A new approach for extreme values in Data Envelopment Analysis

by

Thayendran Arulsivanathan Naidoo

A dissertation submitted in fulfilment of the

requirements for the degree of

Masters of Applied Science (Industrial Engineering)

in the

Faculty of Engineering, Built Environment and

Information Technology

University of Pretoria

January 2015

I would like to thank Lord Muruga, whose many blessings have made me who I am today and who has guided me throughtout my life.

Acknowledgements

I would like to express my deepest appreciation and gratitude to my supervisor Prof Yadavalli for his continued support and guidance throughout my studies. I am thankful for his aspiring guidance, invaluably constructive criticism and friendly advice during the project work.

I would like to convey my gratitude and thanks to my parents Dr. Arulsivanathan Naidoo and Mrs Holly Naidoo for their guidance and support. I am sincerely grateful to them for their love and support during my studies.

A special thanks to my fiancé Karusha Padayachee for all her love, support and understanding during my studies.

To my family Deneshree, Preshnee, Reagan ,Tharushlan and Preshaylan , I thank you for all your support.

I would like to offer my thanks and appreciation to Suben Moodley, for the all the support you have given me during my studies. I value the truthful and illuminating views that he has provided on a number of issues related to the thesis.

Executive Summary

Title: A new approach for extreme values in Data Envelopment Analysis

Author: Thayendran Arulsivanathan Naidoo

Supervisor: Professor V.S.S Yadavalli

Department: Industrial and Systems Engineering

University: University of Pretoria

Degree: Masters of Applied Science (Industrial Engineering)

Data Envelopment Analysis (DEA) is methodology for relative performance measurement and has been extensively utilised over the past few decades. DEA is however sensitive to the presence of outliers in the data and can cause inaccurate reflections of the relative efficiency score and the projections of inefficient Decision Making Units (DMU) onto the efficient frontier. Stochastic frontier analysis can accommodate for the statistical noise but makes certain assumptions on the data. This dissertation introduces an approach to accommodate for outliers in a DEA model without removing observation that would otherwise affect the results. The results on the proposed model are compared to two deterministic and three stochastic models, and have shown an increase in the efficiency score and the number of efficient DMUs and an increase in the overall efficiency scores.

Keywords: Data Envelopment Analysis, Stochastic Frontier Analysis, Outliers, Extreme values, Municipalities

Contents

List of Figures

List of Tables

List of Acronyms

AE	Allocatively Efficient
BCC	Banker Charnes Cooper
CCR	Charnes Cooper Rhodes
CMIP	Consolidate Municipal Infrastructure Programme
CRS	Constant Return To Scale
DEA	Data Envelopment Analysis
DEA-BCC	Data Envelopment Analysis - Charnes Cooper Rhodes
DEA-CCR	Data Envelopment Analysis - Banker Charnes Cooper
DFA	Deterministic Frontier Analysis
DM	District Municipalities
DMU	Decision Making Units
DQAT	Data Quality Assessment Team
DRS	Decreasing Returns Of Scale
EE	Economic Efficiency
FDH	Free-Disposal Hull
IID	Independent And Identically Distributed
IRS	Increasing Returns Of Scale
LGOI	Local Government Output Indicator
LM	Local Municipalities
MAD	Median Absolute Deviation

MM	Metro Municipalities
NSS	National Statistical System
RTS	Returns To Scale
SASQAF	South African Statistical Quality Assessment Framework
SFA	Stochastic Frontier Analysis
SFA-EXP	Stochastic Frontier Analysis Exponential
SFA-HALF	Stochastic Frontier Analysis Half-Normal
SFA-TRUN	Stochastic Frontier Analysis Truncated-Normal
TE	Technical Efficiency
VRS	Variable Return To Scale

Chapter 1

Introduction

1.1 The importance of DEA and Performance measurements

Data Envelopment Analysis (DEA) is methodology of measuring the relative efficiency of Decision Making Units (DMU) (Charnes et.al. 1978). This method has been widely applied to various industries including banking, hospital, libraries and municipalities. The core concept of DEA is that it provides an efficiency score relative to the remaining DMUs in the reference set.

The need for identifying and eliminating outliers in DEA was pointed out by Simar (1996). If an outlier or extreme value exists in the data, it can influence the efficiency score of the other DMUs. Simar (1996) stated that one approach to dealing with an outlier is to make use of a stochastic frontier production model, which was introduced independently by Meeusen and Van den Broeck (1977) and Aigner et.al. (1977). These production models accommodated for the noise in the data. It calculated an efficiency score and an error term to model random noise. Various authors have developed models which can in the process of calculating the efficiency score identify outliers and influential observations (Andersen and Petersen 1993, Wilson 1995 , Simar 1995 , Johnson and McGinnis 2008, Banker and Chang 2006) to name a few.

1.2 Problem Statement

Public sector administrative data around municipal development and service delivery is not readily available. There are several municipalities, especially in rural areas in South Africa, which do not have accurate record keeping. In an attempt to measure the relative efficiency of municipalities, the need for accurate administrative data is essential. The need for quality assessment has also increased in South Africa. Statistics South Africa is currently the sole supplier of official statistics and has embarked on a venture to allow other organisations within South Africa to produce official statistics by means of the South African Statistical Quality Assessment Framework (SASQAF). The demand for quality statistics that can be trusted lead to SASQAF which enables self-assessment for the data producers reviewed by the Data Quality Assessment Team (DQAT). This is to be done in the context of the National Statistical System (NSS) to ensure the integrity and certification of the statistics is as official as stipulated in the Statistics Act (Act No. 6 of 1999). This framework provides a clear and transparent procedure for the evaluation of official statistics and other administrative data (SASQAF 2010).

Until local municipalities have adhered to SASQAF in the process of delivering data regarding their levels of output and operations, the need for robust techniques to accommodate for extreme values is vital. Performance measurements regarding municipal efficiency will be prone to extreme values in the data. Using techniques such as DEA, which is highly sensitive to outliers. The results will yield inaccurate efficiency scores and lead to an efficient frontier which is not a true reflection of reality. The problem statement can be summarised as follow:

As a result of inaccurate administrative data or data with extreme values, the resulting efficiency scores of the DMUs in a relative efficiency analysis will be affected. Since the DEA methodology is based on optimising the weighted sum of outputs over the weighted sum of inputs, the outlier can cause the efficiency score of a DMU to be over or under stated.

1.3 Research Methodology

The aim of this dissertation is to explore alternative approaches to accommodate for outliers in a DEA analysis. This will be accomplished by evaluating the existing methodologies around outlier and extreme values in an efficiency analysis. In order to evaluate all aspects of the efficiency analysis and the effect that outliers can have on the analysis, various factors need to be addressed:

- The impact of outliers in DEA
- The impact of the outlier detection models, with specific considerations to the :
    - Advantages
    - Disadvantages
- The implications of the various approaches on the sample set.

1.4 Research Objectives

The objective of the study is to:

- Calculate the efficiency scores of local municipalities using a DEA and stochastic frontier analysis.
- Propose a new method to accommodate for extreme values in a DEA model.
- Compare the results of the proposed model to the conventional methods of efficiency analysis.
- Analyse the results of the DMUs which had extreme values and compare the results to the proposed model

1.5 Importance of the Research Problem

Given the current state of data quality and the lack of proper data collection procedures in local governments, the need for a robust methodology that can correct and accomodate for outliers and extreme values is essential. Relative efficiency methodologies are powerful techniques that can provide a single measure of efficiency for a homogenous set of DMUs. However when the set of DMUs is not homogenous and outliers or extreme values are present, the results of a DEA analysis can reflect inaccurate measures of relative efficiency. These results can be led to the development of policy and decisions that may not have the

desired effect. The construction of a production frontier based on inaccurate efficiency scores will result in inaccurate projections of inefficient DMUs onto the efficient frontier. These projections influence the areas which a DMU should increase or decrease input and outputs. Inaccurate scores will lead to inaccurate recommendations for a DMU to become efficient. A relative efficiency methodology with outlier or extreme value correction as the main focus can help alleviate these problems.

1.6 Limitations and assumptions of the study

This dissertation aims to calculate the relative efficiency of municipalities and propose a new method that accommodates extreme values in the data which would otherwise reflect inaccurate results. In order to achieve this certain assumptions are made on the data and there exists some limitations of this approach:

1. The results of the efficiency analysis will not conclusively show the magnitude of the improvements a DMU should make in order to become efficient. This study will be comparing actual units or input/outputs with other actual units of inputs/outputs rather than comparing these actual units with industry standards or benchmark values.
2. This study will be able to identify inefficient DMUs given the predetermined mix of inputs and outputs used in each model.
3. The analysis of the municipalities is based on discretionary variables. The non-discretionary factors each municipality is faced with cannot be modelled and incorporated into the analysis.
4. The DEA methodology is unlikely to uncover all inefficiencies in the DMUs as the approach is very sensitive to the mix of inputs and outputs considered in the analysis.

1.7 Concluding remark and Scope of the work

This study proposes an approach to accommodate for extreme values in the input and output data of a DEA model. This research applies the approach to three populations of municipalities in South Africa and compares our results to two conventional DEA models and three Stochastic Frontier Analysis models. The aim of this study is to fully understand the effect of outliers and extreme values in a DEA analysis and propose a new approach to

accommodate for extreme values in the analysis. An extensive literature review has been undertaken into the theory and applications of DEA and outlier analysis.

In summary, DEA is a powerful tool for the measurement of relative efficiency and has been widely applied to various sectors. The sensitivity of DEA to outliers has been one of the limitations of this approach. This study aims to address this problem in the context of South African Local Municipalities.

This dissertation is structured as follows:

Chapter 1 is an introduction into efficiency analysis and provides a brief description on the importance of performance measurement and the limitation associated with a relative efficiency analysis.

Chapter 2 provides a review of the literature and addresses the various approaches taken to deal with outliers and extreme values in an efficiency analysis. A comprehensive literature review is conducted on the outlier detection and elevation techniques. This section also provides a review of the research conducted into local municipality efficiency analysis via deterministic and stochastic approaches.

A detailed review of relative efficiency analysis, conventional DEA models and the Stochastic Frontier Analysis approaches are discussed in Chapter 3.

The proposed approach to accommodate for extreme values in a DEA analysis is discussed in detail in Chapter 4. The data used for this research is also addressed in this chapter. A motivation is given as to why this approach is better, along with the rationale for the proposed model.

Chapter 5 contains the results of the comparison and an interpretation of the findings.

Chapter 6 contains concluding remarks and discusses future research. The conclusion focuses on the results of the proposed model when compared to the conventional models.

Chapter 2

Literature Review

2.1 Introduction

DEA has been widely adopted in the measurement of relative efficiency in both the public and private sectors. The applications of DEA include banking, hospitals, libraries, fisheries, credit markets, investment portfolios and hedge funds. The application of DEA into public service delivery and local government efficiency is also extensive. Numerous studies have been conducted to assess the efficiency of local municipalities and understanding the determinants that drive efficient municipalities Worthington and Dollery (2001), Balaguer-Coll et al.(2002) ,Loikkanen and Susiluoto (2005), Afonso and Fernandes (2006) to name a few. Studies into local government efficiency analysis incorporated not only DEA but SFA. A review of the literature according to each methodology is given in this chapter along with a review of the studies into the effect and correction of outliers in a DEA analysis.

2.2 Municipalities

Government collects revenue in the form of taxes and use the money to provide services to its citizens and improve infrastructure that will benefit the lives of all South Africans. There are three levels of government in South Africa which are governed by the rules set out in the Constitution:

1. National government
2. Provincial government
3. Local government.

This thesis will focus on Local government, which is an important sphere of government and is the closest to the people they serve. Basic service delivery is provided to all South Africans through this channel. Local governments have a clear policy that municipalities and councillors should be sensitive to the communities they serve and be responsive to local problems.

Government is made up of three parts:

1. The elected members who represent the public. These members are responsible for approving laws and policies.
2. The cabinet or executive committee. This component is responsible for creating laws and policies as well as the overseeing of the implementation of government departments.
3. The departments and civil servants. This component is responsible for doing the work of government.

Local government receive grants and some loans through the National Treasury.

There are three different types of municipalities in South Africa:

- Category A: Metropolitan municipalities:

  There are eight metropolitan municipalities which are the biggest cities in South Africa. They have in excess of 500 000 voters. These municipalities are responsible for the co-ordination of the delivery of services to the entire area.

- Category B: Local municipalities:

  All other municipalities that do not fall into the Category A are divided into local municipalities. There are a total of 231 of these local municipalities in South Africa.

- Category C: District municipalities:

  District municipalities are made up of a number of local municipalities that fall in one district. A district a municipality ranges from four to six local municipalities that come together in a district council.

The responsibility of all local services, development and service delivery in the metropolitan area reside with the metropolitan municipalities. Local municipalities share that responsibility with district municipalities. District municipalities will have more responsibility for development and service delivery which are very rural areas.

Municipalities are responsible for the following functions:

- Electrification
- Water for household use
- Sewage and sanitation
- Storm water systems
- Refuse removal
- Fire fighting services
- Municipal health services
- Decisions around land use
- Municipal roads
- Municipal public transport
- Street trading
- Abattoirs and fresh food markets
- Parks and recreational areas
- Libraries and other facilities
- Local tourism

Income received by municipalities is used for the daily operations and to provide all citizens with basics services. Municipalities receive money from various sources:

1. **External loans –** Loans from a bank or other financial institutions. This source of income is very expensive because of the high interest rates in South Africa.

2. **Internal loans –** There are many municipalities which have an internal "savings funds" such as Capital Development Funds or Consolidated Loan Fund. These funds can be made available to other municipalities in the form of an internal loan. These loans are generally used for the purchase or development of capital items. This is a more feasible option as they normally have lower interest rates than a bank or other

financial institutions. The municipality is also paying the interest back to its own "savings fund", which can be utilised at a later date.

3. **Contributions from revenue –** If a municipality is purchasing a small capital item, it can be funded through the operating income in the year of purchase. This source of funding is known as "contributions from revenue". There is no interest is payable, this source of financing is considerably cheaper than external or internal loans.

4. **Government grants –** Grants for infrastructure development are available to municipalities from the National government. There are two main funds available for municipalities to apply for:
   - Consolidate Municipal Infrastructure Programme (CMIP) – available from the Department of Provincial and Local Government
   - Water Services Projects – available from the Department of Water Affairs.
5. **Donations and public contributions -** Disadvantaged areas may receive donations from local and foreign donors. These can be in the form of a capital item or money to be used specifically for the purchase of a capital item.

6. **Public/Private Partnerships -** Partnerships between the private sector and the municipality can exist to fund capital costs. Most instances of these partnerships involve the private sector partner having a profit motive.

7. **Property Rates -** All people and businesses who own fixed property (land, houses, factories, and office blocks) in the municipal area are charged "Property Rates" - a yearly tax based on the value of each property. This rate's income is used for the payment of all general services to the inhabitants. These include the "service charge" for roads, pavements, parks, streetlights, storm water management, etc.

8. **Service Charges / Tariffs –** Certain services are directly charged to a household or business. This charge or "tariff" is for services such as water, electricity or approval of building plans. This is only used in instances where the exact usage of the service can be measured and allocated to a household or business.

9. **Fines** -Traffic fines, penalties for overdue payment of service charges and late library book fines are some examples of fines charged to the public. This is seen more as motivation for the users of services to have a culture of obeying democratic laws, rules and deadlines.

10. **Equitable share** - The equitable share is an amount of money that a municipality gets from national government each year. This is calculated using the Equitable Share formula. This formula relies, in part, on the number of low-income people in the area; therefore more rural municipalities usually get a higher amount. The Constitution of South Africa says that all revenue collected nationally must be divided equitably between national, provincial and local spheres of government. The main purpose of the local government equitable share is to ensure that municipalities can provide basic service and develop their areas.

The income a municipality receives must go towards service delivery for its inhabitants. These services include:

- Water supply
- Sewage collection and disposal
- Refuse removal
- Electricity and gas supply
- Municipal health services
- Municipal roads and storm water drainage
- Street lighting
- Municipal parks and recreation

2.3 Challenges with public service Data

Since the first democratic elections in 1994, the South African government has a strong mandate to transform society. The provisions of services, poverty alleviation and economic development were among the issues identified. The new government, however, inherited a statistical void with regards to reliable information suitable for benchmarking and monitoring

progress in service delivery. The statistics gathered at the time also had little value as a basis for decision making, development of policies and social transformation (SASQAF 2010).

Public sector data around municipal development and service delivery is not readily available. There are several municipalities, especially in rural areas in South Africa, which do not have accurate record keeping. The alignment and evaluation of the data collected by municipalities require a rational, transparent and sustainable framework for assessing the quality of the data. In an attempt to measure the relative efficiency of municipalities, the need for accurate administrative data is essential. The need for such data has lead to the development of the South African Statistical Quality Assessment Framework (SASQAF). The need for quality assessment has increased in South Africa. Statistics South Africa is currently the sole supplier of official statistics and has embarked on a venture to allow other organisations within South Africa to produce official statistics by means of SASQAF.

Within the context of the NSS, SASQAF makes the distinction between national and official statistics. National statistics refers to data in the public domain which has not been certified by the Statistical General as official. Official statistics is data in the public domain which has been certified as official in terms of Section 14(7) of the Statistics Act (Act No. 6 of 1999). The demand for quality statistics that can be trusted lead to SASQAF enabling the self-assessment for the data producers reviewed by the Data Quality Assessment Team (DQAT). This framework provides a clear and transparent procedure for the evaluation of official statistics and other data (SASQAF 2010).

The term “data quality” is defined by Statistics South Africa as “fitness for use”. This can be further defined in terms of the following dimensions:

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Interpretability
- Comparability and Coherence
- Methodological Soundness
- Integrity

2.4 Efficiency Analysis into Municipalities

The applications of efficiency analysis into municipalities have been conducted over a long period of time and across the globe. There have been studies that investigated service delivery as a whole and studies that focused on the delivery of specific services.

