Developing a repeat sales property price index for residential properties in South Africa

H Bester

21575487

Dissertation submitted in partial fulfilment of the requirements for the degree *Master of Science* at the Potchefstroom campus of the North-West University

Supervisor: Prof. H. A. Kruger

Co-supervisor: Prof. P. J. de Jongh

November 2010

Abstract

In South Africa various financial institutions and independent vendors have developed residential property valuation models to estimate the current value of historically traded properties. A natural extension to these models has been to develop historical property price indices. In this dissertation, three of the four approaches to developing property price indices will be examined. Through back-testing and other statistical methods, the most accurate and robust approach will be determined. The four major approaches available are the mean valuation per suburb, the median valuation per suburb, the repeat sales approach and hedonic regression. The mean valuation per suburb approach can be biased because of outliers in property prices. However, outliers in property prices will not influence the median valuation per suburb approach, but in cases where property values in a suburb have a skewed distribution, the valuation amount could be distorted. Neither of the above mentioned shortcomings influences the repeat sales or the hedonic regression approach. To follow the hedonic regression approach, the characteristics of the property need to be known. In South Africa, however, the available property data lacks detailed characteristics of traded properties. This dissertation will therefore focus on the first three methods. The repeat sales approach measures the growth in property prices by applying a generalized linear model to properties that have traded more than once. This approach is only possible if there is a representative amount of repeat sales able to fit a model. The focus of this project will be on the repeat sales approach, but all three the approaches discussed will be analysed to prove that the repeat sales approach is the most accurate in developing a property price index for properties in South Africa.

Keywords:

Repeat sales, property price index, erf key, hedonic regression, property valuation model, heteroskedasticity, data mart, comparable sales, property derivatives and smoothing.

Opsomming

Verskeie finansiële instellings en onafhanklike verkopers in Suid-Afrika het residensiële eiendomswaardasie-modelle ontwikkel ten einde die heersende waarde van eiendomme wat voorheen verkoop was, te bepaal. Daardeur kon ‘n eiendomsprys-indeks ten opsigte van sodanige eiendomme ontwikkel word. In hierdie verhandeling word drie van die vier benaderings ten opsigte van die ontwikkeling van sodanige eiendomsprys-indeks ondersoek om die mees akkurate en betroubare benadering vas te stel. Daar bestaan vier benaderings: die gemiddelde waarde-benadering per voorstad; die mediaan waarde-benadering per voorstad; die herverkoopsbenadering en die hedonistiese benadering. Die gemiddelde waarde-benadering kan deur uiterstes in eiendomspryse beïnvloed word. Daarteenoor sal uiterstes in eiendomspryse nie die mediaan waarde-benadering beïnvloed nie. In gevalle waar eiendomswaardasies in ‘n woonbuurt egter oneweredig is, kan ‘n verwronge beeld ontstaan. Geeneen van bogenoemde tekortkominge beïnvloed egter die herverkoopsbenadering of die hedonistiese benadering nie. Die hedonistiese benadering vereis die beskikbaarheid van inligting oor die kenmerke van die eiendomme, en in Suid-Afrika se beskikbare eiendomsdata is daar ‘n gebrek aan gedetailleerde inligting oor voorheen verkoopte eiendomme. In hierdie verhandeling word dus op die gemiddelde waarde-benadering, die mediaan waarde-benadering en die herverkoopsbenadering gefokus. Die herverkoopsbenadering meet die groei in eiendomspryse deur ‘n veralgemeende lineêre model op eiendomme wat herhaaldelik verkoop was, toe te pas. Hierdie metode is egter slegs moontlik indien ‘n verteenwoordigende aantal herverkoopstransaksies beskikbaar is wat aan die vereistes van die model voorsien. Al drie benaderings word bespreek en ontleed maar die fokus van die projek is op die herverkoopsbenadering. Daardeur sal bewys word dat die herverkoopsbenadering die mees akkurate metode is vir die ontwikkeling van ‘n eiendomsprys-indeks vir die Suid-Afrikaanse eiendomsmark.

Sleutelwoorde:

Herverkope, eiendomsprys-indeks, erfomskrywing, hedonistiese regressie, eiendoms-waardasiemodel, heteroskedastisiteit, dataversameling, vergelykbare verkope, eiendom-afgeleides en uitstryking.

Acknowledgements

Finalising this project has been a journey which first started merely as an idea in 2008 and has proven to be a tremendously fulfilling challenge, providing a great learning curve. I would like to extend my sincerest thanks to the following people who has contributed to the successful completion of this project:

- My heavenly Father for the wisdom and endurance from His hand.
- My supervisor, Prof. H.A. Kruger for the guidance and insight provided, as well as for his willingness to always share knowledge.
- My co-supervisor, Prof. P. J. de Jongh, for his support and comments.
- Dr. Frances Klopper for assisting in proof reading the project and ensuring the quality thereof.
- My family for always keeping up my energy levels through their continuous encouragement, advice and motivation.
- My husband, Mike, for his continuous support, love and understanding.
- A special word of thanks to my father, Gerrit Genis, for his invaluable inputs and suggestions throughout the course of this journey.

Table of Contents

Chapter 1: Introduction and problem statement

1.1 Introduction

Many financial institutions and independent vendors in South Africa have developed property price indices. Property price indices provide a basis for measuring the current values of properties and their growth over time. These indices enable institutions to value the collateral to be held for property portfolios, calculate the current value of a property in order to determine the amount of equity available due to the appreciation or depreciation of property prices, to revaluate the home loan book of a specific financial institution in order to better understand their home loan base and determine which loans to approach first in the collections area when loans default (this applies especially to the banking industry) and to possibly create an opportunity for property price index derivatives.

In general, to evaluate a residential property, an assessor has to be sent out. Their property valuation is based on the amount of the loan granted on the property as well as on the size of the property, the number of rooms, where it is situated, its proximity to various amenities etc. With so many properties trading it becomes very expensive to send out an assessor out to evaluate each property. Models have thus been developed to assist in this process. The major drawback of using a model instead of an assessor is that the model is based on the accuracy and reliability of historical data, whereas an assessor inspects the physical property and based on his knowledge of the area, usually knows the trades so well that a value for the property can be estimated more accurately. An advantage of having a method to calculate the predicted value of a property is, among others, the cost saving of not having to pay for an assessor in all cases whilst speeding up the loan application process.

Worldwide, the most commonly used methods to calculate property price indices are the measures of central tendency based on the mean or median, the repeat sales model, hedonic regression, and hybrids of the latter two (Jansen *et al.,* 2006). Since the property data available in South Africa lacks the detailed characteristics of traded

properties which are required for hedonic regression modelling, this dissertation will only focus on the first three methods.

All three of the basic approaches discussed will be analysed by using various statistical measures and back-testing methods and compared. The most robust and accurate approach will then be further developed into the best proven and practical solution for a national property price index for South African properties.

1.2 Problem statement

Various approaches have been developed to calculate property price indices in South Africa, but they have not been evaluated and compared to determine which is the most robust and accurate approach. According to an article written by Muller (2008), property sellers and buyers are "*confused about what's happening with house prices*" because "*the market has been bombarded by a wide array of housing data in recent weeks"* and *"there's also little correlation between the house price indices*".

1.3 Objectives of the study

The primary objective of this research project is to determine the most robust approach available with which to valuate properties in South Africa in order to develop a residential property price index. This will be accomplished by addressing the following secondary research objectives:

- To gain a clear understanding of and present an introductory overview of property valuation modelling and property price indices and understand the arguments and contributions made by various authors on the various approaches they studied.
- To understand the data sources used for modelling and the necessary clean-up processes, validations and exclusions that need to take place, as well as to clearly understand the model methodology of the three valuation models used in this study.

- To determine the most accurate approach to valuating properties by applying various statistical tests for comparing the three valuation models.
- To improve the most promising model further to become the most robust solution with which to valuate residential properties in South Africa. Using this solution to further derive a national property price index.
- To gain an understanding of various operational model enhancements that will enable adding business value through current and future applications.

1.4 Methodology

The methodology comprises of three main steps:

- A literature review was conducted in order to give an overview of existing property valuation methods, property price indices and techniques;
- An empirical study was performed using three different valuation models and the most accurate approach was determined; and
- This approach was further improved and tested for validity and eventually used in the development of a property price index.

1.5 Layout of study

The project was documented through a set of chapters and this section explains the purpose of each chapter and how it is structured.

In Chapter 2, the results of a literature study in which property evaluations and property price indices are researched, is presented. The necessary background to property valuation modelling and house price indices is presented as well as an overview of the status of what is contributed to the broader topic internationally and in South Africa. Hybrids and variants of the different methods are discussed and compared; and an overview is given of the data sources and data analyses software used to develop these models.

Chapter 3 focuses on a detailed explanation of the data. It addresses the process of data cleansing, the assumptions behind data exclusions and the manipulation of the data. Based on this information, a theoretical overview of the basic methodologies of the three valuation models that are compared in this dissertation is presented.

Following from the methodology explained in Chapter 3, the three models are further developed in Chapter 4. The models are then assessed using back-testing and various statistical measures to determine the most accurate model for predicting the value of a property. The coverage of the models, another important criterion, will also be evaluated.

A number of techniques are discussed with which to improve the most accurate of the three methods in Chapter 5 and the resultant evaluation results for the accuracy of the improved solution, will be presented. Finally, using the most robust solution, the derivation of the property price index will be explained.

In Chapter 6, operational enhancements to the final model are presented and a number of current and future applications to which the study can add business value are discussed.

In conclusion, in Chapter 7, a high level overview is given, summarising the process followed from beginning to end whilst further research that can still be done on the topic is explained.

A project overview is illustrated in Figure 1.1.

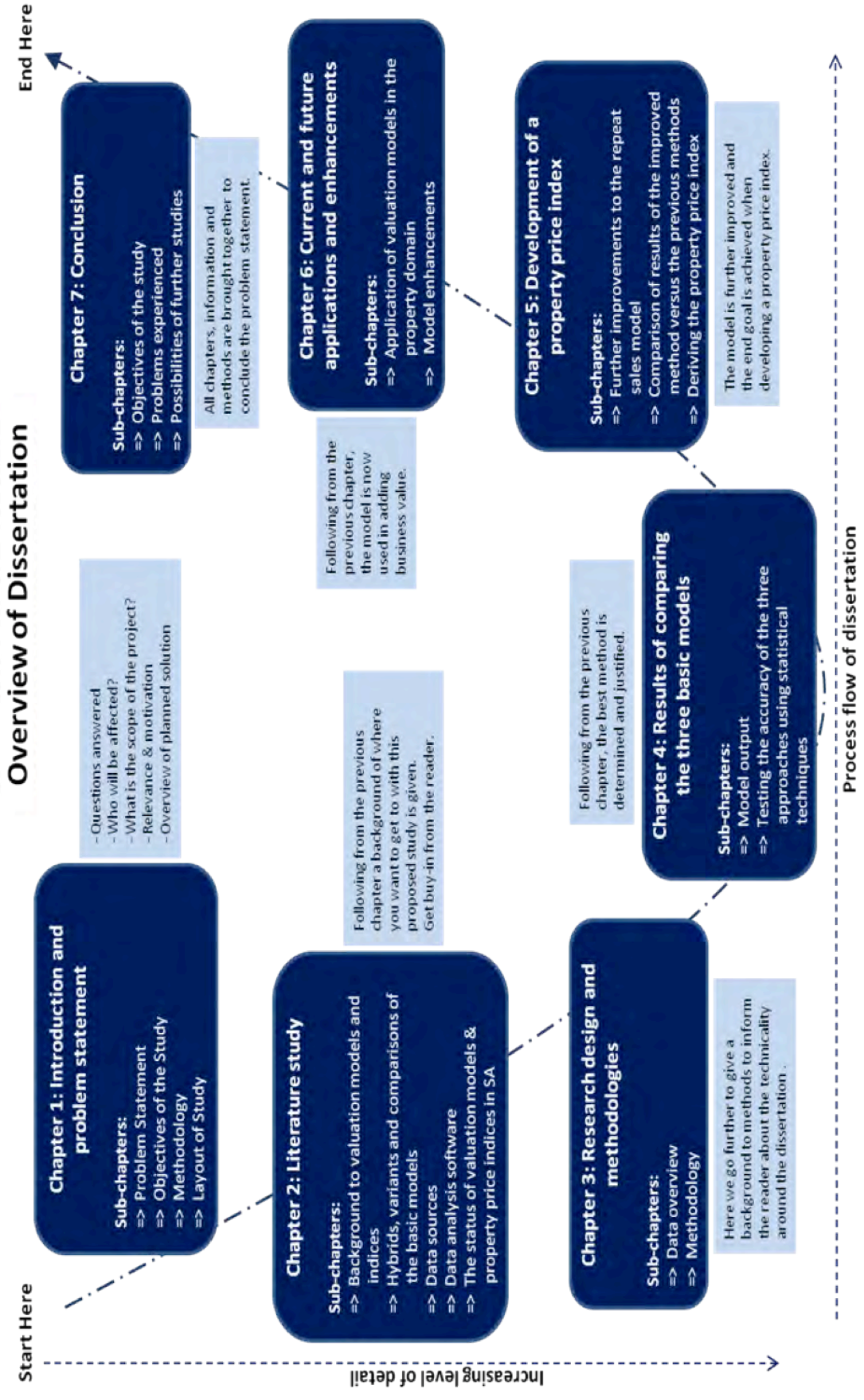


Figure 1.1: *Project Overview (schematically)*

1.6 Conclusion

Chapter 1 served as an introduction and guided the reader into the research project by explaining the problem statement, objectives of the study and the methodology that will be followed. A layout of the study, explaining the purpose of each chapter, was also presented. In the next chapter an overview from the literature of valuation models and property price indices will be presented.

