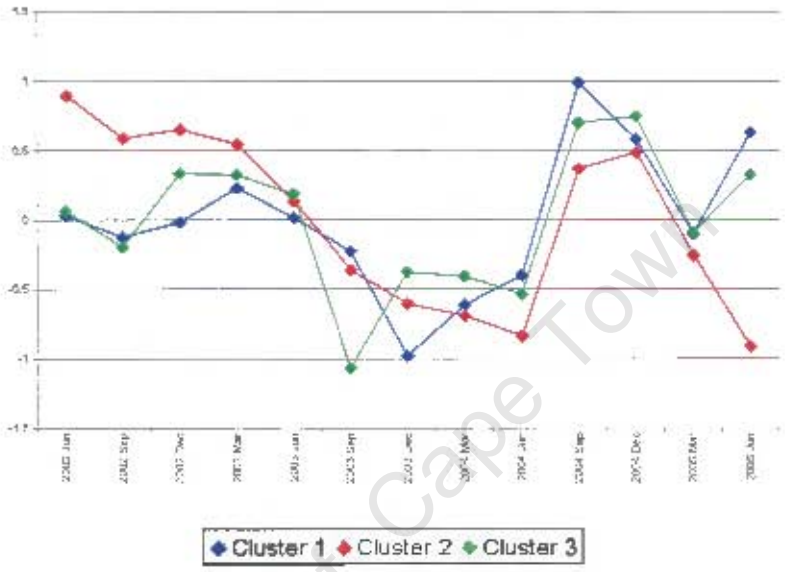



*Illustration 9.7: CDElectricity for Opinion-defined cluster averages across all quarters*

For this particular variable *Matzikama* and *Cape Agulhas* had all the original values missing, which were extrapolated from their relationship to the other municipalities; however for all other municipalities all values were originally there or one or two needed to be interpolated.

These values are far more varied over time; however it is noted that there are sudden drops in values for Cluster 3 in September 2003 and Cluster 1 in December 2003, with a more gradual decline for Cluster 2, which bottoms out just before all the clusters dramatically increase in September 2004. This is consistent with the concern that there was a significant change around this time that affected the profile of the three clusters. Profiles do show a general change over time with Cluster 2 starting with the highest values, but ending with the lowest and Cluster 3 having a general trend of remaining in between the two, and Cluster 1 starting with some of the lowest values, but ending with some of the highest.

It is uncertain whether having high or low debtors would indicate good or bad service delivery for a municipality. The data appears to indicate that the 'best' and 'worst' performing

University of Cape Town

municipalities have very similar profiles for this variable over time and it is the 'average' performing municipalities that show greater variation. It would be interesting to ascertain if there was any major event during September-December 2003 which would explain the sudden dip in values and perhaps aid in understanding the relationship between current debtors and service delivery.

### 9.2.2 General Discriminant Analysis

Now that the cluster profiles have been discussed, it is possible to determine which combination of independent variables may help to better explain the differences between clusters.

The General Discriminant Analysis model using the opinion-defined clusters required five variables to be able to discriminate between cluster with 100% accuracy. This is more than any of the previously defined models, which required at most four variables. However as these clusters are not defined from the data itself, the requirement of extra variables is not surprising. In fact, it is a little surprising that only five variables are required, but this may be due to the fact that there are simply so many variables available to be chosen.

The first two variables chosen by the model as most significant in being able to defined the differences between clusters are *ElecIndMonetary* 2002 December and *PMS%Implemented* 2003 March. The clusters of municipalities simply plotted against these two variables is shown in Illustration 9.8.

It is very clear that neither of these variables is particularly able to separate the three clusters. *ElecIndMonetary* appears able to isolate *Cape Agulhas*, *Swellendam* and *Swartland* to the left of the other municipalities; however they belong to Clusters 3, 2 and 1 respectively. *PMS %Implemented* is only able to isolate *Witzenberg*, from Cluster 3, between two clusters of mixed municipalities. The cluster centroids do appear evenly separated, but not particularly far apart.

University of Cape Town

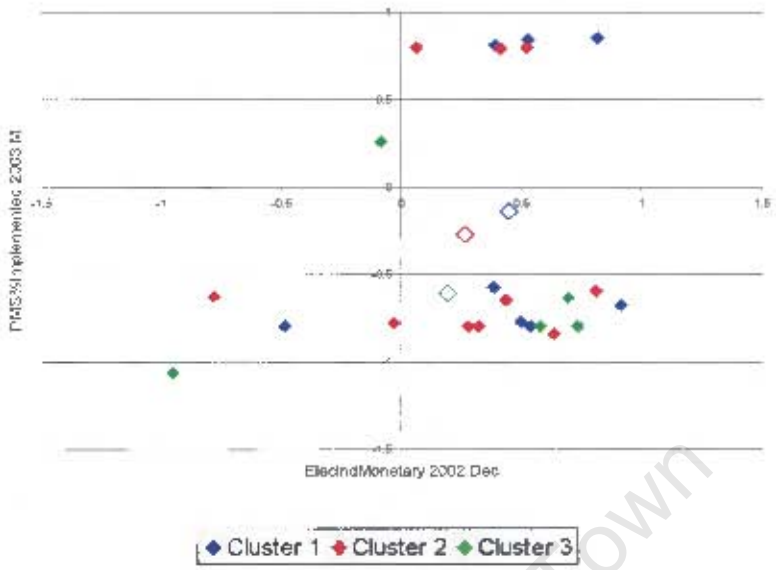



*Illustration 9.8: Municipalities against first two variables from DLGH defined GDA model*

The two dimensions of the model (Appendix XVII), giving a combination of these two variables created as independent transformations that best discriminate between clusters, is shown in Illustration 9.9.

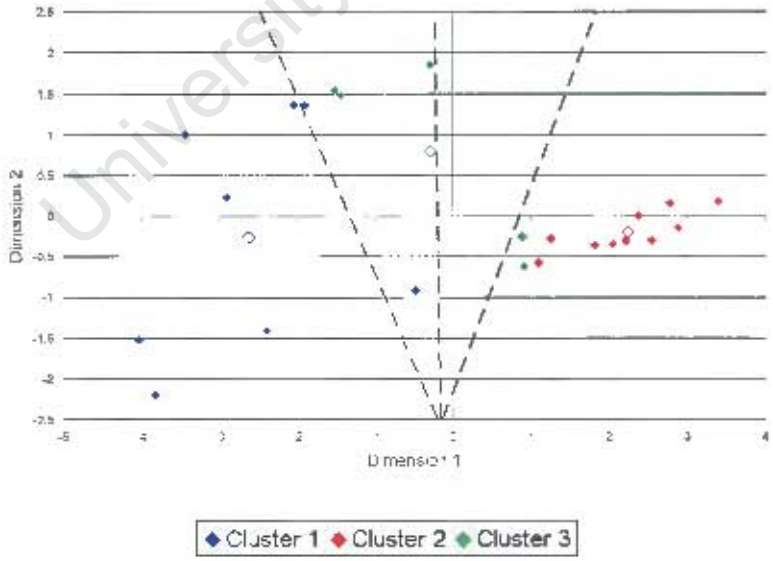



*Illustration 9.9: Municipalities against dimensions of DLGH defined GDA model containing two variables*

The clusters in this graph are far more spread out. Considering both dimensions, no



University of Cape Town

municipalities overlap with clusters they were not assigned to. However, it is not yet possible to use simple cut-off points on these two dimensions to be able to classify municipalities with 100% accuracy. On Dimension 1, *Breede River / Winelands* would be misclassified as part of Cluster 3 and there are no clear cut-off points on Dimension 2. If municipalities were assigned traditionally to the cluster centroid it was closest to, at this point *Overstrand* from Cluster 1 and *Cederberg* and *Kannaland* from Cluster 3 would also be misclassified as belonging to the others' cluster. This indicates that at this point the model is only able to classify municipalities with 82.61% accuracy and therefore requires further variables to improve this classification rate to 100% accuracy.

Once *WaterOthHouse* 2004 December, *CDTotal* 2004 December and *BankBalance* 2003 December are also included in the model, each step shown in Appendix XVII, the municipalities plotted against the two dimensions created by the model are shown in Illustration 9.10.

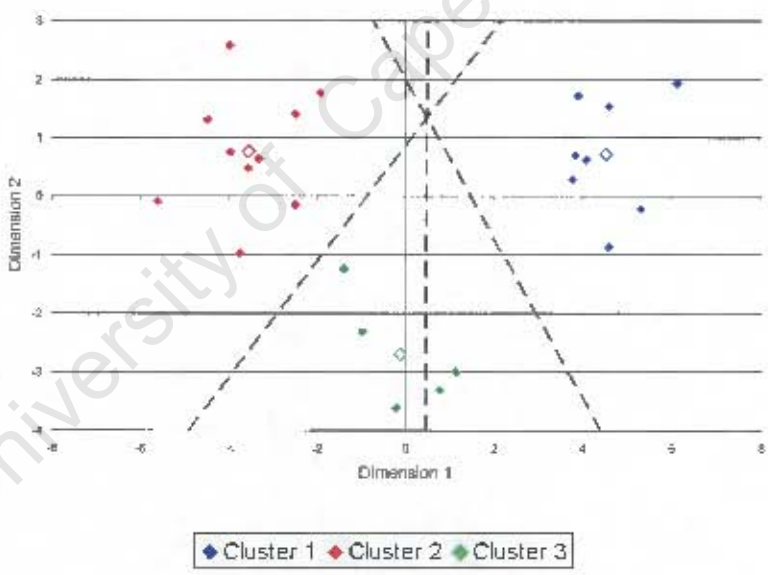



*Illustration 9.10: Municipalities against dimensions of DLGH defined GDA model containing five variables*

It is clear that the clusters of municipalities are now distinctly separated from one another, with no municipalities being misclassified by the model. Any variables added to the model at this point would only further enhance this separation and are therefore considered unnecessary, even though there are still seven other variables that are considered highly significant in being able to discriminate between the three clusters (Appendix XVII).

University of Cape Town

Although it appears that only the first dimension might be necessary to be able to discriminate between clusters, it explains only 86.24% of the model. The 13.76% explained by Dimension 2 should therefore not be considered negligible as it adds a relatively significant portion of explanatory power to the model.

Looking at the graph, there is one interesting placement of a municipality when using the combination of these five variables. This is *Cape Agulhas* which is placed at the very top of Cluster 3, appearing almost closer to Cluster 2, indicating that it may have more in common with this cluster. Although it was not originally chosen unanimously as belonging to the 'bad' performance cluster, it is no less disputed than *Witzenberg*. In fact *Bergriver* was the most disputed of all the municipalities in Cluster 3, with one opinion given that it might belong in the 'good' performance cluster, balanced only by three people indicating that it was a 'bad' performer for the period June 2002 to June 2005.

The five significant variables, all considered separately in terms of the differences between cluster centres, in terms of time rather than the order that they entered the model, are shown in Illustration 9.11.

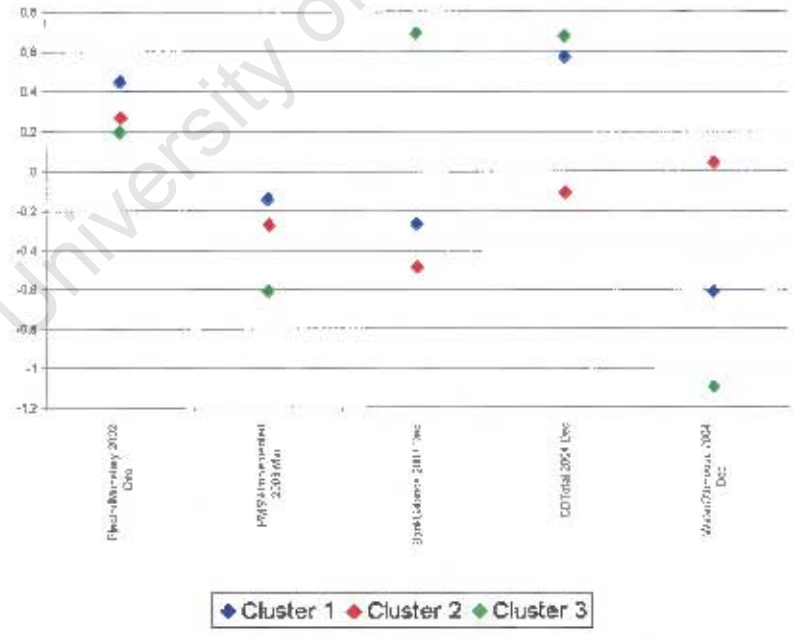



*Illustration 9.11: Three cluster averages across first five significant variables chosen by the DLGH defined GDA model*

The largest differences between clusters are shown by three of the above variables: *BankBalance* 2003 December and *CDTotal* 2004 December single out Clusters 3 and 2 respectively, with the



University of Cape Town

most even spread between the three clusters given by *WaterOthHouse* 2004 December. These first two variables are in fact the last two variables to be chosen by the model and are therefore likely chosen as they separate Cluster 3 and Cluster 2 respectively from the other two clusters and therefore add this information to the model when it is considered needed.

General trends over time show Cluster 1 decreases steadily over time, Cluster 2 dips but then increases again and Cluster 3 has values that vary widely over time. It is interesting to note that for the two financial variables, Cluster 3 shows the highest values, but for the variables relating specifically to service delivery Cluster 3 shows the lowest values of the three clusters.

Although no variable is chosen twice, it is interesting to note that all but one of the significant variables are chosen from December, two of which are this quarter for 2004. This implies that this particular period is important in distinguishing between the profiles of the three clusters. In addition to this, none of the variables chosen are from 2005, indicating that this year is not instrumental in defining the profiles of the clusters. Even when considering all twelve of the highly significant variables that were chosen by the model, only *WaterDisconnection* is chosen from June 2005.

The most interesting of all variables chosen by the model is *PMS%Implemented* as it is not shown as significant in any of the data-defined analyses undertaken, although it is picked up as showing significant differences between clusters in the previous Cluster Analysis. It is interesting that this particular variable is included when attempting to discriminate between clusters of municipalities in terms of performance as determined by people involved in the system. The Performance Management System (PMS) implies a feedback mechanism that gives persons within the local government system an idea of how a municipality might be faring at a particular time.

As shown previously (Illustration 9.11) *PMS%Implemented* 2003 March indicates that Cluster 3 has a lower score than the other two clusters. If this feedback mechanism did in fact influence the opinions that led to grouping the municipalities in such a way, this indicates that municipalities that had the highest percentages of their PMS implemented by March 2003 performed well; with badly performing municipalities having the lowest average percentage of their Performance Management System implemented. This indicates that having the system in place appears to be a good indication of the actual performance of the municipality. However, it does not show causality in terms of whether the system was not implemented because the municipality is

University of Cape Town

struggling, and visa versa.

Finally, it would be ideal to be able to classify either *Laingsburg* or *Oudtshoorn* using this General Discriminant Analysis model and compare these classifications to those given during the group exercise; however the data are simply not available for the variables chosen by the model.

### 9.2.3 Classification Trees

The Classification Tree shown in Illustration 9.12 classifies the municipalities into the clusters determined by the subjective opinion of the members partaking in the clustering exercise.

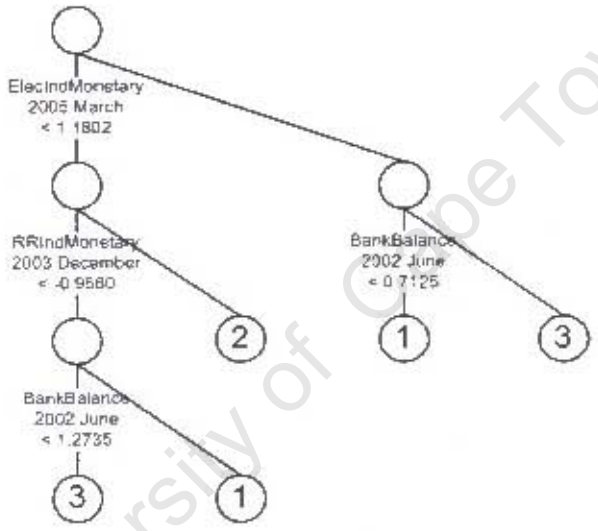



*Illustration 9.12: Classification tree for DLGH defined clusters*

It initially appears as if Clusters 1 and 3 are divided into equal sections in the tree; however it is only *Overstrand* from Cluster 1 that splits down the main left branch of the tree and *Bergrivier* from Cluster 3 that splits down the main right branch of the tree. Therefore the initial split actually separates the majority of Cluster 1 to the right and Cluster 2 along with the majority of cluster 3 to the left.

This is not easily seen in the classification tree and could be misinterpreted. This is a clear example of overfitting the data, but it does indicate the both *Overstrand* and *Bergrivier* should be treated carefully in terms interpreting their profiles as exactly the same as the cluster they are considered to be in.

To make the tree more generalisable, only the initial split along with the second level split in the

University of Cape Town

left branch should be considered, leaving two municipalities misclassified. This misclassification is a little concerning as it takes municipalities that are considered to have a certain performance profile and assigns them to the opposite one.

The final profile of the clusters as defined by the tree is that Cluster 1 has exceptionally high values for *ElecIndMonetary* for the period 2005 March which is confirmed and seen more clearly in Illustration 9.13.

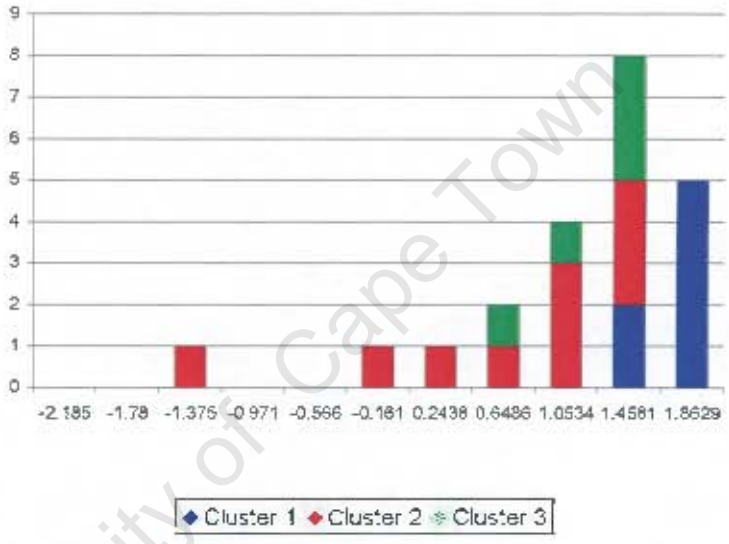



*Illustration 9.13: Municipal values of ElecIndMonetary 2005 March for DLGH defined clusters*

In fact, Cluster 3 generally also has high values for *ElecIndMonetary*, but not so clearly that it can be separated from Cluster 2 on this variable. Cluster 2 does however have average and above values for *RRIndMonetary* 2003 December, whereas Cluster 3 has below average values for this particular variable (Appendix XVIII). It is interesting to note that these major differences between clusters all concern variables relating specifically to service delivery which was not picked up in the clusters as defined by the data itself.

To clearly see how the municipalities lie in relation to each other a different graphical representation is necessary. The major two splits that have been discussed in terms of generalisation and the relationship between clusters is shown in Illustration 9.14.

University of Cape Town

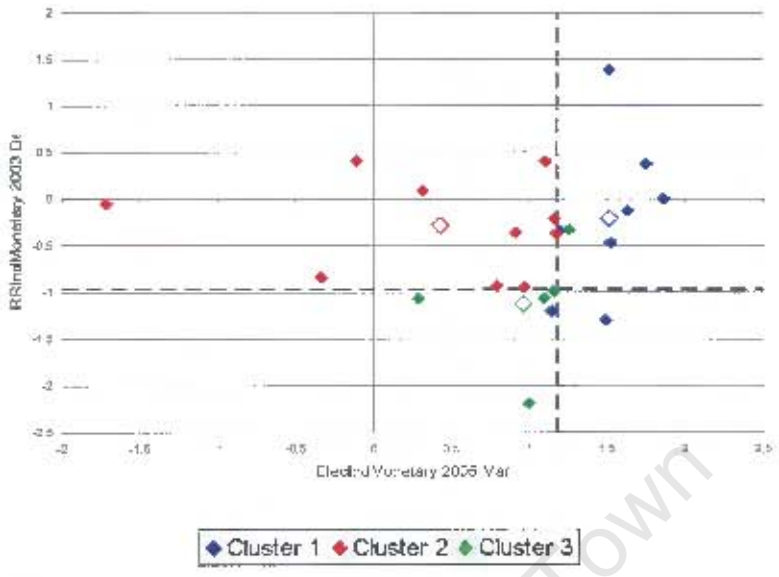



*Illustration 9.14: Municipalities on first two significant variables for DLGH defined clusters*

It is a little more clear now why *Bergrivier* of Cluster 3 splits with Cluster 1, as to push the cut-off point on the x-axis any higher to exclude *Bergrivier* would also exclude a significant portion of Cluster 1. To decrease the cut-off point in the x-axis any further would include all of Cluster 1, therefore reclassifying *Overstrand* correctly; however it would also cause significant portions of Clusters 2 and 3 to be included.

This graph also shows although *Overstrand* from Cluster 1 does appear to sit closest to the centre of Cluster 3 than any other municipality in cluster 1, the distances indicate that its misclassification is not due to it being far from the centre of Cluster 1, but rather that the clusters overlap significantly and are therefore very difficult to separate well. Although this graph represents the two variables that the model indicates will best split the clusters, it is perhaps on a different dimension entirely that the separation of these clusters is best found.