Pina and Torres (2000) compared the efficiency of public and private sectors with regard to urban transportation services. The research showed the results of the empirical dissertation commissioned by the Regional Audit Office of Catalonia (Spain) with the intention of evaluating the efficiency of urban transportation services. The input-output variables were based on the resources needed to carry out the delivery of services. The input variables used in the study were the fuel/100km, cost/km or cost/traveller and subsidiary/traveller. The output variables measured the yield of level of activity of programs and services. These outputs were bus-km/employee, bus-km/year, bus-km/inhabitant, accident rate/frequency and population. Pina and Torres (2000) used these inputs and output in various combination to produce four DEA models. A regression analysis was then performed with efficiency score, the independent variables km/bus, cost/traveller and fuel/100km and non-discretionary variables beyond control of the urban transportation companies such as geographical extensions of the city, population density, the number of cars, income per capita and age of the population. These variables were grouped as the environmental variables.

Pina and Torres (2000) found that no DMU's urban transportation service and public management delivery in the larger cities were more prevalent and that population was not a relevant factor to the urban transportation service. Pina and Torres (2000) came to the conclusion that the urban transport service is organized based on actual demand rather than potential demand. The results of the regression analysis shows that the independent variables; km/bus, cost/traveller and fuel/100km were statistically significant in explaining the efficiency behaviour of urban transport services. The regression analysis also showed that environmental variables have not been significant in any of the models.

Afonso and Fernandes (2008) assessed the relative efficiencies of local municipalities with respect to municipal service delivery using DEA and parametric analysis. In the paper they evaluated and analysed the public spending efficiency of 278 mainland Portuguese municipalities based on 2001 data. The municipalities were split into five clusters. The construction of a Local Government Output Indicator (LGOI) was developed by Afonso and

Fernandes (2008) as a single measure of municipal performance. The relevance of environmental variables and non-discretionary input were also addressed in the study by performing a Tobit analysis with the aim of trying to explain the inefficiency scores. Since the outputs of the DMU being evaluated were a result of discretionary and non-discretionary input, Afonso and Fernandes (2008) applied a two-stage analysis to evaluate the efficiency. The first stage used analysed the controllable input and the second stage analysed the uncontrollable input. The use of a regression model was utilised by Afonso and Fernandes (2008) to determine the impact of environmental variables to local governments' performance. This allowed them to include non-discretionary input in the explanation of efficiency scores.

The input variable used in the DEA model was the total municipal expenditure per inhabitant. The outputs used when creating the composite LGOI were social services, school buildings per capita, education enrolment, cultural services, water supply, waste collection, territory organisation and road infrastructure. The non-discretionary variables used in the study were purchasing power, population with secondary education, population with tertiary education, distance to capital of district, population density and population variation. During the second stage of the analysis a correlation analysis was used to uncover and any relationships between the non-discretionary variables. A Tobit regression model applied to analyse the environmental variable and the efficiency scores Afonso and Fernandes (2008).

The result of the analysis shows that the socio-economic factors: level of education, the purchasing power of per capita income and the wealth of citizens , all had a positive influence on the efficiency scores Afonso and Fernandes (2008). The level of education was found to be a strong positive influence in the efficiency of the local municipalities as higher levels of education, regardless of secondary or tertiary education, would allow citizens to apply more pressure on local governments to improve on the level of service delivery and that they would be able to monitor the progress of their local municipality. Furthermore, Afonso and Fernandes (2008) argued and proved that the wealthier or higher income areas would make more use of the local governments' purchasing power. The premise of this argument was that wealthier areas contributed more to the local governments' income and in return they would expect a higher level of service delivery. Another argument made by Afonso and Fernandes (2008) was that inter-municipal competition would provide the citizens a greater choice and opportunity to move from one jurisdiction to another if they felt that they could gain a higher level of service delivery. This was represented by the geographical distance between the

municipality its capital of district. Afonso and Fernandes (2008) also suggested that the results should be taken into perspective for two reasons. Firstly, that some municipalities did not fall on the production frontier and that being labelled as inefficient did not mean that they could not reach the production frontier. Secondly, the socio-economic factors discussed in the paper could be possible constraints imposed on the municipality from moving towards the production frontier.

Boetti et.al. (2010) wrote a paper in which they assessed the spending efficiency of a sample of Italian municipalities. They investigated the effects of tax decentralisation. The study exploited both parametric and non-parametric method in the form of a stochastic frontier analysis (SFA) and a DEA analysis respectively to study the spending inefficiency and its main determinates. Another aim of the paper was to assess the impact of fiscal decentralisation on local spending efficiency and selected outputs that are proxies for services provided by local governments. These outputs were number of inhabitants, the total length of municipal roads, the amounts of waste collected, the sum of the number of pupils enrolled in nursery, primary and secondary schools and the number of people over age 75.

The inputs used by Boetti et.al (2010) were represented by disaggregated current expenditure in general administration, road maintenance and local mobility, garbage collection and disposal, education, elderly care and other social services. The non-parametric section of the study involved a two-stage analysis. Firstly, a DEA analysis followed by a Tobit regression model; whereas in the parametric section a SFA was used and included the explicative factors for inefficiency directly in the frontier model. This approach was proposed by Battese and Coelli (1995). This enables them to consider the ratio of municipal taxes on current expenditure as a measure of fiscal decentralisation. The results showed that more autonomous municipalities showed less inefficient behaviour. According to Boetti et al. (2010) an autonomous municipality is a municipality with higher share of current spending covered by own taxes. Another factor which influenced spending efficiency was the strictness of the municipalities' budget Boetti et al. (2010). The recent global waves of reform towards tax decentralisation and the movement of municipalities towards a more autonomous existence to improve efficiency and effectiveness of service delivery to the citizens is support in the findings of this paper.

De Borger and Kerstens (1996) investigated the efficiency of 589 Belgium local governments and based the study on data from 1985. The study used a parametric approach in the form of

a SFA and two non-parametric approaches, a DEA analysis and a FDH analysis to evaluate the different DMU's and analysing the correlation coefficients between the efficiency score. A Tobit censored regression model was also used to investigate the explanatory variables in the study.

One of the aims of this study was to add the evolving literature at the time by considering the degree to which the efficiencies can be explained by a set of explanatory variables. The arguments made by De Borger and Kerstens (1996) was that if a set of significant determinants was robust across the various proposed approaches then the explanatory variables are not subject to manipulation and that this would yield information beneficial to the policy makers. De Borger and Kerstens (1996) was the first to use five methods to measure efficiency and compare the differences between them. These methods included two parametric and three non-parametric methods.

The outputs used in this study focused on incorporating the important factors of service delivery by a municipality. These included education, social and recreational services, and overall administrative service. This led to the following variables being chosen as inputs: the number of beneficiaries of minimal subsistence grants, the number of students enlisted in local primary schools, the surface of public recreational facilities, the total population and the fraction of the population older than 65. Total expenditure was the only input as this was a study to measure. The explanatory variables used in the Tobit regression analysis were: municipal property tax rate, per capita block grants, number of coalition parties in government, dummy variable for liberal or socialist ruling party, and proportion of adults with higher education as highest qualification.

The results of the Tobit regression showed that there was a substantial difference in the scores. Fiscal revenue and grants were found as important determinants in the efficiency scores. Overall, the study found that there were large differences in the five methods compared. According to De Borger and Kerstens (1996), the best method or approach also limits the ability to measure efficiency. They proposed that a variety of methods should be adopted and the robustness of the results be checked to ensure that an accurate calculation of the efficiency score is achieved. The study did, however, yield set robust results from the analysis of the explanatory variables. The local tax rates and education were found to have a positive effect on the efficiency and that the per-captia income grants and average income had a negative effect. De Borger and Kerstens (1996) stated that it would be beneficial to

disaggregate the local government service, i.e. police services, civil service, fire brigade etc., and then measure the relative efficiency of each service.

2.4.1 DEA studies into Municipalities

There have been numerous studies into the efficiency analysis of local municipalities across the globe spanning over three decades. Below is a short summary of selected papers and the techniques that were applied. Table 4 and Table 5 provide a review of the literature on DEA and SFA studies respectively within the context of municipalities. Table 1 is an adaptation from Kutlar et al. (2012).

Table 1: Literature review of DEA studies

Year	Authors	Methodology	Sample	Inputs, Output and Explanatory Variables
1991	Deller and Nelson	DEA	446 Illinois, Minnesota and Wisconsin municipalities	**INPUT:** • Number of full-time, • Equivalent labour, • Road graders, • Single-axle trucks, • Amount of purchased surface material, **OUTPUT:** • Price of labour (average annual salary), • Price of capital (fixed proportion of depreciated capital values), • Price of surfacing material (estimates of material requirements for re-surfacing projects), **EXPLANATORY VARIABLES:** • Regional cost-of-living index, • Miles of gravel and low and high bituminous roads.
1991	Vanden Eeckaut, Tulkens and Jamar	DEA	235 Belgium municipalities	**INPUT:** • Total current expenditures, • Total population, **OUTPUT:**

				• Proportion of people who are older than 65, • Number of people who live at the lowest life level, • Number of elementary school students, • Length of roads.
1994a	De Borger, Kerstens, Moesen and Vanneste	FDH	589 Belgian municipalities	**INPUT:** • Number of white-collar and blue-collar municipal employees, • Capital stock, **OUPUT:** • Municipal road surface, • Numbers of beneficiaries of minimal subsistence grants, • Students enrolled in local primary schools, • Surface of area of public recreational facilities, • Ratio of non-residents to residents in municipality, **EXPLANATORY VARIABLES:** • Dummy variable for liberal or socialist party as ruling coalition, • Average personal income, • Block grants, • Proportion of population with higher education, • Total population.

1994b	De Borger, Kerstens, Moesen and Vanneste	FDH	589 Belgium municipalities	**INPUT:** • Number of blue and white coloured workers m2 of buildings, • Length of roads, **OUTPUT:** • Number of people who live at the lowest life level, • Number of elementary school students' unsettled population / log (total employment), • Area of public service spaces.
1997	Rouse ,Putterill and Ryan	DEA	62 New Zealand territorial local authorities	**INPUT:** • Total expenditure on reseals, • Rehabilitation, and general maintenance, • Index of environmental factors, **OUTPUT:** • Kilometres of road resealed and rehabilitated, • General maintenance expenditure, • Annual vehicle kilometres, • Roughness index for urban and rural roads, • Index of road surface defects.

1996b	De Borger and Kerstens	FDH	589 Belgian local governments	**INPUT:** • Total municipal expenditures, **OUTPUT:** • Surface of municipal roads, • Number of beneficiaries of minimal subsistence grants, • Students enrolled in local primary schools, • Surface area of public recreational facilities, • Total population and proportion of population aged over 65 years, **EXPLANATORY VARIABLES:** • Municipal property tax rate, • Per capita block grants, • Number of coalition parties in government, • Dummy variable for liberal or socialist ruling party, • Proportion of adults with higher education as highest qualification.

2001	Prieto and Zofio	DEA	209 Spain municipalities	**INPUT:**
				• Expected budget spending
				  ○ Potable water,
				  ○ Waste,
				  ○ Length of roads,

				OUTPUT: • Number of units that illuminate the roads, • Cultural and sport background.
2001	Worthington and Dollery	DEA	103 New South Wales Municipalities	**INPUT:** • Properties receiving drinking water management system, • Occupancy rate, • Population density, • Population distribution, • Cost of disposal index, • Collection expenditure, **OUTPUT:** • Total garbage collected, • Total recyclables collected, • Implied recycling rate.
2002	Balaguer-Coll Prior-Jimenez and Vela-Bargues	DEA	258 Spain municipalities (Panel data)	**INPUT:** • Total current expenditures, • Number of illumination points, • Total population, **OUTPUT:** • Collected waste (tons),

				• Area of streets backgrounds, • Length of park areas, • Number of voters, • The level of quality.
2005	Loikkanen and Susiluoto	DEA	353 Finland municipalities	**INPUT:** • Total current expenditures, **OUTPUT:** • Daily child care houses, • Child care houses, • Central tooth care, • Older people' home, • Handicapped home, • Schools, • Number of libraries and their users.
2006	Afonso and Fernandes	DEA	51 Lisbon region municipalities	**INPUT:** • Spending per capita, **OUTPUT:** • General administration, • Educational, • Social and cultural services, • Performance of waste collectors.

2008	Afonso and Fernandes	DEA	278 mainland Portuguese municipalities	**INPUT:**
				• Total municipal expenditure per inhabitant,
				OUTPUT:
				• Social services,
				• School buildings per capita,
				• Education enrolment,
				• Cultural services,
				• Water supply,
				• Waste collection,
				• Territory organisation,
				• Road infrastructure,
				EXPLANATORY VARIABLES:
				• Non-discretionary variables:
				  ○ Purchasing power,
				  ○ Population with secondary education,
				  ○ Population with tertiary education,
				  ○ Distance to capital of district,
				  ○ Population density and variation.

2.4.2 SFA studies into Municipalities

Table 2 refers to the literature review conducted on studies which used SFA and the method for measuring the relative efficiency of municipalities.

Table 2: Literature review of Stochastic Frontier Analysis studies

Year	Authors	Methodology	Sample	Inputs, Output and Explanatory Variables
1992	Deller, Nelson and Walzer	Stochastic frontier	435 Illinois, Minnesota and Wisconsin municipal areas	**INPUT**: • Number of full-time equivalent labour, • Road graders, • Single-axle trucks, • Amount of purchased surface material, **OUTPUT**: • Price of labour (average annual salary), • Price of capital(fixed proportion of depreciated capital values), • Price of surfacing material (estimates of material requirements for re-surfacing projects), **EXPLANATORY VARIABLES:** • Regional cost-of-living indexes, • Miles of gravel, • Low and high bituminous roads.
1994	Deller and Halstead	Stochastic frontier	104 Maine, New Hampshire and Vermont	**INPUT**: • Total road costs, • Labour wages,

			municipalities	• Price of grader, • Single-axle-dump truck, Cost of capital (weighted average of new capital items by municipal bond interest rate), **OUTPUT**: • Miles of roads under town jurisdiction, • Chief engineers formal training, • Educational level, • Years of experience and age.
1996a	De Borger and Kerstens	FDH, DEA, deterministic and stochastic frontiers.	589 Belgian local governments	**INPUT:** • Total expenditure, **OUTPUT**: • Number of beneficiaries of minimal subsistence grants and students enlisted in local primary schools, • Surface area of public recreational facilities, • Total population, • Proportion of population over 65 years, **EXPLANATORY VARIABLES:** • Per capita personal income, • Municipal property tax rate,

				• Per capita block grants, • Number of coalition parties in government, • Dummy variable for liberal or socialist ruling party, • Proportion of adults with primary education as highest qualification, • Population density.
1998	Athanassopoulos and Triantis	DEA and stochastic method	172 Greece municipalities	**INPUT**: • Total current expenditures, **OUPUT**: • Number of settled families, • Average area, • Length of spaces, • Length of tourism, • Industrial areas.
2000	Wortinghton	DEA and stochastic method	166 Australia municipalities	**INPUT**: • Number of full time workers, • Financial expenditure, **OUTPUT**: • Total population, • Number of equipment used to collect clean water,

				• Length of rural and urban roads (km).
2010	Boetti, Piancenza and Turati	DEA and SFA	262 Italian Municipalities	**INPUT**: • Disaggregated current expenditure in general administration, • Road maintenance, • Local mobility, • Garbage collection and disposal, • Education, • Elderly care and other social services. **OUTPUT**: • Number of inhabitants, • The total length of municipal roads, • The amounts of waste collected, • The sum of the number of pupils enrolled in nursery, • Primary and secondary schools, • The number of people over age 75.

2.5 Approaches to Outliers

Outliers in DEA models can provide unwanted noise and inaccurate efficiency score. Some researchers have developed methods to remove the *k* most influential observations in from the model to eliminate the outliers. However, DEA is a methodology to provide the relative efficiency scores. If one removes the observations with the outlier then it will change the outcome of the model thus providing an inaccurate reflection of the scores. Other researchers have developed new methods to detect the outliers as part of the DEA model.

If an outlier was present in an observation, Timmer (1971) suggested removing the observation from the sample and calculating the production frontier on the remaining DMUs. This approach will provide an accurate representation of the efficiency score but will not be reflective of the entire sample of DMUs (outlier DMU included). By removing the outlier DMU the remaining DMUs score will inherently change as they will not be contaminated or influenced by the outlier observation.

Andersen and Petersen (1993) used a super-efficiency model which involved using the conventional DEA model after excluding the DMU being evaluated from the reference set. The conventional DEA model calculates the efficiency of a DMU relative to a reference set of all the observations including its own observation. A super-efficiency model removes itself from the reference set allowing for an efficiency score that exceeds one Banker and Chang (2006). The super-efficiency model used by Andersen and Petersen (1993) was based on the super-efficiency model proposed by Banker and Gifford (1988).

An alternative outlier detection method in non-parametric models was proposed by Wilson (1993) and further improved upon by Wilson (1995) by using the super-efficiency model of Andersen and Petersen (1993). Wilson (1995) also introduced a procedure to detect influential observation.

Simar (2003) proposed an alternative method for identifying influential observations and extended this to both input and output orientations. Banker and Chang (2006) conducted simulation experiments to evaluate the performance of super-efficiency models of Banker and Gifford (1989) when it is used to rank the efficient unit and outlier detection. Banker and

Chang (2006) provided an explanation on how the model by Andersen and Petersen (1993) can be used for outlier detection but not for ranking.

Johnson and McGinnis (2008) proposed a two-stage semi-parametric DEA model for outlier detection which involved constructing an inefficient frontier to detect outliers. Yang et al. (2014), Banker and Chang (2006) and Johnson and McGinnis (2008) demonstrated that super efficiency DEA approaches are promising in detecting outliers. Yang et al. (2014) stated that both studies did not compare their approaches with other popular methods developed in statistics and data mining. Furthermore, these studies did not identify the conditions under which the proposed methods would perform well or not and did not examine the predictive performance of the approaches Yang et al. (2014).