Chapter 2: Literature study

2.1 Introduction

Following the global economic crisis of 2008 - 2009, a greater emphasis has been placed on property valuations. A current and accurate valuation for each individual property has become an invaluable item of information for many different audiences, from those providing credit with property as collateral, to investors looking to profit from price shifts, and of any entity involved in the buying and selling of property in South Africa. An opportunity therefore exists for the development and use of a statistically sound tool to accurately calculate the current market value of individual properties. This chapter provides the background to this pursuit.

2.2 Background to valuation models and indices

For each property price index there is an underlying property valuation model stipulating the growth for that segment. Therefore, to attain the end goal of developing a property price index for residential properties in South Africa, the most robust method available to valuate properties first needs to be determined.

There are mainly four different methods (and a number of hybrids or variants of each) that have been investigated and implemented to valuate properties around the world, namely:

- Measures of central tendency – mean;
- Measures of central tendency – median;
- Hedonic regression; and
- Repeat sales.

Each of these basic models have its own limitations and advantages, but all the models have the same purpose – to predict as accurately as possible what the price of a property is or will be at any given time, representing as many properties as possible.

2.2.1 Measures of central tendency (mean and median)

Measures of central tendency are theoretically so straightforward that “no articles that we are aware of have been devoted to their study” (Wang and Zorn, 1997). However, measures of central tendency (also called summary methods) have been mentioned by some authors as points of comparison; for examples, see (Mark and Goldberg, 1984; Hosios, 1991; Crone, 1992; Gatzlaff and Ling, 1994).

The measures of central tendency (mean or median) is the simplest and most obvious way to construct a property price index since the price of a property can be estimated using either the mean or median price of a specific sample of properties. The price index of a portfolio of properties is derived by dividing each period's mean/median price by the base period's mean/median price, while the growth index for the same portfolio is determined by dividing each period's mean/median price by the previous period's mean/median price. These indices are discussed further in Chapter 4.

Although the method features directness and ease of interpretability, its simplicity can also be a weakness. One of the drawbacks of this approach is that, instead of reflecting true price changes in the underlying population, it provides a varying mean price of properties due to the random selection of properties chosen per sample (Wang and Zorn, 1997). According to Jansen *et al.,* (2006), the intrinsic flaw in using the measures of central tendencies is that they are not adjusted for the quality of properties. They are also unable to distinguish between price movements and changes in the composition of properties sold from one period to the next (Bourassa, 2004).

The most significant problem faced by indices using the measures of central tendency, particularly the median purchase price is, according to Case and Schiller (1987), due to the fact that the characteristics of properties sold may change from period to period. For example, if a disproportionate number of expensive properties was sold in a particular period, the mean or median price would rise even if none of the properties appreciated. To correct for this issue, two other methods have been used to derive property price indices, namely hedonic price indices which statistically

"control" for differences in the characteristics of properties and repeat sales property price indices.

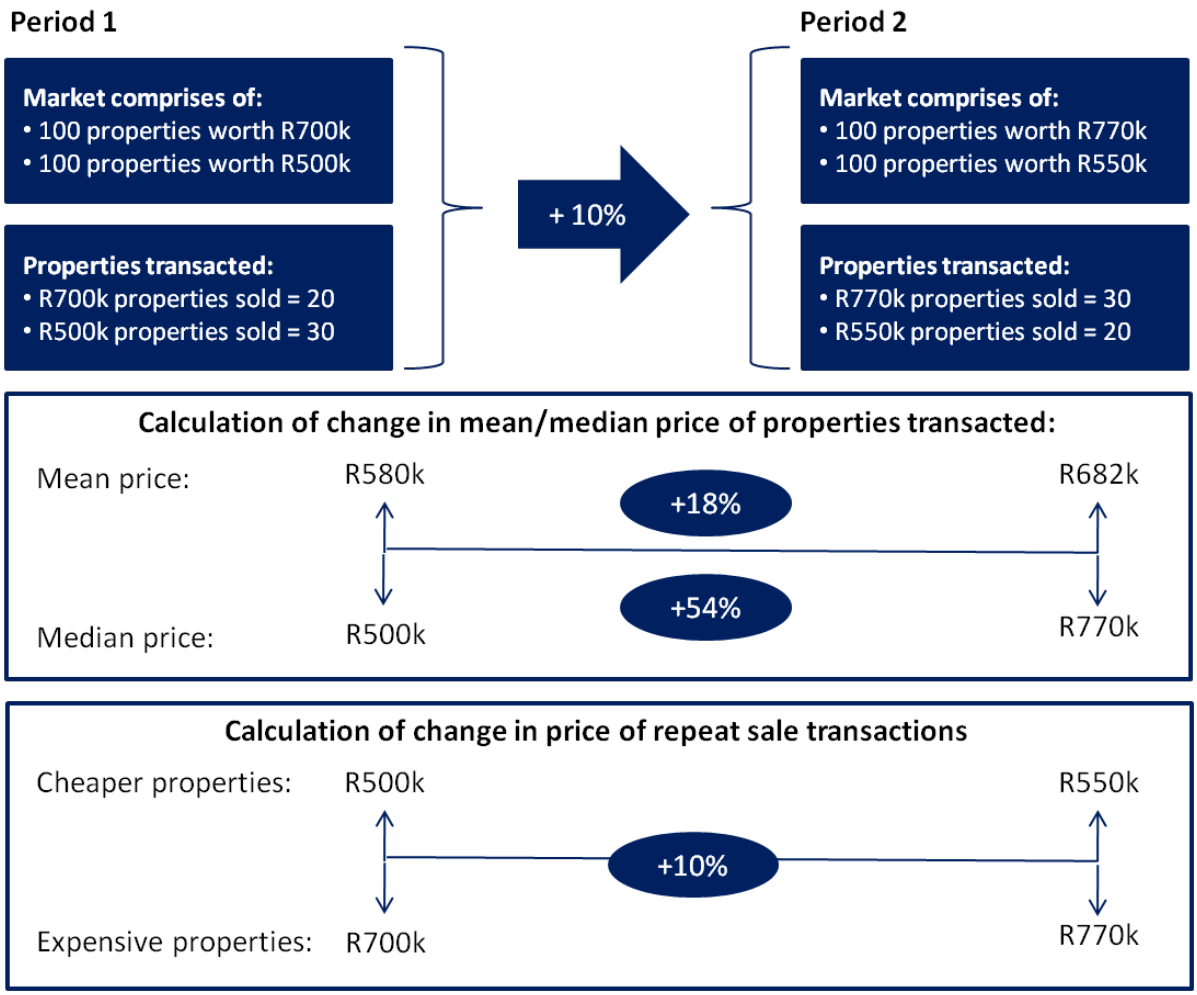


Figure 2.1: *Two hypothetical situations illustrating the advantage of the repeat sales model.*

Assume in the example above (Figure 2.1) that the property market comprises of 100 expensive properties valued at R700 000 and 100 cheaper properties valued at R500 000. In period 1, twenty of the more expensive properties and thirty of the cheaper properties were sold. The mean price of properties that transacted is R580 000 and the median price is R500 000. Assume further that the price of all the properties in the area increases by 10% per year. Thus in period 2, the more expensive properties transact at R770 000 and the cheaper properties at R550 000. If in period 2 only twenty of the cheaper properties and thirty of the more expensive properties were sold, the mean price in period 2 increases to R682 000 and the median price to R770 000. This equates to a mean price increase of 18% from R580

000 in period 1 and a median price increase of 54% from R500k in period 1 whilst the actual price increase of all the properties in the area was actually 10%.

In contrast to using the mean or median price to valuate properties, the repeat sales method provides a measure of inflation of properties over time as opposed to the mean or median property prices from one period to the next. Thus repeat sales are less influenced by the sample of transacting properties in each period than the mean or median equivalents.

2.2.2 Hedonic regression

Although popularised by Griliches (1971), who first applied the regression approach to automobiles in the early 1960s, hedonic price analysis actually dates back to a 1939 article by Court (1939). Rosen (1974) played a major part in establishing its theoretical foundation. Other early studies were conducted by Ferri (1977), whose aim was to prove that hedonic methods are particularly applicable to housing, Goodman and Thibodeau (1995), who studied the age-related heteroskedasticity in hedonic property price equations, and Meese and Wallace (1991, 1997) who used a modern non-parametric regression technique in the hedonic estimation.

To construct a hedonic valuation model, an ordinary least squares regression model is fitted to a set of variables, based on attributes that describe the property, such as the number of rooms, the size of the property, the size of the land and the number of storeys. The regression coefficients are effectively the implicit attribute prices contributions; for example, an additional room in a property or an additional storey may add an additional amount to the property's value.

The hedonic model has the advantage that the price of a property can be accurately estimated, but applying this methodology requires a large amount of data on properties sold, including the characteristics (or attributes) of the properties, such as the number of rooms, the size of the property, the size of the land and the number of storeys.

The hedonic method also allows for the identification of depreciation in properties. Physical deterioration can decrease the price of a property as the property ages and

tastes or preferences in various property characteristics or attributes may change over time. By including a time variable, the hedonic approach may capture the effect of age on a property's value.

There are two approaches to construct a hedonic price index; one way is by using a fixed weighted method in which a separate regression model is run on data from each time period or, alternatively, a single regression model can be run for all time periods. The first approach allows the individual attributes to change for each time period, while the second approach has the disadvantage of constraining attribute prices to be the same over the whole time interval.

The inherent disadvantage of the hedonic approach, namely the large number of variables and assumptions required to accurately model property prices, limits its use to large detailed data sets that are not generally available. The data used for the purpose of this dissertation is very large but does not contain any characteristic information. Thus this method, as well as hybrids or variants based on this method, is beyond the scope of this study.

2.2.3 Repeat sales

The repeat sales methodology is generally used to construct an index of prices or returns for unique, infrequently traded assets such as houses, automobiles, art, and musical instruments which are likely to be prone to exhibit serial correlation in returns (Zanola, 2007). Although Pendleton (1965) was first to apply the regression approach to valuate single-family houses, Bailey *et al.,* (1963) were the first to propose the repeat sales method to develop an index for property prices.

The repeat sales model uses ordinary least squares regression analysis in which the dependent variable is the logarithm of the price relative to the twice-sold property. The dependent variable is then regressed on a set of dummy variables corresponding with time periods ('-1' for the first sale, '+1' for the second sale and '0' otherwise). There is no constant term; the coefficients are estimated only on the basis of change in house prices over time. The estimated coefficients represent the

logarithm of the cumulative price index for each period. The time dummy for the initial period is set at zero to normalise the index at 1.