Looking at the finer details of the tree that would not be used in a generalisation, as with the analysis done on the clusters that are defined from the data alone, *BankBalance* 2002 June is seen as a significant variable. This indicates that this variable is important both from a data perspective as well as the subjective opinion given by the participating members.

Interestingly, when doing this analysis from a purely data perspective none of the non-financial variables are chosen as significant in being able to split the clusters; however in the tree above it

University of Cape Town

can be seen that both *ElecIndMonetary* and *RRIndMonetary* from completely different periods are seen to be more important than financial variables in being able to split clusters. This is a clear indication that the financial variables only tell part of the story concerning performance when it comes to the subjective opinion given by personal knowledge of the municipal context. This is especially true as in the subjective opinion analysis, *BankBalance* is only used to separate out two outlying municipalities rather than clearly separating clusters. Its inability to separate the three main clusters is clearly indicated in Illustration 9.15, where there is a significant overlap seen across all clusters:

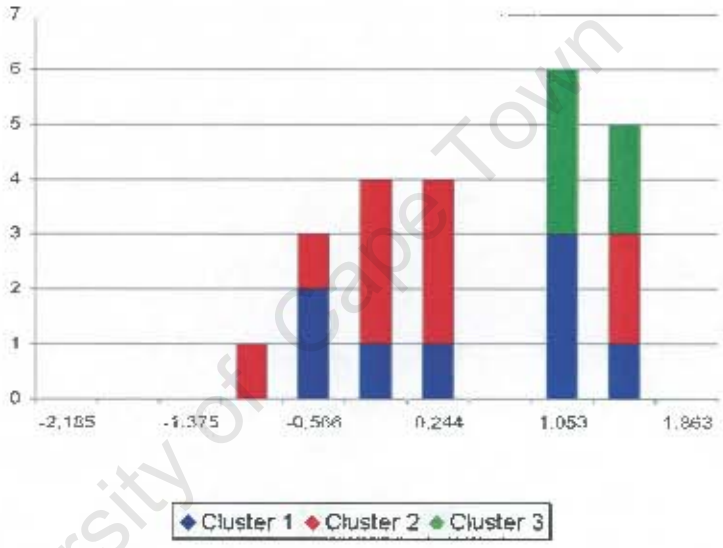



*Illustration 9.15: Municipal values of BankBalance 2002 June for DLGH defined clusters*

The disadvantage of having these particular service delivery type variables as significant in splitting the clusters, rather than financial data, is that it is not possible to classify *Oudtshoorn* or *Laingsburg* using the tree and check this versus the subjective opinion given. The data simply does not exist to be able to do this. *Laingsburg* does have data for *ElecIndMonetary*; however the standardised value splits this municipality down the left branch of the tree and neither municipality contains data for *RRIndMonetary*. The only information that can be gained from attempting this exercise, combined with the knowledge of the structure of the tree, is that it relatively certain that *Laingsburg* does not form part of Cluster 1.

The subjective opinion taken from the exercise is actually that both *Oudtshoorn* and *Laingsburg* belong in Cluster 3 as poor performing municipalities. This does not contradict the above result



University of Cape Town

and if the data were to be relied upon when making decisions, both these municipalities would have been treated as medium to poor performing municipalities – even without all the relevant data available. This does however help to confirm that the actual values given correspond with these municipalities being poor performing ones, and that it is not simply that by not having any data available in the quarterly reports that they are considered to be poor performers.

## 9.3 Discussion

There are various considerations to be made when comparing the results from the three main analyses undertaken with the opinion-defined clusters. All of the analyses identify *ElecIndMonetary* in at least one period as significant in defining the differences, and therefore the profiles of, the three clusters. How this affects the actual profiles is discussed in Section 9.2.1.

Electricity in general is found to show significant differences between the clusters as *ElecIndMonetary* and *CDElectricity* make up ten of the thirteen most significant variables identified in the Cluster Analysis. *ElecIndMonetary* is only picked up as significant once in each of the other analyses; however in both cases it is chosen first, as the most significant. This change is most likely due to the fact that both General Discriminant Analysis and Classification Trees take into account the interaction between variables and therefore all the significant variables in the Cluster Analysis appear to explain the same information present in the data.

Looking at the profiles of the clusters over time, the change in profiles around the time of June-September 2003 is still seen for many of the variables. This indicates that this particular period is important, not just in terms of how the data was reported, but that there may have been some occurrence that caused these changes that also influences the clusters that have been defined by subjective opinion alone.

It is also seen both in the Cluster Analysis and General Discriminant Analysis that across all the periods under consideration, Clusters 1 and 2 show gradual changes over time when analysing the significant variables. Cluster 3 however shows far more variation over time in terms of its values, which raises a query as to why this occurs.

Both the General Discriminant Analysis and Classification Tree models choose *BankBalance* as significant in terms of defining the differences between clusters; however in neither case is it one of the first variables chosen. It appears that both techniques use this variable to define slight changes in profile once various other variables have been taken into account. *BankBalance* is however also seen

University of Cape Town

as a significant variable when analysing the data-defined clusters and therefore could be considered as a proxy of municipal performance if a relationship can be defined.

It is interesting however that as this variable has been standardised by the size of the different municipalities, it is not simply a large bank balance that defines a municipality showing good performance, nor is it this that helped group exercise participants define clusters in terms of their personal opinions. In fact, in the General Discriminant Analysis results, it is seen that Cluster 3 (the 'bad' performing municipalities) clearly has a higher relative bank balance for a particular period than either of the other two clusters.

Finally, in terms of the results both the General Discriminant Analysis and Classification Tree models indicate that when only the first few variables in the model are considered, *Overstrand* from Cluster 1 is likely to be misclassified as Cluster 3. It is concerning that the variables that supposedly discriminate best between clusters tend to misclassify *Overstrand*, requiring other variables for the model to correctly classify this municipality. This is especially curious as *Overstrand* is the only municipality to have been undisputed in opinion that it is a 'good' performing municipality. This is an indication that there may be some discrepancy between the reported values of the municipality and its actual performance.

Neither of the models that is able to predict into which cluster 'new' municipalities would fall is able to classify *Laingsburg* or *Oudtshoorn,* unlike when analysing the data-defined clusters. This is mainly due to the fact that when analysing the opinion-defined clusters there is more of a focus on the service delivery type variables, rather than the financial type variables where there is more data available. This indicates that when attempting to identify the true performance of municipalities, financial data may not necessarily represent the best information and it is vital that more direct indicators are also considered. This does not imply that the financial data are useless, merely that as much data as possible should be analysed, from as many different sources as possible, to be able to define the best profile of clusters.

Concerning the group exercise, with the large number of placements of municipalities in the 'average' category it would have obviously been advantageous to be able to spend time with the participants and allow them to recategorise all the municipalities that they had placed in the 'average' category into either the 'good' or 'bad' category. This would have given an idea of how the municipalities would be placed into two clusters as well as allowing for the possibility of a four clusters comparison.

University of Cape Town

Unfortunately, there were serious time constraints with various members of the meeting not being able to even participate in the exercise as they needed to leave within the first hour. It was felt that the three cluster placement was the easiest for participants to assess, along with being enough data to be able to run an in depth analysis. This was felt to be the best compromise between the needs of the analyst and the constraints of the stakeholders. Again, it was not possible to complete the group exercise earlier as it was necessary for participants to understand the given context which was given through the presentation.

Overall it is felt that this group exercise could be done as part of a larger workshop. This would make it possible not only to gather the individual opinion of participants, but consolidate these results and allow feedback in the group as to why opinions might differ. The benefit of this type of exercise extends to allowing different areas of government to discuss the performance of local municipalities and build plans to improve the situation as part of the whole system.

Concerning the feedback from the presentation itself, although it was felt that a time based analysis would be useful, it is felt that at this particular stage it is simply too complicated. The data set was analysed was far too large and unreliable, with few points across time to be able to pick up any significant trends. However, it may be possible in future to use data analysis techniques to be able to choose those variables that would be most useful for time trend analysis and therefore simplify the process significantly.

Overall, the most important information gained from the feedback process had little to do with the data or results from the analysis. The most significant finding was that results mean little to the people involved if they cannot be presented in a way such that they are easily understood. The group responded best to those techniques that were presented in the simplest form possible, along with much graphical representation. What was considered to be the best way to present the results was rethought and more innovative ways of explaining relationships in the data were found, along with strong graphical representations to be able to support this.

The process that has begun is one of attempting to help the DLGH to gain more knowledge from the significant amount of data that it collects. However, it is vital that this knowledge is presented in such a way that it can be transferred from the analyst to the stakeholders without also having to transfer the expert skills required to understand it. Only then is it truly useful and allows the stakeholders to involve this knowledge in the implementation of plans to help improve the system within which municipalities are encapsulated.

University of Cape Town

# Chapter 10
# Conclusions and Recommendations

The following conclusions, developed through the process of illustrating data mining as a decision tool, discuss the limitations, objectives and results of the process. Data mining is shown to be able to aid in the understanding of a problem context, along with playing a potential role in clarifying data requirements in a development context. Finally the communication of knowledge to decision makers and their role in the process is presented.

University of Cape Town

## Limitations

As with many projects undertaken within the developmental context, there were several limitations placed on the process. Most importantly was the limited access to, and interaction with, the decision makers that would potentially make use of the various methods presented. To gain better understanding as to the context of the problem, along with the results presented, the significant involvement of at least one domain expert would be required. Although some feedback was presented to the DLGH, the scope for interaction was less than ideal.

The data used for the analyses presented in this dissertation were also severely limited in terms of completeness and reliability. In addition to this, the data set was historical, with little relevance to the current problem situations faced by the DLGH. This in turn limits the interest of decision makers in terms of 'buying-in' to the process.

## Problem Context

Using data mining methods to extract knowledge from the data available, even within the limitations of the process, has added to the understanding of the problem context by both analyst and decision makers. In its simplest form, the methods made it possible to identify variables that indicated significant differences between clusters of municipalities that had been defined, both by the data itself and also as part of a group exercise with members of the DLGH. This allowed for the profiling of these clusters, along with the

knowledge as to which variables become important in defining these profiles.

It was found that when clusters were defined through management intuition, rather than from the structure of the data, important variables related more to direct reporting of service delivery rather than to those variables taken from a municipality's financial statements. This indicates that simply assessing the financial profile of a municipality does not fully represent its performance. Although both types of data are important, the more direct service delivery data cannot be ignored in favour of financial data which is generally more reliable and much easier to collect.

A further point is that there exist some major discrepancies between structures mined from the reported data and the opinion of people involved in the processes surrounding municipal service delivery. This indicates that there potentially exists an important gap between what is reported by municipalities compared to what might really be the situation, or that the people involved are unaware of a potential situation that the data might accurately reflect.

University of Cape Town

### Data Requirements: Reduction

Although there are large amounts of data collected by local government, the quality of the data can be limited in terms of reliability. It would be beneficial to decrease the amount of data collected if it could be ensured that the selected variables are able to represent the required aspects of the problem situation. Again, it is vital to therefore involve decision makers in this process, allowing them to decide as to the usefulness of results.

A potential way to reduce the current burden on data reporting without losing valuable information concerns reducing quarterly reporting to annual reporting. This should not be undertaken lightly; however with the use of various data mining techniques it would be possible to compare results using quarterly data to those from annual data. If the results are found to be very similar it indicates that it may be possible to reduce the reporting requirements on a quarterly basis and focus performance analysis to be an annual task. It would be suggested that quarterly reporting continues however, even if at a much reduced capacity, in order to ascertain that the results continue to correspond.

A further way to reduce data requirements is by the use of proxy variables i.e. a particular variable that is able to represent similar information shown by other variables. Concerning the results from the analyses undertaken in this dissertation it was found that of the financial-type variables, various consumer debtors were chosen as significant as was a municipalities bank balance. This indicates that the remainder of the financial-type variables could potentially be ignored as the information they add might already be

represented. It is important to review any decision of this type on a relatively regular basis however, as using variables as proxies might introduce the scenario where municipal reporting focuses on these numbers to the exclusion of all others. This could potentially change the problem context and perhaps the reliability of these proxies as accurate representations.

### Data Requirements: Extrapolation

Although it would be beneficial to reduce data requirements, another significant problem generally exists within the developmental context involving the lack of data. Reducing data requirements may lessen the burden on municipalities and therefore increase the likelihood of reliable data; moreover there exist potential techniques to be able to 'fill-in' missing values instead.

Both General Discriminant Analysis and Classification Trees have the ability to classify 'new' observations into the most appropriate cluster. Without requiring a complete data set, if municipalities that lacked available data simply reported only those variables that were chosen by the data mining model, it would be possible to assign them to a group of municipalities reporting full data.

Within the context of the analyses undertaken, in most cases it was possible to use the models created to assign *Oudtshoorn* to a cluster of municipalities, this particular local municipality having originally been removed from the analysis due to lack of reported data. When using the models developed from the data-defined clusters, *Oudtshoorn* was generally assigned to Cluster 2. During the feedback meeting with the DLGH it was ascertained that, barring *Theewaterskloof*, the data-defined Cluster 2 contained municipalities that, in the opinion of those people involved the meeting, had in actuality performed well over the period.

This result is surprising as *Oudtshoorn* was grouped with those municipalities that were performing 'badly' through the intuitive group exercise done with members of the DLGH. However, the models developed through these opinion-defined clusters, although more difficult to apply as *Oudtshoorn* was lacking data for the variables chosen by the model, tended to classify *Oudtshoorn* accurately as part of the 'bad' cluster. This indicates that the models developed from the opinion-defined clusters may be more accurate in terms of practical use, than those developed from the data-defined clusters.

### Analysis Over Time

During the feedback meeting held with the DLGH it was mentioned that it was perhaps possible to analyse data in terms of time, hoping to predict when a municipality might fail. Although the data analysis undertaken did not focus in this area due to severe data limitations, data mining methods could potentially be used to ascertain which subset of variables should be analysed.

University of Cape Town

What *was* discovered throughout the analysis of the data set in terms of time was that there was a significant change during the period June/September 2003. When analysing the data-defined clusters of municipalities the profiles of these clusters reversed completely from the beginning of the periods under analysis (June 2002) to that at the end of the period (June 2005); and it was during June/September 2003 that this change occurred. Even when analysing the opinion-defined clusters, there was a significant change of clusters profiles during this period, although not the complete reversal seen in the data-defined clusters.

When this phenomenon was discussed with members of the DLGH, although challenging within a historical context, various hypotheses as to what may have caused this change were presented. Although it is difficult to validate these hypotheses it was interesting that data mining methods had been able to objectively determine significant structures in the data that were likely caused by some external influence.

### Communication of Knowledge

In terms of being able to present the results of the analyses to the members of the DLGH in a way that they could understand them, without having to also transfer the knowledge of the techniques themselves, various aspects of this problem were discovered during the feedback meeting.

It was found that Cluster Analysis was thought to be the most useful of the methods presented; however this may have been due to it being the easiest method to understand. It was most difficult to convey the knowledge gained through numerical and text representations, with meeting participants responding best to visual representations of the relationships present in the data.

Due to this response, there was a particular focus on how to 'best' visually represent the relationships in the data. It was discovered that many of the traditional data mining visualisations are useful in this respect, but that they might not offer the 'best' representation in the given context.

The relationships between municipalities and relevant variables can be more clearly represented in General Discriminant Analysis by plotting municipalities as clusters against the first two dimensions of the model, as traditionally offered by the method. However, Classification Trees were originally chosen as a method due to its visualisation of the entire tree, representing the structure of the data without the loss of information. This was however not considered sufficient in being able to transfer the knowledge of these relationships and further visualisations were developed. The clusters of municipalities were plotted against one, or two, of the variables chosen by the model and this was analysed in terms of how well variables were able to distinguish between clusters as well as how the cut-off points affected the split.

University of Cape Town

It is noted that although there exists a need for further visualisations to be developed in order for data relationships to be 'best' represented, many of these may only be discovered within a specific problem context. The analyst will most likely be challenged to develop methods from the analysis of the data itself, as well as through requests from the involved decision makers to visualise relationships in a certain way.

### Involvement of Decision Makers

Although it would have been desirable to have more interaction with the DLGH throughout the analysis process, the little feedback that did occur aided in the understanding of the problem context for the analyst. This initial feedback would become the first stage of a much more involved process, helping to develop research hypotheses and communicate knowledge gained from data analysis to decision makers.

Some indication as to the benefit of involving domain experts as part of the decision making process was discovered in the results of the data mining analysis. It was found that when the opinion-defined clusters were analysed the variables measuring the percentage of a municipality's Performance Management System (PMS) that had been implemented was significant in determining the difference between clusters. This had not been shown to be a significant factor when the data-defined clusters were analysed.

The significance of this particular variable is interesting as the PMS is seen as a feedback mechanism from municipalities to the DLGH. Both the 'good' and 'bad' performing clusters of municipalities indicated high scores for this variable, with the 'average' cluster showing significantly lower scores. It is possible that the members of the DLGH are aware as to how the 'good' and 'bad' municipalities are performing, but are unaware of the performance of the 'average' municipalities as they have less of a feedback mechanism in place.

It would be of benefit to analyse more up-to-date data and present these results to decision makers for discussion. By utilising more relevant data and allowing decision makers to help define the problem context, there would be a better understanding of the relationships between variables in the data and therefore the problem context as a whole. The usefulness of data mining as a decision tool is clear in this sphere. Furthermore, it is suggested that data mining could be used to support and inform a broader decision making workshop approach to managing the the performance of local municipalities.

University of Cape Town

# References


BREIMAN L, FRIEDMAN JH, OLSHEN RA, & STONE CJ, 1984, *Classification and regression trees*, Wadsworth, Belmont (CA).

BURGER D, 2004, *South Africa Yearbook 2003/04*, Government Communications (GCIS), Pretoria, [Online], [Cited 29th July, 2007], Available from http://www.gcis.gov.za/docs/publications/yearbook04/12govern.pdf

BURKHARD R, 2004, *Learning from architects: the difference between knowledge visualisation and information visualisation*, Eight International Conference on Information Visualisation (IV 2004), London, pp. 519-524, [Online], [Cited July 27th ,2007], Available from: http://www.ia.arch.ethz.ch/files/publications/remo_burkhard/2004_Burkhard_Learning from Architects.pdf

CONSTITUTION OF THE REPUBLIC OF SOUTH AFRICA OF 1996, South Africa, [Statute], Government Gazette, **108**, Pretoria.

CROWDER HP, c.2000, *Making Operations Research a Core Business Competency*. The Science of Better Decisions, INFORMS, [Online], [Cited March 15th, 2005], Available from http://www.scienceofbetter.org/whitepapers/INFORMS_WP_Crowder.pdf

ENCYCLOPÆDIA BRITANNICA, 2007, *operations research*, Encyclopædia Britannica Online, [Online], [Cited July 27th, 2007] Available from http://www.britannica.com/eb/article-68172

FIRESTONE JM, 2000, *Data mining and KDD: a shifting mosaic*, DSstar, **4(37)**, [Online], [Cited July 27th, 2007] Available from http://www.taborcommunications.com/dsstar/00/0912/102141.html

GARSON GD, 2007, *Discriminant function analysis*, Statnotes: Topics in Multivariate Analysis, [Online], [Cited June 13th, 2007], Available from http://www2.chass.ncsu.edu/garson/pa765/discrim.htm

HAIR JF, ANDERSON RE, TATHAM RL, & BLACK WC, 1998, *Multivariate data analysis*, 5th Edition, Prentice Hall, Englewood Cliffs (NJ).

INFORMS, 2004, *What operations research is*, Operations research: the science of better, [Online], [Cited July 27th, 2007], Available from http://www.scienceofbetter.org/what/index.htm

KDNUGGETS, 2002, *Polls: Which data mining techniques do you use regularly* [Online], [Cited 29th July, 2007], Available from http://www.kdnuggets.com/polls/2002/data_mining_techniques.htm

LOCAL GOVERNMENT MUNICIPAL SYSTEMS ACT OF 2000, South Africa, [Statute], Government Gazette, **32**, Pretoria.


University of Cape Town

MANLY, BFJ, 1986, *Multivariate statistical methods: a primer*, 3rd Edition, Chapman & Hall, London.

MEISEL S & MATTFELD DC, 2007, *Synergies of Data Mining and Operations Research*, Proceedings of the 40th Hawaii International Conference on System Sciences, pp.1530-1605, [Online], [Cited July 27th,2007], Available from http://csdl2.computer.org/comp/proceedings/hicss/2007/2755/00/27550056c.pdf

MIDGLEY G & REYNOLDS M, 2002, *Systems/Operational Research and sustainable Development: Towards a New Agenda*, Sustainable Development, **12(1)**, [Online], [Cited March 15th, 2005] Available from http://libeprints.open.ac.uk/archive/00000033/01/PDF_Wiley_Tech_eprint.pdf

MUNICIPAL STRUCTURES ACT OF 1998, South Africa, [Statute], Government Gazette, **107**, Pretoria, pp. 16-18

ÓLAFSSON S, 2006, *Introduction to operations research and data mining*, Computers & Operations Research, **33(11)**, pp. 3067-3069, [Online], [Cited July 27th, 2007], Available from http://www.public.iastate.edu/~olafsson/IntroORandDM.pdf

OVENS W, 2005, *Western Cape Municipal Capacity Assessment: Summary Report*, [Government Report], Department of Housing and Local Government, Cape Town.