The need for identifying and eliminating outliers in DEA was pointed out by Simar (1996). Simar (1996) further pointed out that if outliers cannot be identified, a stochastic frontier is recommended. Kuosmanen and Post (1999) proposed a method of removing the *k* most influential DMUs using an empirical specification tests. Tran et al. (2010) proposed a new method of dealing with outliers in DEA based on two scalar measures, relative frequency in which an observation occurs in the frontier and the cumulative weight of the observation in the frontier. Cazals et al. (2002) proposed a nonparametric efficiency score that is robust enough to deal with outliers. This method is based on a concept of expected minimum function or an expected maximum function. This study proposes a new model to cater for outliers in a DEA analysis by substituting specific input and outputs which are considered outliers. Table 3 below, shows an overview of the literature around outlier detection model and approaches over the years.

Table 3: Literature review of Outlier studies, treatment of outliers in DEA

Year	Authors	Methodology	Approach	Findings
1971	Timmer	Probabilistic frontier production frontier using a Cobb-Douglas production frontier	Procedure of removing fixed percentage of the observations until the estimate the production frontier stabilised.	Since the approach removed most of the inefficiency, half of the inefficiency was due to the definitional and measurement problems in the variables.
1993	Andersen and Petersen	DEA super-efficiency	Proposed modified approach to allow for ranking of efficient units and detect influential observations in the sample. This was based on the super-efficiency model.	The approach provides an efficiency rating of efficient units similar to the rating of inefficient units. The study showed that the efficiency score possesses a number of desirable index properties.
1993	Wilson	DEA	Extended the statistic of Andersen and Petersen (1993) to cater for multiple outputs.	The study found that although an observation had a low probability of occurrence; it is not conclusive that it may be an outlier.
1995	Wilson	DEA	Improvements were made to the initial method developed in Wilson (1993). The super-efficiency models used in Andersen and Petersen (1993) was adopted.	The improvement of the model allows for a less computationally expensive approach

1998	Simar and Wilson	DEA with Bootstrapping	Proposed a bootstrap strategy focused around a reasonable Data-Generating Process and approximate the sampling variation of the estimated frontier.	The study showed by focusing on the underlying Date-Generating Processt hey were able to use bootstrap methods to analyse the sensitivity of nonparametric efficiency scores to sampling variation. The bootstrap estiates can also be used to test hypotheses about the structure of the underlying technology, as in Simar and Wilson 1996.
2000	Simar and Wilson	DEA with Bootstrapping	Extended the work of Simar and Wilson 1998.	The proposed method alleviated the restrictive method in Simar and Wilson 1998 by allowing for heterogeneity in the structure of efficiency.
2002	Cazals Florens and Simar	Non Parametric estimator based on the expected minimum function or maximum output function	The approach proposed in the paper is related to DEA/FDH estimators of efficiency but is more robust to outliers, noise and extreme values. The proposed idea is based on the concept of “expected frontier of order-m.”	The approach is found to be more robust to extreme points, noise and outliers. But it does not envelope of all the data points.

2003	Simar	Non parametric frontier model	Based on the work of Cazals et al. (2002), Simar (2003) demonstrates how the procedure can be used to detect outliers. This approach is multivariate and can be applied to a DEA/FDH approach.	The study showed all the steps for computing the order-m inout and output efficiency measures in the general multi-output, multi-input framework. This is useful in an explanatory data analysis phase of any efficiency analysis of firms, with real data, in order to detect any potential outliers.
2006	Banker and Chang	DEA, Super efficiency	Researched two alternative uses of the super efficiency procedure. The first was in detecting outliers and the second was for ranking efficient DMUs.	It was shown that the ranking procedure does not perform satisfactorily. The correlationsbetween the true efficiency and the estimated super-efficiency are negative for the subset of efficient observations. The evidence supports the use of Banker and Gifford 1988 and Banker et al 1989.
2008	Johnson and McGinnis	Two-stage semi parametric DEA. Super efficiency and bootstrapping	Proposed using the efficient frontier and the inefficient frontier to identify outliers. An iterative outlier detection approach is implemented in the super-efficiency method (efficient and inefficient frontier) and uses a semi-parametric bootstrapping method as the	The results show that the conclusion drawn can be different when outlier identification includes consideration of the inefficient frontier.

			second stage	
2010	Tran, Shively and Preckel	DEA	Proposed a new method of detecting outliers based on two scalar measures. The first measure, when testing the efficiency of other observations, is the relative frequency with which an observation appears in the construction of the frontier. The second measure is the cumulative weight of an observation in the construction of the frontier.	The approach was found to be computationally inexpensive. The method will find the greatest weight every time a weight is calculates and remove the observation.
2010	Chen and Johnson	DEA	Identified a set of axioms and developing an approach consistent with the axioms.	This approached allowed for detection of both efficient and inefficient outliers that would have otherwise influenced post analysis procedures.
2010	Yang, Wang and Sun	DEA and bootstrapping	The proposed approach introduced two parameters: probability level and tolerance. Both parameters must be specified externally. A bootstrap was also proposed to approximate the true distribution.	The study showed the existence of outliers may contaminate the measured efficiency. After some of the DMUs are detected as potential outliers, removing real outliers is justified only if they can be identified.
2012	Bellini	DEA super-efficiency and forward search	The approach proposed in this papers merged a super-efficiency model with a forward search and introduced a distance to be monitored	The approach developed allows for the avoidance of subjectivity in outlier dectection by performing a comparison of

			along the search. The distance was obtained by integrating a regression model with the super-efficiency DEA model.	the maximum distance within the subset and the minmum distance outside the subset.
2014	Bahari and Emrouznejad	DEA , bootstrapping	An alternative approach to Yang et al. (2010) is proposed and is applied to a sample of hospitals.	A new method was developed to detect whether a specific DMU is truly influential. An application in measuring hospital efficiency was used to show significant advancements in outlier detection.
2014	Yang, Wang Zheng	Bi-super DEA, Predictive DEA	A super DEA based method that constructs an efficient and inefficient frontier is proposed to detect outliers. To evaluate the performance of the method a predictive DEA procedure is also proposed. Conducted simulation experiments to examine the performance of the outlier detection methods.	It was shown that under linearity, normality and homogeneity conditions, Robust Regression stands out. However when the underlying Data-Generating Process isnonlinear, the Bi-Super DEA procedure is superior.

2.6 Conclusion

In this chapter, the focus was on the literature available regarding the problem statement. The overview of municipalities and the role they play in service delivery has been discussed in detail. The studies involving DEA and SFA in the context of municipal efficiency have been provided. The approaches to deal with data that have outliers and the methodology used to accommodate extreme values have also been discussed. This chapter provides the building blocks for the dissertation and foundation for the proposed outlier method. The methodologies of these models are addressed in the next chapter.

Chapter 3

Relative Efficiency Analysis

3.1 Introduction

"*Efficiency can be simply defined as the ratio of output to input. More output per unit of input reflects relatively greater efficiency. If the greatest possible output per unit of input is achieved, a state of absolute or optimum efficiency has been achieved and it is not possible to become more efficient without new technology or other changes in the production process*." Sherman and Zhu (2006).

The traditional form of efficiency measurements in any industry is measured as a score of the ratio of outputs to inputs.

$$Effieciency = \frac{Output}{Inputs} \tag{3.1}$$

Although this traditional form is a simple measure, it has some limitations:

1. The accommodation of multiple inputs and outputs cannot be incorporated into the measurement.
2. Both the input and output variables must be measured in the same unit.

Over the years there have been three types of efficiency measures that have been developed. The first is technical or productive efficiency. This refers to the use of productive resources available in the best technologically efficient manner. Technical efficiency aims at

identifying the maximum obtainable level of outputs given a set of inputs. In the private sector this would typically equate to trying to achieve maximum profit whilst minimizing the cost. The second is allocative efficiency, and this refers to the distribution of the productive resources with the aim of achieving the optimal mixture of outputs. It accomplishes this by choosing between the different technically efficient combinations of outputs. The degree of economic efficiency is when both the technical and allocative efficiencies are evaluated. In other words, if a DMU is technically and allocatively efficient, then DMU is considered to be economically efficient. Conversely, if the aforementioned DMU is neither technically nor allocatively efficient then that DMU is operating at less than the total economic efficiency. The third measure of efficiency is dynamic efficiency. Dynamic efficiency which refers to an economically efficient usage of resources such that appropriately balances short term concerns with long term concerns. Allocative and technical efficiency are the basis of an empirical measurement of efficiency of a DMU.

Consider the following examples: if one wishes to measure the efficiency of two municipalities with respect to the refuse removal, it can be done by measuring the amount of refuse collected per number of refuse removal trucks. Municipality A is stated to collect refuse with a maximum of 16 tons per truck. Municipality B operates with a maximum of 14 tons per truck. If Municipality A actually obtained 12 tons per truck then its efficiency will conclude that Municipality A is running inefficiently at 75% (12 tons/truck / 16 tons/truck). This represents technical inefficiency in that extra input (trucks) is used to get the output (tons of refuse). Municipality A should either lower the number of trucks used or increase the number of tons of refuse removed per truck to achieve 100% efficiency.

Similarly, if Municipality B achieved 11 tons per truck it runs at an inefficiency of 78%. Again, Municipality B can either lower the number of trucks used or increase the number of tons removed. Since both municipalities are not the same, they could have different areas to cover. These areas could be densely/sparsely populated and the trucks could have been travelling in urban or rural areas with steep inclines and declines. The reliability of the trucks, speed at which the workers load the trucks and the traffic between pic-up and drop-off locations also need to be taken into consideration. If these factors were incorporated into the calculation as input/outputs, then they can be classified into two main groups. The first is discretionary and non-discretionary input/outputs. Discretionary variables are variables that are in the control of the municipality, i.e. the municipality can choose how many workers

they assign to a truck and how many trucks they designate to an area. Non-discretionary variables will be the reliability of the trucks and the amount of traffic, as the municipality has no control if the roads are congested or if a truck breaks down.

One can also measure the price efficiency of two of the municipalities. If Municipality A operates at a maximum of 16 tons per truck at a cost of R100.00 per trucks' trip and Municipality C also operates at a maximum of 16 tons per truck at a costs R80.00 per trucks' trip. One can now measure the efficiency of one municipality relative to the other. It will cost Municipality A R6.25 per ton and R5.0 per ton for Municipality C. Municipality A is then R5.0/ R6.25 or 80% as Municipality C. This 20% inefficiency is not due to the area coverage or any other differences in the municipalities but more at the cost of the removal of refuse per ton. Municipality C can utilise the same truck at a lower price and therefore is more efficient than Municipality A. However, Municipality C may not be absolutely efficient as one may be able run the truck at a lower price than R80.00, but it is clear that Municipality A is inefficient due to the higher running cost of the truck.

Let us assume that there are two municipalities, Municipality 1 and Municipality 2. Each municipality has Truck A and Truck B at their disposal for the refuse removal. Truck A operates at a maximum of 16 tons per trip and Truck B operates at a maximum of 14 tons per trip. Both trucks cost R100.00 per trip. If both municipalities use the trucks in different proportions, each refuse removed and cost will be different. If we assume that Municipality 1 only uses Truck A and Municipality 2 only uses Truck B but both municipalities use the trucks equally, Truck A then costs R6.25 per ton and Truck B cost R7.14 per ton. We can then state that Truck B is R6.25 / R7.14 or 87% efficient as Truck A or that Municipality 2 is 87% efficient to Municipality 1. This is an example of allocative efficiency, which is a result of an inefficient mix of inputs used to produce a certain number of outputs Sherman and Zhu (2006).

Technical frontiers aim at creating an envelope of efficient entities by making use of the combination of outputs. When a DMU fails to fall on the technical frontier, it is said that the DMU is technically inefficient. In a similar fashion if a DMU fails to provide a combination of outputs that do not fall on the technical frontier then that DMU is said to be allocatively inefficient.

DEA uses the concept of the efficiency that is defined as the following:

$$Effieciency = \frac{Weights\ sum\ of\ Outputs}{Weights\ sum\ of\ Inputs} \tag{3.2}$$

This concept is discussed in more detail in Section 3.2.

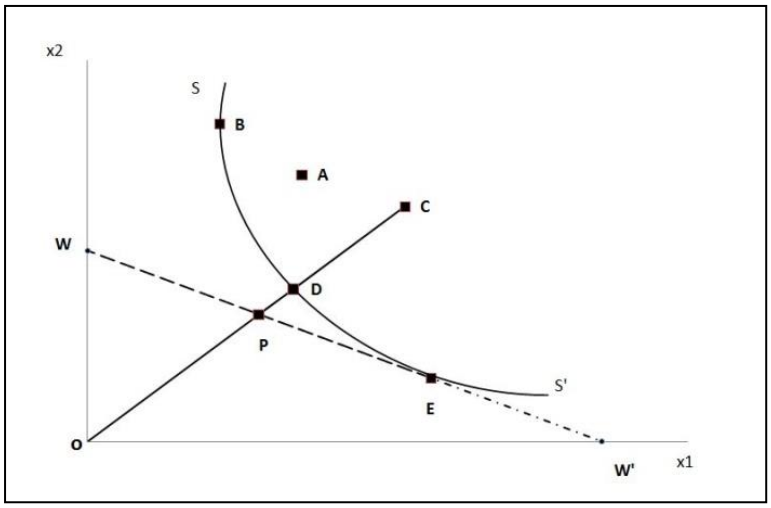

Figure 1: Efficiency Frontier

Figure1, adapted from Cooper et al. (2010), illustrates graphically the efficiency frontier. In this example there are five DMUs: A, B, C, D and E with $x1$ and $x2$ as the two inputs used to produce one unit of output. By connecting the three DMUs B, D and E the curve formed SS' is referred to as the efficiency frontier in DEA literature. These three DMUs are efficient as they used the least amount of inputs to provide the one unit of output. The efficiency frontier is a representation of the least cost combination of inputs used to produce a given amount of output. The DMUs that do not fall on the efficiency frontier (A and C) are considered to be inefficient. This efficiency score is the technical efficiency (TE) trying to achieve maximum output while using minimal input. They use a higher quantity of inputs to produce the same amount of outputs. If these inefficient DMUs were to reduce either input $x1$ or $x2$ or both, they would improve their efficiency and move closer to the efficiency frontier, until they lie on the frontier. The estimated efficiency of each DMU is given by the symbol j, and is the ratio of the distance from the origin to the efficiency frontier and the distance from the origin to the DMU. For example, DMU C would have an efficiency score of $j = \frac{0D}{0C}$. If a DMU has an efficiency score $j = 1$, then that DMU is efficient, if the efficiency score is $j < 1$ then that DMU is considered to be inefficient. For these inefficient DMUs, the value of $(1 - j)$ provides an indication as to the proportion by which that DMU should reduce the inputs to

achieve the required amount of output Farrell (1957). For example, if DMU A has an efficiency score of 0.75, then DMU A would have to reduce its inputs by 25% to achieve a relative efficiency score of 1.

The TE values for the five DMUs do not, however, take into consideration the costs associated with the two inputs. By adding the costs to the graph we can construct the cost line WW'. We can now see that DMU E falls on both the cost line and the efficiency frontier, i.e. the point of tangency , therefore we can conclude that DMU is allocatively efficient (AE) as well as technically efficient. DMU E maximizes the amount of output while using the minimum amount of inputs as well as using an optimal combination of inputs so that the costs are minimized. Hence it is economically efficient, opposed to DMU B and D which are only technically efficient.

The AE score is estimated by the ratio of the distance from the origin to the cost line and the origin to the DMU. For example if we estimate the AE score for DMU C we will get: $score = \frac{0P}{0C}$.

The total efficiency indicator, Economic efficiency (EE), can then be estimates by combining the TE and AE:

$$EE = TE * AE$$

$$= \left(\frac{OD}{OC}\right) * \left(\frac{OP}{OC}\right)$$

$$= \frac{OD}{OP} \qquad (3.3)$$

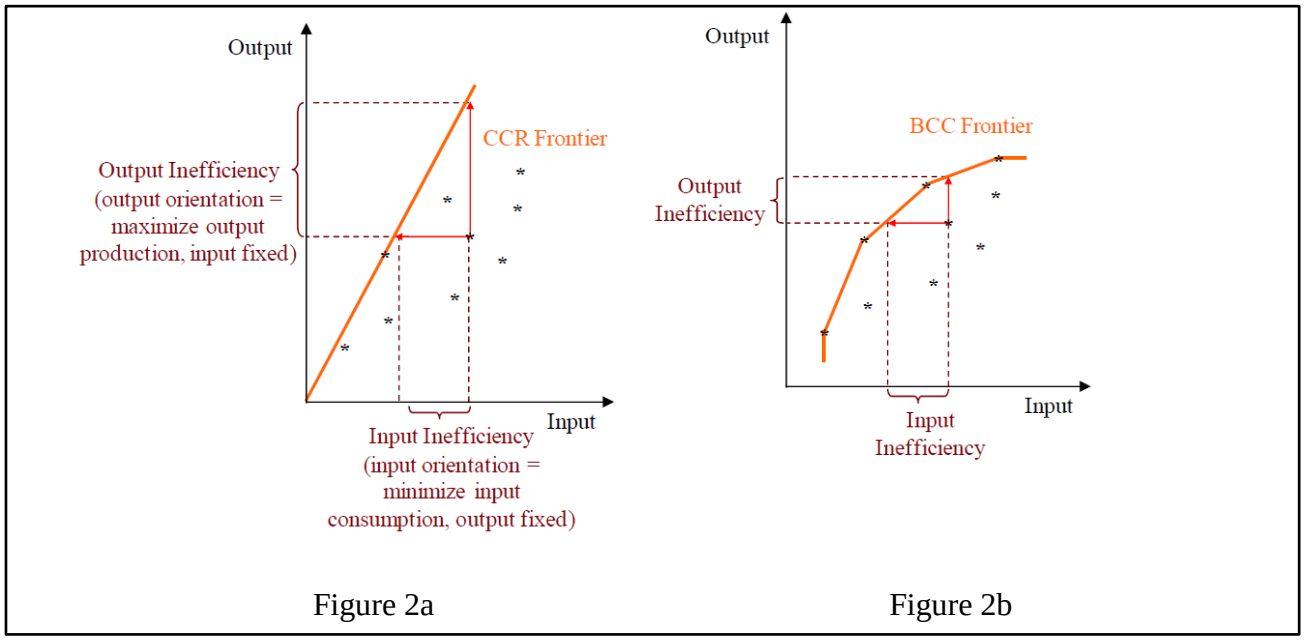

Figure 2a Figure 2b

Figure 2: Projections onto the Efficient Frontier

Figure 2 above shows the projections of the inefficient DMUs on to the efficiency frontier. This is based on a single input single output case Kingyens (2012).