Case and Shiller (1987) published an adapted version of the repeat sales model in which they argue that the variance in property prices widens as the period between sales increases (this is known as heteroscedasticity) which undermines efficiency when the variance of the index values becomes to great (Wang and Zorn, 1997). They managed to minimise this effect by using a weighted repeat sales model which comprising the following steps:

- Step 1: The logarithm of the price relative from the twice-sold property is regressed on a set of dummy variables corresponding to time periods.
- Step 2: A regression analysis is performed on the squared residuals from step 1. Time is incorporated as an independent variable in the model and a constant term (estimate of the variance of twice the property-specific random error variance – once for the first sale and once for the second sale) is included. The increase in variance for each additional period mentioned above is estimated by using a 'Gaussian Random Walk'.
- Step 3: A weighted regression analysis (general least squared regression) is applied where the weights are the reciprocals of the square roots of the fitted values of the second-stage regression. This procedure minimises the impact that houses with a relatively long period between sales have on the regression analysis.

According to Case and Shiller (1987) the logarithm of the price of the *i*-th property at time *t* is given by:

$$P_{it} = C_t + H_{it} + N_{it} \qquad (2.1)$$

where

C_t = the log of the city-wide level of property prices at time *t,*

H_{it} = Gaussian random walk that represents the drift in individual housing value through time,

N_{it} = a house-specific random error that has zero mean and equal variance and is serially uncorrelated.

When used in the property domain, the repeat sales method focuses on price changes rather than prices themselves, directly measuring these changes by examining only properties that have been sold at least twice. It is more sophisticated than the simplistic idea of taking samples and finding the central tendency of the price and avoids having to correctly specify the critical characteristics determining a property's value or their mathematical relationships to price. By only using properties that have been sold at least twice, other contributing factors to variation in price growth are controlled.

A disadvantage of the repeat sales method is that it is wasteful of data, especially if only a small amount of properties transacted more than once. The properties that sold more than once may not be representative of the entire population of properties. Obviously the extent of this problem depends on the coverage of the data used in creating the indices.

Advocates of the repeat sales methodology argue that it “controls” the characteristics of properties more accurately than the hedonic methodology, since it is based on the observed appreciation of actual properties. Repeat sales does not require the measurement of quality specifically, only that the quality of a sample of properties be constant over time.

According to Butler *et al.,* (2005), repeat sales property price indices are the most widely used measures of changes in property values by researchers, business analysts, regulators and fraud investigators. According to them, the most widely cited repeat sales property price indices are the Conventional Mortgage Home Price Index (CMHPI) and the OFHEO Home Price Index (OHPI) in the United States, which are both based on the valuations of properties with home loans from Freddie Mac and Fannie Mae, two of the largest mortgage finance companies in the USA (Nielsen, 2008).

The repeat sales methodology forms the basis of this study and will be discussed in more detail in the following chapters.

2.3 Hybrids, variants and comparisons of the basic models

2.3.1 Hybrids

A mixed, hybrid model that utilizes three different equations to apply to three different groups of transactions was proposed by Case and Quigley (1991):

- Hedonic regression is applied to all properties that transacted only once during the sample period;
- Repeat sales regression is applied to properties that transacted more than once during the sample period but had no change in property attributes (an attempt to keep quality constant); and
- A modified repeat sales regression is applied to properties that transacted more than once during the sample period but had some change in property attributes.

The hybrid formulation uses the repeat sales idea whenever possible, thereby both exploiting the control of variation inherent in repeat sales and staying clear of the problems of possible misspecification inherent in the hedonic methodology, but avoids the inefficiency of using pure repeat sales, because it also uses information from properties that were sold only once. Owing to its partially hedonic-like structure, hybrid models share the criticism regarding complexity associated with hedonic models.

2.3.2 Variants

In this section a short overview is given of a few studies on extensions and variations of existing valuation models.

Wang and Zorn (1997) investigated many statistical properties of various property price indices focusing on a particular feature of indices, namely “revision volatility”, the tendency of previously estimated values for prior time periods to change with a new run of the model. Their repeat sales index is developed using a regression

model and when any additional data is added to a regression model, the data will affect all the parameter estimates and introduce sampling selection bias.

Indices based on property prices are not easy to construct as there is great variation in quality and features amongst properties. Illustrating this point, Bailey *et al*., (1963) commented on the bias introduced in the indices when using the average sales prices of all properties sold in a specific period. The variation in quality of properties sold from one period to the next will cause the index to vary more than the value of any given property, and the change in quality for properties that were sold at different time periods will cause the index to become biased over time. A way to avoid these quality constraints is to use regression analysis instead of mean or median sales prices.

The basic repeat sales method was improved by Case and Shiller (1987, 1989) and Shiller (1991) as discussed in Section 2.2.3. They investigated the relationship between the increases in errors with the increase in time between sales and suggested the approach of weighing it back in the model, resulting in the weighted repeat sales method. Various authors have contributed additions and corrections to the weighted repeat sales methodology. Abraham and Schauman (1991) argued that the variance of the error term associated with any repeat sales transaction pair will not indefinitely increase linearly (proportional to the time between sales). Instead, they proposed a quadratic model, to model the initial increase in variance to start decreasing at some time as the period between sales increased. Based on their empirical estimates, they determined the maximum variance to be at between twenty to thirty years.

Lastly, Case and Shiller (S&P/Case-Shiller, 2009), used a three-month moving average algorithm for which property prices are accumulated in rolling three-month periods to which the repeat sales methodology is applied. Each index point is then based on that month and the preceding two months, which helps to offset inevitable delays that may occur in the deeds data and to keep the sample sizes large enough to create meaningful price change averages.

2.3.3 Comparisons

The following articles have appeared in the real estate literature comparing the various methodologies:

- Mark and Goldberg (1984) compared eleven models of which five were favoured: a mean series, a median series and three variations of the hedonic model. They rejected the repeat sales method, finding that the index values showed much smaller increases in home prices compared to some of the other models.

- Case *et al,* (1991) compared 14 models and variants representing the repeat sales, hedonic, and hybrid methodologies. In agreement with Mark and Goldberg (1984), they found that repeat sales price estimates increased more slowly than those of the other methodologies. In disagreement with Case and Quigley (1991), they did not find any clear efficiency gains using the hybrid methodology.

- Crone and Voith (1992) compared measures of central tendency, hedonic and repeat sales methods. They concluded that measures of central tendency were generally less accurate than hedonic or repeat sales methods. Comparing the two types of measures of central tendency, they found to their surprise that means were better than medians.

- Hosios and Pesando (1991) compared repeat sales and measures of central tendency, deciding in favour of repeat sales. Clapp and Giacotto (1992) compared their assessed value variant to a pure repeat sales approach and based on efficiency preferred the assessed value method.

- Meese and Wallace (1997) studied hedonic, repeat sales and hybrid approaches. Their work is of particular interest because they used a modern, non-parametric regression technique in the hedonic estimation. They rejected repeat sales as being very sensitive to small samples in favour of either the hedonic or hybrid methodology.

- Finally, Gatzlaff and Ling (1994) compared median, repeat sales, hedonic and assessed value methods, preferring repeat sales as the benchmark. They

found that all of their models produced precise estimates of the index and growth rates.

Wang and Zorn (1997) argued that although valuation methodologies are generally compared in terms of bias and efficiency, or against some stated benchmark, much of the disagreement over preferred procedures above arises because the targets and aims of the underlying estimation methods have never been precisely or explicitly established.

2.4 Data sources

Data is a very important component when constructing a property price index. As will be discussed in Chapter 3, the property data can be very volatile and if the data is not representative of the properties in South Africa, a model based on the data may lack predictability. Cleaning the property data and getting it to be functional, is the first step in constructing a robust property price index. This will be discussed in Chapter 3. This section provides a brief overview of data sources as well as where they can be obtained.

In South Africa, the Department of Land Affairs is the owners of all the deeds data for local properties. In 2001 they developed a web-enabled database system of Deeds registration information (Deeds, 2001) which allows clients to access Deeds information.

Many vendors use the Deeds data to develop client-specific business intelligence solutions for various financial service providers. One of these vendors is AfriGIS, who specialises in spatial datasets (AfriGIS, n.d.). They are a provider of the National Address Database (NAD), national coverage of Cadastral boundaries and Street Centre Lines (SCL) for South Africa as well as a preferred supplier to various government departments and parastatals, State Information Technology Agency (SITA), and many corporate clients. Monthly bonds and transfers registered with the deeds office, have been provided by AfriGIS since 1997. All deeds datasets are linked to the NAD and Cadastral datasets.

Strategis Consulting (founded in early 2002) is another vendor that provides skills in Geographic Information Systems (GIS) and Statistical Analytics (Footprint, 2009). In 2007, Lightstone and Strategis merged to become Lightstone, which now offers a range of web-based information systems to various financial service providers. Strategis assists in the provision of customised data cleansing, standardising, geo-coding and enrichment solutions. Their services are delivered by means of Dataflux, the global leader in data quality software (Footprint, 2009).

The deeds data used in this study is provided by Knowledge Factory (KF) (Knowledge Factory). KF is a leading marketing insights company which helps its clients to understand their customers, channels and markets better in order to leverage and enhance business performance. KF has developed comprehensive market and marketing datasets and has extensive experience in both the building of spatial and statistical models, data mining and insight delivery portals. A variety of data sets are obtained from numerous different sources by KF and integrated into a single cohesive framework. The most comprehensive deeds information database in South Africa belongs to KF, with a full history dating back to 1993. This deeds information is sourced from the Deeds Office and includes ownership details, the price paid, the date of transfer and mortgage information, (including the name of the financial institution holding the bond and the original value of the bond) for each property in South Africa. The complete transfer history of the past 15 years of every property in South Africa is thus available.

When considering property data in the construction of a property price index, two important aspects must be taken into account. Firstly, the data needs to be as clean as possible. Secondly, the data needs to be representative of the South African property portfolio. Both of these requirements are discussed in more detail in the following chapter.

2.5 Data analyses software

It is important to use a sufficiently-capable software analysis package able to deal with large datasets and that can be used to perform a range of analytic procedures in constructing a property price index. The software used in this dissertation was

Statistical Analysis Software, better known as SAS (SAS, 1976). SAS is an international leader in business analytics software and services and the largest independent vendor in the Business Intelligence (BI) market. SAS has been developed to be able to analyze huge quantities of data to make discoveries and solve complex problems. It provides an integrated environment for predictive and descriptive modelling, from dynamic visualisation to performing predictive modelling, model deployment and process optimisation. SAS provides a range of techniques for the collection, classification, analysis and interpretation of data to reveal patterns, anomalies, key variables and relationships.

2.6 The status of valuation models and property price indices in South Africa

There are four major financial institutions in South Africa of which three publish a monthly property price index for residential properties. Standard Bank publishes a median property price index and ABSA and FNB publish a mean average price index, all three of these indices are constructed by using the financial institution's own financed property transactions.

ABSA's property price index (ABSA) dates back to 1966 and is based on the mean purchase price of properties in the 80m²- 400m² size categories. ABSA's index is smoothed to exclude distorting effects of seasonal factors and outliers in the data.

Standard Bank (Standard Bank) maintains that due to the way in which house prices are measured, it will always be inherently volatile. The data available to valuate properties consists of properties being sold in a particular period instead of representative data of all properties, further complicated by the heterogeneity of properties. Changes in property prices may be a result of the general price level of changes in the distribution of the houses being sold, or the changes may be completely random in nature. Thus, Standard Bank reasons that the median price is more accurate than the mean price, as half of all properties are more expensive and half is less expensive. According to them it is substantially less volatile and less sensitive to the typical problems found in property data.

The outliers in FNB's property data (FNB) is eliminated by only including property transactions above 70% and below 130% of FNB's valuations of the property. A statistical smoothing function is applied to the data, to further assist in eliminating outliers.

Lightstone (Lightstone) is a financial service provider that focuses on the property market and publishes a repeat sales property price index. According to Lightstone, in contrast to the measures of central tendency mean property price indices, repeat sales indices provide a measure of the actual price inflation of properties that have transacted twice within a particular period of time and are less influenced by the mix of transacting properties.