POPULATION CENSUS, 2001, Digital Census Atlas, Statistics South Africa, [Census Data], Census South Africa, [Online], [Cited March 8th, 2006], Available from http://www.statssa.gov.za/census2001/digiatlas/index.html

PROVINCIAL TREASURY, 2007, *Perceptions of Government Performance and service delivery: a population survey of the Western Cape*, [Government Report], Provincial Government of the Western Cape, Cape Town, [Online] [Cited July 27th, 2007] Available from http://www.capegateway.gov.za/Text/2007/5/population_survey_treasurymay2007_budget.pdf

ROMEU JL, 2001, *Operations research/statistics techniques: a key to quantitative data mining*, Paper presented at the Federal Committee on Statistical Methodology (FCSM) Research Conference, Arlington (VI). [Online], [Cited June 13th, 2007] Available from http://www.fcsm.gov/01papers/Romeu.pdf

STATSOFT INC, 1997, *Electronic statistics textbook*, Statsoft, Tulsa (OK), [Online], [Cited June 13th, 2007], Available from http://www.statsoft.com/textbook/stathome.html

TWAIN M, 1906, *Chapters from my autobiography*, North American Review, DXCVIII, [Online], [Cited July 27th, 2007], Available from http://www.gutenberg.org/etext/19987

VOS W & EVERS L, 2006, *Statistical data mining*, AIMS Review Course, [Course Notes], University of Oxford, Oxford.

University of Cape Town

# Appendix I

# Cluster Analysis: 2-Means Clustering

Table 1 shows how the municipalities have clustered into two groups and their distances to their respective cluster centroid:

*Table I.1: Membership of two cluster solution*

| Cluster 1 | Distance | Cluster 2 | Distance |
|---|---|---|---|
| Matzikama | 0.6292 | Saldanha | 0.8222 |
| Cederberg | 0.6617 | Drakenstein | 0.9043 |
| Bergrivier | 0.5518 | Stellenbosch | 0.8785 |
| Swartland | 0.7195 | Breede Valley | 0.7853 |
| Witzenberg | 0.7687 | Theewaterskloof | 0.8376 |
| Breede River/Winelands | 0.8262 | Overstrand | 0.7764 |
| Cape Agulhas | 0.7013 | Mossel Bay | 0.8250 |
| Swellendam | 0.5241 | George | 0.8078 |
| Kannaland | 0.5940 | | |
| Langeberg | 0.6315 | | |
| Bitou | 0.8897 | | |
| Knysna | 0.8445 | | |
| Prince Albert | 0.4928 | | |
| Beaufort West | 0.8109 | | |
| City of Cape Town | 0.9556 | | |

Most of the municipalities are a similar distance to their group centroid; however *Prince Albert* is the closest to the average of Cluster 1 and *City of Cape Town*, *Bitou* and *Drakenstein* are the furthest away from their respective cluster centres. This implies that these three municipalities have the least in common with the other municipalities in their respective clusters and are most likely to split away if more than two clusters are considered. Illustration 1 shows how this clustering relates in geographical terms:

University of Cape Town

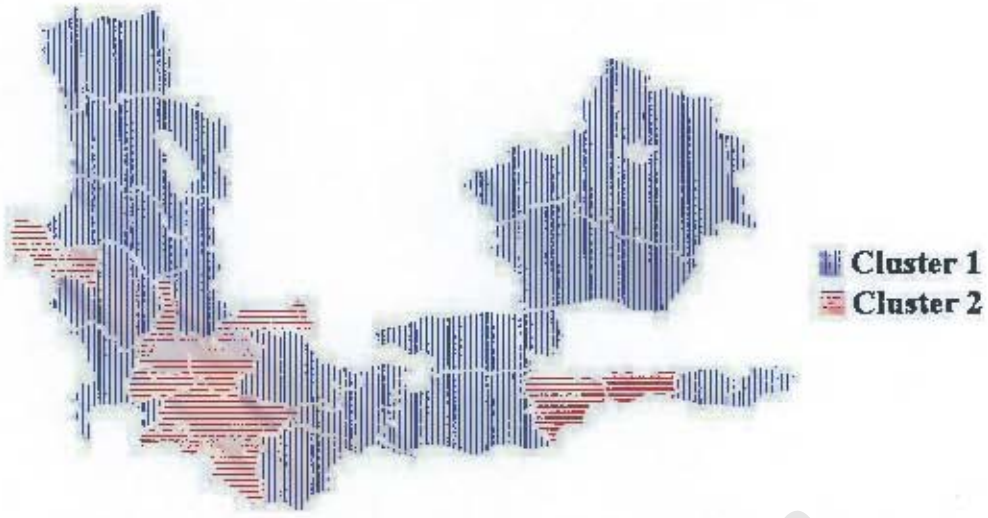



*Illustration I.1: Geographical two cluster membership*

Cluster 2 is the smaller cluster and it formed from three geographical regions. At least two municipalities are grouped together in two areas; however *Saldanha* is unique as although it is near to other municipalities in Cluster 2 it does not border with any of them. The data indicates however that *Saldanha* is one of the closest to the cluster centroid and therefore, although it may not have much in common geographically with the other municipalities in Cluster 2, it definitely has a strong relationship in terms of reported data. Illstrations 2 and 3 show what the relationship between the two clusters of municipalities in terms of the 39 variables that have significantly different means at the 0.1% level.

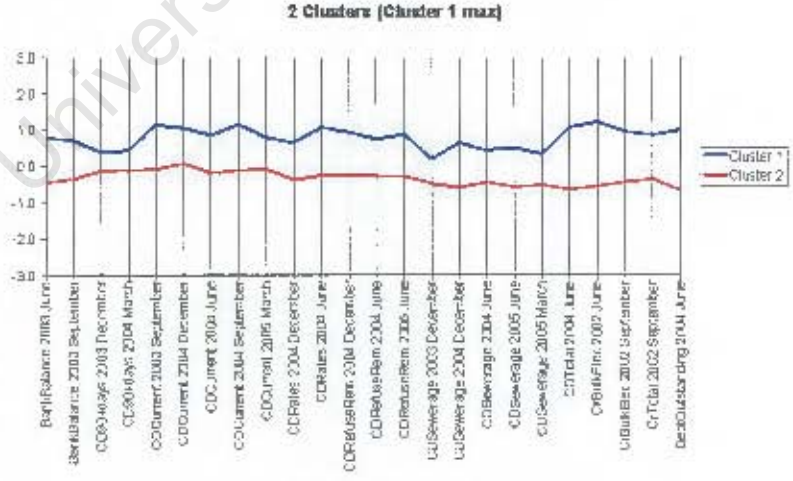



*Illustration I.2: Significant variables for 2 clusters - Cluster 1 highest*

University of Cape Town

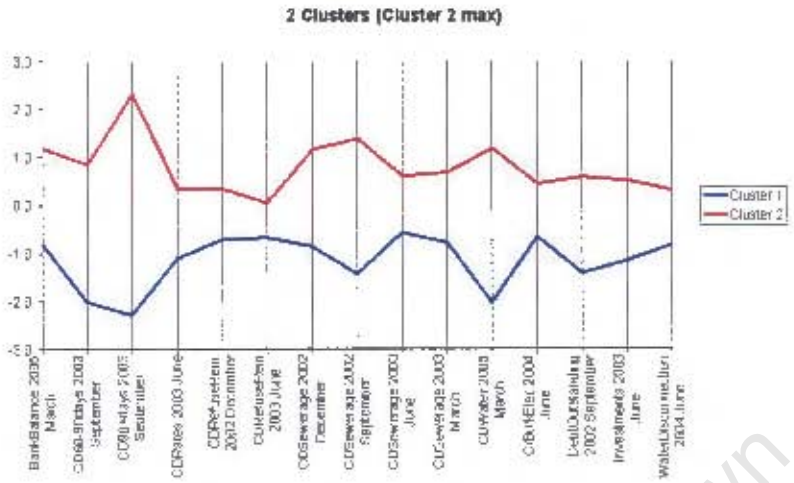



*Illustration I.3: Significant variables for 2 clusters - Cluster 2 highest*

Illustration 2 shows those variables where Cluster 1 has the higher mean score for a particular variable and Illustration 3 where Cluster 2 has the higher mean score. There are substantially more variables where Cluster 1 has a higher mean score (24 variables) as opposed to the 15 variables for Cluster 2. As the two graphs are represented on the same scale, the variables that have a higher mean score in Cluster 2 also appear to vary around the mean far more than those variables where Cluster 1 has a higher mean score.

Almost all variables relate to financial data, with *WaterDisconnection* 2004 June being the only exception, having a significantly higher mean score in Cluster 2. There are significant differences between the means scores of the clusters for *BankBalance* and *CrBulkElec* where Cluster 1 has the higher mean score in early periods, but Cluster 2 has the higher mean score in later periods. The opposite occurs with *CDRefuseRem*, *CDSewerage* and *DebtOutstanding* where initially Cluster 2 has a higher mean score which moves in later periods to Cluster 1 having the higher mean score. Another result of note is that the mean score *CDCurrent* in the later periods is higher in Cluster 1 than in Cluster 2, but there is no significant difference in earlier periods. Finally, the variables that vary the most about the mean between the two clusters are *CD90+Days* 2003 September, *CD60-90Days* 2003 September and *CDWater* 2005 March, indicating that there may have been some event concerning long-term debtors for the quarter ending September 2003.

University of Cape Town

# Appendix II

# Cluster Analysis: 3-Means Clustering

Table 2 shows how the municipalities have clustered into three groups and their distances to their respective cluster centroid:

*Table II.1: Membership of three cluster solution*

| Cluster 1 | Distance | Cluster 2 | Distance | Cluster 3 | Distance |
|---|---|---|---|---|---|
| Matzikama | 0.6215 | Saldanha | 0.7954 | Drakenstein | 0.3677 |
| Cederberg | 0.6541 | Stellenbosch | 0.8655 | Bitou | 0.5377 |
| Bergrivier | 0.5440 | Breede Valley | 0.7833 | | |
| Swartland | 0.7372 | Theewaterskloof | 0.7991 | | |
| Witzenberg | 0.7741 | Overstrand | 0.7830 | | |
| Breede River/Winelands | 0.8279 | Mossel Bay | 0.8213 | | |
| Cape Agulhas | 0.8935 | George | 0.3350 | | |
| Swellendam | 0.5171 | | | | |
| Kannaland | 0.5907 | | | | |
| Langeberg | 0.6771 | | | | |
| Knysna | 0.8391363 | | | | |
| Prince Albert | 0.4690694 | | | | |
| Beaufort West | 0.7997783 | | | | |
| City of Cape Town | 0.9633108 | | | | |

Clustering with 3 clusters is very similar to that of when 2 clusters is used. As mentioned previously, *Drakenstein* and *Bitou* were considered to be two of the municipalities with the least in common to the other municipalities in the 2 cluster analysis and this follows through to the result above where they now form a cluster on their own. Surprisingly, *City of Cape Town*, which had a larger distance to its cluster centroid than any other municipality, does not split away; although it continues to have the largest distance to its cluster centroid. As in the 2 cluster analysis *Prince Albert* is still the closest to its cluster average, as are *Swellendam* and *Bergrivier*. The municipalities furthest away from the Cluster 1 centroid are *City of Cape Town*, *Beaufort West* and *Knysna*. Although *Stellenbosch*, *George* and *Mossel Bay* are furthest from the Cluster 2 centroid, the municipalities in Cluster 1 show a larger distance for *City of Cape Town* and are therefore more likely to split away if more than 3 clusters are considered. Illustration 1 shows how this clustering relates in geographical terms.

University of Cape Town

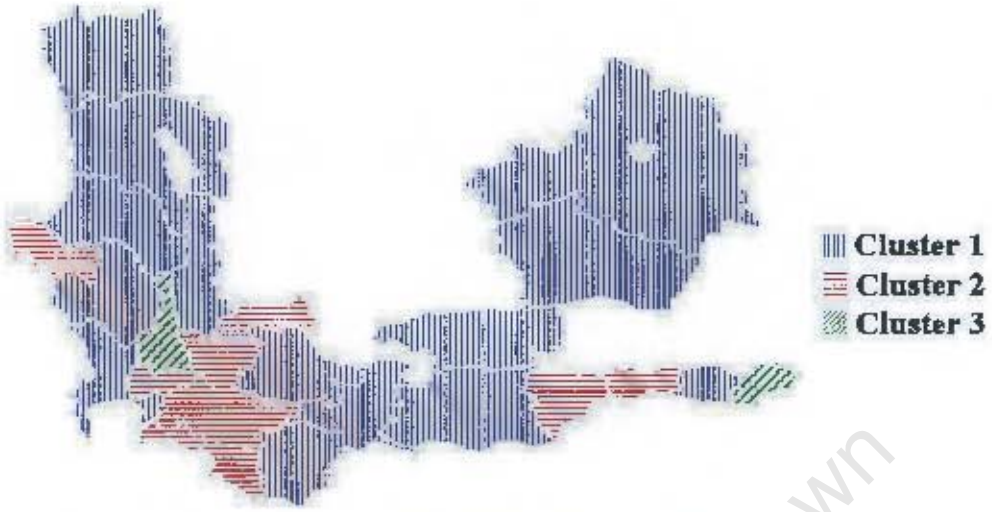



*Illustration II.1: Geographical three cluster membership*

Cluster 2 continues to be formed from three geographical regions, with *Saldanha* not bordering with any other municipalities in the same cluster. It is still one of the closest to the cluster centroid and therefore continues its strong relationship with the other municipalities in the cluster in terms of reported data. It is also noted that although *George* and *Mossel Bay* form part of Cluster 2 it has been mentioned above that, besides *Stellenbosch*, they have the least in common with the other municipalities in this cluster, confirmed by the fact that they are separated from the main group of municipalities that make up Cluster 2. What is surprising is that *Drakenstein* and *Bitou* are almost at opposite ends of the Western Cape geographically; however the data indicates that they have the most in common in terms of reported data. Illustrations 2, 3 and 4 show what the relationship between the three clusters of municipalities in terms of the 38 variables that have significantly different means at the 0.1% level.

University of Cape Town

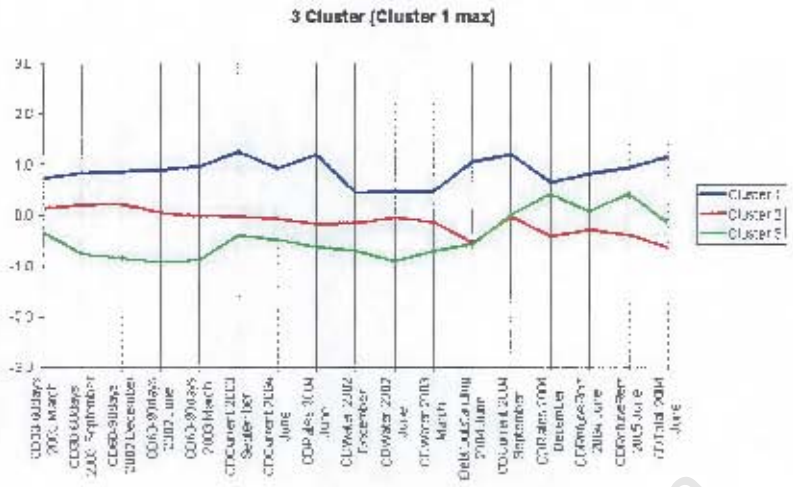



*Illustration II.2: Significant variables for 3 clusters - Cluster 1 highest*

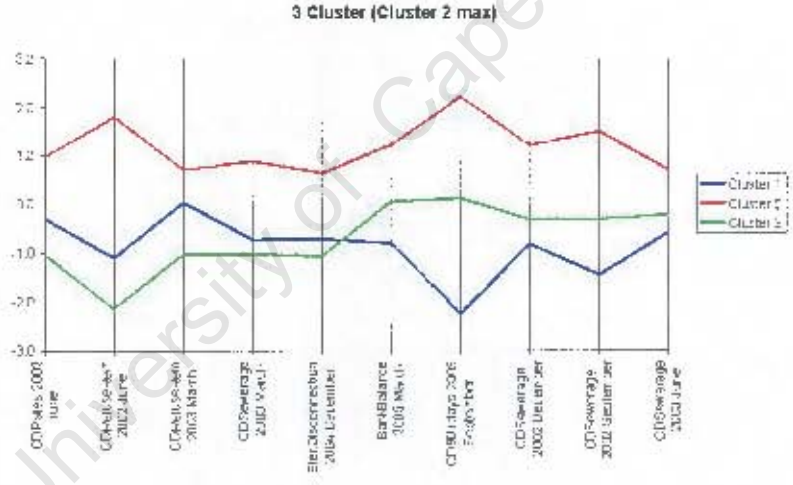



*Illustration II.3: Significant variables for 3 clusters - Cluster 2 highest*

University of Cape Town

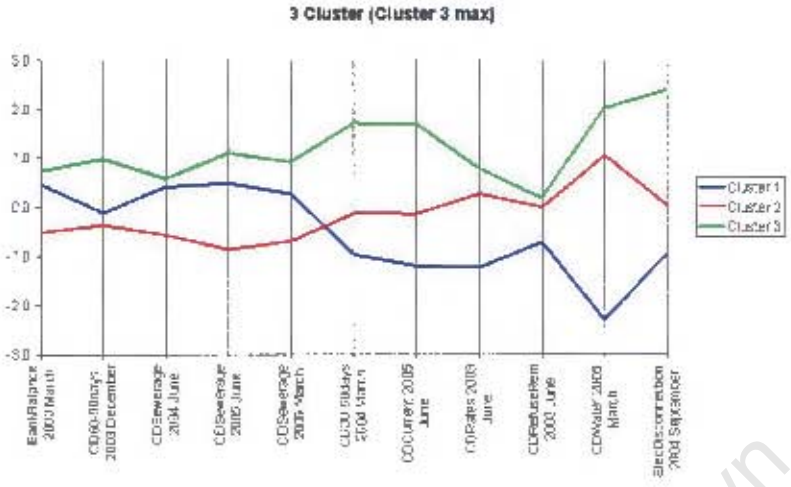



*Illustration II.4: Significant variables for 3 clusters - Cluster 3 highest*

Illustration 2 shows those variables where Cluster 1 has the highest mean score for a particular variable, Illustration 3 where Cluster 2 has the highest mean score and Illustration 4 where Cluster 3 has the highest mean score as compared to the other two clusters.. There are slightly more variables where Cluster 1 has a higher mean score (17 variables) as opposed to the 10 variables for Cluster 2 and 11 variables for Cluster 3. As the three graphs are represented on the same scale, the variables that have a highest mean score in Cluster 2 appear to vary the most around the mean, followed by the graph where Cluster 3 has the highest mean score, leaving the graph where Cluster 1 has a higher mean score as the one with the least variation around the mean.

Again, almost all variables relate to financial data, with *ElecDisconnection* 2004 September and December being the only exception. The relationship between Clusters 1 and 2 for *ElecDisconnection* is the same for both of these quarters with the mean score of Cluster 2 being higher than that of Cluster 1, but the mean score of Cluster 3 is higher than both other clusters in September, dropping substantially to be just below that of Cluster 1 in December.

There are also significant differences between the means scores of the clusters for *CD60-90Days* where Cluster 1 has the highest mean score and Cluster 3 the lowest for the first few periods under analysis, but this changes significantly in December 2003 where the mean for Cluster 3 suddenly becomes much larger for Cluster 3 than the other two clusters which maintain their relationship. *CDRates* also shows significant differences with Cluster 2 having the highest mean score in June 2002, but by June 2003 Cluster 3 has the highest mean score and by June and December 2004 Cluster 1 has the highest mean score for *CDRates*. Cluster 1 also has the highest mean score for *CDWater* in some of the earliest periods, with Cluster 3 having



University of Cape Town

the lowest mean score, but in March 2005 these positions are completely reversed. *CDRefuseRem* appears to have significant differences at similar times in the year with Cluster 2 having the highest mean score in June 2002 and March 2003, but the lowest mean scores in June 2004 and June 2005, with Cluster 1 always having a slightly higher mean score than Cluster 3 for all of these quarters. Finally, in some of the earlier periods Cluster 2 has the highest mean score for *CDSewerage*, with Cluster 3 having a higher mean score than that of Cluster 1, but is overtaken by Cluster 1 in March 2003. However in some of the latest periods Cluster 3 has the highest mean score for *CDSewcarge* and Cluster 1 the lowest.

Finally, the variables that vary most around the mean between the three clusters are *CDRefuseRem* 2002 June, *CD90+Days* 2003 September, *CDWater* 2005 March and *ElecDisconnection* 2004 September, indicating that there may have been some event particular to these periods to cause this wide variation between clusters.