The basic DEA model is a Charnes et.al. (1978) model or CCR model which assumes constant returns to scale (CRS). The Banker et.al. (1984) model or BCC assume variables returns to scale (VRS).

There are two main types of technical efficiency analysis in the literature, a parametric approach and a non-parametric approach. The parametric approach makes use of a stochastic frontier analysis (SFA) or deterministic frontier analysis (DFA) and the non-parametric approach uses data envelopment analysis (DEA). The SFA is a method of economic modelling. The development of the SFA is due to the fact that if a DMU is technically or allocatively inefficient, it can be the result of external factors beyond its control. A SFA approach takes these external factors into account when measuring the efficiency, as opposed to a deterministic frontier analysis (DFA). DFA and SFA both attempt to measure the absolute efficiency which can be used to compare a DMU with an industry benchmark.

3.2 Data Envelopment Analysis

DEA is a mathematical programming procedure, developed by Charnes et al. (1978), which can be used to measure the relative efficiency of DMU. Each production unit of a set of comparable producing units can be regarded as a DMU. DEA aims at evaluating the relative efficiencies of entities in the same industry and provides a measure of relative performance rather than comparing it with an idealised benchmark or level of performance. Free-disposal hull (FDH) is an important variant of DEA and is commonly used in the measure of public sector efficiency. FDH has the advantage of taking fewer observations and being able to determine what the industry's best practise is and it does not assume multiple methods of delivery a service or producing goods.

The main difference between SFA and DEA is that SFA creates a stochastic frontier with a probability distribution; whereas DEA has a non-stochastic frontier. DEA analysis uses a non-parametric approach to produce an efficiency frontier. It does not; impose any assumptions on the functional form i.e. it does not take into consideration the influence of external factors. It is a non-statistical approach which disregards statistical noise but because it is non-parametric, there are few underlying assumptions that need to be taken. On the other hand, SFA takes into account the statistical noise but then it then requires certain assumptions to be made.

The core concept of measuring efficiency can be dated back to Farrell (1957). DEA evaluates the relative efficiency of homogenous DMUs which have no known relationship between the incorporation of inputs used and the outputs generated from the unit. The vital characteristic of DEA is the ability to transform a multi-input and multi-output unit into a single unit of measurement. However DEA can formulate the relative efficiency of any number of DMUs. The relative efficiency of DMUs can be calculated in relation to all other DMUs.

The objective function of the model is to maximize the relative efficiency scores, subject to the set of weights required for each DMU, which must be feasible. The variables for any DEA model are as follows: assume that there are n DMUs to be analysed, each DMU has m inputs which contribute to the s outputs produced. DEA assigns a set of weights to the inputs and outputs of a DMU with the aim of yielding the best possible efficiency.

The weights placed on each DMU reflect the importance of each input and output and is used to emphasise certain input and outputs. DEA simultaneously gives the other DMUs the same weight and compares the resulting efficiency with the DMU in question.

The decision variables for the model are the set of weights assigned to each input and output.

Let:

x_{ij} be defined as the amount of input *i* used for DMU *j*.

y_{rj} be defined as the amount of output *r* produced by DMU *j*.

v_i be defined as the weight assigned to input *i*

u_r be defined as the weight assigned to output *r*.

"We can interpret the CCR construction as the reduction of the multiple-output /multiple-input situation (for each DMU) to that of a single 'virtual' output and 'virtual' input. For a particular DMU the ratio of this single virtual output to single virtual input provides a measure of efficiency that is a function of the multipliers. In mathematical programming parlance, this ratio, which is to be maximized, forms the objective function for the particular DMU being evaluated, so that symbolically"

Cooper et al. (2010)

$$\max \; h_o(u,v) = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}} \tag{3.4}$$

Subject to:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \qquad \forall\, j \in \{1,..,n\})$$

$$u_r \geq 0 \qquad \forall\, r \in \{1,..,s\}$$

$$v_i \geq 0 \qquad \forall\, i \in \{1,..,m\} \tag{3.5}$$

The h_o value calculated is a ratio of the weighted sum of outputs divided by the weighted sum of inputs, such that it is maximized for DMU_o the DMU to be evaluated.

According to Charnes and Cooper (1982) the model above can be transformed into a linear model by changing the objective function and adding a constraint and the variables (μ, v) to changes to(μ, v).

This resulting model is known as the Charnes-Cooper (1982) transformation:

$$\max \ z = \sum_{r=1}^{s} \mu_r y_{ro} \tag{3.6}$$

Subject to:

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \ \leq 0 \quad \forall j \in \{1,..,n\}$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$\mu_r \ \geq 0 \qquad \forall r \ \in \{1,..,s\}$$
$$v_i \ \geq 0 \qquad \forall i \ \in \{1,..,m\} \tag{3.7}$$

The dual form of this model is often implemented as follows:

$$\min \ \theta = \theta^* \tag{3.8}$$

Subject to:

$$\sum_{j=1}^{n} x_{ij} \lambda_j \ \leq \theta x_{io} \quad \forall i \in \{1,..,m\}$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j \ \leq \theta y_{ro} \quad \forall r \in \{1,..,s\}$$

$$\lambda_j \ \geq 0 \qquad \forall j \in \{1,..,n\} \tag{3.9}$$

In the economics section of DEA literature, due to the fact that the model ignores the presence of non-zero slack, the model is said to conform to the assumption of "strong disposal" Cooper et al. (2010). This is referred to as a "weak efficiency" in the operations research section of the DEA literature Cooper et al. (2010).

In the dual theorem of linear programming we have $z^* = \theta^*$. We can use the model to solve for an efficiency score, by setting j =1 and $\lambda_k^* = 1$ with $\lambda_k^* = \lambda_o^*$ and all other $\lambda_j^* = 0$. In doing so, we can ensure that we will always have solutions Cooper et al. (2010). This also implies that $\theta^* \leq 1$. Therefore the optimal solution θ^* will yield the efficiency score for the DMU being evaluated Cooper et al. (2010). The process will be repeated for each DMU where $(X_o, Y_o) = (X_k, Y_k)$, where (X_k, Y_k) is the vectors with components x_{ik} , y_{rk} . Similarly

(X_o, Y_o) have the vector components x_{io} , y_{ro} . The DMUs which have an efficiency score $\theta^* = 1$ are considered to be efficient and are referred to as boundary points. DMUs that have efficiency scores $\theta^* < 1$ are found to be inefficient Cooper et al. (2010).The boundary points can have a "weak efficiency" due to the presence of non-zero slacks. This can lead to some DMUs being efficient with non-zero slacks and others that are also efficient with zero slacks. This problem can be avoided by modifying the linear programme whereby the slacks are taken to their maximum values.

$$\max \ \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \tag{3.10}$$

Subject to :

$$\sum_{j=1}^{n} x_{ij}\lambda_j + \ s_i^- \leq \theta^* x_{io} \quad \forall\, i \in \{1,..,m\}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - \ s_r^+ \leq y_{ro} \quad \forall\, r \in \{1,..,s\}$$

$$\lambda_j \ \geq 0 \qquad \forall\, j \in \{1,..,n\}$$

$$s_r^+ \ \geq 0 \qquad \forall\, r \in \{1,..,s\}$$

$$s_i^- \ \geq 0 \qquad \forall\, i \in \{1,..,m\} \tag{3.11}$$

Where the choice of s_i^- and s_r^+ do not affect the optimal θ^*. According to the definition of DEA efficiency, the performance of a DMU is said to be fully (100%) efficient, if and only if, the following apply Cooper et al. (2010):

(i) $\theta^* = 1$ (3.12)

(ii) $s_r^{+*} = \ s_i^{-*} = 0$ (All slack variables are zero) (3.13)

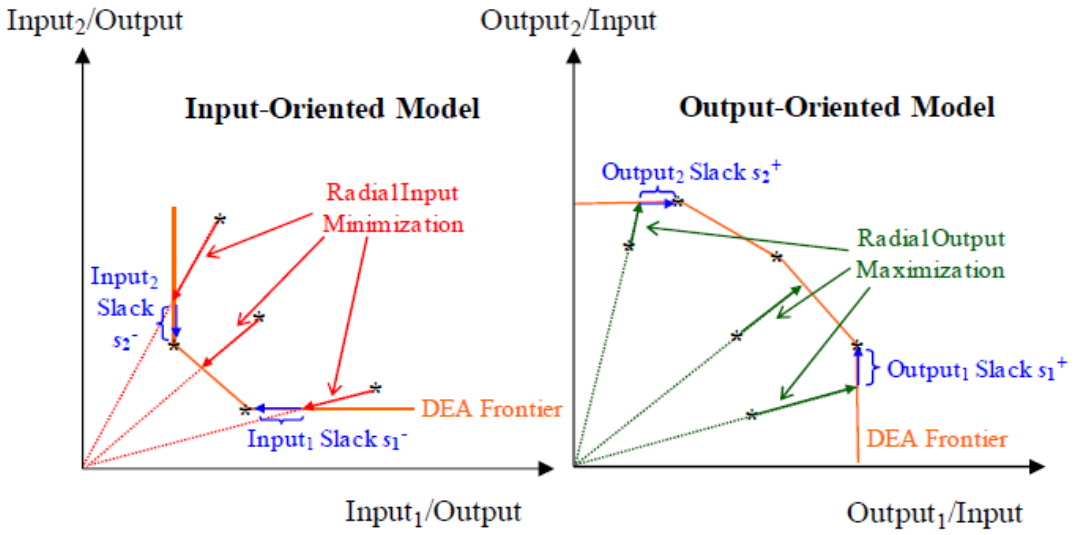

Figure 3a Figure 3b

Figure 3: Slack based DEA models

Figure 3 graphically shows the input minimisation, output maximisation and the slacks Kingyens (2012).

The model in (3.11) amounts to solving the problem in two steps. By maximizing $\theta^* = \theta$ and then using that value in the slack modification model (4) to solve for the values of $s_r^+ \; and \; s_i^-$. This is overcome by introducing ε, a so-called non-Archimedean element defined to be smaller than any positive real number. We then get the following model:

$$\max \; \theta - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \qquad (3.14)$$

Subject to :

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta x_{io} \quad \forall\, i \in \{1,..,m\}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ = y_{ro} \quad \forall\, r \in \{1,..,s\}$$

$$\lambda_j \geq 0 \qquad \forall\, j \in \{1,..,n\}$$

$$s_r^+ \geq 0 \qquad \forall\, r \in \{1,..,s\}$$

$$s_i^- \geq 0 \qquad \forall\, i \in \{1,..,m\}$$

$$\varepsilon > 0 \qquad (3.15)$$

By adding the non-Archimedean element ε to the initial model (3.4) we get the following model which has reoriented the objective function from max to min:

$$\min \frac{\sum_{i=1}^{m} v_i x_{io}}{\sum_{r=1}^{s} u_r y_{ro}} \tag{3.16}$$

Subject to :

$$\frac{\sum_{i=1}^{m} v_i x_{ij}}{\sum_{r=1}^{s} u_r y_{rj}} \geq 1 \qquad \forall j \in \{1,..,n\}$$

$$u_r \geq 0 \qquad \forall r \in \{1,..,s\}$$
$$v_i \geq 0 \qquad \forall i \in \{1,..,m\}$$

$$\varepsilon > 0 \tag{3.17}$$

By performing a Charnes-Cooper (1982) transformation we get the following linear model:

$$\min \; q = \sum_{i=1}^{m} v_i x_{io} \tag{3.18}$$

Subject to:

$$\sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} \mu_r y_{rj} \geq 0 \quad \forall j \in \{1,..,n\}$$

$$\sum_{r=1}^{s} \mu_r y_{ro} = 1$$

$$\mu_r \geq \varepsilon \qquad \forall r \in \{1,..,s\}$$
$$v_i \geq \varepsilon \qquad \forall i \in \{1,..,m\} \tag{3.19}$$

And the corresponding dual problem:

$$\max \; \phi + \varepsilon(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+) \tag{3.20}$$

Subject to:

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = x_{io} \quad \forall i \in \{1,..,m\}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \leq \phi y_{ro} \quad \forall r \in \{1,..,s\}$$

$$\lambda_j \geq 0 \qquad \forall j \in \{1,..,n\}$$

$$s_r^+ \geq 0 \qquad \forall r \in \{1,..,s\}$$

$$s_i^- \geq 0 \qquad \forall i \in \{1,..,m\} \tag{3.21}$$

Table 4 below contains the conventional DEA models used in this study.

Table 4: Conventional DEA models

	CCR Model	BCC Model
Objective Function	$\max \ h_o(u,v) = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}$	$\max \ h_o(u,v) = \frac{\sum_{r=1}^{s} u_r y_{ro}}{\sum_{i=1}^{m} v_i x_{io}}$
Subject to	$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \ \ \forall j \in \{1,..,n\}$ $u_r \geq 0 \ \forall r \in \{1,..,s\}$ $v_i \geq 0 \ \ \forall i \in \{1,..,m\}$	$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \ \forall j \in \{1,..,n\}$ $u_r \geq 0 \forall r \in \{1,..,s\}$ $v_i \geq 0 \ \forall i \in \{1,..,m\}$ $\sum_{j=1}^{n} u_j = 1$

The h_o value calculated is a ratio of the weighted sum of outputs divided by the weighted sum of inputs, such that it is maximized for DMU_o the DMU to be evaluated.

The non-parametric approach to measuring efficiency has been mainly focused around Data Envelopment Analysis; where a number of DMUs are analysed using a multi-input multi-output production technology. The specification of a functional form is not required in DEA. The basic DEA models are deterministic.

3.3 Returns to Scale

When one examines economic literature, Returns to Scale (RTS) is normally defined for single output scenarios. The RTS is said to be increasing if all the inputs are increased by a certain proportion resulting in an increase of the singular output. For example, let us assume that γ represents the amount which the inputs are increased and δ represents the change in the amount of outputs. If $\gamma > \delta$ then we say that an increasing returns of scale (IRS) is present and if $\gamma < \delta$ then a decreasing returns of scale (DRS) is present Cooper et al. (2010).

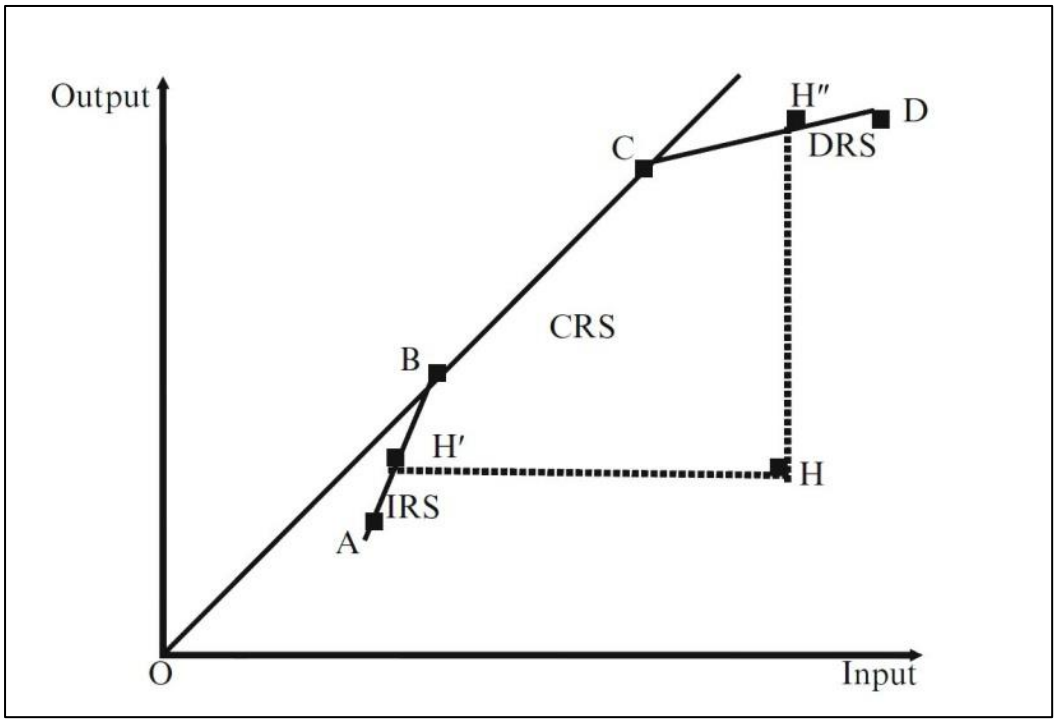

Figure 4: Returns to Scale: Constant versus Variable

Figure 4 above illustrates the differences between a CRS and VRS. There are five DMUs plotted A, B, C, D and H. The line OBC represents the CRS frontier. The lines AB, BC and CD represent the VRS frontier. DMUs A, B, C and D are considered to be efficient under the VRS but only DMU B and C are efficient under CRS. The line segment of AB has an increasing return of scale (IRS), BC has a constant return of scale and CD has a decreasing return of scale (DRS) to the right of C. When we look at point H and plot a projection, H', onto the frontier at line segment AB we will get an IRS and if we plot the projection, H", using the BBC output-oriented model on the line segment CD we will get IRS. This is due to the fact that the input-orientation and output-orientation will yield different projection points on the frontier resulting in different RTS Zhu (2009).

3.4 Limitations of DEA

According to Kingyens (2012), there are several limitations to the DEA methodology:

1. DEA does not account for random error or noise in the analysis, i.e. any deviation of the efficiency score from the efficient frontier is seen as inefficiency.