Lastly, Ooba (Oobarometer) is a financial services provider that focuses on property finance. They publish a mean property price index. Information on the data and history behind Ooba's property price index is limited.

2.7 Conclusion

Chapter 2 was devoted to background information of property valuation models and indices. Four major well known methods and their variants were presented. A succinct comparison of literature resources, comparing the four methodologies, was also given. Brief mention was made of the various data sources and data analysis software available for constructing a property price index. The chapter was concluded with a short summary of the status of valuation models and property price indices in South Africa. In the next chapter the research methodology will be discussed, followed in Chapter 4 by an evaluation of the basic models.

Chapter 3: Research design and methodologies

3.1 Introduction

The goal of the research project presented in this dissertation is to evaluate various modelling approaches for property valuation; to identify the drivers contributing to better predictions; to determine the models' dependencies on the data, both quality and quantity, and finally, to propose a best model for constructing a robust property price index for residential properties in South Africa.

In this chapter, the challenges of selecting quality data from all the available data, to be used for the property valuation modelling with which to derive a property price index, as well as the construction of a data mart will first be discussed. This is followed by an introduction and analysis of the different physical variables that drive the model. Two measures of central tendency and a repeat sales approach is subsequently introduced and evaluated in Chapter 4 as possible valuation models. In Chapter 5, a number of improvements to the latter with variations thereof are investigated and a best model for valuating South African residential properties proposed and evaluated by means of back testing. Finally, the most robust valuation model is implemented to develop a comprehensive property price index.

3.2 Data overview

The development of a suitable valuation model is, as expected, dependent on sourcing reliable and accurate property price information. Appropriate data validation and correction procedures are required to ensure that the data used is suitable as a basis for generating the property price index. The data sourcing and validation procedures are described in more detail below.

All property transactions in South Africa are registered in the Deeds Office in accordance with the South African Deeds Registries Act (Act No. 47 of 1937) and the Sectional Titles Act (Act No. 7 of 2005). These transactions include residential

property sales, land sales, farm sales, commercial property sales and property transfers[1].

As discussed, this study is based on property transactional data for residential properties sourced from Knowledge Factory (KF). KF supplies raw and geographically-linked derived property and market insight data and has the most comprehensive deeds information database in South Africa, with a full history dating back to 1993[2]. The following transactional data is available for each property in South Africa:

- Details of the seller and the buyer;
- The purchase price of the property;
- The transfer date and registration date of each transaction;
- Geographical information of the property and the socio-economic groups residing in the area;
- The size of the land on which the property is built;
- The deeds office at which the property was registered;
- The name of the financial institution holding the bond; and
- The original value of the bond.

Approximately 14.7 million property transactions (defined as a property sold by one owner to another), involving around 7 million physical properties transacted all over South Africa, have been recorded on the database of the Deeds Office and taken up in the KF database in April 2009. The data from 1993 and onwards still requires extensive cleaning before use. Even after cleaning the data by applying validation and correction procedures and limiting the data to transactions that have taken place since 1993, only a fraction of this data is suitable for use in developing the valuation

[1] Property transfers refer to the transfer of property ownership without there being a sale, such as in divorces or parent to child transfers

[2] Although Knowledge Factory's transfers data dates back to 1753, there are not many transactions per year for the initial period, most records lack purchase prices, and the records before 1993 are notoriously dirty and not of the same quality as that after 1993.

model[3]. The resultant model can, however, be applied to any property for which the appropriate input data to the model is available.

3.2.1 Detailed overview of the data

In developing a statistical model based on vast amounts of data, it is common practise to collate all the data in a data mart. To decide which specific data sets should be used and which data sets provide the most accurate and robust information regarding deeds data, it is important to consider the following:

- What is the strict definition of a property?
- How are the various spatial levels of the deeds data defined?
- How are the various geographical levels of deeds data defined?

For the purposes of this study, a property is defined for the purposes of this study, as an object that may be transacted or sold from one person or entity to another (see Figure 3.1).

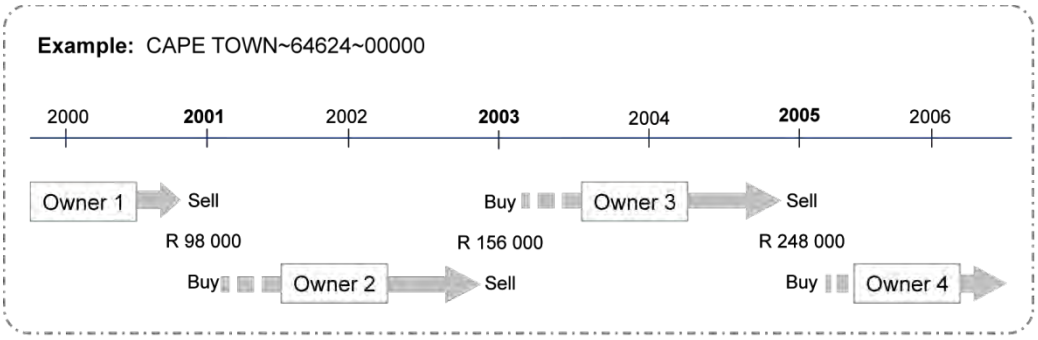


Figure 3.1: *Example of a series of property transactions involving a specific property*

The first person or entity is called the seller and the second person or entity is called the buyer. Each transaction involves a property (erf key[4]), a seller (e.g. owner 1), a buyer (e.g. owner 2), a purchase price, a date when the property was registered and a deeds office where it was registered. Each deeds office has a deeds office name

[3] Of the 14.7 million transactions in the post 1993 database, only 1,325 million transaction pairs (9%) remain after cleaning and validation.

[4] See Appendix A for a detailed explanation of an Erf Key

and a deeds office number.

The term “spatial levels” refer to generic areas of land (e.g. provinces, towns), while geographic elements refer to specific locations (e.g. Gauteng, North West Province, Pretoria and Potchefstroom).

The hierarchical spatial levels of property data used in this study are illustrated in Figure 3.2 below. The highest spatial level is defined as the national level which represents the whole of South Africa. The second spatial level is the provincial level. South Africa has 9 provinces divided into major and minor provinces based on the number of actual property transactions. The three “major” provinces are Gauteng, Western Cape and Kwa-Zulu Natal[5]. The rest of the provinces are the “minor” provinces, namely Northern Cape, Eastern Cape, Free State, Limpopo, Mpumulanga and North West. These provinces have the lowest number of property transactions or trades per province.

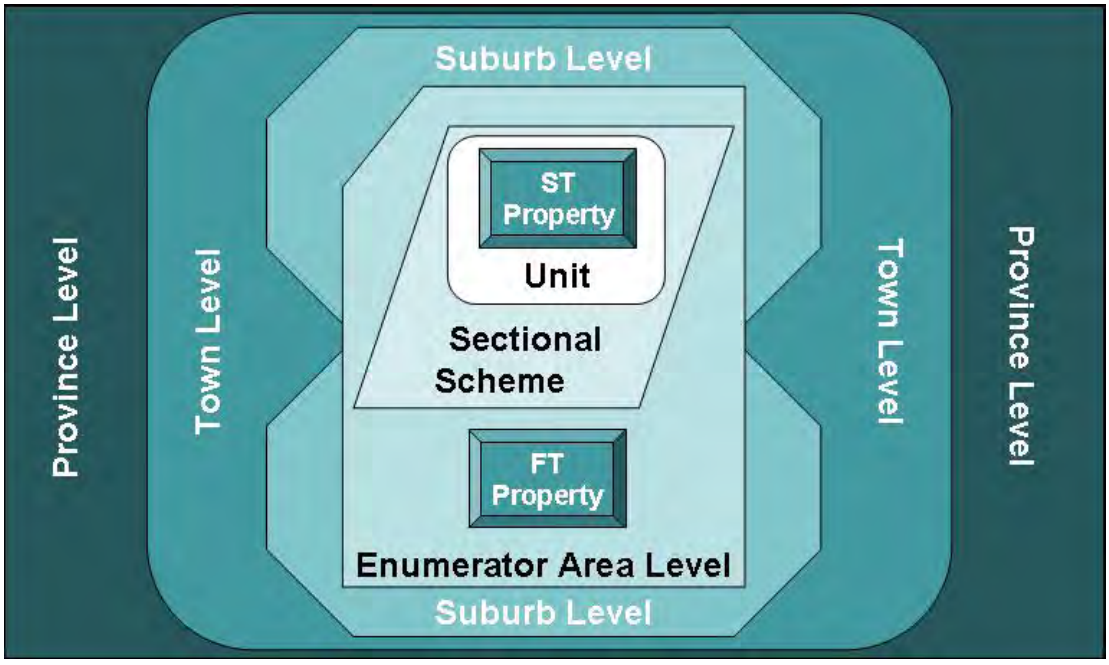


Figure 3.2: *Hierarchical spatial levels of data*

Each province in South Africa has one or more metro-towns; such as, Johannesburg being a metro-town in the province of Gauteng. There are a total of 5,080 metro-towns in South Africa which make up the next level of the data. Each metro-town in

[5] These provinces are not necessarily the largest in surface area, but have the highest number of property transactions or trades per province.

return is made up of one or more suburbs; Houghton is, for example, a suburb of Johannesburg. There are 9,955 suburbs (with suburb codes) in South Africa representing the next level of spatial data.

Enumeration areas (EA) is a small-area spatial layer that is defined as a small unit of manageable size (in terms of population and land area) assigned to a single person to enumerate during a census count (Figure 3.3). A total of 80,787 EAs were demarcated for Census 2001[6]. For the purposes of this study, EAs represent an even lower spatial level that can be utilised in a comparable sales method in the same way that the concept of a housing complex is used when comparing sectional title properties[7].

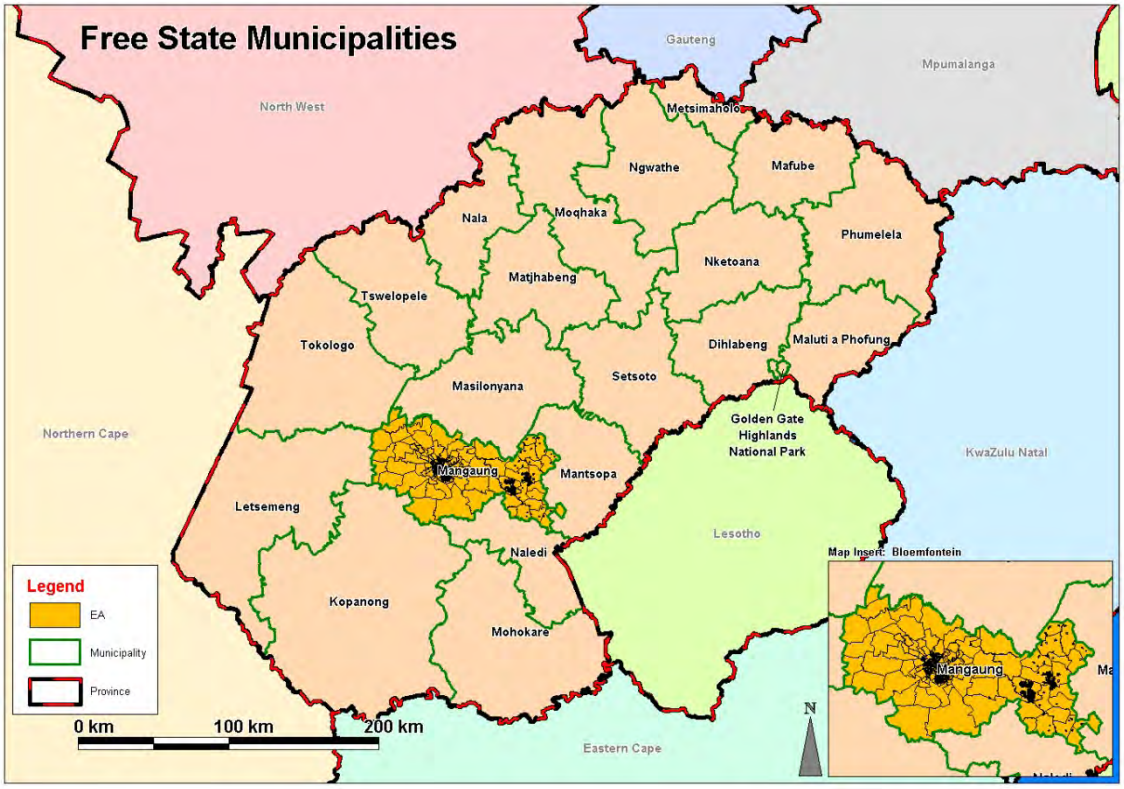


Figure 3.3: *Municipalities of the Free State and EAs in Mangaung municipality*

EAs cover a subset of residential properties in one or more suburbs within one metro-town. These areas are differentiated by EA codes which are assigned to

[6] There are, however, only 52 377 EAs defined in the deeds data for South Africa.