University of Cape Town

# Appendix III

# Cluster Analysis: 4-Means Clustering

Table 3 shows how the municipalities have clustered into four groups and their distances to their respective cluster centroid:

*Table III.1: Membership of four cluster solution*

| Cluster 1 | Distance | Cluster 2 | Distance | Cluster 3 | Distance | Cluster 4 | Distance |
|---|---|---|---|---|---|---|---|
| Matzikama | 0.6185 | Saldanha | 0.7954 | Drakenstein | 0.5977 | Knysna | 0.6713 |
| Cederberg | 0.6265 | Stellenbosch | 0.8655 | Bitou | 0.5977 | Beaufort West | 0.6553 |
| Bergrivier | 0.5040 | Breede Valley | 0.7830 | | | City of Cape Town | 0.7645 |
| Swartland | 0.6943 | Theewaterskloof | 0.7931 | | | | |
| Witzenberg | 0.7625 | Overstrand | 0.7630 | | | | |
| Breede River /Winelands | 0.8181 | Mossel Bay | 0.8213 | | | | |
| Cape Agulhas | 0.6376 | George | 0.8350 | | | | |
| Swellendam | 0.5154 | | | | | | |
| Kannaland | 0.5846 | | | | | | |
| Langeberg | 0.6436 | | | | | | |
| Prince Albert | 0.4762 | | | | | | |

University of Cape Town

Clustering with 4 clusters is very similar to that of when 3 clusters is used. As mentioned previously, *Drakenstein* and *Bitou* were considered to be two of the municipalities with the least in common to the other municipalities in the 2 cluster analysis and this follows through again in the results as they remained in their own cluster. The municipalities that were seen to be furthest away from the Cluster 1 centroid, *City of Cape Town*, *Beaufort West* and *Knysna* have now split away from Cluster 2 in the 3 cluster analysis, as expected, and formed their own cluster. *Prince Albert*, *Bergrivier* and *Swellendam* continue to be closest to the average of Cluster 1 with *Breede River / Winelands* and *Witzenberg* now being the furthest from the centroid of Cluster 1. *Stellenbosch*, *George* and *Mossel Bay* also continue to have the least in common with the other municipalities in Cluster 2 and would be the most likely to split away if more than 4 clusters were considered. Illustration 1 shows how this clustering relates in geographical terms:


*Illustration III.1: Geographical four cluster membership*

Cluster 2 has not changed geographically as the municipalities in the cluster have not changed. The surprising fact that that *Drakenstein* and *Bitou* are almost at opposite ends of the Western Cape geographically; but that the data indicates that they have the most in common in terms of reported data is seen again with the newly formed Cluster 4. *City of Cape Town* is literally situated at the opposite end of the Western Cape to *Beaufort West* and *Knysna*, but the data indicates that they have the most in common with each other, with *Beaufort West* and *Knysna* being closest to the cluster centroid which is not surprising as they are the closest together geographically, separated only by *Oudtshoorn* which was removed from the analysis at the beginning. Illustrations 2, 3, 4 and 5 show what the relationship between the four clusters of municipalities in terms of the 34 variables that have significantly different means at the 0.1% level:

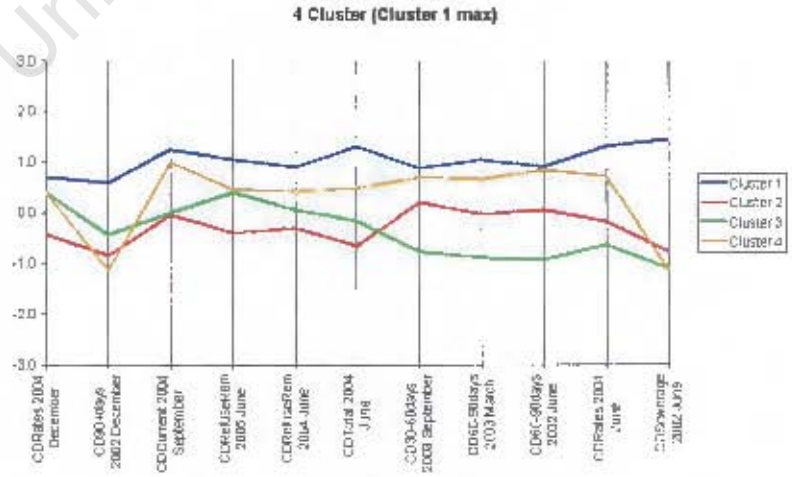



*Illustration III.2: Significant variables for 4 clusters - Cluster 1 highest*



University of Cape Town

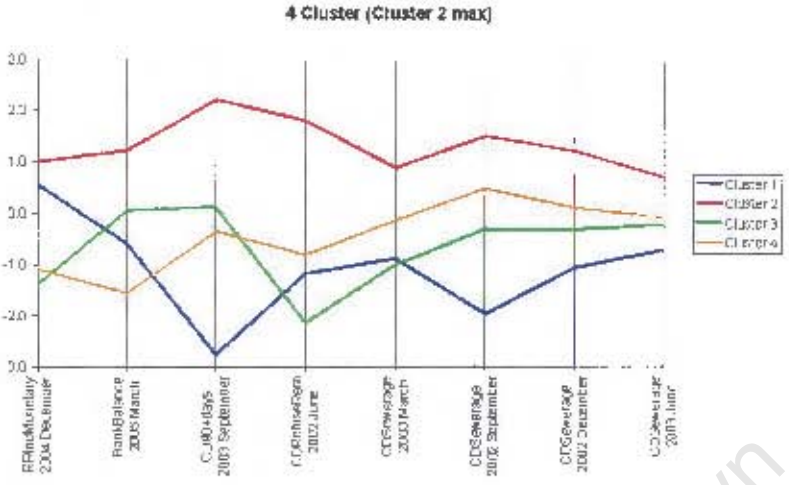



*Illustration III.3: Significant variables for 4 clusters - Cluster 2 highest*

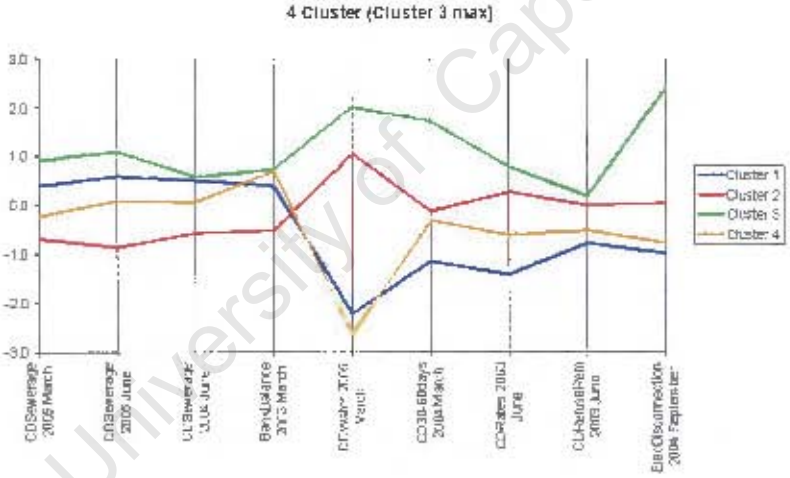



*Illustration III.4: Significant variables for 4 clusters - Cluster 3 highest*

University of Cape Town

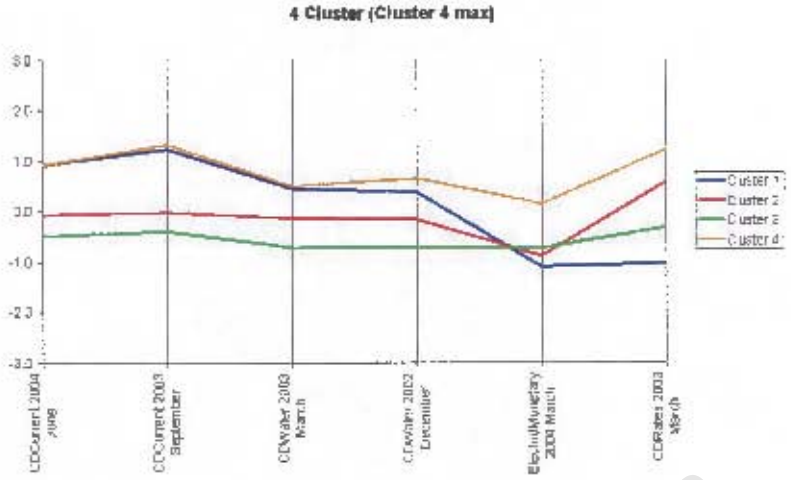



*Illustration III.5: Significant variables for 4 clusters - Cluster 4 highest*

Illustration 2 shows those variables where Cluster 1 has the highest mean score for a particular variable, Illustration 3 where Cluster 2 has the highest mean score, Illustration 4 where Cluster 3 has the highest mean score and Illustration 5 where Cluster 4 has the highest mean score as compared to the other two clusters. Cluster 1 continues to have the highest mean score for the most variables (11 variables) with 8, 9 and 6 variables for Clusters 2, 3 and 4 respectively. As the four graphs are represented on the same scale, the variables that have a highest mean score in Cluster 2 appear to vary the most around the mean, followed by the graph where Cluster 3 has the highest mean score, leaving the graph where Cluster 1 has a higher mean score with little variation around the mean, but the graph where Cluster 4 has the highest mean score appears to vary the least around the mean.

Again, almost all variables relate to financial data, with *ElecDisconnection* 2004 September, *ElecIndMonetary* 2004 March and *RRIndMonetary* 2004 December being the only exceptions. The periods of September and December 2004 are the same as in the 3 cluster analysis; however the variables is only the same for September where Cluster 3 again has the highest mean score by far. In March 2004 Cluster 4 has the highest mean score and in December 2004 Cluster 2 has the highest mean score, with Cluster 4 just below that with Clusters 3 and 1 significantly lower and close together.

There are also significant differences between the means scores of the clusters for *CDRates* where Cluster 3 has the highest mean score in June 2003 and Cluster 1 the lowest, with Cluster 2 being slightly higher than Cluster 4. In June 2004 the clusters have reversed position completely in terms of their mean scores and in December 2004 the results are similar, but Clusters 2 and 4 have almost the same score. This occurs again for CD30-60Days where in September 2003 Cluster 1 has the highest mean score and Cluster 3 the lowest, with

University of Cape Town

Cluster 3 being slightly above Cluster 3. By March 2004 these clusters have reversed positions completely with Cluster 3 now having a substantially higher mean score than the other three clusters.

For the earliest period, Cluster 2 has the highest mean score for *CDRefuseRem* with Cluster 4 being slightly above Cluster 1 which are both only slightly above Cluster 1. In June 2003 this relationship is very similar except that Cluster 3 now has the highest means score and Cluster 2 does not exhibit such an extreme value. By some of the latest periods Cluster 1 has the highest mean score for *CDRefuseRem* and Cluster 2 the lowest with Cluster 4 being slightly higher than Cluster 3. For *CDWater* Cluster 1 and 4 have the highest mean score, which are almost the same, for two of the earliest periods followed by Cluster 2 and then Cluster 3. In March 2005 Cluster 3 has the highest mean score for *CDWater*, with the mean scores for Cluster 1 and 4 again being almost the same, but now substantially lower than the other two clusters.

*CDSewerage* is shown as having at least one significantly different mean score across cluster for 8 out of the 13 periods which is substantially more than has been seen for any other variable throughout the analyses. In June 2002, Cluster 1 has the highest mean score for *CDSewerage* Cluster 2 below this with Clusters 3 and 4 having almost that same lowest mean score. In the early periods after this Cluster 1 moves to have the lowest means score for *CDSewerage*, with Cluster 2 having the highest and Cluster 4 being only slightly higher than Cluster 3. In the latest periods, Cluster 1 still has the lowest mean score, but the mean score for Cluster 3 becomes much higher than that of Cluster 2 and 4.

Finally, the variables that vary most around the mean between the four clusters are *ElecDisconnection* 2004 September, *CDWater* 2005 March, *CD90+Days* 2003 September, *CDRefuseRem* 2002 June and *CDSewerage* 2002 September, indicating that there may have been some event particular to these periods to cause this wide variation between clusters.

University of Cape Town

# Appendix IV

## General Discriminant Analysis: 2-Means Clustering

The variables that are included in the model at the 1% level of significance when attempting to discriminate between the two clusters are shown in Illustration 1. They have been presented such that they are shown in the order in which they entered the model:

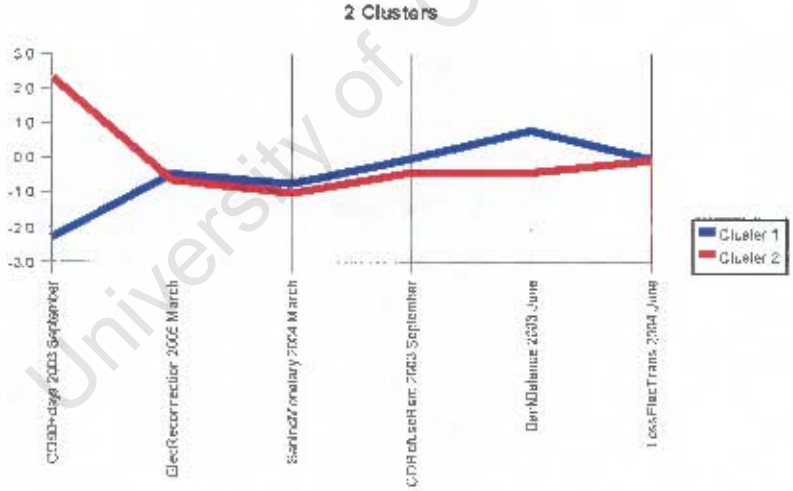



*Illustration IV.1: Significant Variables in 2-Cluster GDA*

It is clear that the greatest deviation between means occurs for *CD90+Days* 2003 September, where the mean of Cluster 2 is that much higher than the mean of Cluster 1. Once the effect of this is taken into consideration, all other significant variables relate to relationship where Cluster 1 has a higher mean than Cluster 2. If the analysis is considered at each step it is determined that only the first three variables are required to be able to classify the municipalities with 100% accuracy, indicating that anything above this

University of Cape Town

would lead to over-fitting of the model. At each step the two clusters of municipalities, along with the mean score of each cluster, is plotted against their relevant one-dimensional discriminant function (Appendix VII) shown in Illustration 2.

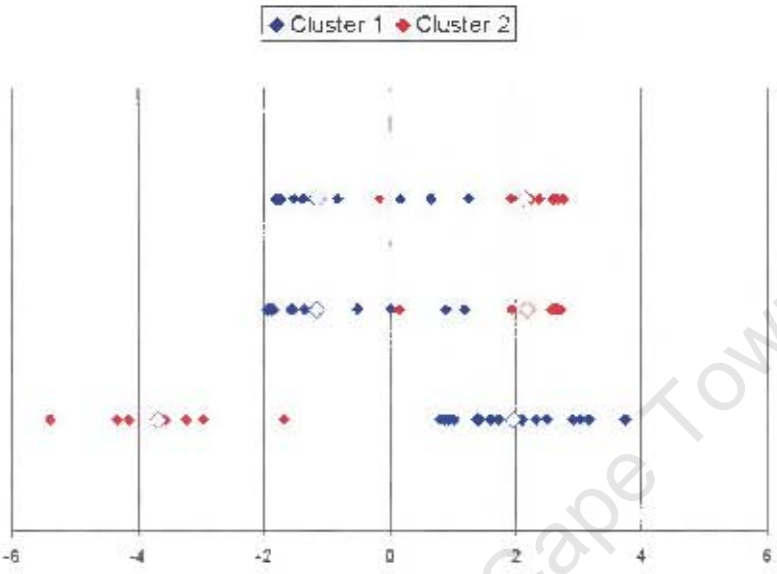



*Illustration IV.2: Stepwise Discriminant Function Plot for 2-Cluster GDA*

University of Cape Town

It can be seen in the top line of the graph at the first step in the model, when only *CD90+Days* 2003 September is included in the model, that the clusters are quite separated and it should be possible to discriminate fairly well between the clusters. However *Saldanha* is misclassified into Cluster 1 and *Beaufort West* misclassified into Cluster 2, which leads to an overall accuracy of prediction in the model of 91.30%. At the second step, shown in the second line of the graph, where *ElecReconnection* 2005 March is now also included in the model, the two clusters move further away from each other, but the model's overall accuracy of prediction actually decreases with *Saldanha* and *Beaufort West* again being misclassified, but the *City of Cape Town* also being misclassified as Cluster 2. This leads to an overall accuracy of prediction in the model of 86.96%. It is only when *SanIndMonetary* 2004 March joins the model that the clusters very clearly separate away from one another and there is 100% accuracy of prediction of the municipalities into their respective clusters. However, this also results in the discriminant score of the clusters being reversed and the municipalities in Cluster 2 now have lower discriminant scores than those in Cluster 1.

It is necessary to look at the factor structure coefficients (Appendix VII) to determine the relationship between the variables used in the model and the respective clusters. As *CD90+Days* has a negative relationship with the discriminant function, but both *ElecReconnection* and *SanIndMonetary* have positive

relationships, Cluster 1 has high electricity reconnections over the three months ending in March 2005 and high provision of free basic sanitation in terms of its monetary value in March 2004, but low long-term debtors for the period September 2003. Cluster 2 obviously has the opposite profile.

University of Cape Town

# Appendix V

## General Discriminant Analysis: 3-Means Clustering

The variables that are included in the model at the 1% level of significance when attempting to discriminate between the three clusters are shown in Illustration 1. They have been presented such that they are shown in the order in which they entered the model:

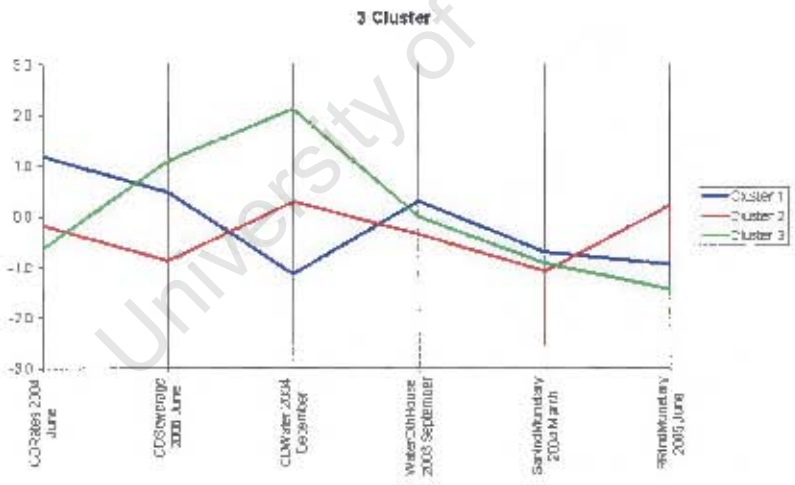



*Illustration V.1: Significant Variables for 3-Cluster GDA*

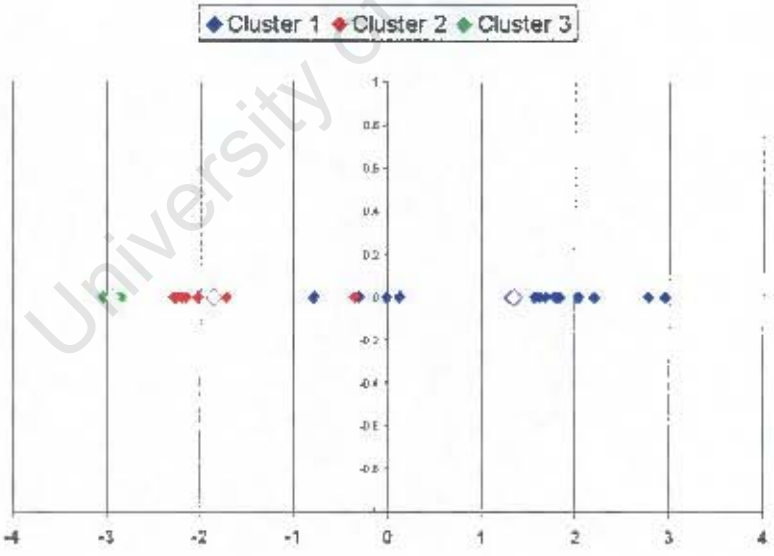
It is clear that the greatest deviation between means is for *CDWater* 2004 December; however it is not as significant as *CDRates* 2004 June in terms of discriminatory power between the three clusters as this is the first variable to enter the model. It appears that the mean of Cluster 1 is significantly higher than the other two clusters for *CDWater* 2004 June, which is not seen again until *WaterOthHouse* 2003 September and *SanIndMonetary* 2004 March. There is also larger deviations between the clusters for *CDSewerage* 2005

University of Cape Town

June, however in this case it is Cluster 2 that has a significantly lower mean than the other two clusters, confirming that the first two variables explain different variation in the model. The large deviation between all means that *CDWater* 2004 December provides allows for further clarity in terms of being able to discriminate between the three clusters, however if the analysis is considered at every step if is determined that the model requires *WaterOthHouse* 2003 September to further clarify this discrimination and allow for the model to predict classification of the municipalities with 100% accuracy.

It is interesting to note that although not all the variables are necessary to significantly discriminate between clusters, both financial and service delivery type variables are seen in terms of the provision of water and sewerage/sanitation, indicating either that these variables do not correspond to each other as expected or that there were significant changes over the periods to create variables that explain independent information between periods even if they are consistent over the reporting of financial and service delivery type data.