2. DEA is sensitive to small sample sizes. When the recommended number of DMUs is not met, it can lead to higher efficiency scores with several DMUs appearing on the efficient frontier.
3. The efficiency score calculated in the DEA analysis is specific to the DMUs that form the reference set. If certain DMUs were to be removed, it would affect the score of the remaining DMUs in the set.
4. There are no restrictions in DEA for the assignment of the weights, i.e. the model can assign zero or near zero weights to certain inputs/outputs in order to maximize the efficiency score.

3.5 Super-Efficiency

The main difference between the conventional DEA models and the super-efficiency models is that when the super-efficiency is employed, the DMU under evaluation is not part of the reference set for the constraints Banker and Chang (2006). The exclusion of the DMU under evaluation from the reference set does not ensure that a convex combination of the remaining DMUs can be created to envelop the DMU under evaluation for its inputs and outputs Banker and Chang (2006).

Banker and Gifford (1988) proved that a feasible solution will always exist for a super-efficient CCR model but they will not always exist for a super-efficient BCC model for certain extreme values.

3.6 Stochastic Frontier Analysis

There have been various approaches to measure the efficiency over the years. The non-parametric DEA methods have been more popular but the use of econometric methods is also evident. The econometric approach involved developing a stochastic frontier model, which is based on the deterministic parameter frontier of Aigner and Chu (1968). The advantage of the Stochastic Frontier Analysis (SFA) is in its ability to differentiate between the random noise and the estimated production frontier.

The Stochastic frontier production models were introduced independently by Aigner et al. (1977) and Meeusen and Van den Broeck (1977). Once the functional form had been chosen for the development of the production frontier, the authors proposed the model.

$$y_i = f(x_i, \beta) + \varepsilon_i \tag{3.22}$$

Where y_i is the output obtained by DMU, x_i is the vector of selected inputs. β is the vector of parameters to be estimated and ε_i is the composed error term. The composite error term comprises of the two elements: $\varepsilon_i = v_i + u_i$, where v_i represents the symmetric disturbance that encapsulate the random variation in the production frontier. This disturbance is due to factors such as random errors, errors in the observation and measuring of data and chance, which is assumed to be identically and independently distributed following a $N(0, \sigma_v^2)$ distribution. The error component u_i is asymmetric and encapsulates the technical inefficiency. This component is assumed to be distributed independently of v_i and satisfies the condition $u_i \leq 0$. The statistical distribution of error component u_i is assumed. In the paper by Aigner et al. (1977), the cases for a half-normal and exponential distribution was analysed, and Meeusen and Van den Broeck (1977) only considered the exponential distribution.

This model assumes that the production function takes a log-linear Cobb-Douglas form; hence the stochastic frontier production model can be written as:

$$\ln(y_i) = \beta_0 + \sum_n \beta_n \ln(x_{ni}) + \varepsilon_i \tag{3.23}$$

3.6.1 The Normal-Half Normal Model

In case of the normal-half normal model we assume the following:

1. $v_i \; is \; iid \; with \; N(0, \sigma_v^2)$

2. $u_i \; is \; iid \; with \; N^+(0, \sigma_u^2)$

3. $v_i and \; u_i \; independent \; of \; each \; other, and \; of \; the \; regressor$

Given the independence of error terms, the joint density of v and u can be written as:

$$f(u,v) = \frac{2}{2\pi\sigma_u\sigma_v} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\} \quad (3.24)$$

Substituting$v = (\varepsilon + u)$ into the preceding equation gives:

$$f(u,v) = \frac{2}{2\pi\sigma_u\sigma_v} exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\} \quad (3.25)$$

Integrating u out to obtain the marginal density function of ε results in the following form:

$$f(\varepsilon) = \int_0^\infty f(u,\varepsilon)du$$

$$= \frac{2}{\sqrt{2\pi\sigma}}\left[\mathrm{I} - \phi\left(\frac{\varepsilon\lambda}{\sigma}\right)\right] exp\left\{-\frac{\varepsilon^2}{2\sigma^2}\right\}$$

$$= \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(-\frac{\varepsilon\lambda}{\sigma}\right) \quad (3.26)$$

Where $\lambda = \frac{\sigma_u}{\sigma_v}$ and $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ (3.27)

In the case of a stochastic frontier cost model, $v = \varepsilon - u$ and

$$f(\varepsilon) \ \frac{2}{\sigma}\phi\left(\frac{\varepsilon}{\sigma}\right)\Phi\left(\frac{\varepsilon\lambda}{\sigma}\right) \quad (3.28)$$

The log-likelihood function for the production model with N producers is written as:

$$lnL = constant - Nln\sigma + \sum_i ln\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_i \varepsilon_i^2 \quad (3.29)$$

3.6.2 The Normal-Exponential Model

Under the normal-exponential model, we assume the following:

1. $v_i is\ iid\ N(0, \sigma_v^2)$

2. $u_i is\ iid\ exponential$

3. $v_i and\ u_i\ independent\ of\ each\ other, and\ of\ the\ regressor$

Given the independence of error term components $v_i and\ u_i$, the joint density of v and u can be written as:

$$f(u,v) = \frac{1}{\sqrt{2\pi}\,\sigma_u\sigma_v} exp\left\{-\frac{u}{\sigma_u} - \frac{v^2}{2\sigma_v^2}\right\} \tag{3.30}$$

The marginal density function of ε for the production function is:

$$f(\varepsilon) = \int_0^{\infty} f(u,\varepsilon)du$$

$$= \left(\frac{1}{\sigma_u}\right)\Phi\left(-\frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right) exp\left\{\frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}\right\} \tag{3.31}$$

The log-likelihood function for the normal-exponential production model with N producers is:

$$lnL = constant - Nln\sigma_u + N\left(\frac{\sigma_v^2}{2\sigma_u^2}\right) + \sum_i \frac{\varepsilon_i}{\sigma_u} + \sum_i ln\Phi\left(\frac{\varepsilon_i}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right) \tag{3.32}$$

3.6.3 The Normal-Truncated Normal Model

The normal-truncated normal model is a generalization of the normal-half normal model by allowing the mean of u_i to differ from zero. Under the normal-truncated normal model, we assume the following:

1. The error term component $v_i\ is\ iid\ with\ N\ (0, \sigma_v^2)$
2. $u_i\ is\ iid\ with\ N^+(u, \sigma_u^2)$
3. $v_i and\ u_i\ independent\ of\ each\ other, and\ of\ the\ regressor$

The joint density of $v_i and\ u_i$ can be written as:

$$f(u,v) = \frac{1}{\sqrt{2\pi}\,\sigma_u\sigma_v\,\Phi\left(\frac{\mu}{\sigma_u}\right)} exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\} \tag{3.33}$$

The marginal density function of ε for the production function is:

$$f(\varepsilon) = \int_0^\infty f(u,\varepsilon)du$$

$$= \frac{1}{\sqrt{2\pi}\sigma\Phi(\frac{\mu}{\sigma_u})}\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right) exp\left\{-\frac{(\varepsilon+\mu)^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sigma}\phi\left(\frac{\varepsilon+\mu}{\sigma}\right)\Phi\left(\frac{\mu}{\sigma\lambda} - \frac{\varepsilon\lambda}{\sigma}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1} \tag{3.34}$$

The log-likelihood function for the normal-truncated normal production model with N producers is:

$$lnL = constant - Nln\sigma - Nln\Phi\left(\frac{\mu}{\sigma_u}\right)$$

$$+\sum_i ln\Phi\left(\frac{\mu}{\sigma\lambda} + \frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{1}{2}\sum_i\left(\frac{\varepsilon_i+\mu}{\sigma}\right)^2 \tag{3.35}$$

Since the random error u_i cannot be directly observed, it poses a problem with the above mentioned models. The suggestion to overcome this problem was given by Jondrow et al. (1982) to estimate the inefficiency using a conditioned distribution of u_i given ε_i. The specification of the distribution of *ui*, given it is exponential, is as follows:

$$E(u_i|\varepsilon_i) = \sigma_v\left\{\frac{f^*\left[\frac{\left(\frac{\varepsilon_i}{\sigma_v}\right)}{\left(\frac{\sigma_v}{\phi}\right)}\right]}{1-F^*\left[\frac{\left(\frac{\varepsilon_i}{\sigma_v}\right)}{\left(\frac{\sigma_v}{\phi}\right)}\right]} - \left[\frac{\left(\frac{\varepsilon_i}{\sigma_v}\right)}{\left(\frac{\sigma_v}{\phi}\right)}\right]\right\} \tag{3.36}$$

Where f^* is the standard normal density function and F^* is the cumulative distribution function. Once $E(u_i|\varepsilon_i)$ is known the efficiency of each observation can be estimated as $\exp[-E(u_i|\varepsilon_i)]$.

3.7 Conclusion

In this chapter the objective was to provide a detailed understanding of the various relative efficiency analysis models and approaches. The conventional DEA models have been extensively discussed along with the various variations of the core model. The concept of RTS and the limitations of DEA, with respect to the sensitivity of the data, have led to the discussion of the SFA. The SFA approach, with the three distributions to model the error terms, have provided the basis of the methodologies used in the dissertation.

Chapter 4

Outlier Correction Model

4.1 Introduction

Statistical theory regarding outlier analysis and detection is comprehensive. There are a numbers of approaches to detecting outliers. One such method is a clustering analysis, which creates clusters of similar values and very small clusters are potential outliers. A simpler method is using the generally accepted "68-95-99.7" rule. This rule suggests that an outlier's values that are greater than the mean plus three times the standard deviation; or smaller than the mean minus three times the standard deviation. All of these approaches are accurate and robust.

In DEA literature there has been a significant amount of research conducted in identifying outliers and influential DMUs Timmer (1971), Andersen and Petersen (1993), Wilson (1995), Simar (1996), Banker and Chang (2006) and Johnson and McGinnis (2008) to name a few.

Most of these studies focus around using DEA to identify and eliminate outliers and extreme values from the model. Some studies remove the influential DMUs from the reference set, which alters the efficiency score of the remaining DMUs due to the nature of the relative efficiency calculation.

This study focuses on harnessing the already defined and trusted methods of outlier detection to identify outliers in the data and then to compensate for the effect of these outliers in the DEA analysis. This allows for a more accurate view of the relative efficiency scores, as all of the DMUs in the reference set, with and without outliers, is still part of the analysis.

4.2 Proposed approach

The proposed model in this thesis harnesses these techniques and identifies extreme values in the data and then is accommodated for in the model. Relevant accommodation is made to the model. The model uses the identified extreme values and substitutes the extreme values when calculating the score for a particular DMU. In doing this, the score of the DMU being calculated is not influenced by the extreme values. When the score for the DMU with the extreme value is being calculated the actual value is used in the objective function.

The rationale for the proposed model is that removing the observation from the analysis is not always the best approach as DEA is a relative measure of efficiency. Removing the observation will inherently affect all remaining DMUs and the DMU with the outlier will have an efficiency score. The ability to correct for the unintended influence of the extreme values will provide a more accurate reflection of the relative efficiency and provide a more accurate projection onto the efficient frontier.
The proposed model is defined as follows:

$$\text{Max } \theta \tag{4.1}$$

Subject to:

$$\sum_{i=1}^{m} u_r[\,(y_{rj} * (1 - f_{rj}) + (a_{rj} * f_{rj})\,] \geq \theta y_{ro} \quad \forall\, j \in \{1,..,n\}$$

$$\sum_{i=1}^{m} v_r[\,(x_{rj} * (1 - g_{rj}) + (b_{rj} * g_{rj})\,] \geq x_{ro} \quad \forall\, j \in \{1,..,n\}$$

$$u_r \geq 0 \;\; \forall\, r \in \{1,..,s\}$$
$$v_i \geq 0 \;\; \forall\, i \in \{1,..,m\}$$

$$f_{rj} \in \{1,0\} \;\; \forall\, i \in \{1,..,s\}$$

$$g_{rj} \in \{1,0\}\, \forall\, i \in \{1,..,m\} \tag{4.2}$$

Where:

f_{rj} and g_{rj} are flags for extreme values in the outputs and inputs respectively

a_{rj} and b_{rj} are alternative values for the extreme values in the outputs and inputs respectively

The flags for the output and inputs are predetermined binary variables resulting from an outlier analysis of the data. The alternative values for these extreme values are median values for each output and input.

The common approach to identify extreme values in a data set is to use the interval of the mean plus/minus a specified coefficient times the standard deviation. This coefficient is normally 2, 2.5 or 3 representing a poor conservative, mild conservative and very conservative respectively. According to the generally accepted "68-95-99.7" rule, by taking the mean plus three times the standard deviation, one will include 99.73% of the observations. This means that only 0.27% of the data will be considered to be outliers.

Miller (1991) stated that there are three problems which arise when using this approach. The first is that data is assumed to follow a normal distribution which includes the extreme values. Secondly, the statistical measures used in the approach, the mean and the standard deviation are heavily influenced by the extreme values. Thirdly, Cousineau and Chartier (2010) stated that this approach is not likely to be effective in smaller population.

Leys et al. (2013) proposed using the Median Absolute Deviation (MAD) as an alternative to the generally accepted approach of the mean plus minus three standard deviations. Their approach states that the median plus/minus a specific coefficient times the MAD will provide a more robust measure for outlier detection. The MAD is defined as Huber (1981):

$$MAD = b\, M_i\left(\left|x_i - M_j\left(x_j\right)\right|\right) \tag{4.3}$$

According to Leys et al. (2013) the MAD is the most robust dispersion/scale measure in the presence of outliers, yet the decision making concerning the exclusion criteria of outliers (specific coefficient of 2 , 2.5 or 3) is subjective. Leys et al. (2013) suggested that 2.5 would be a reasonable choice. This paper uses the approach proposed by Leys et al. (2013) to identify extreme values in the data.

4.3 Data used for the Model

The data used in the models are provided by Statistics South Africa. The variables relating to the number of people that have the services in the study were obtained from the Census 2011 data. These services included the number of households with:

- Access to piped water
- Access to flushing toilets
- Access to electricity for lighting.

From the Financial Census of Municipalities 2011 one value was obtained:

- The total expenditure.

Descriptive Statistics on the input data Models are given in the Table 5 below. The distributions of the various groups are fairly symmetrical.

Table 5: Descriptive Statistics of input/output variables

	Variable	No. of households with access to Water	No. of households with access to Toilets	No. of households with access to Electricity	Total expenses
District Municipalities	**Mean**	161 985	81 786	153 349	1 830 650
	Std. Deviation	95 952	61 648	95 905	1 266 595
	Minimum	18 966	11 256	17 046	303 487
	Maximum	452 949	247 740	422 460	4 450 560
Metropolitan Municipalities	**Mean**	757 110	635 854	686 883	17 309 817
	Std Deviation	446 147	410 106	411 078	10 907 561
	Minimum	217 932	145 182	180 915	3 677 488
	Maximum	1 415 004	1 282 011	1 303 044	32 046 907
Sample of Local Municipalities	**Mean**	28 995	13 851	27 855	308 917
	Std Deviation	25 542	17 649	25 481	367 837
	Minimum	1 668	444	1 326	19 919
	Maximum	113 922	98 538	126 045	2 001 525

The input for the model is the total expenditure and the outputs of the model are the number of households with access to: electricity, water and toilets. This model was applied to three sets of DMUs:

- The 44 District municipalities (DC)
- The eight Metropolitan municipalities (MM)
- Sample of 80 Local municipalities (LM)

4.4 Conclusion

The objective of this study was to formulate a new model based on the conventional DEA model to correct for the effect of extreme values in the analysis of relative efficiency. The rationale for the proposed model has defined the need for such a model. The data used in the study to compare the proposed model with the two conventional deterministic DEA models and the three SFA models have been presented. This comparison will be applied to the three samples defined in this chapter. The next chapter deals with the results of the five models and the proposed model.

Chapter 5

Comparisons of conventional and proposed models

5.1 Introduction

The results of the five conventional models (two deterministic and three stochastic) are shown in Table 6 below. The descriptive statistics of the various models indicate that the proposed model has a higher mean than that of the two deterministic models and more DMUs are found to be efficient. According to Bahari and Emrouznejad (2014), three questions were asked in the evaluation of the performance of the proposed method:

1. How much the efficiency of removed DMU is changed?
2. How many DMUs are affected by the removed DMU?
3. How much is total change of the efficiencies?

Since the proposed method in this thesis does not remove observation but merely corrects for extreme values, this study poses the following questions to evaluate the performance of the proposed model:

1. How many municipalities had extreme values?
2. How many municipalities are affected by the correction for extreme values?
3. How much is total change of the efficiencies?

These questions are addresses in the sections that follow.

5.2 Overall Results

The results in general of the conventional models on the three groups show that there are significant differences in the means and standard deviations of the each approach. Table 6 shows the descriptive statistics of the conventional models per group. The deterministic models have higher standard deviations with varying mean values indicating that the efficiency score of the municipalities are fairly scattered across the range. The SFA-EXP and SFA-HALF models have much higher means and smaller standard deviations indicating that the efficiency score of these municipalities are clustered around the mean. This is the stochastic attempt to detecting random noise in the data. However, this approach is not necessarily better.

The two deterministic models are fairly similar in their results. The range for the DEA-CCR and DEA-BCC models differ by only 0.01 for the District municipality group and 0.03 for the Local municipality group. There is a 0.16 difference in the range for the deterministic models which can be attributed to the small sample size.