[7] The area of the residential property in m^2 is also used in comparative sales.

individual EAs. Each EA may contain both full title properties[8] (where the owner owns the property as well as the land it is built on) and sectional title properties (characterised by separate ownership of units or sections within a complex or security area development). Although there is additional spatial data such as Group Codes[9], the full title and sectional title levels referred to above are the lowest level of area-linked spatial property data referred to in this study.

KF's Clusterplus (Knowledge Factory) is a geo-demographic segmentation system which provides insight into the behaviours, characteristics, lifestyles and locations of the people of South Africa. It is developed at a suburb and sub-place level and modelled utilising primarily Deeds Office and Census information. Clusterplus is based on comprehensive datasets and provides coverage of the entire South African population. Distinguishing itself in terms of specificity, Clusterplus segments the South African population into 11 main groups, which are further divided into 38 clusters[10]. These groups and clusters have been defined in terms of the following variables:

- Socio-economic rank – income, property value, education and occupation;
- Life stage – age, household and family structure; and
- Dwelling type – size, type and age of structure.

The socio-economic groups are ranked or classified, with group A (Silver Spoons) being the top end of the South African population and group J (Below the Breadline) being the bottom end. Group S is classified as special cases such as golf courses, cemeteries etc.

3.2.2 Understanding and selecting the underlying data

All the data sources used to construct the data mart from which the models are

[8] There were 12.375 million full title and 2,322 million sectional title transactions registered in South Africa in April 2009.

[9] Group Codes are related to the Living Standards Measures (or LSM) developed by the SA Advertising Research Foundation as a market research tool (SAARF).

[10] See Appendix B for more detail on groups and clusters. The cluster level is not used for the purposes of this study, due to limited data.

developed, are from KF. For the purposes of this study, five data sets that contain the necessary data were identified and sourced.

3.2.2.1 KF Transfers table

The Transfers table (Table 3.1), consists of all the transactions that have taken place on properties in South Africa. Most of the fields required for the final data mart are populated in this table, namely title deed number, deeds office code, registration date, purchase price, erf key, sectional scheme id, unit, extent sqm (size of the erf in m^2) and the buyer's name.

Sectional Title

Title Deed Number	Deeds Office Code	Resitration Data	Purchase Prive	Erf Key	Sectional Scheme ID	Unit	Estent Sqm	Buyer Name
ST9102/1998	J	19980217	240000	BLACKHEATH JHB~318~00000	28256	4	148	BENEKE SANDRA GWENDOLINE

Full Title

Title Deed Number	Deeds Office Code	Resitration Data	Purchase Prive	Erf Key	Sectional Scheme ID	Unit	Estent Sqm	Buyer Name
T83033/2003	J	20031201	1198100	BLACKHEATH JHB~31~00000	0	0	1983	ALLRO INV PTY LTD

Table 3.1: *Example of KF® Transfers Table*

3.2.2.2 KF Cad to Sub table

The Cad to Sub table (Table 3.2) is a suburb cadastre and consists of all the suburb information in South Africa. It is unique by erf key, which means there should not be any duplicate erf keys in this table. This table contains the suburb code, metro town and province.

Erf Key	Suburb Code	Metro Town	Province
STELLENBOSCH RD~701~00000	0081-0129-000	STELLENBOSCH RD	WESTERN CAPE

Table 3.2: *Example of KF® Cad to Sub Table*

3.2.2.3 KF Clu Nat table

The Clu Nat table (Table 3.3), consists mainly of all the socio-economic group information (Appendix B). The table is unique by suburb code and consists of suburb code, suburb name, socio-economic group code, socio-economic cluster name and

socio-economic cluster code.

Suburb Code	Suburb Name	Socio Economic Group Code	Socio Economic Cluster Name	Socio Economic Cluster Code
1603-0990-000	SHERWOOD	B	Terracotta Terraces	7M

Table 3.3: *Example of KF® Clu Nat Table*

3.2.2.4 KF Scheme table

The Scheme table (Table 3.4), consists of most of the sectional schemes in South Africa. The fields used in this table are deeds office code, erf key, sectional scheme ID and sectional scheme name. This table is unique by deeds office code and erf key, thus there are no duplicate erf keys in a deeds office code.

Deeds Office Code	Erf Key	Sectional Scheme ID	Sectional Scheme Name
N	FT~15152~00136	32713	WATERFALL PARK

Table 3.4: *Example of KF® Scheme Table*

3.2.2.5 KF Cad EA table

Lastly, to add the enumerator area code the Cad EA table is used (Table 3.5), which is an enumerator area cadaster and contains all the enumerator area information. This table is unique by suburb code and erf key (see Appendix A) and thus there are no duplicate erf keys per suburb.

Suburb Code	Erf Key	Enumerator Area Code
1611-0000-000	UMHLANGA ROCKS~2517~00015	57204074

Table 3.5: *Example of KF® Cad EA Table*

All the tables mentioned above need to be collated to construct a data mart on which to develop the valuation models.

3.2.3 Data mart construction

To construct a data mart, all the data needs to be integrated into a single table. The *Transfers table* is joined to the *Cad to Sub table* by erf key so that the suburb information for all transactions is populated. This data mart is then further integrated with the *Clu Nat table* by suburb code to populate all the socio-economic group information. In the Transfers table there were a few of the sectional scheme IDs that were populated inaccurately, for example, a sectional title property having a unit number but no sectional scheme ID. To clean-up this data, two versions of the Scheme table from KF (Table 3.6) were derived, consisting of all the sectional scheme information.

KF Scheme Table (1)

Deeds Office Code	Erf Key	Sectional Scheme ID
N	FT~15152~00136	32713

KF Scheme Table (2)

Deeds Office Code	Erf Key	Sectional Scheme ID	Sectional Scheme Name
N	FT~15152~00136	32713	WATERFALL PARK

Table 3.6: *Example of KF® Scheme Table (1) and (2)*

Firstly, from the Scheme (1) table in Table 3.6, the fields for deeds office code, erf key and sectional scheme ID are kept and the duplicates are removed. When it is integrated into the data mart above by deeds office code and erf key, the missing sectional scheme IDs from the data mart is replaced with the populated sectional scheme IDs from the Scheme (1) table. The Scheme (2) table is then used to add the field sectional scheme name to the data mart. The duplicate records are removed by inspecting the deeds office code, erf key, sectional scheme ID and sectional scheme name fields. Lastly the data mart is joined to the *Cad EA table* to add the enumerator area code information. This step completes construction of the data mart.

In an ideal world, one would expect all data fields to be completely populated with accurate and reliable data, but because the data is notoriously dirty as mentioned above, this is not the case. Quite a lot of data was missing. In SAS, if a variable is missing and it is part of the required variables to develop a model, the whole record is excluded from the data, thus it is important to deal with missing information in a

clever way. Missing values must be populated with a text string or number that does not already represent something else to be able to distinguish between missing values and non-missing values after the model development. One can deal with missing information in a number of ways (SAS 1976), but for the purpose of this study it was decided that missing numeric values will be populated with a zero and text fields populated with five Z's ('ZZZZZ'). Except for modelling purposes, this clean-up process also helps to identify missing information and makes it easier to join tables in order to end up with one data mart.

Even after the final data mart has been constructed, all the missing values have been replaced and the sectional scheme IDs have been corrected, more clean-up activities were required which will be detailed in the next section.

3.2.4 Cleaning the data

In validating the data, a further few data problems were identified and corrected.

3.2.4.1 Removing garages indicated as residential properties

If a unit in a sectional title has a garage, it is seen as a separate property in the deeds data and because a garage is not sold separately as in the case of a property this may skew the data and thus needs to be removed. Garages are identified as having an erf size ("extent sqm") less than or equal to 36m^2 (see Table 3.7), being a sectional title and having the same province, group code, suburb code, erf key, sectional scheme id and registration date as one of the other units in the sectional scheme. All garages were subsequently removed from the final data mart.

Example	Province	Group Code	Erf Key	Sectional Scheme ID	Registration Date	Extent Sqm
Property	Gauteng	A	Sandown~23450~00000	212	10/14/2009	160
Garage	Gauteng	A	Sandown~23450~00000	212	10/14/2009	20

Table 3.7: *Example of a property with a garage*

3.2.4.2 Multiple property purchases on the same date

If one owner buys a few properties with the same erf key on the same day at the same deeds office, for example a developer buying a few units in a sectional scheme, the total price he paid for all of the units is displayed as the price paid for each unit respectively. To correct this, the prices therefore have to be split amongst all the units in the sectional scheme, and the assumption made that the buyer bought all the units for the same price. If this is not corrected, it can cause the data to be biased and the predictions will be less accurate for that area.

Title Deed Number	Deeds Office Code	Registration Date	Purchase Price	Buyer Name	Extent Sqm	New Purchase Price
ST35216/2008	N	20080724	750000	AITKEN BRONWEN ANGELA	95	375000
ST35216/2008	N	20080724	750000	AITKEN BRONWEN ANGELA	77	375000

Table 3.8: *Example of a duplicate property – possibly a developer*

The Transfers Adj table (Table 3.8) is created with fields for title deed number, deeds office, registration date, purchase price, buyer name and extent sqm and new purchase price. The purchase price is divided by the number of duplicated records for these fields. The data mart is then joined with the Transfers Adj table by buyer name, title deed number, deeds office and registration date and where the new purchase price was calculated, the old purchase price was replaced with the new one.

3.2.4.3 Invalid dates in registration date field

Finally, the last clean-up is the registration dates which are populated with a few invalid dates which could be due to typing errors and could influence the data. The empirical study behind this dissertation was based on data from January 1993 up until April 2009. All dates later than April 2009, because of capturing errors or dates that are populated with a blank or a zero, have been replaced with a missing value.

3.2.5 Exclusions and validations

In this section the exclusions and validation of the data will be explained. The data exclusions focus mainly on removing 'dirty' data or data that can't be used to develop valuation models. The data validation details statistical techniques used to identify outliers, arms length transactions etc which need to be taken into account so as not to develop a valuation model based on biased information.

3.2.5.1 Data exclusions (phase 1)

The first round of data exclusions is based on a number of assumptions around the main input variables to the model, namely province, group code, suburb code, registration date and purchase price (see Figure 3.4).