At each step the three clusters of municipalities, along with the mean score of each cluster, is plotted against their relevant discriminant functions (Appendix VIII). At the first step there is only one discriminant function, leading to a one-dimensional plot, but in the other three steps there are two discriminant functions allowing the graphs to be considered in the two dimensions shown in Illustration 2




University of Cape Town

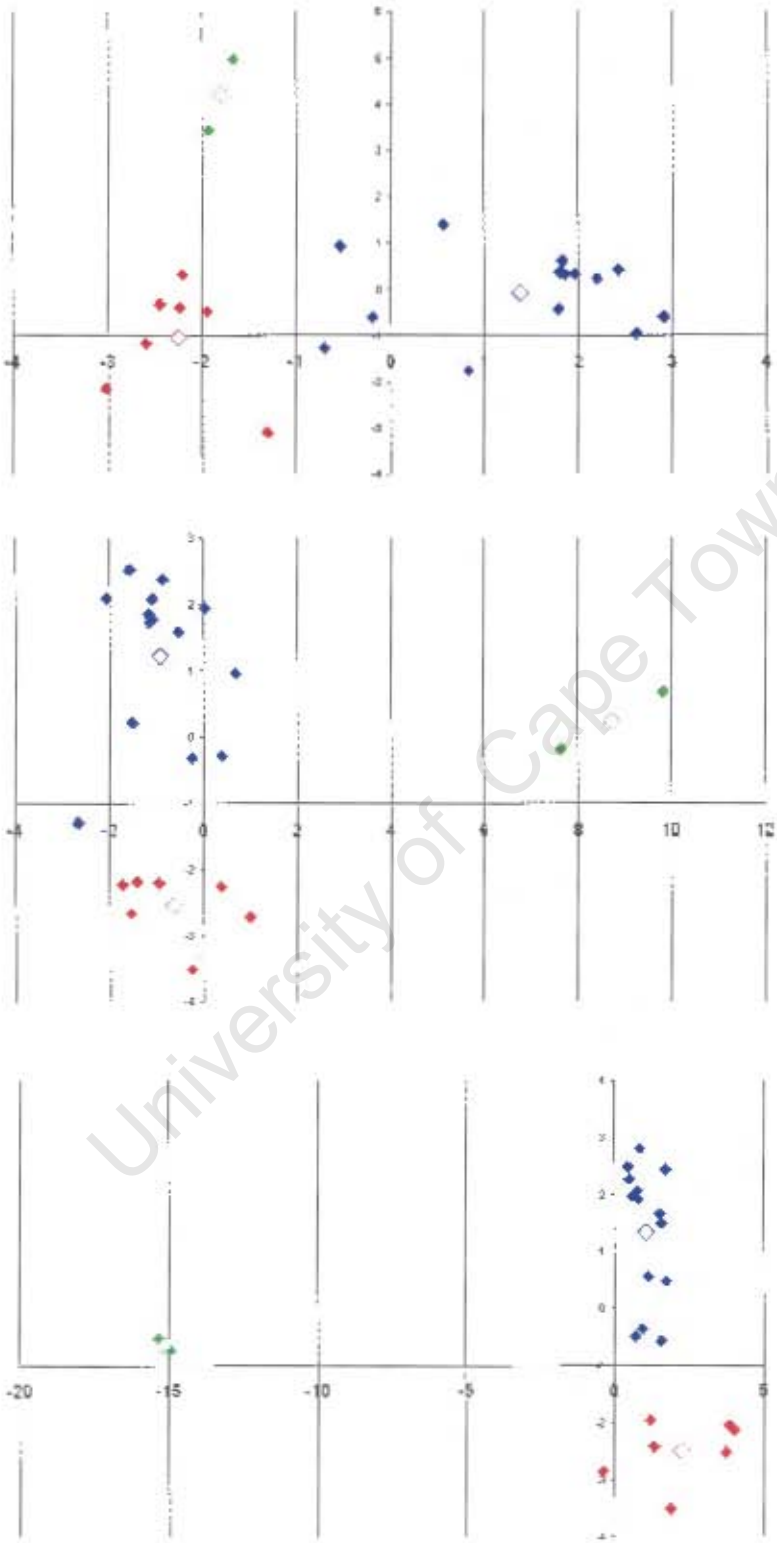


*Illustration V.2: Stepwise Discriminant Function Plots for 3-Cluster GDA*

University of Cape Town

It can be seen in the first graph at the first step in the model, when only *CDRates* 2004 June is included in the model, that the clusters are quite separated and it should be possible to discriminate fairly well between the clusters, especially for Cluster 3 which is completely separate from the other two clusters. However, *Overstrand* from Cluster 2 is misclassified as Cluster 1 and *Drakenstein*, *Bitou* from Cluster 3 and *City of Cape Town* from Cluster 1 are all misclassified as Cluster 2, which leads to an overall accuracy of prediction in the model of 82.61%. This is not surprising as in the 2-Means Cluster Analysis it is seen that *City of Cape Town* has the least in common from the other municipalities in Cluster 1, however *Overstrand* is the closest to the average of Cluster 2 when all variables are considered indicating that this relationship is significantly different when only *CDRates* 2004 June is considered. The misclassification of *Drakenstein* and *Bitou* is unsurprising as they both belong to Cluster 3 and from the plot of means it was seen that *CDRates* 2004 June discriminates well between Cluster 1 and the other two clusters, but that it is very likely to treat Clusters 2 and 3 in a similar manner.

At the second step, when *CDSewerage* 2005 June enters the model, there is a much clearer separation of clusters over the two dimensions. The first dimension (x-axis) appears to discriminate well between Cluster 1 and Cluster 2 and 3, as expected from the above discussion and explains 61.56% of the variation in the model. The second dimension (y-axis) appears to discriminate well between Cluster 2 and Cluster 3, explaining the remaining 38.44% of the variation in the model, again expected as when *CDSewerage* 2005 June is considered in the plot of means Cluster 2 has a much lower mean than that of the other two clusters. Recall that the second dimension already accounts for the variation explained by the first dimension and therefore although Cluster 2 also has a much lower mean than Cluster 1 for *CDSewerage*, the model is already able to discriminate Cluster 1 from the other two clusters and this additional information is not necessary. At point the model is able to classify municipalities with 95.65% accuracy, with *Beaufort West* from Cluster 1 being misclassified as Cluster 2.

When *CDWater* 2004 December is added to the model the clusters again separate further, but the model's prediction power remains stable with 95.65% accuracy. Although there appears to be more separation between the clusters, with the first discriminant function now explaining 71.66% of the variation in the model, *Beaufort West* continues to be misclassified and it is necessary to add *WaterOthHouse* 2003 September to the model for 100% accuracy in prediction. As with the 2-Means Clustering, this final step reverses the discriminant scores of the municipalities in each cluster.

It is necessary to look at the factor structure coefficients (Appendix VIII) to determine the relationship between the variables used in the model and the respective clusters. As the first discriminant function explains 88.01% of the variation in the data and the second explains the remaining 11.99%, the focus of the

University of Cape Town

interpretation will be on the first dimension, particularly as the two dimensions conflict in terms of interpretation at some points, but dimension two already takes dimension 1 into account when explaining the variation in the data. *CDRates* has a positive relationship with the first discriminant function and *CDSewerage*, *CDWater* and *WaterOthHouse* have a negative relationship, but *CDWater* has a negative relationship with the second discriminant function and the remaining variables have positive relationships.

Cluster 1 has high consumer debtors relating specifically to rates and low consumer debtors relating specifically to water services when compared to Clusters 2 and 3. It also appears to have higher consumer debtors relating specifically to sewerage/sanitation services along with a higher number of non-indigent households with access to free basic water than Cluster 3, but these are both lower than Cluster 2.

Cluster 2 has low consumer debtors relating specifically to water services along with a low number of non-indigent households with access to free basic water when compared to Clusters 1 and 3. It also appears to have higher consumer debtors relating specifically to rates and lower consumer debtors relating specifically to sewerage/sanitation services than Cluster 3, but the opposite relationship to Cluster 1.

Finally, Cluster 3 has high consumer debtors relating specifically to sewerage/sanitation and water services along with a high number of non-indigent households with access to free basic water when compared to Clusters 1 and 2. It also appears to have lower consumer debtors relating specifically to rates when compared to Cluster 1, but this is higher when compared to Cluster 2.

University of Cape Town

# Appendix VI

## General Discriminant Analysis: 4-Means Clustering

The variables that are included in the model at the 1% level of significance when attempting to discriminate between the four clusters are shown in Illustration I. They have been presented such that they are shown in the order in which they entered the model:

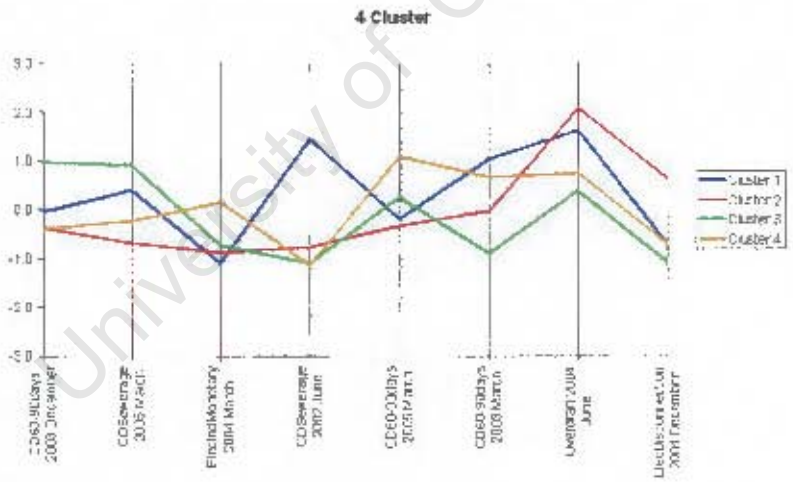



*Illustration VI.1: Significant Variables for 4-Cluster GDA*

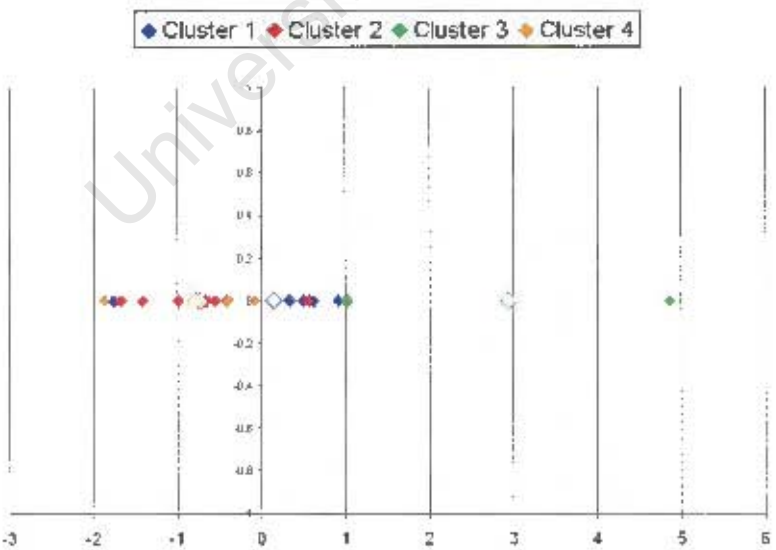
It is clear that the greatest deviation between means is for *CDSewearge* 2002 June; however it is not as significant as *CD60-90Days* 2003 December in terms of discriminatory power between the three clusters as this is the first variable to enter the model. It appears that the mean of Cluster 3 is significantly higher than the other three clusters for *CD60-90Days* 2003 December, which is not seen again throughout the significant variables, although *Overdraft* 2004 June and *ElecDisconnection* 2004 December are both able to

University of Cape Town

discriminate Cluster 3 from the other three clusters. The second variable that enters the model is *CDSewerage* 2005 March which appears to be able to discriminate between all four clusters, but both *ElecIndMonetary* 2004 March and *CDSewearge* 2002 June are needed by the model in order to clarify the discrimination further and allow for the model to predict classification of the municipalities with 100% accuracy. These two additional variables appear to discriminate Cluster 4 and Cluster 1 respectively from the other clusters and therefore when the four are combined, it is not surprising that it leads to 100% accuracy of prediction.

It is interesting to note that although not all the variables are necessary to significantly discriminate between clusters, *CDSewerage* appears twice in the significant variables, indicating that there are completely different profiles of clusters in these two periods. However, this is not that surprising as the one is taken from the very beginning of the period under analysis and the other from the second to last quarter.

At each step the four clusters of municipalities, along with the mean score of each cluster, is plotted against their relevant discriminant functions (Appendix IX). At the first step there is only one discriminant function, leading to a one-dimensional plot, but in the second step there are two discriminant functions allowing the graphs to be considered in two dimensions. The final two steps have three discriminant functions, but as the first two explain the largest amount of variation in the data only they will be considered in the analysis, giving again the two dimensions shown in Illustration 2.




University of Cape Town

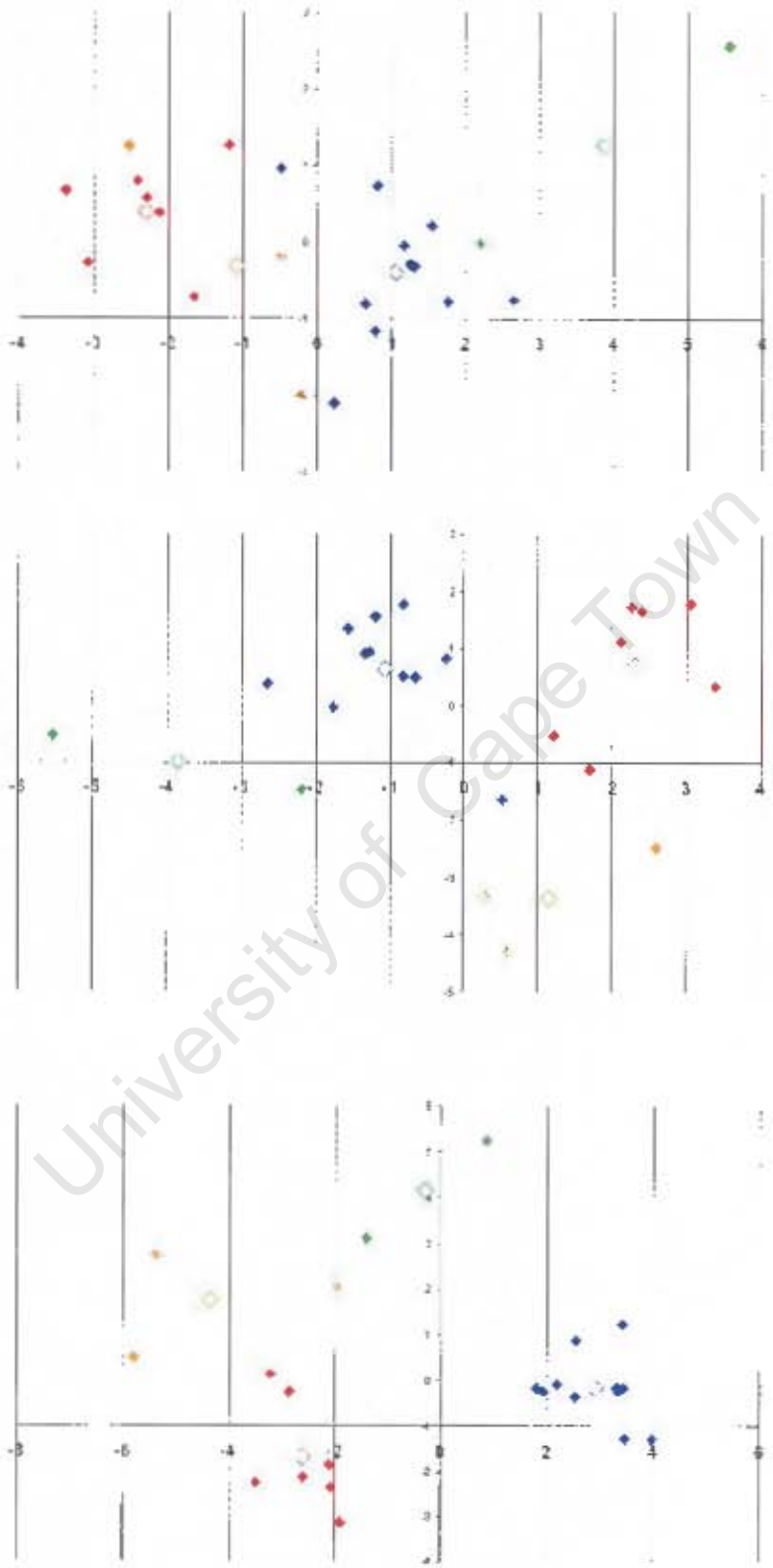


*Illustration VI.2: Stepwise Discriminant Function Plots for 4-Cluster GDA*

As expected, Cluster 3 is separated relatively successfully from the other three clusters in the first graph,

University of Cape Town

where only *CD60-90Days* 2003 December is included in the model. The accuracy of the model at this point in terms of prediction power is 60.88% with nine municipalities misclassified, with *Cederberg* from Cluster 1 and *City of Cape Town* from Cluster 4 being misclassified as Cluster 2, *Breede Valley*, *Theewaterskloof* and *Overstrand* from Cluster 2, *Drakenstein* from Cluster 3 and *Knysna* and *Beaufort West* from Cluster 4 being misclassified as Cluster 1. This classification rate improves dramatically at the next step when the clusters are now considered in two dimensions and *CDSewerage* 2005 March enters the model. Cluster 3 is still quite separate from the other clusters, however Clusters 1 and 2 have separated further, with Cluster 4 being situated between them. Only *Drakenstein* from Cluster 3 and *Knysna* and *City of Cape Town* from Cluster 4 are misclassified as Cluster 1 and *Beaufort West* from Cluster 4 is misclassified as Cluster 2, leading to an accuracy in the model of 82.61% in terms of correct classifications of the municipalities.

At the third step there are three discriminatory functions, however only the first two will be considered as together they explain 95.28% of the overall model and the clusters are easier to interpret in two dimensions. In the third step, *ElecIndMonetary* 2004 March enters the model and Cluster 4 is now well separated from the other clusters as shown in the third graph. The only municipalities misclassified now are *Drakenstein* from Cluster 3 and *Witzenberg* from Cluster 1 which are misclassified as Cluster 1 and 4 respectively, leading the model to have an accuracy of prediction of 91.30%. It is only when *CDSewerage* 2002 June enters the model at the fourth step that the model is able to predict with 100% accuracy and the clusters are clearly separated from one another when considering only the first two of three dimensions.

It is necessary to look at the factor structure coefficients (Appendix IX) to determine the relationship between the variables used in the model and the respective clusters. As the first discriminant function explains 70.17% of the variation in the data, the second discriminant function 22.81% and the third 7.05%, the focus of the interpretation will be on the first dimension, but the second dimension will also be considered as it adds substantially to the explanation, with the first two dimensions explaining 92.95% of the variation in the data. Keeping in mind that dimension 2 already takes into account the variation explained by dimension 1, *ElecIndMonetary* 2004 March has a negative relationship with the first discriminant function, but *CD60-90Days* 2003 December, *CDSewerage* 2005 March and *CDSewerage* 2002 June exhibit a negative relationship, whereas *CDSewerage* 2002 June has a negative relationship with the second discriminant function and the remaining variables have a positive relationship.

Cluster 1 has high medium-term debtors for December 2003, but low monetary value for the provision of free basic electricity in March 2004. When considering consumer debtors relating specifically to sewerage/sanitation services for both March 2005 and June 2002, Cluster 1 is higher for both than any other cluster, however it is higher for the period March 2005 than in the period June 2002.

University of Cape Town

Cluster 2 has below average medium term debtors for December 2003, but above average monetary value for the provision of free basic electricity in March 2004. When considering consumer debtors relating specifically to sewerage/sanitation services for both March 2005 and June 2002, Cluster 2 is low, but not the lowest compared to the other clusters, but has the lowest values in March 2005 and the highest in June 2002.

Cluster 3 has above average medium term debtors for December 2003, but high monetary value for the provision of free basic electricity in March 2004. When considering consumer debtors relating specifically to sewerage/sanitation services for both March 2005 and June 2002, Cluster 3 is average, but is highest in March 2005 and lowest in June 2002.

Finally, Cluster 4 has below average medium term debtors for December 2003, but high monetary value for the provision of free basic electricity in March 2004. When considering consumer debtors relating specifically to sewerage/sanitation services for both March 2005 and June 2002, Cluster 4 has the lowest debtors than any other cluster, however in the cluster itself, it is higher for the period March 2005 than in the period June 2002.