Table 6: Results of the model per municipal group

	Variable	DEA-CCR	DEA-BCC	SFA-EXP	SFA-TRUN	SFA-HALF
Type=District municipality	**Mean**	0.68	0.77	0.89	0.58	0.84
	Standard Deviation	0.16	0.18	0.03	0.14	0.06
	Range	0.61	0.60	0.13	0.63	0.25
	N	44	44	44	44	44
	No. Efficient	4	9	-	1	-
Type=Metro municipality	**Mean**	0.87	0.97	0.85	0.87	0.85
	Standard Deviation	0.11	0.06	0.09	0.09	0.09
	Range	0.32	0.16	0.25	0.26	0.25
	N	8	8	8	8	8
	No. Efficient	1	6	1	2	1
Type=Local municipality	**Mean**	0.56	0.62	0.81	0.54	0.79
	Standard Deviation	0.19	0.22	0.07	0.10	0.07
	Range	0.73	0.70	0.29	0.42	0.28
	N	80	80	80	80	80
	No. Efficient	5	12	-	-	-

5.2.1 Metro Group

The results of the DEA-CCR model found one metro to have an efficiency score of 1 and the second nearest efficiency score was 0.92. The range for the eight metros was found to be 0.32 with mean of 0.68 and a standard deviation of 0.1. The DEA-BCC model found six Metros having an efficiency of 1 with a higher mean of 0.97 and lower standard deviation of 0.059. The range for this model was halved to 0.16 when compared to the DEA-CCR model. The variable returns to scale model, tends to group the efficiency score of the metros together.

The results of the three SFA (EXP, TRUN, HALF) models in the Metro municipality group are almost identical. The SFA-TRUN model found two metros with efficiency scores of 1. The SFA-EXP model for the Metro municipality group found one metro with an efficiency score of 1, yet the range was 0.25 with a mean of 0.84 and a standard deviation of 0.088

A Pearson correlation analysis supports that conclusion that they are statistically the same. This in indicated in Table 7 below. The correlation coefficients for the SFA models are all 0.95 with p-value < 0.0002. Since the three SFA models are statistically indifferent, only one was used, SFA-EXP, when the DEA-CCR model was compared with the DEA-BCC model using Spearman's Correlation to analyse the ranking of the eight metros. The correlation coefficient between the DEA-CCR and DEA-BCC model was 0.1 and the p-value was 0.81. Similarly the DEA-CCR and SFA-EXP coefficient was 0.58 with a p-value of 0.12. This indicates that there is a close relationship between the two deterministic models and a statistically difference between the stochastic models.

In all five models the Ethekwini was shown to have and efficiency score of 1. This is evidence of the data exhibits consistency throughout the various models. The stochastic models clustered the efficiency scores. By analysing the descriptive statistics of the metro municipalities we find a trend in the range and numbers of metros with an efficiency score of 1. The DEA-BCC model had a higher range, 0.32, with one metro having an efficiency score of 1. The DEA-CCR model's range half at 0.16 but found six metros with an efficiency score of 1. The three stochastic models SFA-EXP, SFA-TRUN and SFA-HALF had ranges of 0.25, 0.26 and 0.25 respectively. The numbers of metros with an efficiency score of 1 in three stochastic models SFA-EXP, SFA-TRUN and SFA-HALF were one, two and one. The DEA-BCC model was the best performer in the metro municipality group, which found the most metros to have an efficiency score of one. The stochastic models efficiency scores were clustered. The DEA-CCR model provided only one municipality with an efficiency score of

1. This indicates that the variable returns to scale model is the better approach for the metro municipalities.

Table 7: Pearson Correlation: Metro Group

Pearson Correlation Coefficients, N = 8 *Prob > |r| under H0: Rho=0*					
	DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
DEA-BCC	1	-0.00219 *0.9959*	0.25867 *0.5362*	0.22341 *0.5948*	0.25867 *0.5362*
DEA-CCR		1	0.56866 *0.1413*	0.69641 *0.055*	0.56866 *0.1413*
SFA-EXP			1	0.95911 *0.0002*	1 *<.0001*
SFA-TRUN				1	0.95911 *0.0002*
SFA-HALF					1

5.2.2 District Group

The results of the district municipalities show that the DEA-CCR found four districts to have an efficiency score of 1, with a mean of 0.68 and a standard deviation of 0.16. The range for this model was 0.61. The DEA-BCC model found nine districts to have an efficiency score of 1, with a mean of 0.77 and a standard deviation of 0.18. The range for this model was 0.60.

Only one of the stochastic models, SFA-TRUN, found districts with an efficiency score of 1. The SFA-TRUN found one district with the efficiency score of 1 with a mean of 0.57 and a standard deviation of 0.13. The range for this model was 0.63. The SFA-EXP model found no municipalities with an efficiency score of 1, and had a mean of 0.88 with a standard deviation of 0.13. The SFA-HALF also had no districts with an efficiency score of 1 with a mean of 0.84 and standard deviation of 0.056. The ranges for the SFA-EXP and SFA-HALF were 0.13 and 0.25 respectively. This shows that the relationships between the three stochastic models are not as close as noted in the metro municipalities' group. These models have no statistical difference. Table 8 shows the results of the Pearson Correlation analysis.

Statistically there is no difference in the five models. All the correlation coefficients are very close to 1 with p-values <0.0001.

The results of the models in the district municipalities are different to the results from metro municipalities. In the metro group we found models with smaller ranges yielded more districts with efficiency scores of 1. In the district group we do not observe the same trend. The DEA-BCC and DEA-CCR model have a difference of 0.01 in their ranges yet, the DEA-BCC model found five more districts with an efficiency score of 1.

There were also statistically differences in the deterministic and stochastic models in the metro municipality groups which do not exist in the district municipality group.

Table 8: Pearson's Correlation: District Group

Pearson Correlation Coefficients, N = 8 Prob > |r| under H0: Rho=0					
	DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
DEA-BCC	1	0.86529 <.0001	0.83444 <.0001	0.73221 <.0001	0.82371 <.0001
DEA-CCR		1	0.90021 <.0001	0.89943 <.0001	0.89645 <.0001
SFA-EXP			1	0.87379 <.0001	0.97754 <.0001
SFA-TRUN				1	0.89708 <.0001
SFA-HALF					1

A rank correlation of the models in the district municipality group was conducted and the results of the Spearman's correlation analysis are found in Table 9. There not statistically differences found in the five conventional models. All p-values are < 0.0001.

Table 9: Spearman Correlation: District Group

Pearson Correlation Coefficients, N = 44 Prob > |r| under H0: Rho=0					
	DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
DEA-BCC	1	0.86857 <.0001	0.80975 <.0001	0.77418 <.0001	0.80815 <.0001
DEA-CCR		1	0.89601 <.0001	0.89112 <.0001	0.89312 <.0001
SFA-EXP			1	0.9713 <.0001	0.95709 <.0001
SFA-TRUN				1	0.9721 <.0001
SFA-HALF					1

5.2.3 Local Group

The results of the models in the Local municipality group are quite similar to the district municipality group. The DEA-CCR model found five local municipalities with an efficiency score of 1, a mean of 0.56 and a standard deviation of .19. The range for this model was the highest out of the entire set of models at 0.73. The DEA-BCC model found 12 local municipalities with an efficiency score of 1, a mean of 0.62 and standard deviation of 0.22. The range for this model was 0.7.

The stochastic models found no local municipality with an efficiency score of 1. The SFA-EXP model had a mean of 0.8 and standard deviation of 0.072 with a range of 0.29. The SFA-TRUN model had a mean of 0.53 and standard deviation of 0.1 with a range of 0.42. The SFA-HALF model had a mean of 0.79 and standard deviation of 0.07 with a range of 0.28.

The three stochastic models are statistically equal. The DEA-BCC model correlated with the DEA-CCR model has a coefficient of 0.93 and a p-value of <0.0001. This correlation is similar to the DEA-BCC model and the SFA-EXP model where the coefficient is 0.86 and the p-value is <0.0001. The full results of the Pearson Correlation analysis are given in Table 10 below.

Table 10: Pearson Correlation: Local Group

Pearson Correlation Coefficients, N = 80 Prob > |r| under H0: Rho=0					
	DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
DEA-BCC	1	0.93034 <.0001	0.86738 <.0001	0.91097 <.0001	0.88363 <.0001
DEA-CCR		1	0.8699 <.0001	0.9423 <.0001	0.88643 <.0001
SFA-EXP			1	0.92441 <.0001	0.99649 <.0001
SFA-TRUN				1	0.94068 <.0001
SFA-HALF					1

A rank correlation of the models in the local municipality group yielded the same result, that the entire set of models are statistically the same. There are strong rank correlations between the deterministic and stochastic models. The results of the correlation are given in Table 11 below.

Table 11: Spearman Correlation: Local Group

Spearman Correlation Coefficients, N = 80 Prob > |r| under H0: Rho=0					
	DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
DEA-BCC	1	0.96135 <.0001	0.95345 <.0001	0.93301 <.0001	0.95082 <.0001
DEA-CCR		1	0.97175 <.0001	0.96433 <.0001	0.96947 <.0001
SFA-EXP			1	0.97836 <.0001	0.99752 <.0001
SFA-TRUN				1	0.97712 <.0001
SFA-HALF					1

5.2.4 Data Imputation

Table 12 below shows the correlation analysis conducted on the five conventional models when the data was modified. This modification involved substituting the outlier values found with the criteria limit (median -/+ 2.5* standard deviation).

The results of the correlation analysis with the imputed data suggest that the imputation had little effect on the model. Although the models are statistically equal to one another the relationship is much weaker. The correlation coefficients are closer to 0. The biggest impact was between the SFA-EXP model, when compared to the SFA-EXPA imputed results yields a correlation coefficient of 0.54 and a p-value <0.0001.

Table 12: Data Imputation

		Pearson Correlation Coefficients, N = 44				
		Prob > |r| under H0: Rho=0				
				Imputed Data		
		DEA-BCC	**DEA-CCR**	**SFA-EXP**	**SFA-TRUN**	**SFA-HALF**
Original Data	**DEA-BCC**	0.99229	0.7703	0.35358	0.61509	0.4234
		<.0001	<.0001	0.0185	<.0001	0.0042
	DEA-CCR	0.87064	0.93881	0.56686	0.81309	0.62345
		<.0001	<.0001	<.0001	<.0001	<.0001
	SFA-EXP	0.83845	0.84061	0.54614	0.74023	0.59551
		<.0001	<.0001	0.0001	<.0001	<.0001
	SFA-TRUN	0.74257	0.85535	0.50657	0.73127	0.52367
		<.0001	<.0001	0.0005	<.0001	0.0003
	SFA-HALF	0.82366	0.8248	0.52301	0.70492	0.55701
		<.0001	<.0001	0.0003	<.0001	<.0001

5.2.5 Proposed Model Results

Since the proposed model in this study was based on the conventional BCC model, at first glance the results of the proposed model are very similar to the DEA-BCC model. Table 13 shows the results of the proposed model per municipal group. To fully understand the impact of the results of the proposed model it must be broken up into the municipalities with extreme values and municipalities without extreme values.

Table 13: Results of the proposed model per municipal group

	Variable	Proposed model
Type=District municipality	**Mean**	0.77
	Standard Deviation	0.18
	Range	0.6
	N	44
	No. Efficient	8
Type=Metro municipality	**Mean**	0.97
	Standard Deviation	0.059
	Range	0.16
	N	8
	No. Efficient	6
Type=Local municipality	**Mean**	0.65
	Standard Deviation	0.22
	Range	0.70
	N	80
	No. Efficient	16

5.3 Comparison

The comparisons of the results from the proposed model and the conventional models are given in Table 14 below. In each case, the movement of the efficiency score of the conventional models are given relative to the proposed model. Each comparison is broken up by the extreme values flag indicating that municipality had an extreme value in one of its values. There are significant increases in the efficiency scores of the municipalities without extreme values and a decrease in the scores with extreme values. Some municipalities with extreme values also have an increase in the efficiency score; this can be attributed to the influence of the other municipalities with extreme values.

In the District municipality group, there were 4 municipalities with extreme values. This led to an average increase of 0.08 affecting 35 municipalities when compared to the DEA-CCR model. The total change resulted in 2.81. This is evidence that the effect of the 4 municipalities with extreme values had understated the efficiency scores of 35 municipalities.

Additionally there were 4 municipalities without extreme values that were found to be efficient in the proposed model that were not efficient in the DEA-CCR model.

The exponetial model overstated 30 municipalities, by an average of 0.2. This equated to a total change of 6.11. The half-normal model overstated 28 municipalities with an average change of 0.16 and a total change of 4.4.

The SFA-TRUN model had understated the efficiency score and the proposed model showed an average increase of 0.18. This equates to a total change of 6.84, affecting 39 municipalities.

The Metro municipality group had only 8 municipalities, no extreme values were found and the results of the proposed model are identical to its parent model DEA-BCC.

The Local municipality group had 80 municipalities, of which 7 were found to have an extreme value. The DEA-CCR model understated 59 municipalities with an average of 0.09 and a total change of 5.09. The DEA-BCC and SFA-TRUN models also understated the efficiency score by a total of 1.93 and 7.69, affecting 31 and 44 municipalities respectively. Similarly the exponential model overstated the efficiency score of 61 municipalities with an average of 0.24 and a total change of 14.36. The half-normal model overstated the efficiency score of 61 municipalities with an average of 0.22 and a total change of 13.49. The proposed model found 4 more efficient municipalities than the DEA-BCC model and 11 when compared to the DEA-CCR.

The proposed model decreased the efficiency score of these two SFA models in municipalities flagged as having no extreme values. It can be argued that the SFA–EXP and SFA-HALF models have been significantly influenced by extreme values resulting in a higher average score and smaller standard deviation. The comparison of the proposed model to both SFA-EXP and SFA-HALF models have corrected for the extreme values influence reducing the efficiency.

Table 14: Movement of the conventional model relative the proposed model

			DEA-CCR			DEA-BCC			SFA-EXP			SFA-HALF			SFA-TRUN		
Type	Change Relative to Proposed model	Extreme Value	Sum	Count	Average	Sum	Count	Average	Sum	Count	Average	Sum	Count	Average	Sum	Count	Average
District municipality	Down	No				-0.05	3.00	-0.02	-6.11	30.00	-0.20	-4.40	28.00	-0.16			
	Same	No	0.00	5.00	0.00	0.00	36.00	0.00							0.00	1.00	0.00
	Up	No	2.81	35.00	0.08	0.04	1.00	0.04	0.58	10.00	0.06	0.81	12.00	0.07	6.84	39.00	0.18
	Down	Yes				-0.01	1.00	-0.01									
	Same	Yes				0.00	3.00	0.00									
	Up	Yes	1.05	4.00	0.26				0.35	4.00	0.09	0.51	4.00	0.13	1.68	4.00	0.42
Metro municipality	Down	No															
	Same	No	0.00	3.00	0.00	0.00	8.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	2.00	0.00
	Up	No	0.83	5.00	0.17				0.99	7.00	0.14	0.99	7.00	0.14	0.78	6.00	0.13
	Down	Yes															
	Same	Yes															
	Up	Yes															
Local municipality	Down	No							-14.36	61.00	-0.24	-13.49	61.00	-0.22	-1.19	26.00	-0.05
	Same	No	0.00	14.00	0.00	0.00	42.00	0.00							0.00	3.00	0.00
	Up	No	5.09	59.00	0.09	1.93	31.00	0.06	1.35	12.00	0.11	1.51	12.00	0.13	7.69	44.00	0.17
	Down	Yes							-0.09	1.00	-0.09	-0.08	1.00	-0.08			
	Same	Yes	0.00	1.00	0.00	0.00	3.00	0.00									
	UP	Yes	1.92	6.00	0.32	0.52	4.00	0.13	0.71	6.00	0.12	0.78	6.00	0.13	2.53	7.00	0.36

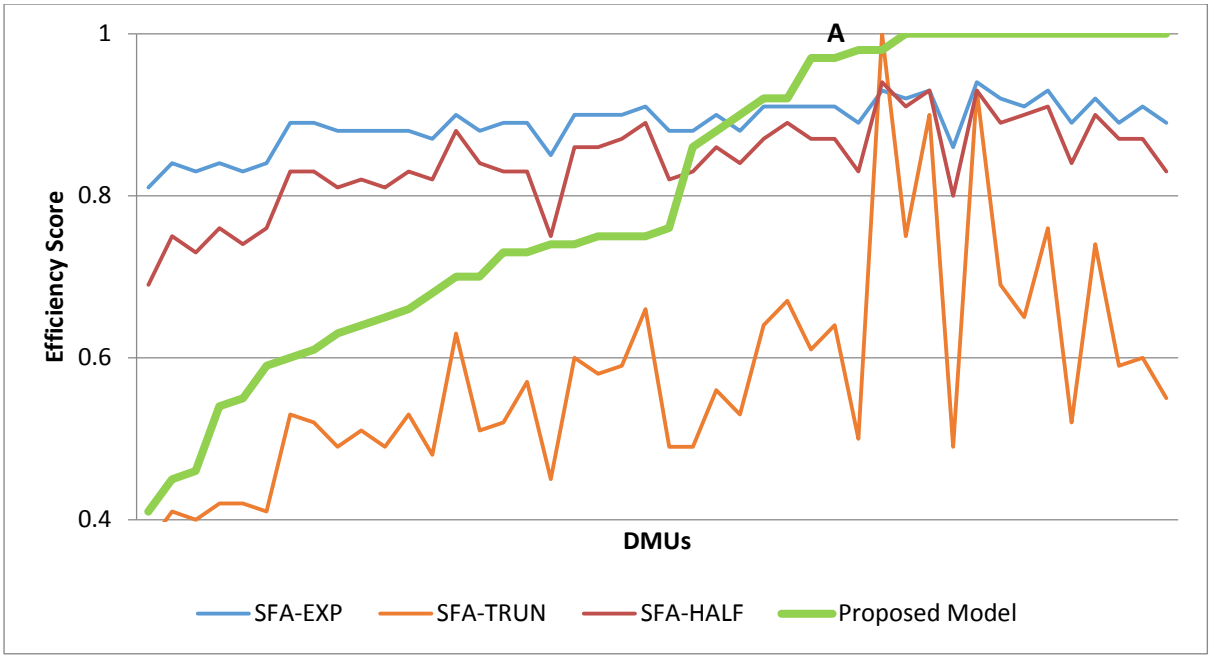

Figure 5: District municipality Group: Proposed model vs. SFA Models

Based on the above table, it is difficult to see the results of the proposed model compared to the conventional models. Figure 5 above shows a graphical representation of the results. It compares the proposed model to the stochastic models for the District municipality group.