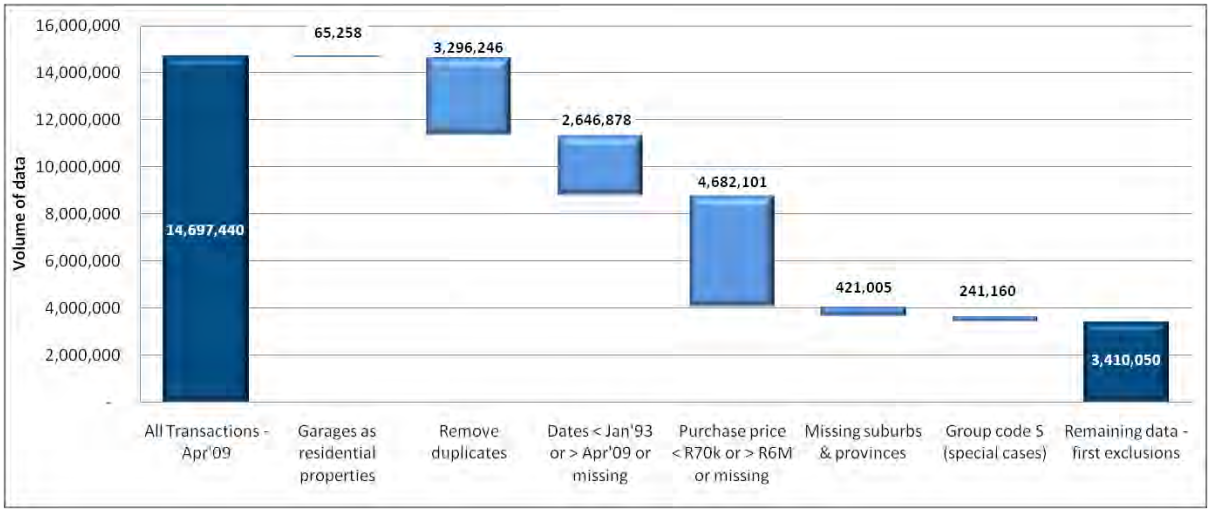


Figure 3.4: *Exclusions in the data – phase 1*[11]

Firstly, all duplicates in the data are removed by property type (full title or sectional title), province, socio-economic group code, cluster name, cluster code, suburb code, erf key, sectional scheme ID, unit, registration date, purchase price, extent sqm, title deed number and deeds office.

[11] It is important to note that the cleaning actions and the removal of exclusions were done as serial processes, i.e. each action was executed on the resulting and remaining data of the previous process. The number of transactions identified to be removed for each process is therefore not the number of such transactions in the original database.

Secondly, there were exclusions around the registration date. The property transactions in the data mart whose registration dates are before 1993 was part of a data dump that took place to start off the *Transfers table* (KF). The data before 1993 is thus very thin and inaccurate and had to be excluded because it will skew the data. The missing or zero registration dates need to be excluded as the registration date is one of the key inputs to the model and thus needs to be as accurate and clean as possible. Transactions with registration dates after April 2009 also had to be excluded as these are obviously incorrect registration dates, given that the data available was for up to April 2009.

Thirdly, there were exclusions around the purchase price. If the purchase price of a property is smaller than a certain value, which is dependent on the area in which it is situated, this could mean that the transaction involved a section of vacant land and not a residential unit and these transactions may influence the model incorrectly and need to be excluded. For this study a value of R70 000 was chosen for this lower limit. Similarly, there is an upper price limit for which the volume of properties with purchase prices larger than the limit is too small to be able to accurately predict the current valuation. A value of R6 000 000 was selected as upper limit and transactions with purchase prices exceeding this limit were thus excluded from the final data mart. Records with missing purchase prices are also excluded as purchase price is another one of the key variables required for the valuation models.

All transactional data with a ZZZZZ value in either the province or suburb fields in the final data mart also had to be excluded, since suburb and province are two key variables required for developing the valuation models.

Finally, to keep only residential properties in the final data mart, an exclusion is made when the group code field in the final data mart is S (special cases) and the cluster code is not 1S (agricultural) or 16S (small holdings), in other words, all special cases that are not agricultural or small holdings are excluded.

3.2.5.2 Exclusions due to validation (phase 2)

The next round of data exclusions were based more on the statistical measures and business input to the data (see Figure 3.5).

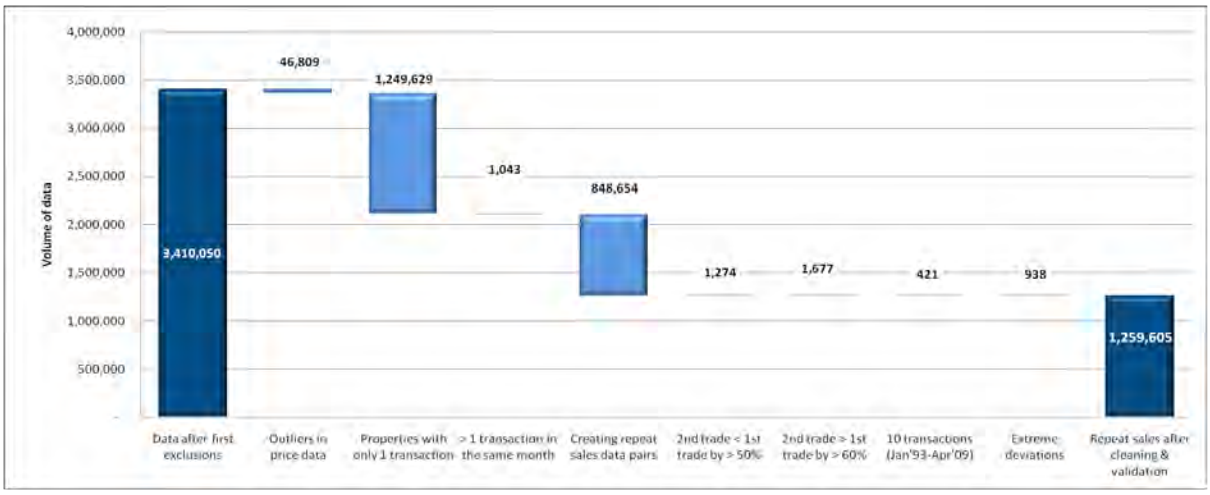


Figure 3.5: *Exclusions in the data – phase 2*

Firstly, outliers were removed from the data based on an interval of two standard deviations from the purchase price per province, group code, cluster code, suburb code and registration year. This implies that in order to deal with noise and dirty data, prices that deviate outside the bounds of two standard deviations from the mean price per suburb, were excluded (see Figure 3.6).

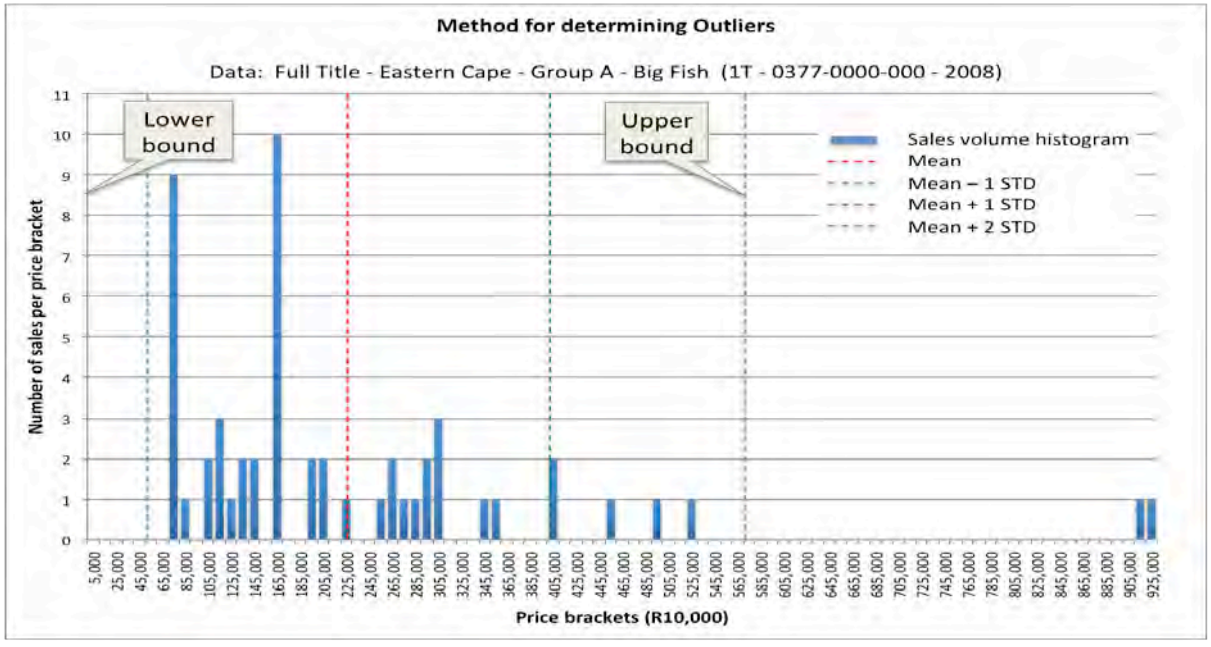


Figure 3.6: *Removal of outliers*

Secondly, in cases where more than one transaction was registered per property in the same month, the data of the most recent transaction is used, because the time resolution of the model is in months, and only one transaction per month can be used (see Figure 3.7 below between owner 1 and owner 2).

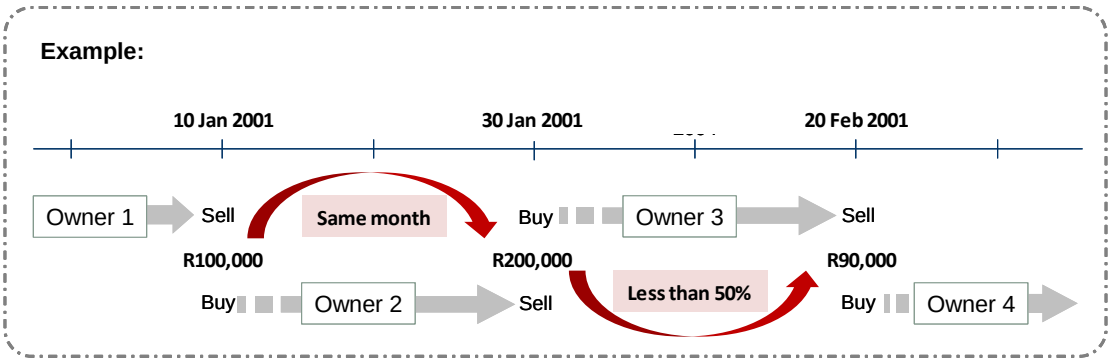


Figure 3.7: *Data exclusions: (i) Transactions in the same month and (ii) Transaction price < 50% of the price it was bought for.*

Before implementing the next data exclusion, the transactional data was grouped into pairs and transposed. This is a necessary step for developing a repeat sales valuation model. The models based on central tendency can also be evaluated using the same data.

Thirdly, if a property is traded more than once and the last trade is less than the previous trade by more than 50% per month for the number of months in between the two trades then the property transaction pair is excluded from the data mart because of too severe depreciation of property value (it is classified as abnormal value depreciation of property values). For an example, see Figure 3.7 between owner 3 and owner 4.

If a property is traded more than once and the growth in value from the first trade to the second trade is more than 60% per month, then the property transaction pair is also excluded from the data mart. The rapid growth could be due to renovations or extra development on the property and would therefore skew the model output if it was included because this behaviour cannot be generalised for the area in which it exists. For an example, see Figure 3.8 illustrating a transaction between owner 1 and owner 2.

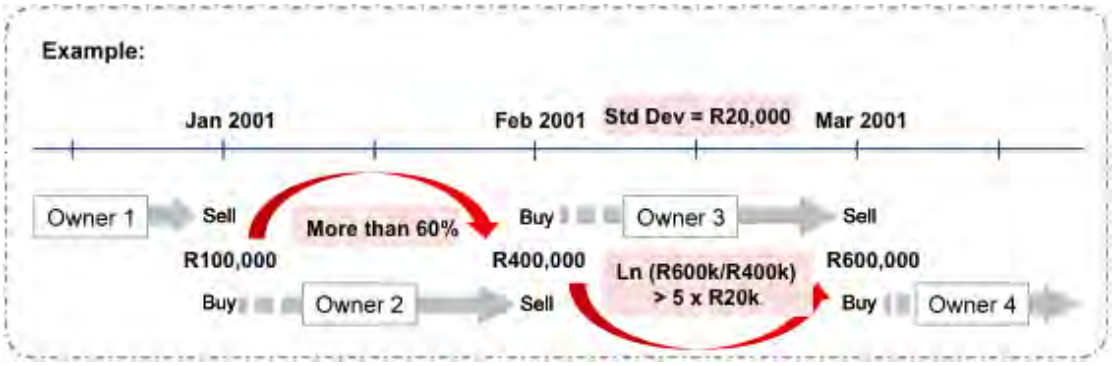


Figure 3.8: *Data exclusions: (i) A transaction is > 60% per month of the price it was bought for and (ii) The natural log of the last price divided by the previous price is more than 5 standard deviations.*

If there are more than ten trades on the same property between 1993 and 2009 then the property is also excluded from the data. Numerous trades on the property can be due to something being wrong with the property or the area that the property is in and this can skew the data.