University of Cape Town

# Appendix VII

# General Discriminant Analysis: 2-Means Tables

Variables in the model

| | | F to remove | P to remove |
|---|---|---|---|
| CD90+days | 2003 September | 460.1781 | 0.0000 |
| ElecReconnection | 2005 March | 23.0560 | 0.0002 |
| SanIndMonetary | 2004 March | 174.7375 | 0.0000 |
| CDRefuseRem | 2003 September | 27.4619 | 0.0001 |
| BankBalance | 2003 June | 85.1313 | 0.0000 |
| LostElecTrans | 2004 June | 16.2114 | 0.0010 |

Discriminant Function

Step 1

| | | Function 1 |
|---|---|---|
| Intercept | | 0.1692 |
| CD90+days | 2003 September | 0.7130 |
| Eigenvalue | | 2.6824 |
| Cum. Prop. | | 1.0000 |

Step 2

| | | Function 1 |
|---|---|---|
| Intercept | | 0.5604 |
| CD90+days | 2003 September | 0.7295 |
| ElecReconnection | 2005 March | 0.1501 |
| Eigenvalue | | 2.7660 |
| Cum. Prop. | | 1.0000 |

Step 3

| | | Function 1 |
|---|---|---|
| Intercept | | 0.6865 |
| CD90+days | 2003 September | -1.1271 |
| ElecReconnection | 2005 March | -0.7577 |
| SanIndMonetary | 2004 March | 2.2165 |
| Eigenvalue | | 7.9537 |
| Cum. Prop. | | 1.0000 |

University of Cape Town

Final Step

| | | Function 1 |
|---|---|---|
| Intercept | | -1.8773 |
| CD90+days | 2003 September | 2.5995 |
| ElecReconnection | 2005 March | 1.3682 |
| SanIndMonetary | 2004 March | -6.1002 |
| CDRefuseRem | 2003 September | -3.2148 |
| BankBalance | 2003 June | -3.8489 |
| LossElecTrans | 2004 June | -0.9142 |
| Eigenvalue | | 92.6616 |
| Cum. Prop. | | 1.0000 |

Factor Structure Coefficients

| | | Function 1 |
|---|---|---|
| CD90+days | 2003 September | -0.5807 |
| ElecReconnection | 2005 March | 0.0249 |
| SanIndMonetary | 2004 March | 0.0756 |

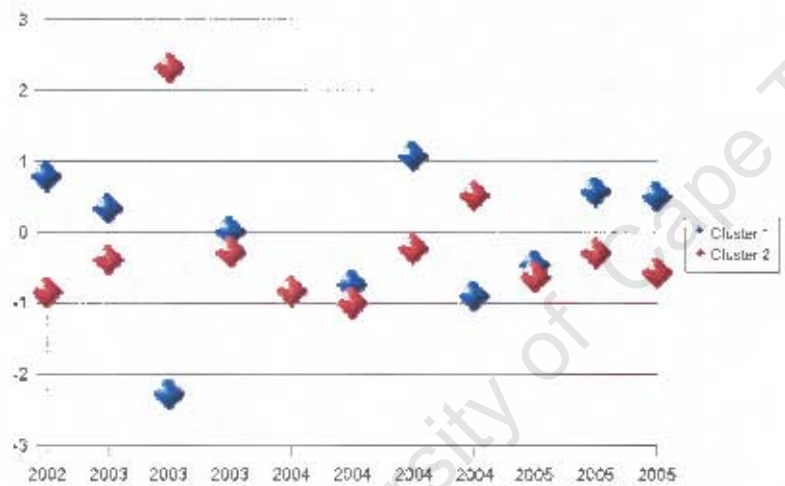



*Illustration VII.1: Two cluster averages across all significant GDA variables*

University of Cape Town

# Appendix VIII

# General Discriminant Analysis: 3-Means Tables

Variables in the model

| | | F to remove | P to remove |
|---|---|---|---|
| CDRates | 2004 June | 1407.4406 | 0.0000 |
| CDSewerage | 2005 June | 506.9992 | 0.0000 |
| CDWater | 2004 December | 100.9842 | 0.0000 |
| WaterOthHouse | 2003 September | 31.2770 | 0.0000 |
| SanIndMonetary | 2004 March | 1788.7259 | 0.0000 |
| RRIndMonetary | 2005 June | 373.7404 | 0.0000 |
| CDTotal | 2002 June | 223.3656 | 0.0000 |
| CDRefuseRem | 2003 December | 120.0041 | 0.0000 |
| CD60-90days | 2004 December | 46.6671 | 0.0000 |
| CDElectricity | 2004 March | 10.9388 | 0.0024 |

Discriminant Functions

Step 1

| | | Function 1 |
|---|---|---|
| Intercept | | -1.4385 |
| CDRates | 2004 June | 2.3561 |
| Eigenvalue | | 3.3326 |
| Cum. Prop. | | 1.0000 |

Step 2

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | -1.3010 | 0.6667 |
| CDRates | 2004 June | 1.9852 | -1.5729 |
| CDSewerage | 2005 June | 0.7011 | 2.3132 |
| Eigenvalue | | 3.4413 | 2.1492 |
| Cum. Prop. | | 0.6156 | 1.0000 |

Step 3

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 1.4596 | -0.9462 |
| CDRates | 2004 June | -2.3830 | 1.2938 |
| CDSewerage | 2005 June | 3.2853 | 1.6134 |
| CDWater | 2004 December | 1.0178 | 0.1170 |
| Eigenvalue | | 8.3521 | 3.3033 |
| Cum. Prop. | | 0.7166 | 1.0000 |

University of Cape Town

Step 4

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | -2.1796 | -1.0946 |
| CDRates | 2004 June | 3.5651 | 1.5459 |
| CDSewerage | 2005 June | -5.6388 | 1.2116 |
| CDWater | 2004 December | -2.1441 | 0.0575 |
| WaterOthHouse | 2003 September | -2.2705 | 0.2793 |
| Eigenvalue | | 25.4156 | 3.4624 |
| Cum. Prop. | | 0.8801 | 1.0000 |

Final Step

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 15.1533 | -3.2369 |
| CDRates | 2004 June | 34.7452 | 4.0409 |
| CDSewerage | 2005 June | 27.5434 | -7.8891 |
| CDWater | 2004 December | 2.8864 | -2.8856 |
| WaterOthHouse | 2003 September | 3.8720 | -3.8375 |
| SanIndMonetary | 2004 March | 32.9880 | -0.8763 |
| RRIndMonetary | 2005 June | 8.4403 | -0.5560 |
| CDTotal | 2002 June | 12.2777 | 0.4166 |
| CDRefuseRem | 2003 December | 11.7444 | 0.7364 |
| CD60-90days | 2004 December | -5.2285 | 0.3373 |
| CDElectricity | 2004 March | 2.2836 | 0.0804 |
| Eigenvalue | | 1687.8861 | 39.8999 |
| Cum. Prop. | | 0.9769 | 1.0000 |

Factor Structure Coefficients

| | | Function 1 | Function 2 |
|---|---|---|---|
| CDRates | 2004 June | 0.1570 | 0.8849 |
| CDSewerage | 2005 June | -0.1744 | 0.7359 |
| CDWater | 2004 December | -0.1176 | -0.3268 |
| WaterOthHouse | 2003 September | -0.0024 | 0.2482 |

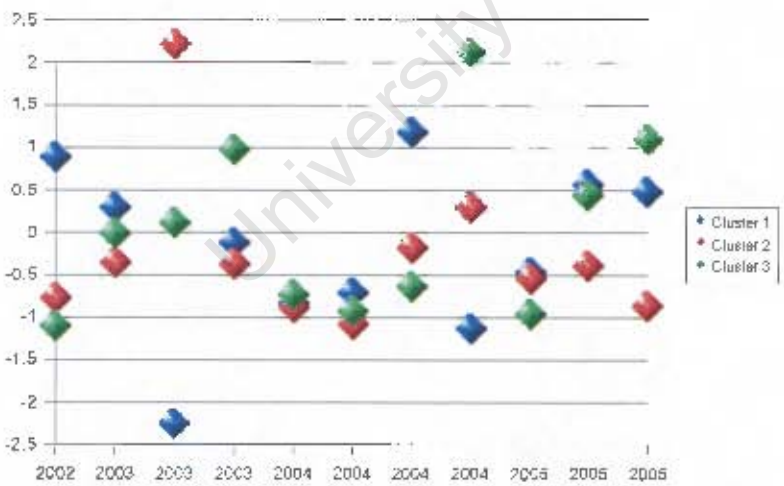



*Illustration VIII.1: Three cluster averages across all significant GDA variables*

University of Cape Town

# Appendix IX

# General Discriminant Analysis: 4-Means Tables

Variables in the model

| | | F to remove | P to remove |
|---|---|---|---|
| CD60-90days | 2003 December | 60.5579 | 0.0000 |
| CDSewerage | 2005 March | 64.1900 | 0.0000 |
| ElecIndMonetary | 2004 March | 29.6049 | 0.0000 |
| CDSewerage | 2002 June | 21.3836 | 0.0000 |
| CD60-90days | 2005 March | 5.7222 | 0.0114 |
| CD60-90days | 2003 March | 14.5358 | 0.0003 |
| Overdraft | 2004 June | 7.7562 | 0.0038 |
| ElecDisconnection | 2004 December | 6.7691 | 0.0064 |

Discriminant Functions

Step 1

| | | Function 1 |
|---|---|---|
| Intercept | | 0.2725 |
| CD60-90days | 2003 December | 2.7210 |
| Eigenvalue | | 1.2192 |
| Cum. Prop. | | 1.0000 |

Step 2

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 0.0639 | 0.2744 |
| CD60-90days | 2003 December | 1.3919 | 2.3419 |
| CDSewerage | 2005 March | 2.6855 | -1.4212 |
| Eigenvalue | | 4.3886 | 0.3315 |
| Cum. Prop. | | 0.9298 | 1.0000 |

Step 3

| | | Function 1 | Function 2 | Function 3 |
|---|---|---|---|---|
| Intercept | | -0.0033 | -2.9590 | 0.2587 |
| CD60-90days | 2003 December | -1.3946 | 0.1117 | 2.3426 |
| CDSewerage | 2005 March | -2.6690 | -0.8276 | -1.4255 |
| ElecIndMonetary | 2004 March | 0.0731 | -3.5686 | -0.0190 |
| Eigenvalue | | 4.3895 | 2.2965 | 0.3314 |
| Cum. Prop. | | 0.6255 | 0.9528 | 1.0000 |

University of Cape Town

Step 4

| | | Function 1 | Function 2 | Function 3 |
|---|---|---|---|---|
| Intercept | | -2.2796 | 2.0838 | -1.3814 |
| CD60-90days | 2003 December | -0.3872 | 1.4693 | 2.0496 |
| CDSewerage | 2005 March | 2.2479 | 2.1372 | -0.8433 |
| ElecIndMonetary | 2004 March | -2.2682 | 2.2915 | -2.0505 |
| CDSewerage | 2002 June | 1.2246 | -0.3511 | -0.5169 |
| Eigenvalue | | 10.4603 | 3.4014 | 1.0515 |
| Cum Prop. | | 0.7014 | 0.9295 | 1.0000 |

Final Step

| | | Function 1 | Function 2 | Function 3 |
|---|---|---|---|---|
| Intercept | | -0.9335 | -5.3927 | -1.9910 |
| CD60-90days | 2003 December | 3.8097 | -5.6613 | 0.3755 |
| CDSewerage | 2005 March | 8.0083 | -2.1068 | -1.4770 |
| ElecIndMonetary | 2004 March | 1.9612 | -5.1942 | -2.4421 |
| CDSewerage | 2002 June | 3.4743 | -0.0219 | 0.5266 |
| CD60-90days | 2005 March | -5.9703 | 3.6608 | -1.0787 |
| CD60-90days | 2003 March | -0.2178 | 4.9137 | -0.6425 |
| Overdraft | 2004 June | 1.0578 | -1.1502 | 0.4819 |
| ElecDisconnection | 2004 December | -1.0055 | 0.1427 | 1.3240 |
| Eigenvalue | | 96.8557 | 22.3744 | 7.7203 |
| Cum. Prop. | | 0.7629 | 0.9392 | 1.0000 |

Factor Structure Coefficients

| | | Function 1 | Function 2 | Function 3 |
|---|---|---|---|---|
| CD60-90days | 2003 December | 0.1354 | 0.4466 | 0.5762 |
| CDSewerage | 2005 March | 0.4246 | 0.6511 | 0.0408 |
| ElecIndMonetary | 2004 March | -0.3586 | 0.3801 | -0.6854 |
| CDSewerage | 2002 June | 0.4243 | -0.1516 | -0.3656 |

University of Cape Town

# Appendix X

## Classification Trees: 2-Means Clustering

Illustration 1 shows the classification tree that is able to successfully classify all municipalities into the two clusters that were identified in the K-Means Cluster Analysis:

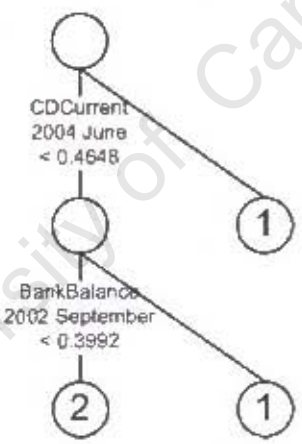



*Illustration X.1: Classification tree for 2 clusters*

University of Cape Town

Although there need only be one split to separate a group in two, there is no one variable that is identified in being able to do this for this particular analysis. Two levels in the tree are required in order to completely split the municipalities into their respective clusters, identifying *CDCurrent* 2004 June and *BankBalance* 2002 September as the most significant variables able to do so. The decision rules that will classify a municipality into a particular cluster are:

**Cluster 1**

*CDCurrent* 2004 June > 0.4648 OR

*CDCurrent* 2004 June < 0.4648 AND *BankBalance* 2002 September > 0.3992

**Cluster 2**

*CDCurrent* 2004 June < 0.4648 AND *BankBalance* 2002 September < 0.3992

As there are only two variables that are necessary to be able to classify municipalities with 100% accuracy when considering the clusters created from the 2-Means Cluster Analysis, it is possible to map these decision rules. In Illustration 2 the most significant split using *CDCurrent* 2004 June shown on the x-axis and *BankBalance* 2002 September on the y-axis. The relevant cut-off points for each is also shown.

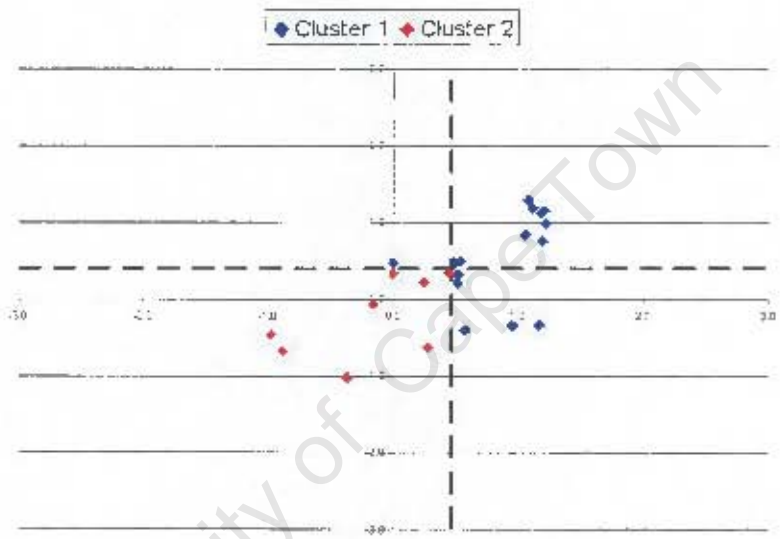



*Illustration X.2: Municipalities plotted on significant variables for 2 clusters*

It is now easy to see the decision rules taken from the classification tree that split the municipalities between the two clusters. Initially, current consumer debtors are seen as most significant in splitting into the two clusters, with Cluster 1 having higher consumer debtors than Cluster 2 for June 2004. However, this does not completely separate the municipalities into the two clusters and *Bitou* would be misclassified as Cluster 2 if only this split was used. This is not that surprising as *Bitou* has the least in common with the rest of Cluster 1, to the point where it splits off from this cluster when more than 2 clusters are considered and forms a separate cluster with *Drakenstein*.

In terms of *CDCurrent* 2004 June specifically, *Bitou* has more in common with Cluster 2, however once this variable is accounted for, the Classification Tree determines that *BankBalance* 2002 September is the most significant in being able to split *Bitou* from the Cluster 2 and classify all municipalities correctly.

University of Cape Town

# Appendix XI

# Classification Trees: 3-Means Clustering

Illustration 1 shows the classification tree that is able to successfully classify all municipalities into the three clusters that were identified in the K-Means Cluster Analysis:

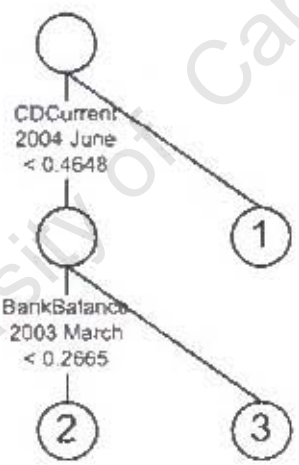



*Illustration XI.1: Classification tree for 3 clusters*

It is only necessary to split twice to be able to separate a group into three and the classification tree is able to do this for this particular analysis. The two levels in the tree are required in order to completely split the municipalities into their respective clusters, again identifying *CDCurrent* 2004 June as the most significant variable to be able to do so, but to split between Cluster 2 and 3 after this *BankBalance* 2003 March is seen as the most significant variable. The decision rules that will classify a municipality into a particular cluster are:

**Cluster 1**

*CDCurrent* 2004 June > 0.4648



University of Cape Town

**Cluster 2**

*CDCurrent* 2004 June < 0.4648 AND *BankBalance* 2003 March < 0.2665

**Cluster 3**

*CDCurrent* 2004 June < 0.4648 AND *BankBalance* 2003 March > 0.2665

As again there are only two variables that are necessary to be able to classify municipalities, but this time into the three clusters, it is possible to map this. In Illustration 2 the variables showing the most significant split, *CDCurrent* 2004 June, is shown on the x-axis and *BankBalance* 2003 March on the y-axis. The relevant cut-off points for each is also shown:

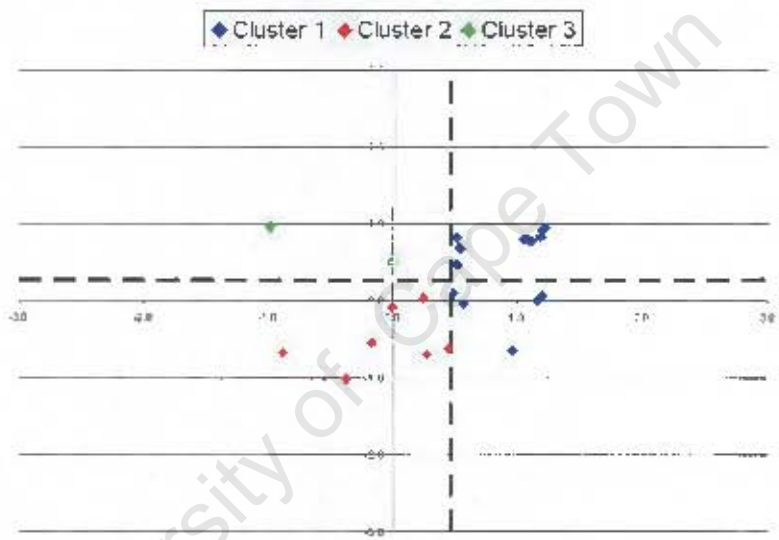



*Illustration XI.2: Municipalities plotted on significant variables for 3 clusters*

University of Cape Town

It is now easy to see the decision rules taken from the classification tree that split the municipalities between the three clusters. Initially, current consumer debtors are seen as most significant in splitting Cluster 1 off from Clusters 2 and 3. There is no longer the problem of trying to then split *Bitou* to classify it as Cluster 1 as there was in the 2 cluster analysis, as it now forms its own cluster with *Drakenstein*. There is obviously still the need to split Cluster 2 and 3 after taking the effect of *CDCurrent* 2004 June into account and it is *BankBalance* 2003 March that most successfully manages this, allowing the classification tree to now classify municipalities with 100% accuracy. The profile of the 3 clusters can then be seen as Cluster 1 having high consumer debtors in June 2004 whereas Clusters 2 and 3 have low consumer debtors in June 2004, but Cluster 2 has below average bank balance in March 2003 whereas Cluster 3 has above average bank balance in this period.

# Appendix XII

## Classification Trees: 4-Means Clustering

Illustration 1 shows the classification tree that is able to successfully classify all municipalities into the four clusters that were identified in the K-Means Cluster Analysis:

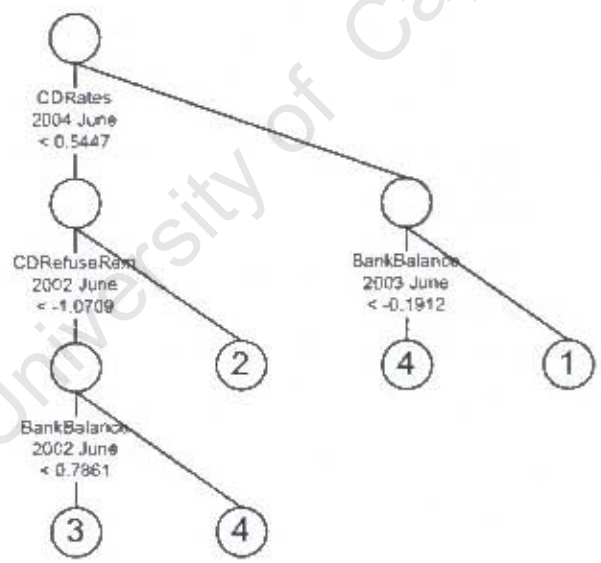



*Illustration XII.1: Classification tree for 4 clusters*

Although there need only be three splits to separate a group into four, no subset of variables are identified in being able to do this for this particular analysis. Three levels in the tree are required in order to completely split the municipalities into their respective clusters, identifying *CDRates* 2004 June, *BankBalance* 2003 June, *CDRefuseRem* 2002 June and *BankBalance* 2002 June as the most significant variables able to do so. This is slightly different to the previous two results where *CDCurrent* 2004 June had been seen as the most significant to create the first split between municipalities, indicating that the creation of the fourth cluster

University of Cape Town

significantly influences the analysis. The decision rules that will classify a municipality into a particular cluster are:

**Cluster 1**

*CDRates* 2004 June > 0.5447 AND *BankBalance* 2003 June > -0.1912

**Cluster 2**

*CDRates* 2004 June < 0.5447 AND *CDRefuseRem* 2002 June > -1.0709

**Cluster 3**

*CDRates* 2004 June < 0.5447 AND *CDRefuseRem* 2002 June < -1.0709 AND
*BankBalance* 2002 June < 0.7861

**Cluster 4**

*CDRates* 2004 June > 0.5447 AND *BankBalance* 2003 June < -0.1912 OR
*CDRates* 2004 June < 0.5447 AND *CDRefuseRem* 2002 June < -1.0709 AND
*BankBalance* 2002 June > 0.7861

There are now four variables that are necessary to be able to classify municipalities, unlike the previous two in both other analyses. This does not mean that the decision rules cannot be represented graphically, just that the variables must be chosen carefully to represent them in the best way. For this particular classification tree it is not necessary to analyse each pair of variables which would create 6 different graphs, but due to the structure of the tree these splits are easy to see using only the two graphs shown on the next page (Illustration 8.6). The first graph shows the variable with the most significant split, *CDRates* 2004 June on the x-axis and *BankBalance* 2003 June on the y-axis, which represents the main right branch in the tree. The second graph shows *CDRefuseRem* 2002 June on the x-axis and *BankBalance* 2002 June on the y-axis, representing the lower left branch of the tree.