When examining Figure 5, we find that the SFA-EXP and SFA-HALF models are very similar in the calculation of the efficiency scores; this was also found when interpreting Table 8. What is not clear in Table 8, but evident in the graphical representation, is the improvement of the efficiency score of the proposed model when compared to the SFA-TRUN model.

The proposed model reflects a higher score in 99% of the observations when compared to the SFA-TRUN model. Approximately 45% of the observations had a higher score in the proposed model when compared to the remaining stochastic models. The proposed model also found more efficient municipalities than all three of the stochastic models. Point A is the only instance where the SFA-TRUC yielded a higher score than the proposed model.

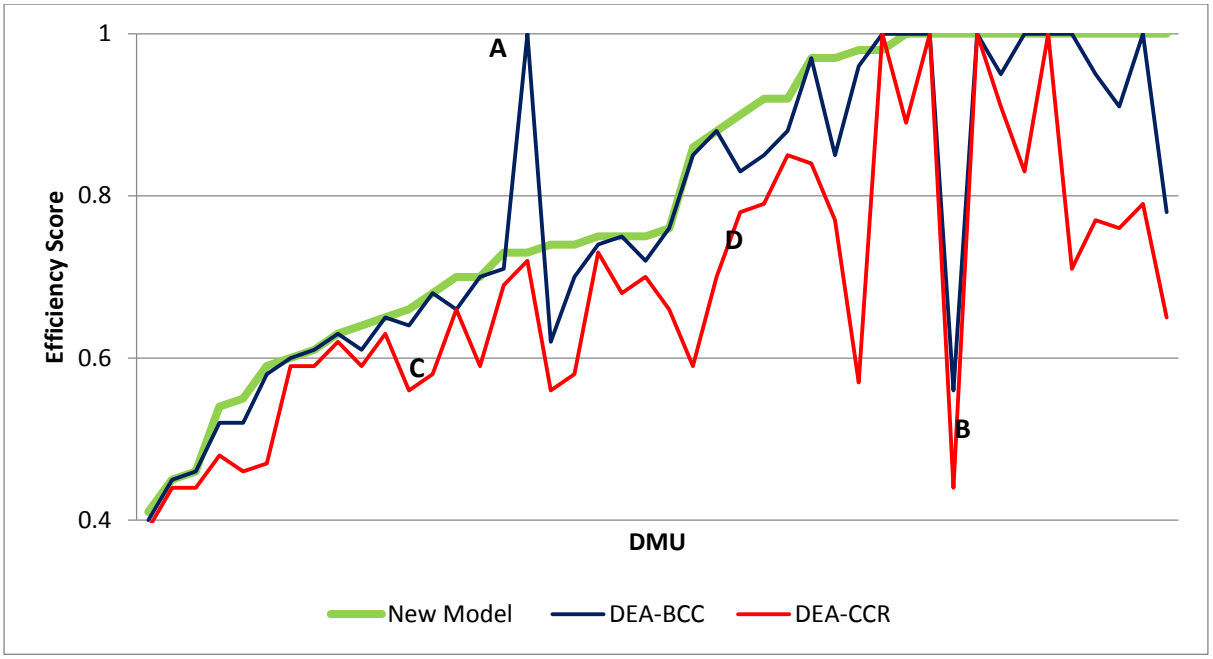

Figure 6: District municipality Group: Proposed model vs. DEA Models

Figure 6 above, shows the proposed model compared to the deterministic models for the District municipality group. There are significant variations in the different models. The proposed model performs better than the DEA-BCC model with two major differences. Point A on the DEA-BCC curve shows that the model was found to be efficient, but the proposed model yielded only 75% efficiency. Conversely, point B shows that both deterministic models found the observation to have an efficiency score of less than 60%, yet the proposed model found the same observation to be fully efficient.

Point C is an example of both the DEA-BCC and DEA-CCR models finding the same efficiency scores as the observations, but the proposed model found a higher efficiency score.

The point D, is one of many instances on the graph that show that the DEA-BCC has improved on the DEA-CCR model score but the proposed model found a higher efficiency score than both models.

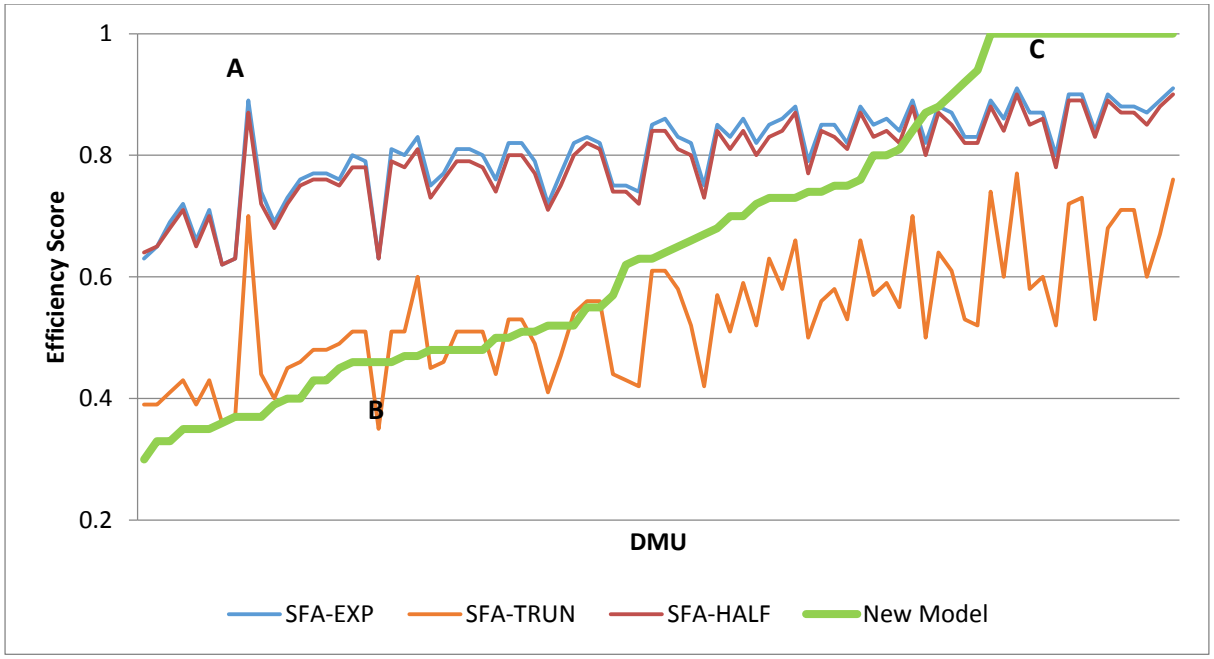

Figure 7: Local municipality Group: Proposed model vs. SFA Models

Figure 7 above is the graphical illustration of the proposed model and the stochastic models in the Local municipality group. Point A in the figure shows an instance where the observation had a value which was deemed as an extreme value, but all three stochastic models over stated the efficiency score. The proposed model was the lowest score for this municipality.

The municipality at point B had no extreme value but the stochastic models for this observation had two distinct scores. The SFA-EXP and SFA-HALF models scored this municipality at 63% and the SFA-TRUN model was 35%. The proposed model scored this municipality at 46%. The effect of the extreme values in the group had affected this municipality and yielded an inaccurate score.

Point C is an example of numerous instances where the proposed model found the municipality to be efficient and the three stochastic models varied in their scores.

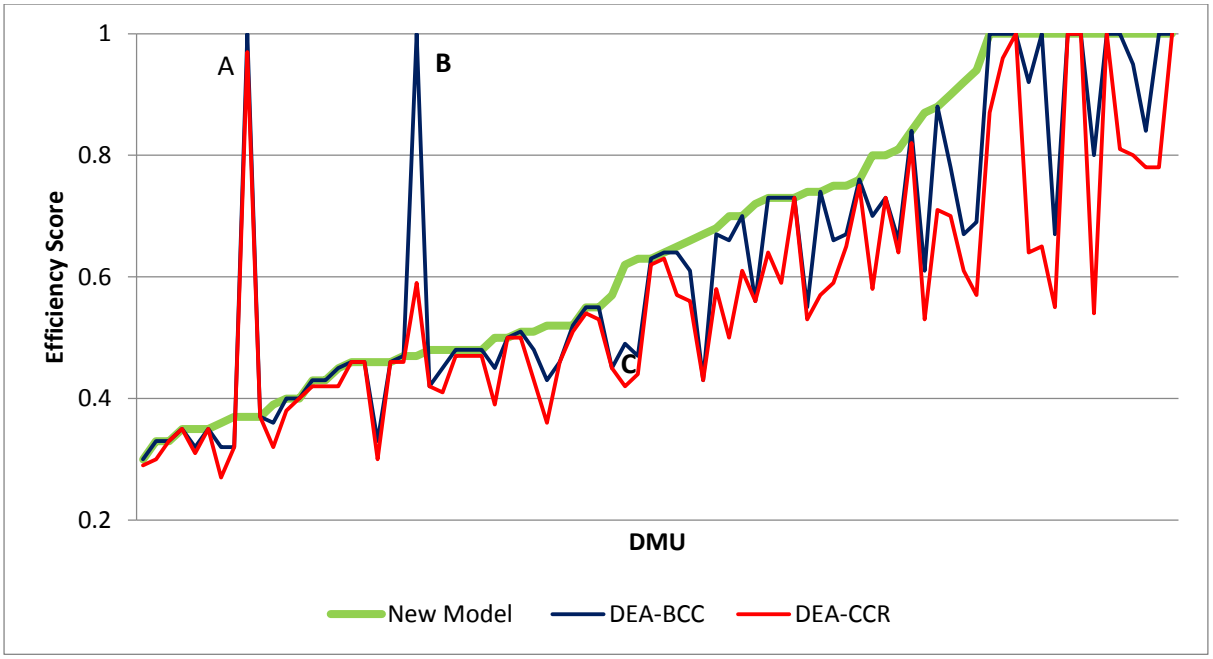

Figure 8: Local municipality Group: Proposed model vs. DEA Models

Figure 8 below shows the proposed model with the two deterministic models in the Local municipality group. Point A refers to a municipality that had an extreme value in the data. The DEA-BCC and DEA-CCR model scored this municipality at 100% and 97% respectively. The proposed model corrected for the effect of this extreme value and scored the municipality at 37%.

Point B is another example of where a municipality with an extreme value was scored at 100% in the DEA-BCC model and 59% in the DEA-CCR model, but was given a more accurate score in the proposed model with 47%.

Point C is an example of an instance where the municipality had no extreme value but the DEA-CCR and DEA-BCC models both scored the municipality at 43%. The proposed model, taking into account the effect of other municipality with extreme values, scored the municipality at 63%.

5.4 Conclusion

The results of the study have shown that the proposed model is effective in correcting for extreme values in a relative efficiency analysis. There are numerous instances where an extreme value was found in the data for a municipality and the conventional methods either over stated or understated the efficiency score. The proposed model has compensated for this extreme value and has allowed for a more accurate reflection of the relative efficiency. The graphical illustrations of the proposed model against the conventional models have shown that there are significant differences in the scores of municipalities with extreme values. The results have shown that the DEA-BCC model and DEA-CCR model have taken the extreme value effects into the analysis. This is evident in point A of figure 8.

The general trend observed as the sample size increased is the stronger correlation between the conventional models. The metro municipality group showed that there exists a statistically difference in the deterministic and stochastic models. Between the stochastic models there exhibited no statistically difference in the efficiency scores, on both product and rank correlations.

The number of municipalities that were found to be efficient in the District municipality group for the DEA-CCR and DEA-BCC models is 4 and 9 respectively. There was a major shift in this number in the Metro municipality group, 1 for the DEA-CCR model and 6 for the DEA-BCC model. In the Local municipality group, 5 municipalities were found to be efficient for the DEA-CCR model and 12 for the DEA-BCC model.

The district municipality group showed that there was no statistically difference in the deterministic and stochastic models.

In the conventional DEA-BCC model for the Local municipality group, seven municipalities were found to have an extreme value. This has resulted in a change of efficiency score to 34 municipalities. All 34 of these municipalities experienced an increase in efficiency score.

The results from the stochastic models are very different across the three groups. For the District municipality group the SFA-EXP model had a range of 0.13, with a mean of 0.88 and standard deviation of 0.029 with no municipalities found to be efficient. There were no efficient municipalities in stochastic models in the Local municipality group. The results are of the SFA-EXP and SFA-HALF models are almost identical.

The SFA-TRUN model showed an improvement on the efficiency analysis, when compared to the remaining stochastic models. This model found 1 municipality to be efficient in the District municipality group, 2 in the Metro municipality group and none in the Local municipality group.

The proposed model showed a significant increase in the efficiency score in all three groups. The effect of the outliers dropped certain municipality scores and increased the scores of the remaining municipalities. Although the proposed model may not have always found more municipalities to be efficient, it did provide an increased efficiency score.

This is an important finding in this study. The proposed model has shown that outliers can cause inaccurate efficiency scores. These scores will lead to inaccurate projections onto the efficient frontier. The increase of the efficiency scores in the proposed model and the correction for the effect of the outliers have allowed for a more accurate projection onto the efficient frontier.

Chapter 6

Conclusions

One of the aims of this research was to propose a new approach to accommodate for extreme values in the input and output data of a DEA model. This study applied the approach to three populations of municipalities in South Africa and compares our results to two conventional DEA models and three Stochastic Frontier Analysis models. The sensitivity of DEA to extreme values has been one of the limitations of this approach. Another research aim was to address this problem in the context of South African Local Municipalities.

This study conducted an extensive review of the literature into relative efficiency analysis, focusing on data envelopment analysis, stochastic frontier analysis and outlier detection analysis in a DEA model. A review of the studies conducted into municipal relative efficiency has also been provided. This dissertation addresses the need for quality data and the approach that should be undertaken when there is insufficient administrative data of acceptable quality available. An analysis into the relative efficiency can still be conducted using the proposed method.

The explanation of the various methodologies used in this study has been provided along with an assessment of the strengths and weaknesses of approaches. The formulation of the proposed approach has been clearly defined and applied to the data along with the five conventional methods.

The results of the research have shown the impact of extreme values in a relative efficiency analysis. The results of the correlation analysis of the three groups indicate a specific trend. The bigger the sample size, the more the models are correlated with one another. The eight

metro municipalities had distinct statistical differences in the deterministic and stochastic models, the strength of this correlation increased as the sample size increased. As the sample size increased in each group, the strength of the correlations also increased. The ranges for the models started to increased and fewer municipalities were found having an efficiency score of 1. Only one municipality in the metro municipality group was consistent across all five conventional models and the proposed model obtaining an efficiency score of 1.

The test conducted to impute the outlier values showed no statistically difference in the models. The results of the correlation analysis with the imputed data suggest that the imputation had little effect on the model. Although the models are statistically equal to one another the relationship is much weaker. This could be attributed to the extreme value moving to the upper or lower limit of the criteria. The difference in the extreme value and the criteria value may be negligible.

The proposed model which accommodates for the outlier has shown to vastly improve the efficiency score across the entire set of model and groups. The graphically illustrations provided in chapter 5, clearly show the impact of the outliers, not only on the outlier municipality but for all the remaining municipalities. There are significant changes in the efficiency score.

In the Local municipality group, seven municipalities with outliers influenced the scores of 34 municipalities in the same group. municipalities which were given lower efficiency scores, under the conventional models with extreme values. The proposed method has shown that the compensation of the extreme values in the group has improved the accuracy of the efficiency scores.

The effect of the outliers would have caused inaccurate projections and requirements for a municipality to become efficient.

Future reseach into this field can include the use of varying techniques for outlier detection. Compliment the Median Absolute Deviation method with more intense analysis such as the use of a cluster analysis to detect anomalies in the data. One can conduct further validation of this proposed model in other context. There could have existed nuiances in the sphere of municipalities and service delivery that could negatively affect the results of a validation. For example, an area could have had piped water for the last 25 years, yet this municipality may

be shown as efficient when compared to a municipality that has only recently provided piped water to all its residences.

Bibliography

1. Afonso, A & Fernandes, S, 2008. Assessing and explaining the relative efficiency of local government. The Journal of Socio-Economics 37, 1946–79.
2. Afonso, A., & Fernandes, S. (2006). Measuring local government spending efficiency: evidence for the Lisbon region. Regional Studies, 40(1), 39-53.
3. Aigner, D.; Lovell, C. A. K.; Schmidt, P. (1977): Formulation and estimation of stochastic frontier production function models. In: Journal of Econometrics 6, 1, 21–37.
4. Aigner, Dennis J. and S. F. Chu (1968), "On Estimating the Industry Production Function," American Economic Review, 17, 826-39.
5. Andersen, P., & Petersen, N. (1993). A procedure for ranking efficient units in data envelopment analysis. Management Science, 39, 1261–1264.
6. Athanassopoulos, A & Triantis, K, 1998. Assessing aggregate cost efficiency and the related policy implications for Greek local municipalities. INFOR 36(3), 66–83.
7. Bahari, A. R., & Emrouznejad, A. (2014). Influential DMUs and outlier detection in data envelopment analysis with an application to health care. *Annals of Operations Research*, 1-14.
8. Balaguer-Coll, M, Prior-Jimenez, D & Vela-Bargues, J, 2002. Efficiency and Quality in Local Government Management. The Case of Spanish Local Authorities. Universitat Autonoma de Barcelona.
9. Banker R, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science. 1984;30:1078–92.
10. Banker, R. D., & Chang, H. (2006). The super-efficiency procedure for outlier identification, not for ranking efficient units. European Journal of Operational Research, 175(2), 1311–1320.
11. Banker, R. D., & Gifford, J. L. (1988). A relative efficiency model for the evaluation of public health nurse productivity. Working paper, Fox School of Business, Temple University, Philadelphia, PA.
12. Bellini, T. (2012). Forward search outlier detection in data envelopment analysis. *European Journal of Operational Research*, *216*(1), 200-207.