Finally, transactions where the price inflation differs dramatically from the norm for similar properties are also excluded. Such extreme deviations could be due to:

- An unidentified non-arms-length transaction (i.e. where it is a parent to child transaction);
- The previous purchase was for a piece of land which has subsequently been developed;
- The property has undergone major renovations; and
- Township transactions where the price inflation is extremely high due to artificially low initial prices resulting from housing allocations made by, for example, the state or parastatals.

Based on 18 months' cumulative experience with the data and by referring to other studies, an empirical criteria was developed for extreme deviations: If the logarithm of the last trade divided by the second to last trade (normalised growth from the second to last trade to the last trade) is more than five standard deviations, the property is excluded from the data (Jansen *et al.,* 2006). For an example illustrating

this, see Figure 3.8 between owner 3 and owner 4. After the data exclusions and validations there were 1 260 964 records left in the data set.

Upon inspection, the volume of transactions per month indicated a very clear decrease for the months of December 1998 to February 1999 for both the raw data as well as the filtered data to be used for modelling (Figure 3.9). This dip in transaction volumes could not be explained, but its effect is visible in most of the graphs involving transaction volumes in the remainder of this dissertation.

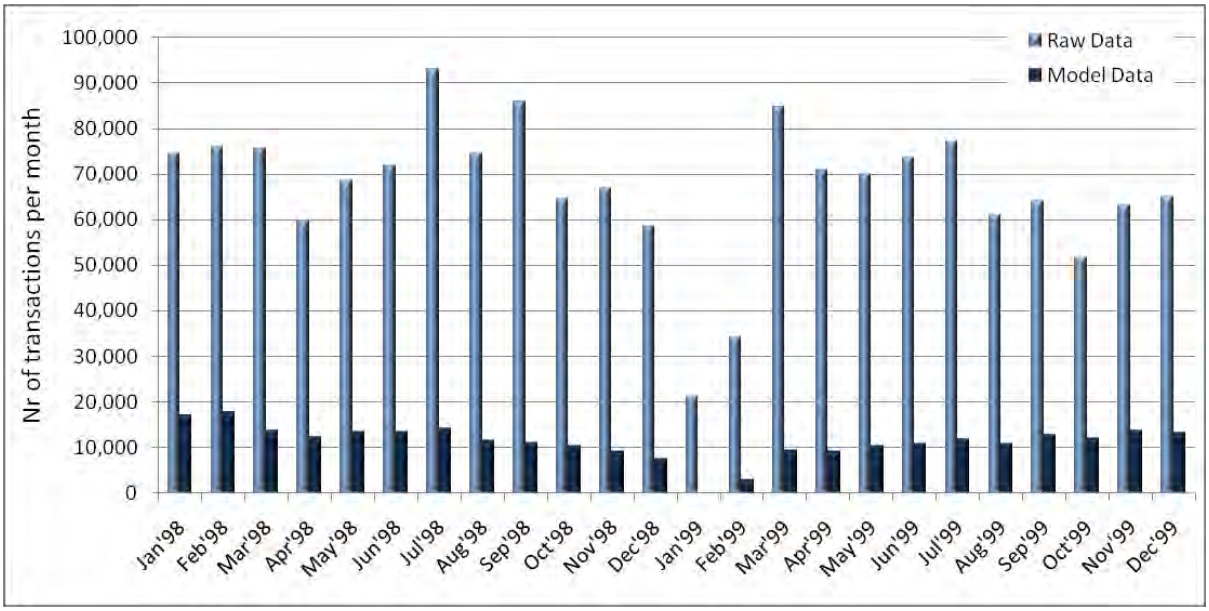


Figure 3.9: *Monthly transaction volume distribution of raw and model data*

3.3 Methodology

The central tendency (mean and median) approach to valuation modelling is deemed adequate by three of the four major South African banks who use it for generating their annual property price index[12]. Only recently an independent vendor in South Africa developed a repeat sales model which they use to calculate property price indices[13].

[12] ABSA House Price Index (ABSA)
FNB House Price Index (FNB)
Standard Bank House Price Index (Standard Bank)

[13] Lightstone House Price Index (Lightstone)

The inherent benefit of the repeat sales approach for property valuation is that the characteristics of a property are “locked in” the moment the property transacts. As an example, in the same area, a big house with good finishes will cost more than a small house with average finishes when sold at the same time. Thereafter, the model will differentiate between the two properties without needing to know information on size or finishes. The repeat sales methodology accounts for the quality characteristics of the property such as the number of bedrooms or quality of finishes (hedonic information) by comparing the trading of the same property over time.

The repeat sales methodology is based on a regression method wherein the logarithm of the price inflation for a property between two transactions is modelled as a function of the period between the two transactions. The underlying assumption is that prices are driven solely by inflation, in other words, two similar properties alongside each other will have the same growth in value over time. The method does not differentiate for the possibility that, for example, one property may be owned by someone who cannot afford to maintain it to the same level as the owner of another property.

3.3.1 Measures of central tendency (mean and median)

The valuation models using the measures of central tendency is very simplistic. As mentioned above, the variables required for the measures of central tendency is the purchase price and the registration date. Using this information after the data has been cleaned, the mean and median valuation models can be constructed.

Firstly, for the valuation model using the measures of central tendency, the mean and median purchase prices have to be calculated per month. A growth index is calculated using one of the following formulas below:

Mean growth index:

$$GI\left(Mean\right)_t = \frac{P_{t+1}}{P_t} \tag{3.1}$$

where

$GI\left(Mean\right)$ = Mean growth index for time *t,*

P_{t+1} = Mean price for time *t+1,*

P_t = Mean price for time *t.*

Median growth index:

$$GI\left(Median\right)_t = \frac{P_{t+1}}{P_t} \tag{3.2}$$

where

$GI\left(Median\right)_t$ = Median growth index for time *t,*

P_{t+1} = Median price for time *t+1,*

P_t = Median price for time *t.*

Multiplying the growth indices (GI) from one trade to the next for each month in between trades with the original purchase price the value of a property can be predicted for any month using the formula below:

$$PV_t = P_1 GI_1 GI_2 \ldots GI_t \tag{3.3}$$

where

PV_t = Predicted value for time *t,*

P_1 = Median price for the first time period *(t=1),*

GI_t = Growth index (mean or median) for time *t.*

3.3.2 Repeat sales approach

The repeat sales model uses a collection of the prices paid for single properties at different points in time to estimate a vector of numbers that 'best' explains the observed changes in price over the sample period (Abraham and Schauman, 1991).

The repeat sales approach uses the same variables as the measures of central tendency, namely the purchase price and registration date but this approach is a little more complex. Similar to the hedonic regression model, this approach is based on

general linear regression modelling, which requires explanatory variables (purchase price, registration date and grouping information) as well as a response variable, the target.

In practice, according to Jansen *et al.,* (2006), the repeat sales model uses ordinary least squares regression analysis in which the dependent variable is the logarithm of the price relative from the twice-sold property. The logarithm price relatives are then regressed on a set of dummy variables corresponding with the time periods. There is no constant term in the analysis; the coefficients are estimated only on the basis of changes in house prices over time. The estimated coefficients represent the logarithm of the cumulative price index for each period. The time dummy for the initial period is set at zero to normalise the index at 1.

The aim of the repeat sales model is to determine the growth index from one trade to the next thus the percentage growth from the 1st trade to the 2nd trade would be equal to the 2nd trade divided by the 1st trade. If the data is split for example into full title and sectional title properties, several property transactions which have traded more than once (repeat sales), could each form part of each split, thus the growth indices can differ quite a bit for full title and for sectional title properties. For this reason the target needs to be normalised to get a standardised growth index for each property transaction. The target can be normalised by taking the natural logarithm of the target.

The target (growth index) that will be used in repeat sales for modelling is:

$$GI_t = \ln\left(\frac{P_{t+1}}{P_t}\right) \tag{3.4}$$

with

$$\ln\left(\frac{P_{t+1}}{P_t}\right) = \sum_{i=1}^{t} \alpha_i m_i + \varepsilon_i \tag{3.5}$$

where

GI_t = Repeat sales growth index for time *t,*

P_{t+1} = Purchase price for time *t+1,*

P_t = Purchase price for time *t,*

α_i = Model output (unknown index numbers to be estimated) for time *i,*

m_i = Monthly dummy variables populated (-1, 0, 1) indicating an occurrence of P_t,

ε_i = Residuals in log form with zero mean, equal variances and uncorrelated with each other,

i = Jan 1993 ... Apr 2009.

Dummy variables are created for each month except the first month, the base period. For each property, the dummy variable for the first sale has a value of '-1' and the dummy variable for the next sale has a value of '+1'. The dummy variables for all the other months have a value of '0' (see Figure 3.10).

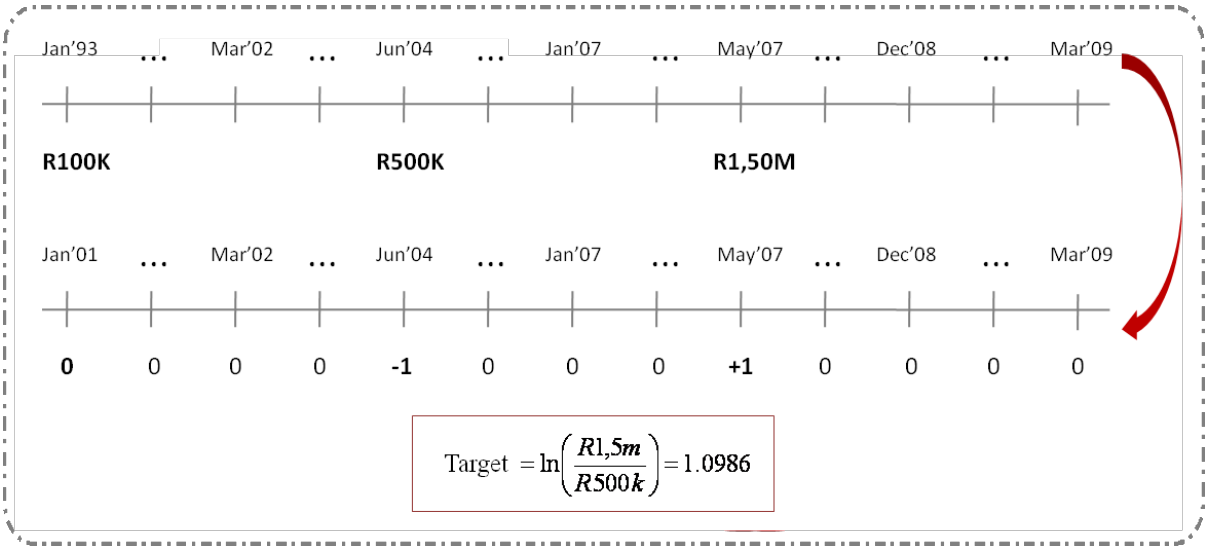


Figure 3.10: *Dummy variables created for each property transaction*

The target (growth index) of the property is defined as the response variable and the dummy variables are defined as the explanatory variables. The explanatory variables are then regressed to fit the response variable, in other words the explanatory variables are used to predict the target (growth index) of a property, the response variable.

The formula for the derivation of the predicted price at any future time period from the repeat sales formula is given by $P_{t+1} = P_t e^{\sum_{i=1} \alpha_i m_i}$ and is derived as follows:

$$\ln\left(\frac{P_{t+1}}{P_t}\right) = \sum_{i=1}^{t} a_i m_i + \varepsilon_i$$

$$e^{\left(\ln\left(\frac{P_{t+1}}{P_t}\right)\right)} = e^{\sum_{i=1}^{t} a_i m_i} \qquad (3.6)$$

$$\frac{P_{t+1}}{P_t} = e^{\sum_{i=1}^{t} a_i m_i}$$

and

$$\frac{P_{t+1} - P_t}{P_t} = \frac{P_{t+1}}{P_t} - \frac{P_t}{P_t} = \frac{P_{t+1}}{P_t} - 1 \qquad (3.7)$$

set equation 3.6 equal to 3.7

$$\frac{P_{t+1}}{P_t} - 1 = e^{\sum_{i=1}^{t} a_i m_i} - 1 \qquad (3.8)$$

$$P_{t+1} = P_t e^{\sum_{i=1}^{t} a_i m_i}$$

where

GI_t = Repeat sales growth index for time *t,*

P_{t+1} = Purchase price for time *t+1,*

P_t = Purchase price for time *t,*

α_i = Model output (unknown index numbers to be estimated) for time *i,*

m_i = Monthly dummy variables populated (-1, 0, 1) indicating an occurrence of P_t,

ε_i = Residuals in log form with zero mean, equal variances and uncorrelated with each other,

i = Jan 1993 ... Apr 2009.