It is now much easier to see the decision rules taken from the classification tree that split the municipalities between the four clusters. Considering the first graph which represents the main right branch in the classification tree which splits Cluster 1 from the other clusters and part of Cluster 4, it can be seen that consumer debtors for services relating specifically to rates is the most significant in terms of splitting Cluster 1 from the other clusters. However if only this rule was used *Knysna* would be misclassified as Cluster 1 and therefore once this is taken into account the classification tree chooses bank balance for the period June 2003 as the most significant in being able to split *Knysna* from Cluster 1 and classify it correctly as belonging to

University of Cape Town

Cluster 4.

The remainder of Cluster 4 joins Clusters 2 and 3 in the main left branch of the tree, which is considered in the second graph where *CDRates* 2004 June has already been taken into account. Ignoring Cluster 1 this time, it is seen that Cluster 2 is split successfully from the other two clusters using *CDRefuseRem* 2002 June. Again, it appears that part of Cluster 4 is misclassified if only this rule is used, but once again it is *Knysna* which has already been classified using the other main split of the tree and can therefore be ignored here with Cluster 1. The only split that still needs to occur is between Cluster 3 and the remainder of Cluster 4, for which the classification tree identifies *BankBalance* 2002 June as creating the most significant split once the other two variables are taken into account.

University of Cape Town

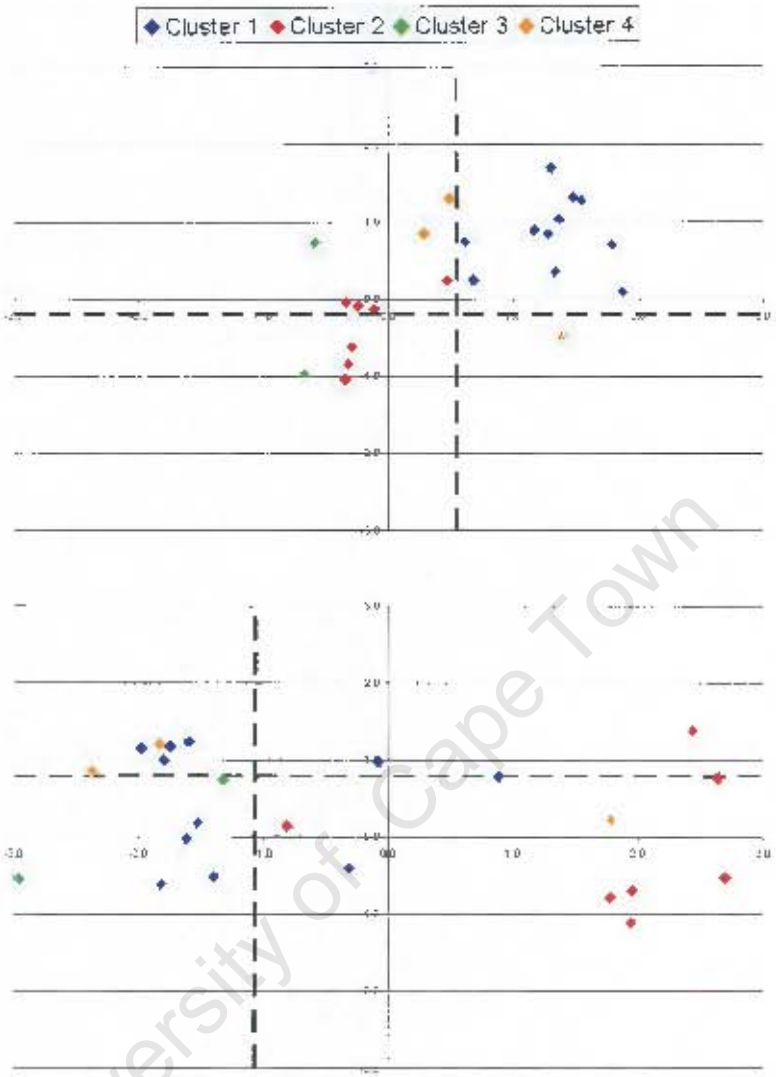



*Illustration XII.2: Municipalities plotted on significant variables for 4 clusters*

University of Cape Town

The profile of the 4 clusters can finally be seen with Cluster 1 having high consumer debtors for services relating specifically to rates in June 2004 and an above average bank balance in June 2003. Clusters 2 and 3 have low consumer debtors for services relating specifically to rates in June 2004, but Cluster 2 has higher consumer debtors for services relating specifically to refuse removal in June 2002 and Cluster 3 is low for this particular variable, along with having a low bank balance in June 2002. Cluster 4 is the most complicated to profile as part of the cluster has high consumer debtors for services relating specifically to rates in June 2004 and a below average bank balance in June 2003, but the other part of cluster has low consumer debtors for services relating specifically to rates in June 2004, low consumer debtors for services relating specifically to refuse removal in June 2002 and a high bank balance in June 2002.

# Appendix XIII

## Classification Trees: Tables

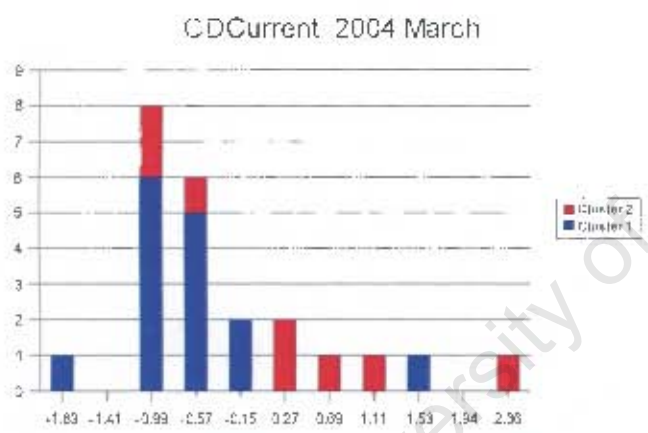



*Illustration XIII.1: Municipal values of CDCurrent 2004 March for 2 clusters*

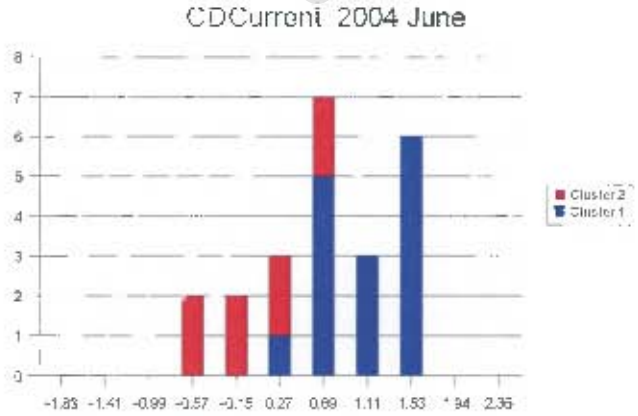



*Illustration XIII.2: Municipal values of CDCurrent 2004 June for 2 clusters*

University of Cape Town

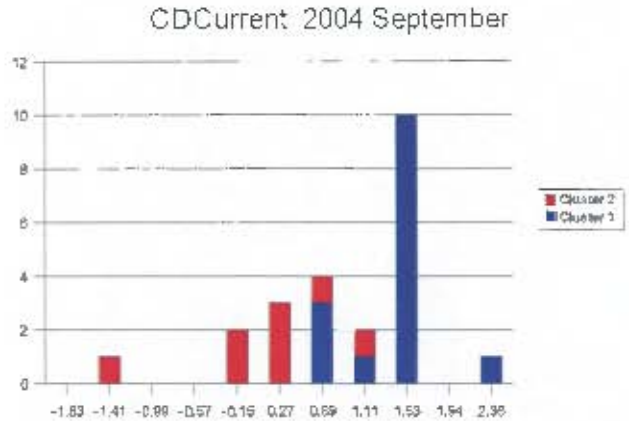



*Illustration XIII.3: Municipal values of CDCurrent 2004 September for 2 clusters*

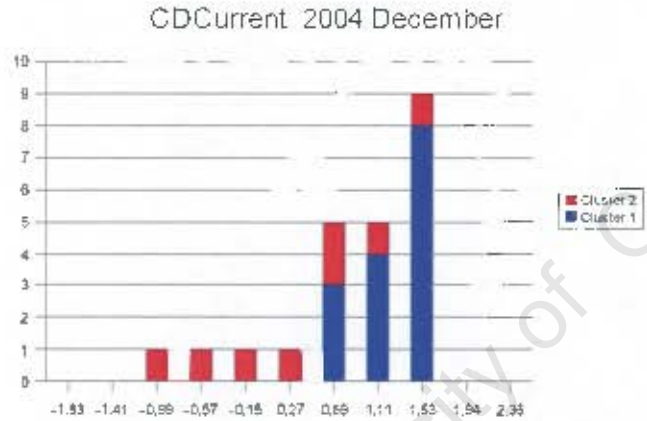



*Illustration XIII.4: Municipal values of CDCurrent 2004 December for 2 clusters*

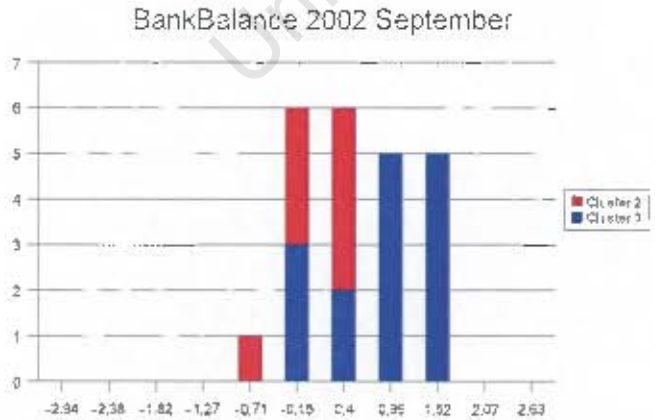



*Illustration XIII.5: Municipal values of BankBalance 2002 September for 2 clusters*

University of Cape Town

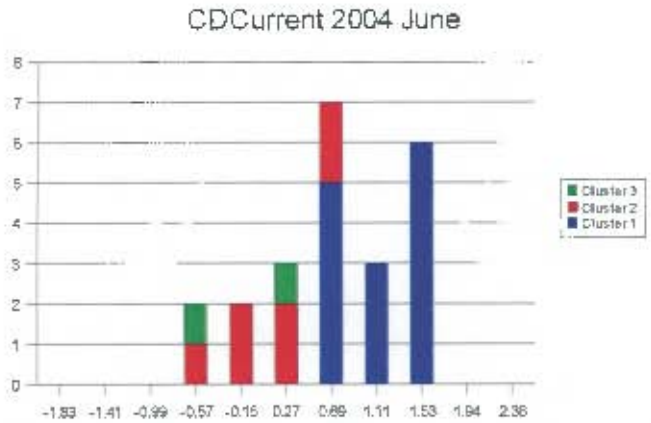



*Illustration XIII.6: Municipal values of CDCurrent 2004 June for 3 clusters*

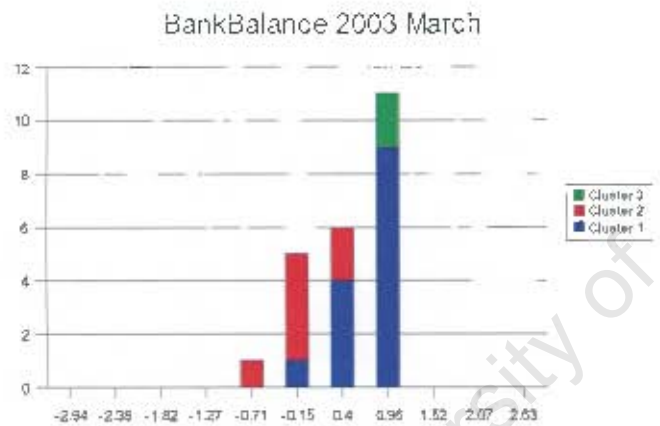



*Illustration XIII.7: Municipal values of BankBalance 2003 March for 3 clusters*

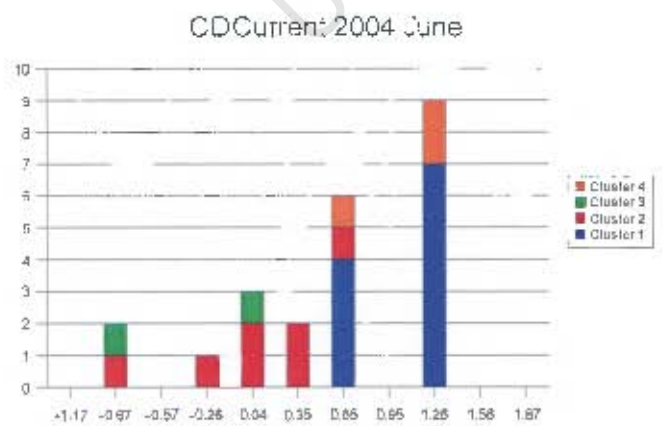



*Illustration XIII.8: Municipal values of CDCurrent 2004 June for 4 clusters*

University of Cape Town

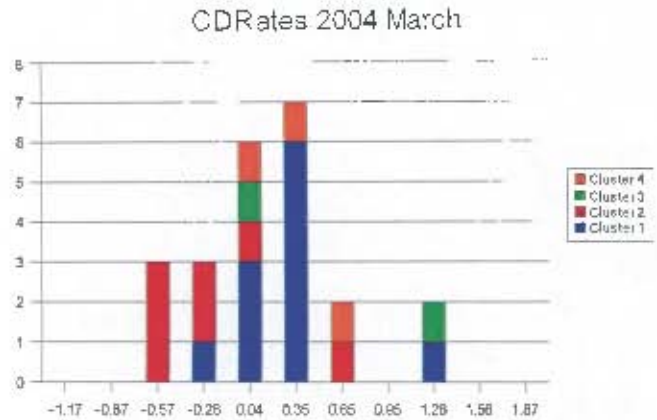



*Illustration XIII.9: Municipal values of CDRates 2004 March for 4 clusters*

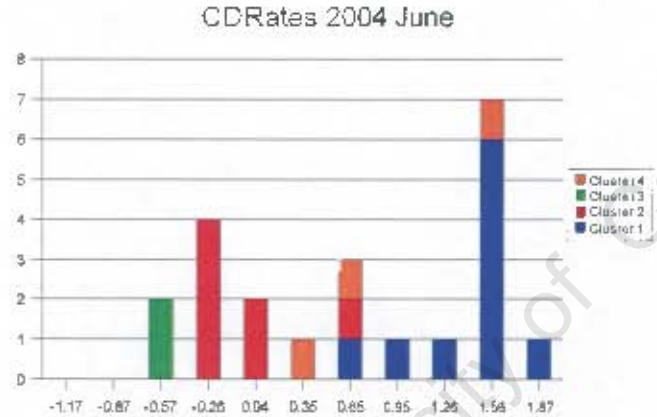



*Illustration XIII.10: Municipal values of CDRates 2004 June for 4 clusters*

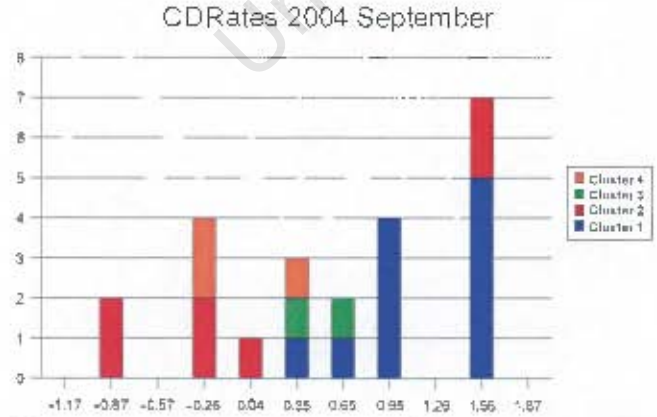



*Illustration XIII.11: Municipal values of CDRates 2004 September for 4 clusters*

University of Cape Town

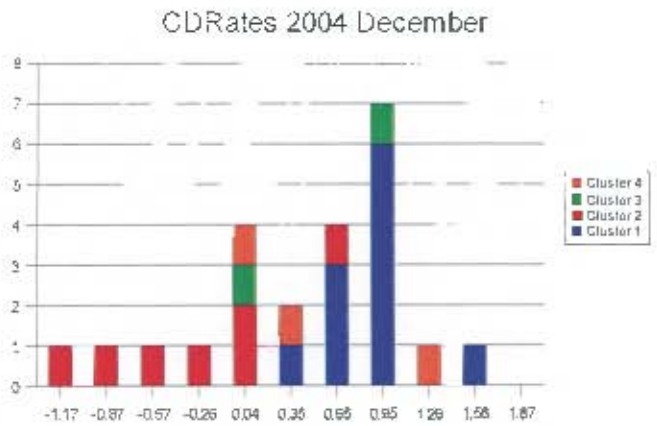



*Illustration XIII.12: Municipal values of CDRates 2004 December for 4 clusters*

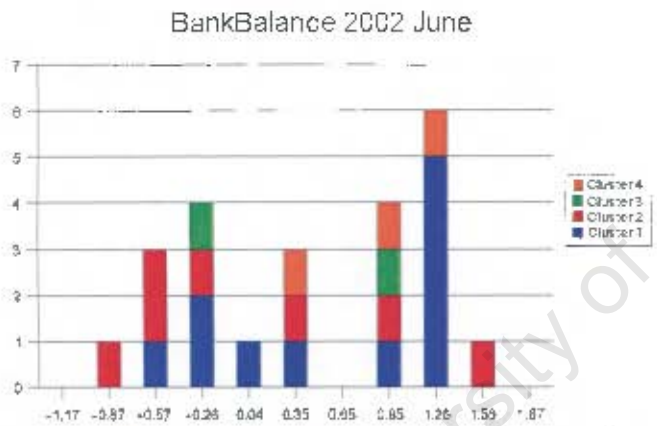



*Illustration XIII.13: Municipal values of BankBalance 2002 June for 4 clusters*

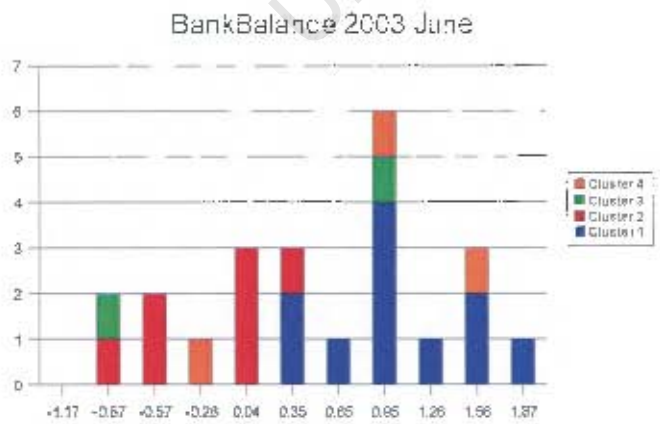
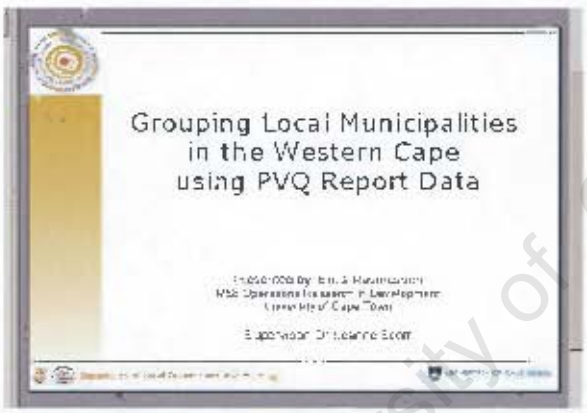


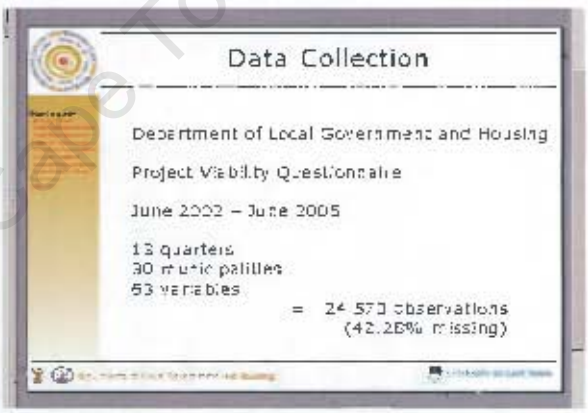

*Illustration XIII.14: Municipal values of BankBalance 2003 June for 4 clusters*

University of Cape Town

# Appendix XIV

# Government Opinion: Feedback Presentation






University of Cape Town









University of Cape Town

## Missing Data

13 quarters
23 municipalities
36 variables
= 10 764 observations
(41.52% missing)

Substitution

- Smaller sections
- Entire variables for a municipality

## Data Management

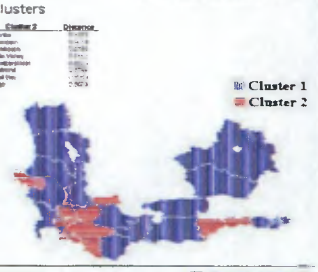
Standardisation
- By Households
- By Population
- 2001 Census

Quarterly

Annualised ?