13. Boetti, L, Piacenza, M & Turati, G, 2010. Decentralization and local governments' performance: How does fiscal autonomy affect spending efficiency? Paper presented at the 66th Congress of the International Institute of Public Finance (IIPF), 23–26 August 2010, Uppsala, Sweden.
14. Cazals C, Florens J-P, Simar L (2002) Nonparametric frontier estimation: a robust approach. Journal of Economics 106:1–25
15. Cazals, C., Florens, J. P., & Simar, L. (2002). Nonparametric frontier estimation: a robust approach. Journal of econometrics, 106(1), 1-25.
16. Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European Journal of Operations Research. 1978; 2:429–44.
17. Chen, W. C., & Johnson, A. L. (2010). A unified model for detecting efficient and inefficient outliers in data envelopment analysis. Computers & Operations Research, 37(2), 417-425.
18. Coelli, T., D.P. Rao, and G. Batteese. An introduction to efficiency and productivity analysis. London, Kluwer Academic Publishers. 1998.
19. Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). Handbook on data envelopment analysis (Vol. 164). Springer.
20. Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. International Journal of Psychological Research, 3(1), 58–67.
21. Datar, S., Banker, R. D., & Das, S. (1990). Analysis of Cost Variances for Management Control in Hospitals. In Research in Governmental and Nonprofit Accounting, 5, 269–291.
22. De Borger B. and Kerstens K. (1996a) Cost efficiency of Belgian local governments: a comparative analysis of FDH, DEA, and econometric approaches, Regional Science and Urban Economics 26, 145–170.
23. De Borger, B., & Kerstens, K. (1996). Cost efficiency of Belgian local governments: A comparative analysis of FDH, DEA, and econometric approaches. Regional Science and Urban Economics, 26(2), 145-170.
24. De Borger, B., Kerstens, K., Moesen, W., & Vanneste, J. (1994a). Explaining differences in productive efficiency: An application to Belgian municipalities. Public Choice, 80(3-4), 339-358.
25. De Borger, B., Kerstens, K., Moesen, W., & Vanneste, J. (1994b). A non-parametric free disposal hull (FDH) approach to technical efficiency: an illustration of radial and

graph efficiency measures and some sensitivity results. Swiss Journal of Economics and Statistics, 130(4), 647-667.

26. Deller, S. C., & Halstead, J. M. (1994). Efficiency in the production of rural road services: The case of New England towns. Land Economics, 247-259.
27. Deller, S. C., & Nelson, C. H. (1991). Measuring the economic efficiency of producing rural road services. American Journal of Agricultural Economics,73(1), 194-201.
28. Deller, S. C., Nelson, C. H., & Walzer, N. (1992). Measuring managerial efficiency in rural government. Public Productivity & Management Review, 355-370.
29. Fare, R., Grosskopf, S., Kirkley, J. L., & Squires, D. (2001). Data envelopment analysis (DEA): a framework for assessing capacity in fisheries when data are limited.
30. Farrell, M. (1957), 'The measurement of productive efficiency', Journal of the Royal Statistical Society, Series A General 120, 253–281.
31. Huber, P. J. (1981). Robust statistics. New York: John Wiley.
32. Johnson, A. L., & McGinnis, L. F. (2008). Outlier detection in two-stage semiparametric DEA models. European Journal of Operational Research, 187(2), 629–635.
33. Jondrow, J.; Lovell, C. A. K.; Materov, I.; Schmidt, P. (1982): On the estimation of technical inefficiency in the stochastic frontier production function model. In: Journal of Econometrics 19, 2–3, 233–238.
34. Kingyens, A. T. Y. T. (2012). Bankruptcy Prediction of Companies in the Retail-apparel Industry using Data Envelopment Analysis (Doctoral dissertation, University of Toronto)
35. Kuosmanen, T., & Johnson, A. L. (2010). Data envelopment analysis as nonparametric least-squares regression. Operations Research, 58(1), 149-160.
36. Kuosmanen, T., & Post, G. T. (1999). Robust efficiency measurement. Rotterdam Institute for Business Economic Studies (RIBES), Rotterdam. Report, 9911.
37. Kutlar A., Bakirci, F., & Yüksel, F. (2012). An analysis on the economic effectiveness of municipalities in Turkey. African Journal of Marketing Management, 4(3), 80-98.

38. Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology.
39. Loikkanen, H & Susiluoto, I, 2005. Cost efficiency of Finnish municipalities in basic service provision 1994–2002. Paper prepared for the 45st Congress of the European Regional Science Association in Amsterdam, the Netherlands, 23–7 August.
40. Meeusen, W.; van den Broeck, J. (1977): Efficiency estimation from Cobb-Douglas production functions with composed error. In: International Economic Review 18, 2, 435–444.
41. Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. The Quarterly Journal of Experimental Psychology, 43(4), 907–912
42. Mingwen Yang, Guohua Wan & Eric Zheng (2014) A predictive DEA model for outlier detection, Journal of Management Analytics, 1:1, 20-41,
43. Prieto, A. M., & ZofIo, J. L. (2001). Evaluating effectiveness in public provision of infrastructure and equipment: the case of Spanish municipalities. Journal of productivity Analysis, 15(1), 41-58.
44. Rouse, P., Putterill, M., & Ryan, D. (1997). Towards a general managerial framework for performance measurement: A comprehensive highway maintenance application. Journal of Productivity Analysis, 8(2), 127-149.
45. Simar, L. (1996).Aspects of statistical analysis in DEA-type frontiers models. Journal of Productivity Analysis, 7, 117–185.
46. Simar, L. (2003). Detecting outliers in frontier models: A simple approach. Journal of Productivity Analysis, 20, 391–424.
47. Simar, L., &Wilson, P.W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science, 44(1), 4961.
48. Simar, L.,&Wilson, P.W. (2000).Ageneral methodology for bootstrapping in non-parametric frontier models. Journal of Applied Statistics, 27(6), 779–802.
49. Timmer, C. P. (1971). Using a probabilistic frontier production function to measure technical efficiency. *The Journal of Political Economy*, 776-794.
50. Tran, N. A., Shively, G., & Preckel, P. (2010). A new method for detecting outliers in data envelopment analysis. Applied Economics Letters, 17(4), 313-316.
51. Vanden Eeckaut, P., Tulkens, H., & Jamar, M. A. (1991). A study of cost-efficiency and returns to scale for 235 municipalities in Belgium (No. 1991058). Université

catholique de Louvain, Center for Operations Research and Econometrics (CORE).

52. Wilson, P. W. (1995). Detects influential observations in data envelopment analysis. Journal of Productivity Analysis, 6, 27–45.
53. Wilson, P.W. (1993). Detecting outliers in deterministic nonparametric frontier models with multiple outputs. Journal of Business and Economic Statistics, 11, 319–323.
54. Worthington, A. C. (2000). Cost Efficiency in Australian Local Government: A Comparative Analysis of Mathematical Programming and Econometrical Approaches. Financial Accountability & Management, 16(3), 201-223.
55. Worthington, A. C., & Dollery, B. E. (2001). Measuring efficiency in local government: an analysis of New South Wales municipalities' domestic waste management function. Policy Studies Journal, 29(2), 232-249.
56. Yang, Z., Wang, X., & Sun, D. (2010). Using the bootstrap method to detect influential DMUs in data envelopment analysis. Annals of Operations Research, 173(1), 89-103.
57. Yang M, Wan G, Zheng E. (2014) A predictive DEA model for outlier detection, Journal of Management Analytics, 1:1, 20-41, DOI: 10.1080/23270012.2014.889911

Appendix A

SAS Code

```
/*Standard DEA CCR Model */

data c.inputs;
input input $9.;
datalines;
Water
Toilet
Elec
;
Run;

data c.outputs;
input output $11.;
datalines;
Total_exp
;
Run;


proc optmodel;
set <str> INPUTS;
read data c.inputs into INPUTS=[input];
set <str> OUTPUTS;
read data c.outputs into OUTPUTS=[output];
set <num> MunicS;
str Code {MunicS};
num input {INPUTS, MunicS};
num output {OUTPUTS, MunicS};
read data raw.Munic_data_DC into MunicS=[_N_] Code
{i in INPUTS} <input[i,_N_]=col(i)>
{i in OUTPUTS} <output[i,_N_]=col(i)>;
num k;
num efficiency_number {MunicS};
num weight_sol {MunicS, MunicS};


var Weight {MunicS} >= 0;
var Inefficiency >= 0;
max Objective = Inefficiency;
con Input_con {i in INPUTS}:
sum {j in MunicS} input[i,j] * Weight[j] <= input[i,k];
con Output_con {i in OUTPUTS}:
sum {j in MunicS} output[i,j] * Weight[j] >= output[i,k] *
Inefficiency;
```

```
do k = MunicS;
solve;
efficiency_number[k] = 1 / Inefficiency.sol;
for {j in MunicS}
weight_sol[k,j] = (if Weight[j].sol > 1e-6 then Weight[j].sol else
.);
end;

set EFFICIENT_MunicS = {j in MunicS: efficiency_number[j] >= 1};
set INEFFICIENT_MunicS = MunicS diff EFFICIENT_MunicS;

print Code efficiency_number;
create data c.efficiency_data from [Munic] Code efficiency_number;

create data c.weight_data_dense from
[inefficient_Munic]=INEFFICIENT_MunicS
Code
efficiency_number
{efficient_Munic in EFFICIENT_MunicS} <col('w'||efficient_Munic)
=weight_sol[inefficient_Munic,efficient_Munic]>;
create data c.weight_data_sparse from
[inefficient_Munic efficient_Munic]=
{g1 in INEFFICIENT_MunicS, g2 in EFFICIENT_MunicS: weight_sol[g1,g2]
ne .}
weight_sol;


quit;

proc sort data=c.efficiency_data;
by descending efficiency_number;
run;
proc print;
run;

proc sort data=c.weight_data_dense;
by descending efficiency_number;
run;
proc print;
run;

proc print data=c.weight_data_sparse;
run;



/*Standard DEA BCC Model */
data d.inputs;
input input $9.;
datalines;
Water
Toilet
Elec
;
Run;
```

```
data d.outputs;
input output $11.;
datalines;
Total_exp
;
Run;


proc optmodel;
set <str> INPUTS;
read data d.inputs into INPUTS=[input];
set <str> OUTPUTS;
read data d.outputs into OUTPUTS=[output];
set <num> MunicS;
str Code {MunicS};
num input {INPUTS, MunicS};
num output {OUTPUTS, MunicS};
read data raw.Munic_data_DC into MunicS=[_N_] Code
{i in INPUTS} <input[i,_N_]=col(i)>
{i in OUTPUTS} <output[i,_N_]=col(i)>;
num k;
num efficiency_number {MunicS};
num weight_sol {MunicS, MunicS};


var Weight {MunicS} >= 0;
var Inefficiency >= 0;
max Objective = Inefficiency;
con Input_con {i in INPUTS}:
sum {j in MunicS} input[i,j] * Weight[j] <= input[i,k];
con Output_con {i in OUTPUTS}:
sum {j in MunicS} output[i,j] * Weight[j] >= output[i,k] *
Inefficiency;

Con Weight_con :
Sum {j in MunicS} Weight[j] = 1;

do k = MunicS;
solve;
efficiency_number[k] = 1 / Inefficiency.sol;
for {j in MunicS}
weight_sol[k,j] = (if Weight[j].sol > 1e-6 then Weight[j].sol else
.);
end;

set EFFICIENT_MunicS = {j in MunicS: efficiency_number[j] >= 1};
set INEFFICIENT_MunicS = MunicS diff EFFICIENT_MunicS;

print Code efficiency_number;
create data d.efficiency_data from [Munic] Code efficiency_number;

create data d.weight_data_dense from
[inefficient_Munic]=INEFFICIENT_MunicS
Code
```

```
efficiency_number
{efficient_Munic in EFFICIENT_MunicS} <col('w'||efficient_Munic)
=weight_sol[inefficient_Munic,efficient_Munic]>;
create data d.weight_data_sparse from
[inefficient_Munic efficient_Munic]=
{g1 in INEFFICIENT_MunicS, g2 in EFFICIENT_MunicS: weight_sol[g1,g2]
ne .}
weight_sol;


quit;

proc sort data=d.efficiency_data;
by descending efficiency_number;
run;
proc print;
run;

proc sort data=d.weight_data_dense;
by descending efficiency_number;
run;
proc print;
run;

proc print data=d.weight_data_sparse;
run;



/*-- Stochastic Frontier Production Model Half
Normal --*/
proc qlim data=raw.Munic_data_DC;
  model tot_exp_log = Water_log Elec_log Toilet_log ;
  endogenous tot_exp_log ~ frontier (type= half production);
  OUTPUT OUT=s.predicted_Half_prod TE1 TE2 PREDICTED EXPECTED
RESIDUAL XBETA PROB PROBALL CONDITIONAL ERRSTD MARGINAL MILLS;
     NLOPTIONS technique=none ;
run;


/*-- Stochastic Frontier Production Model
Truncated Normal--*/
proc qlim data=raw.Munic_data_DC;
model tot_exp_log = Water_log Elec_log Toilet_log ;
  endogenous tot_exp_log ~ frontier (type=truncated production );
  OUTPUT OUT=s.predicted_trun_prod TE1 TE2 PREDICTED EXPECTED
RESIDUAL XBETA PROB PROBALL CONDITIONAL ERRSTD MARGINAL MILLS;
  NLOPTIONS technique= congra ;
run;


/*-- Stochastic Frontier Production Model
Expoential --*/
```

```
proc qlim data=raw.Munic_data_DC;
model tot_exp_log = Water_log Elec_log Toilet_log ;
  endogenous tot_exp_log ~ frontier (type=exponential production );
     OUTPUT OUT=s.predicted_exp_prod TE1 TE2 PREDICTED EXPECTED
RESIDUAL XBETA PROB PROBALL CONDITIONAL ERRSTD MARGINAL MILLS;
     NLOPTIONS technique= congra ;
run;


/*Proposed Proposed model for Outlier
Correction*/

data c.inputs;
input input $50.;
datalines;
Water
Toilet
Elec
;
Run;

data c.iflags;
input iflag $50.;
datalines;
waterf
toiletf
elecf
;
Run;

data c.iflagsV;
input iflagV $50.;
datalines;
ColumnWater
ColumnToilet
ColumnElec
;
Run;


data c.outputs;
input output $50.;
datalines;
Total_exp
;
Run;

data c.oflags;
input oflag $50.;
datalines;
expf
;
Run;

data c.oflagsV;
```

```
input oflagV $50.;
datalines;
ColumnTotal_exp
;
Run;



proc optmodel;
set <str> INPUTS;
read data c.inputs into INPUTS=[input];

set <str> OUTPUTS;
read data c.outputs into OUTPUTS=[output];

set <num> MunicS;
str Code {MunicS};

num input {INPUTS, MunicS};
num output {OUTPUTS, MunicS};



set <str> IFlags;
read data c.IFlags into IFlags=[IFlag];
set <str> OFlags;
read data c.OFlags into OFlags=[OFlag];

num IFlag {IFlags, MunicS};
num OFlag {OFlags, MunicS};



set <str> Iflagsv;
read data c.Iflagsv into Iflagsv=[Iflagv];
set <str> Oflagsv;
read data c.Oflagsv into Oflagsv=[Oflagv];

num Iflagv {Iflagsv, MunicS};
num Oflagv {Oflagsv, MunicS};



read data WORK.QUERY_FOR_MUNIC_DATA_DC_0006 into MunicS=[_N_] Code
{i in INPUTS} <input[i,_N_]=col(i)>
{i in OUTPUTS} <output[i,_N_]=col(i)>
{f in IFlags} <IFlag[f,_N_]=col(f)>
{g in IFlagsv} <IFlagv[g,_N_]=col(g)>
{o in oFlags} <oFlag[o,_N_]=col(o)>
{m in oFlagsv} <oFlagv[m,_N_]=col(m)>

;
num k;
```

```
num efficiency_number {MunicS};
num weight_sol {MunicS, MunicS};


var Weight {MunicS} >= 0;
var Inefficiency >= 0;

max Objective = Inefficiency;
con Input_con {i in INPUTS, f in IFlags, g in IFlagsV }:
sum {j in MunicS} (input[i,j]*( 1- Iflag[f,j]) + Iflagv[g,j]
*Iflag[f,j] )* Weight[j] <= input[i,k];

con Output_con {i in OUTPUTS, o in oFlags, m in oFlagsV }:
sum {j in MunicS} (output[i,j]* (1- oflag[o,j]) +
oflagv[m,j]*oflag[o,j] ) * Weight[j] >= output[i,k] * Inefficiency;

Con Weight con :
Sum {j in MunicS} Weight[j] = 1;

do k = MunicS;
solve;
efficiency_number[k] = 1 / Inefficiency.sol;
for {j in MunicS}
weight_sol[k,j] = (if Weight[j].sol > 1e-6 then Weight[j].sol else
.);
efficiency_number[k] =(if efficiency_number[k] <= 1 then
efficiency_number[k] else 1);

end;

set EFFICIENT_MunicS = {j in MunicS: efficiency_number[j] >= 1};
set INEFFICIENT_MunicS = MunicS diff EFFICIENT_MunicS;

print Code efficiency_number;
create data c.efficiency_data_new from [Munic] Code
efficiency_number;

create data c.weight_data_dense from
[inefficient_Munic]=INEFFICIENT_MunicS
Code
efficiency_number
{efficient_Munic in EFFICIENT_MunicS} <col('w'||efficient_Munic)
=weight_sol[inefficient_Munic,efficient_Munic]>;
create data c.weight_data_sparse from
[inefficient_Munic efficient_Munic]=
{g1 in INEFFICIENT_MunicS, g2 in EFFICIENT_MunicS: weight_sol[g1,g2]
ne .}
weight_sol;


quit;

proc sort data=c.efficiency_data_new;
by descending efficiency_number;
run;
```

```
proc sort data=c.weight_data_dense;
by descending efficiency_number;
run;
```

Appendix B

List of Publications

1. “A new approach for extreme values in Data Envelopment Analysis”, communicated to the South Africa Statistical Journal 2015.

2. “Evaluating efficiency analysis methodologies in the context of Municipalities in South Africa” , Communicated to the Conference proceedings International Statistical lnstitute Brazil 2015.