## Cluster Analysis

Example 2 Clusters




## Cluster Analysis

*Which municipalities naturally group together?*

Can pick the number of clusters beforehand

Indicates which variables show significantly different means across clusters

## Analysis Techniques

Principal Component Analysis

Cluster Analysis

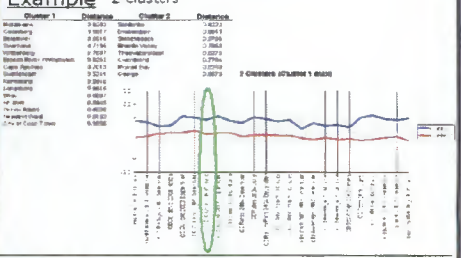
General Discriminant Analysis

Classification Trees

*Examples will use Quarterly Data standardised by Number of Households*

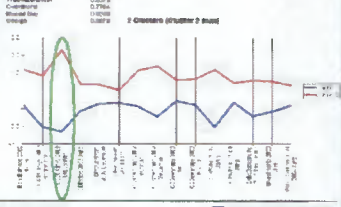
## Cluster Analysis

Example 2 Clusters



## Cluster Analysis

Example 2 Clusters



## General Discriminant Analysis

*Which variables discriminate well between clusters?*

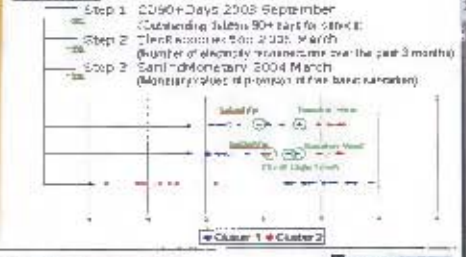
Variables that are best able to discriminate

Variables explain independent information

Can classify new cases/municipalities

## General Discriminant Analysis

Example



## Classification Trees

What levels of variables classify municipalities into a particular cluster?

Variables that are best able to discriminate

Levels at which they discriminate

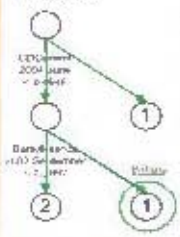
Can classify new cases/municipalities

Can prune tree for 'best' generalisation

## Classification Trees

Example

**Decision Rules:**



## Questions?

## Questions

June 2002 - June 2005

Were there any major events / significant changes?

Are any variables more important than others?

Are any variables more reliable than others?

How would the municipalities actually cluster?

## Exercise

To gather personal opinions as to how the local municipalities in the Western Cape might actually cluster in terms of performance.

## Exercise

University of Cape Town

# Appendix XV

# Government Opinion: Feedback Meeting

This meeting was held
with the Department of Local Government and Housing
on 16 May 2007
at 14h00 until 15h30.

Meeting Participants

| Name | Department / Section |
|---|---|
| Arendse H.W. | Legislation Support (Dir) |
| Harris P. | Government & Integration (Dir) |
| Hoza N. | Municipal Support & Capacity Building (Dir) |
| Ismail M. | Capacity Building (Dir) |
| Jacobs L. | Monitoring & Evaluation |
| Kock G. | Government & Integration (Dir) |
| Riet F. | Municipal Support & Capacity Building (Dir) |
| Shepherd J. | Municipal Support (Dir) |
| Smit C. | Government & Integration (Dir) |
| Welderman P. | Municipal Performance (Dir) |

Group Exercise Participants

| Name | Department / Section |
|---|---|
| Arendse H.W. | Legislation Support (Dir) |
| Jacobs L. | Monitoring & Evaluation |
| Kock G. | Government & Integration (Dir) |
| Shepherd J. | Municipal Support (Dir) |
| Welderman P. | Municipal Performance (Dir) |

University of Cape Town

# Appendix XVI

# Government Opinion: Cluster Analysis

| Good | Average | Bad |
|---|---|---|
| Bitou | Beaufort West | Bergrivier |
| Breede River / Winelands | Drakenstein | Cape Agulhas |
| Breede Valley | Knysna | Cederberg |
| City of Cape Town | Langeberg | Kannaland |
| George | Matzikama | Witzenberg |
| Overstrand | Mossel Bay | |
| Stellenbosch | Prince Albert | |
| Swartland | Saldanha | |
| | Swellendam | |
| | Theewaterskloof | |

*Illustration XVI.1: Three cluster data-defined municipal membership*

| Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|
| Beaufort West | Breede Valley | Bitou |
| Bergrivier | George | Drakenstein |
| Breede River / Winelands | Mossel Bay | |
| Cape Agulhas | Overstrand | |
| Cederberg | Saldanha | |
| City of Cape Town | Stellenbosch | |
| Kannaland | Theewaterskloof | |
| Knysna | | |
| Langeberg | | |
| Matzikama | | |
| Prince Albert | | |
| Swartland | | |
| Swellendam | | |
| Witzenberg | | |


*Illustration XVI.2: Three cluster opinion-defined municipal membership*

University of Cape Town

# Appendix XVII

# Government Opinion: General Discriminant Analysis

Variables in the model

| | | F to remove | P to remove |
|---|---|---|---|
| ElecIndMonetary | 2002 December | 5587.9324 | 0.0000 |
| PMS%Implemented | 2003 March | 859.3432 | 0.0000 |
| WaterOthHouse | 2004 December | 309.5840 | 0.0000 |
| CDTotal | 2004 December | 1711.2278 | 0.0000 |
| BankBalance | 2003 December | 20.1966 | 0.0005 |
| WaterDisconnection | 2005 June | 735.7477 | 0.0000 |
| CDSewerage | 2003 June | 212.7556 | 0.0000 |
| CDCurrent | 2002 December | 300.0439 | 0.0000 |
| RRindMonetary | 2002 December | 185.2205 | 0.0000 |
| RRindMonetary | 2003 June | 59.5803 | 0.0000 |
| CrTotal | 2002 June | 19.0396 | 0.0006 |
| CDTotal | 2004 March | 15.3014 | 0.0013 |

Discriminant Functions

Step 1

| | | Function 1 |
|---|---|---|
| Intercept | | -0.9342 |
| ElecIndMonetary | 2002 December | 2.9687 |
| Eigenvalue | | 1.3575 |
| Cum. Prop. | | 1.0000 |

Step 2

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | -1.9909 | 0.0168 |
| ElecIndMonetary | 2002 December | 4.3114 | 1.0394 |
| PMS%Implemented | 2003 March | -2.1359 | 1.1581 |
| Eigenvalue | | 5.2699 | 0.2072 |
| Cum. Prop. | | 0.9622 | 1.0000 |

University of Cape Town

## Step 3

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | -2.4104 | 0.1698 |
| ElecIndMonetary | 2002 December | 5.0744 | -0.5940 |
| PMS%Implemented | 2003 March | -1.9320 | -1.3283 |
| WaterOthHouse | 2004 December | -0.5552 | 0.8729 |
| Eigenvalue | | 6.6737 | 0.8944 |
| Cum. Prop. | | 0.8818 | 1.0000 |

## Step 4

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 2.9729 | 0.1679 |
| ElecIndMonetary | 2002 December | -7.0067 | -0.6141 |
| PMS%Implemented | 2003 March | 2.4184 | -1.3141 |
| WaterOthHouse | 2004 December | 1.0152 | 0.8688 |
| CDTotal | 2004 December | 1.2979 | -0.0312 |
| Eigenvalue | | 13.7158 | 0.8950 |
| Cum. Prop. | | 0.9387 | 1.0000 |

## Step 5

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 2.9111 | 0.6254 |
| ElecIndMonetary | 2002 December | -7.0070 | -1.1284 |
| PMS%Implemented | 2003 March | 2.7593 | -1.4852 |
| WaterOthHouse | 2004 December | 0.8137 | 1.2042 |
| CDTotal | 2004 December | 1.3429 | -0.0071 |
| BankBalance | 2003 December | -0.4093 | 1.2359 |
| Eigenvalue | | 14.5339 | 2.3195 |
| Cum. Prop. | | 0.8624 | 1.0000 |

## Final Step

| | | Function 1 | Function 2 |
|---|---|---|---|
| Intercept | | 15.8043 | -1.1087 |
| ElecIndMonetary | 2002 December | -116.2142 | 3.4838 |
| PMS%Implemented | 2003 March | 27.5963 | 3.8132 |
| WaterOthHouse | 2004 December | 9.7003 | -2.1248 |
| CDTotal | 2004 December | 32.9813 | -0.7230 |
| BankBalance | 2003 December | -3.2508 | -1.7352 |
| WaterDisconnection | 2005 June | -22.7516 | -0.0745 |
| CDSewerage | 2003 June | 11.3969 | -1.0568 |
| CDCurrent | 2002 December | -27.5007 | -0.8756 |
| RRIndMonetary | 2002 December | -25.4701 | -5.7337 |
| RRIndMonetary | 2003 June | 23.6162 | 9.4208 |
| CrTotal | 2002 June | -3.7123 | -2.7624 |
| CDTotal | 2004 March | -3.8485 | 0.9716 |
| Eigenvalue | | 4141.3047 | 16.2190 |
| Cum. Prop. | | 0.9961 | 1.0000 |

## Factor Structure Coefficients

| | | Function 1 | Function 2 |
|---|---|---|---|
| ElecIndMonetary | 2002 December | -0.2952 | -0.1981 |
| PMS%Implemented | 2003 March | 0.1303 | -0.3207 |
| WaterOthHouse | 2004 December | 0.0248 | 0.3254 |
| CDTotal | 2004 December | 0.0486 | -0.1468 |
| BankBalance | 2003 December | -0.0256 | 0.1898 |

University of Cape Town

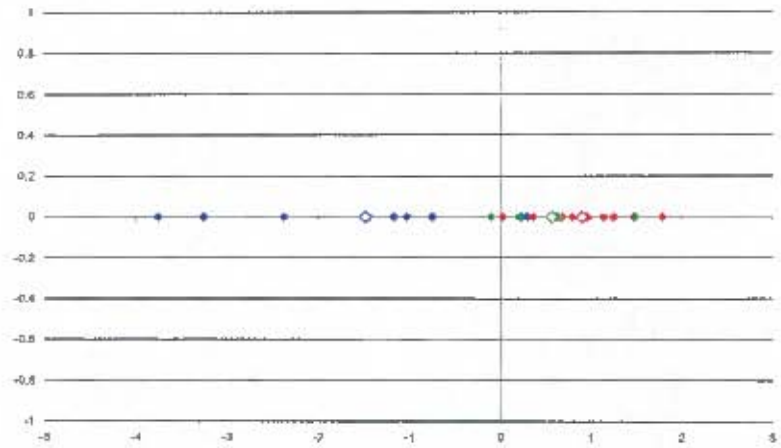


*Illustration XVII.1: Municipalities against first two dimensions in government opinion GDA model containing one variable*

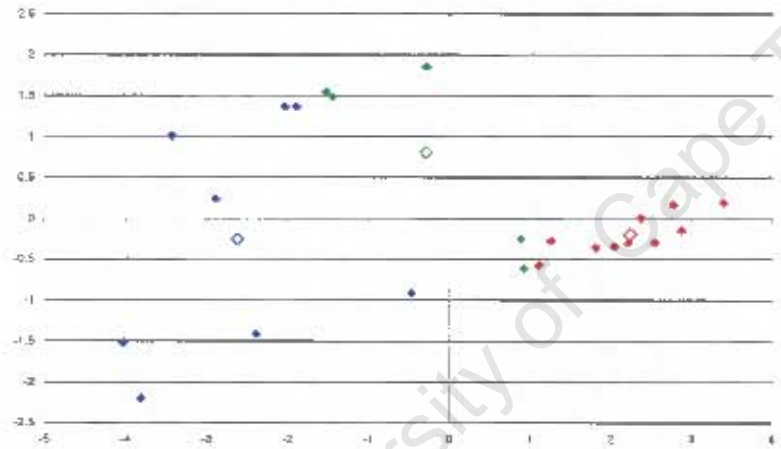


*Illustration XVII.2: Municipalities against first two dimensions in government opinion GDA model containing two variables*

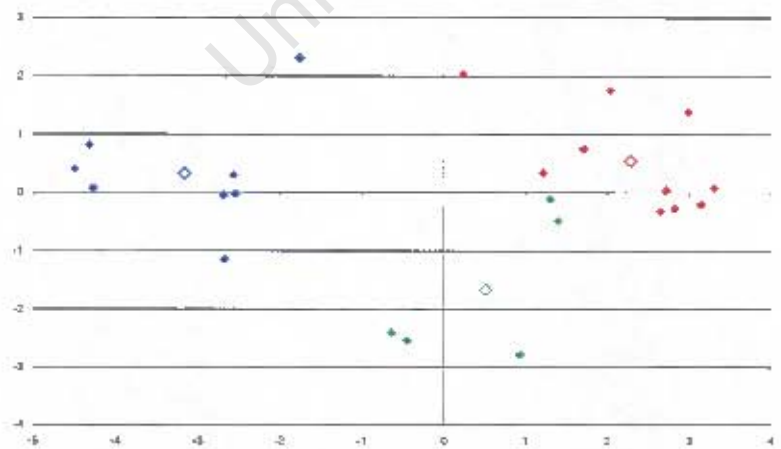


*Illustration XVII.3: Municipalities against first two dimensions in government opinion GDA model containing three variables*

University of Cape Town

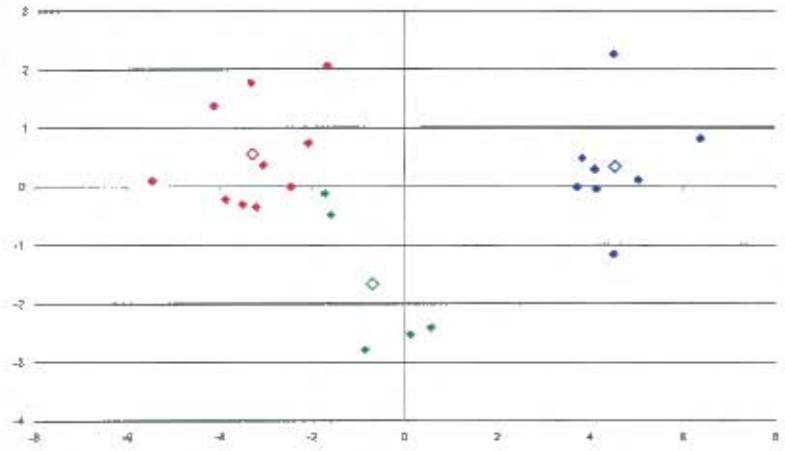


*Illustration XVII.4: Municipalities against first two dimensions in government opinion GDA model containing four variables*

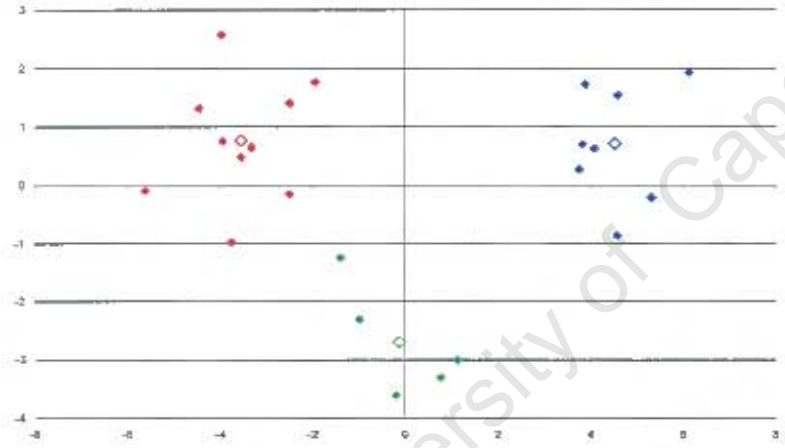


*Illustration XVII.5: Municipalities against first two dimensions in government opinion GDA model containing five variables*

University of Cape Town

# Appendix XVIII

## Government Opinion: Classification Trees

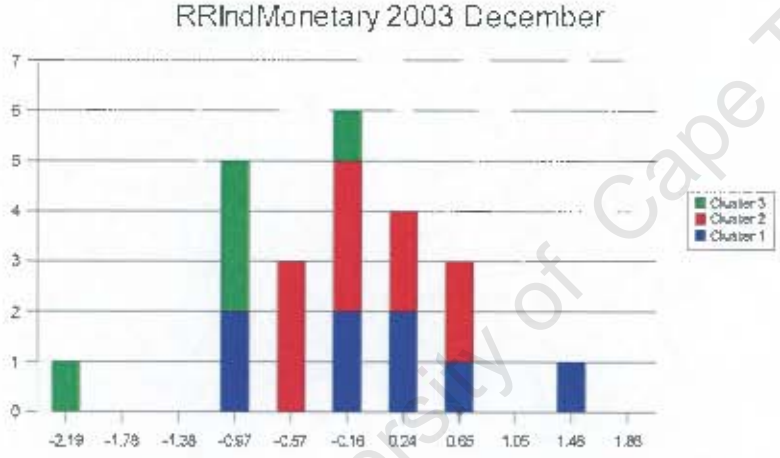



*Illustration XVIII.1: Municipal values of RRIndMonetary 2003 December for opinion defined clusters*

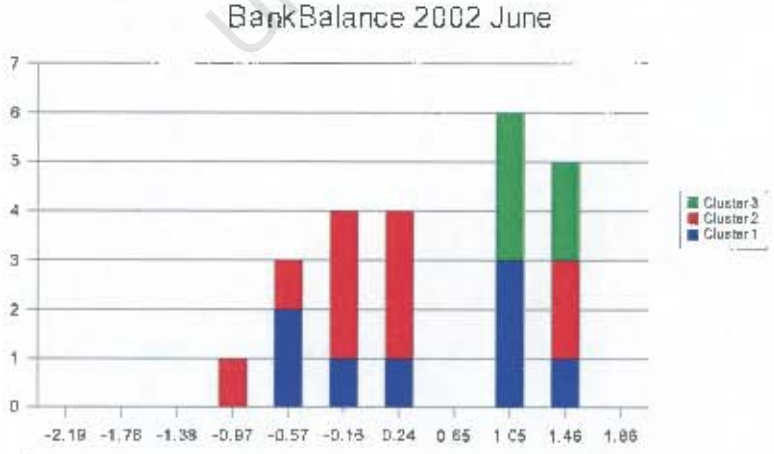



*Illustration XVIII.2: Municipal values of BankBalance 2002 June for opinion defined clusters*

University of Cape Town

The objectives of local government as set out in Chapter 7 of the *Constitution of the Republic of South Africa* (1996) are as follows:

a. to provide democratic and accountable government for local communities;
b. to ensure the provision of services to communities in a sustainable manner;
c. to promote social and economic development;
d. to promote a safe and healthy environment; and
e. to encourage the involvement of communities and community organisations in the matters of local government.

These objectives allow local government to not only act in the role of service delivery to communities, but also play a key role in terms of economic development (Burger 2004). It is noted that these "municipal functions are regarded as critical in improving the quality of life of individual and communities" (Ovens 2005, p.5) within the Western Cape and the country as a whole. In order to be able to carry out these responsibilities laid out by the constitution, the *Local Government Municipal Systems Act* (2000) sets out a framework containing "planning, performance management systems, effective use of resources and organisational change in a business context" (Burger 2004, p. 350). Along with this the Act created a system in which municipalities can report on their performance and allows residents to compare this performance with others (Burger 2004).

In terms of the types of municipalities that can be found within the local government structure, there are three categories of municipalities as laid out by the *Municipal Structures Act* (1998). Category A municipalities may only be formed in metropolitan areas and are therefore referred to as metropolitan municipalities. There are six metropolitan municipalities in South Africa, of which Cape Town is the only one located in the Western Cape.

Both category B and C municipalities form non-metropolitan areas, referred to as local and district municipalities respectively (*Municipal Structures Act* 1998). Of these district councils are "primarily responsible for capacity building and district-wide planning" (Burger 2004, p.349). Each local municipality forms a subdivision of a district municipality, sharing the authority with the district municipality into which they fall.

Finally, there exist District Management Areas, which are areas of a district municipality that do not qualify for a local municipality, therefore requiring local government services to be provided directly by the district municipality itself (*Municipal Structures Act* 1998).

University of Cape Town

# Chapter 5
# Data Collection

The data set was created from data made available through the Department of Local Government and Housing (DLGH) in the Western Cape. It is contained in quarterly Project Viability Questionnaires (PVQ) and is restricted to the period of June 2002 through to June 2005. Those reports dating after December 2003 were available electronically; however the remaining reports required data capturing from printed reports. The final set of data was the most recent available at the time of constructing this particular collection of data and the June 2002 report was the earliest report available from the department. Prior to this the municipalities in the Western Cape numbered 136 and the questionnaire was developed to coincide with the amalgamation of these into 5 District Municipalities, 24 local municipalities and the City of Cape Town, totalling 30 municipalities. Any data collected before this amalgamation cannot be used due to the significant changes in restructuring the system. There were further changes made to the questionnaire in December 2003; however there being much in common between the two it was possible to extrapolate observations throughout the period to create a complete data set.

The initial data set chosen from these reports included all 30 municipalities and 63 variables across all 13 quarters, resulting in 24 570 observations. There were two main types of variables apparent in the data: those that were taken from the financial statements of a municipality and are mainly measured in Rands (henceforth referred to as *financial type* variables) and those that directly represent the service delivery of a municipality measured mainly in terms of counts or percentages (henceforth referred to as *service delivery type* variables).

## 5.1 Missing Data

There was a substantial amount of data missing from the original data set. This is due to various reasons: including some sections not always being compulsory to complete; a learning period as municipalities adapted to the new system; and sometimes data simply not being available. Various techniques were utilised to create a complete data set, keeping in mind the large amount of data and

University of Cape Town