When the general linear regression model fits the dummy variables to the given target, the parameter outputs are cumulative growth factors for each time period. With this cumulative growth factor, the predicted price of a property can now be

calculated as per the example given in Figure 3.11. This growth factor is also further used to calculate the property price index.

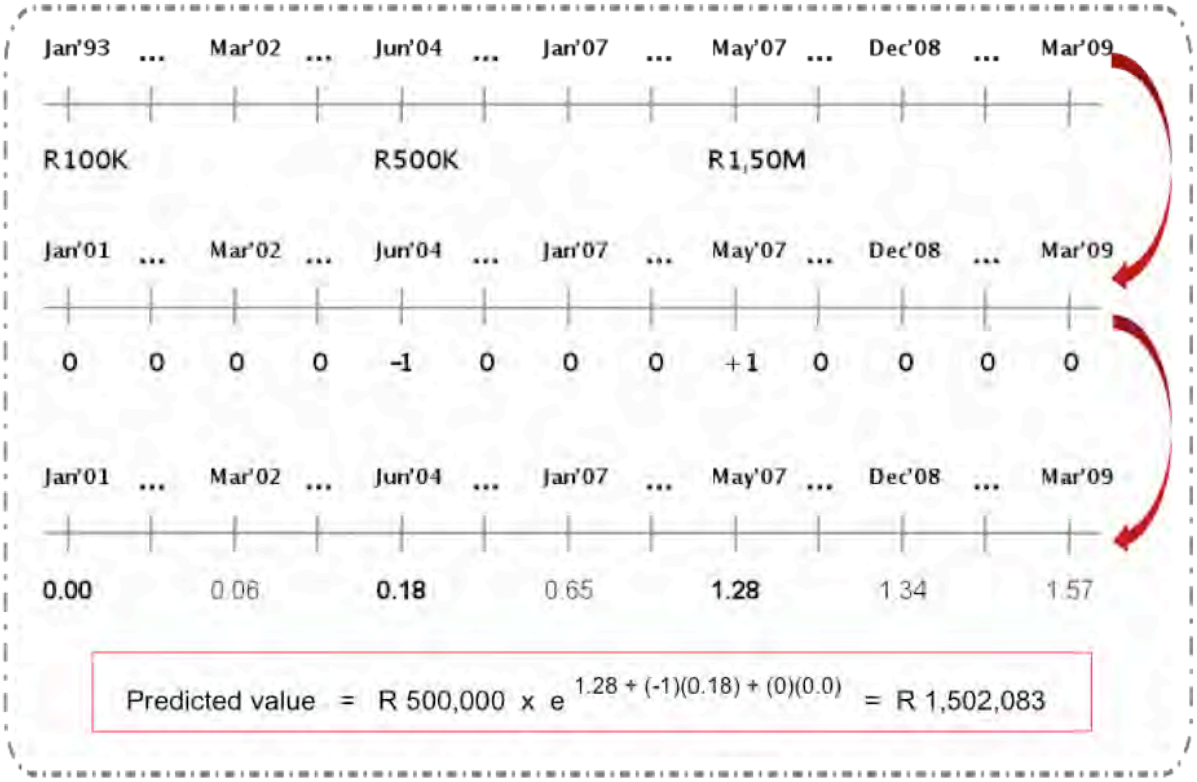


Figure 3.11: *Example of the calculation of the predicted price*

In SAS the procedure used to run the repeat sales model is called GLM. According to the SAS help file (SAS, 1976) the GLM procedure uses the method of least squares to fit general linear models and analyzes data within the framework of general linear models. The procedure handles models relating one or several continuous dependent variables to one or several independent variables. The independent variables may be either classification variables, which divide the observations into discrete groups, or continuous variables.

The least squares approach provides estimates of the linear parameters that are unbiased and have minimum variance among linear estimators. Under the further assumption that the errors have a normal (or Gaussian) distribution, the least squares estimates are the maximum likelihood estimates and their distribution is known. All of the significance levels ("*p* values") and confidence limits calculated by the GLM procedure require this assumption of normality in order to be exactly valid, although they are good approximations in many other cases.

3.4 Conclusion

In this chapter a detailed overview was given on the data used behind the developing of a property valuation model and the methodology of the valuation models discussed further in this study. The overview included insight into the clean-up process, validations and assumptions and exclusions made as well as the formulas used to develop the model based on the measures of central tendency (mean and median) and the repeat sales model.

In the next chapter the three models are developed based on the methodology discussed in Chapter 3 and various statistical tests are performed to determine the best of the three models with which to predict property prices.

Chapter 4: Results of comparing the three basic models

4.1 Introduction

In this chapter, the three property valuation models that were identified and explained in the previous chapter, namely the two measures of central tendency and the repeat sales model, will be evaluated to determine the most suitable one for developing a comprehensive property price index. A number of practical criteria identified in the literature will be applied to a large set of representative data to determine the most robust model. In the next chapter a few improvements to the most robust model will be presented and justified and from this 'improved' model the property price index will be derived.

4.2 Model output

4.2.1 Measures of central tendency (mean and median)

In the previous chapter, the various processes involved in cleaning the data and combining it into a single data mart, a necessary step before starting to develop and optimise models, were described. The measures of central tendency for the mean and median are the first two models that will be applied to the data. All three models will be applied to a random sample of 700 000 properties out of a total of 1 260 964 'cleaned' records (see Section 3.2).

The process starts by calculating the mean and median purchase prices for each month (see Section 3.3 on the methodology). From the results of this process, the growth index can be derived for both measures of central tendency, by applying formulas 3.1 and 3.2 respectively. The results in terms of the mean and median price and resultant growth index are illustrated in Figure 4.1 and 4.2 for full title properties in group A for major provinces.

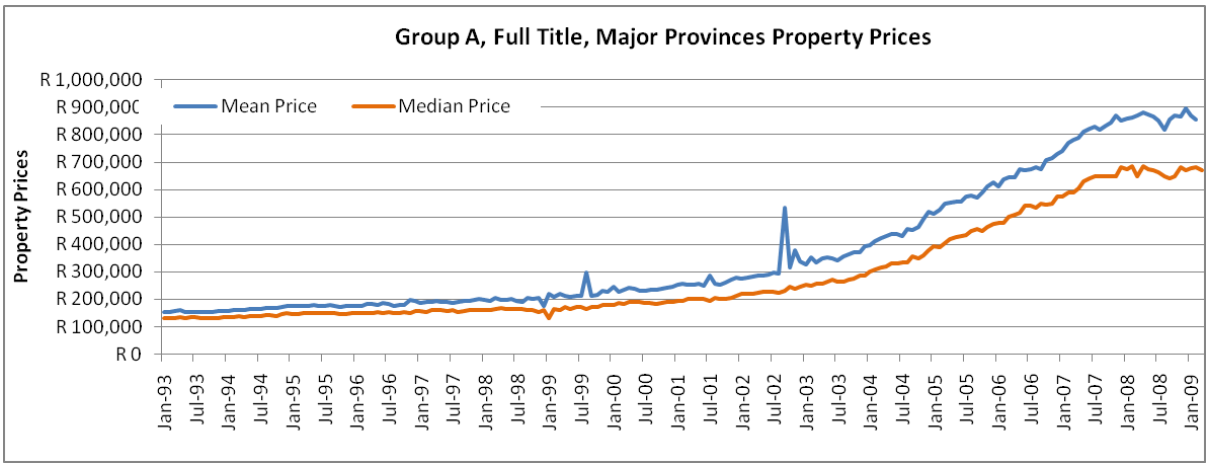


Figure 4.1: *Mean property prices for group A, full title, major provinces using the measures of central tendency - mean and median*

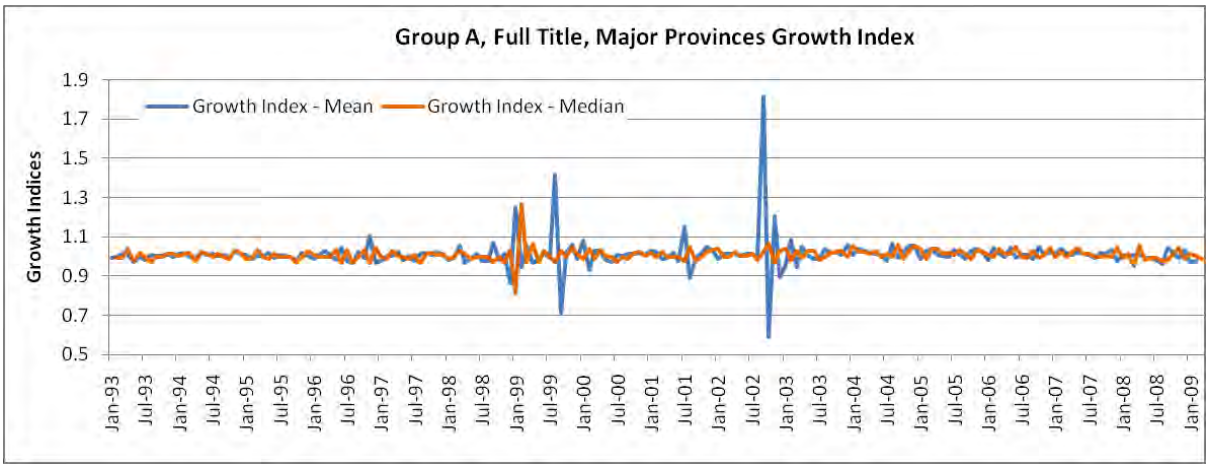


Figure 4.2: *Price growth index for group A, full title and major provinces using the measure of central tendency – mean and median*

From Figure 4.1 and 4.2 above it can be seen that the median property price is lower than the mean property price, especially from 2005 onwards.

4.2.2 Repeat sales approach

In this section the repeat sales model will be applied to the same random sample (700 000 properties) used for the measures of central tendency. The repeat sales model will be derived using formula 3.5.

Dummy Variables

As discussed in Section 3.3, the first step to developing a repeat sales model is creating dummy variables ('1' for the 2nd transaction, '-1' for the first transaction and '0' otherwise) as the independent variables for each transaction across each period as seen in Table 4.1:

Erf Key	Transactions								
	Dec'07	...	Feb'08	Mar'08	Apr'08	...	Aug'08	...	Apr'09
MIDSTREAM ESTATE~513~00008		...				...	1,800,000	...	1,900,000
MONUMENT PARK~1439~00004		...			521,345	...		...	1,300,000
PAULSHOF~349~00000		...			1,300,000	...		...	1,300,000
DOUGLASDALE~2258~00000		...		400,000		...		...	400,000
WILLOW ACRES~161~00000		...	1,600,000			...		...	1,770,000
DOWERGLEN~4~00000	2,100,000	...				...		...	2,850,000

Erf Key	Dummy Variables								
	Dec'07	...	Feb'08	Mar'08	Apr'08	...	Aug'08	...	Apr'09
MIDSTREAM ESTATE~513~00008	0	...	0	0	0	...	-1	...	1
MONUMENT PARK~1439~00004	0	...	0	0	-1	...	0	...	1
PAULSHOF~349~00000	0	...	0	0	-1	...	0	...	1
DOUGLASDALE~2258~00000	0	...	0	-1	0	...	0	...	1
WILLOW ACRES~161~00000	0	...	-1	0	0	...	0	...	1
DOWERGLEN~4~00000	-1	...	0	0	0	...	0	...	1

Table 4.1: *Example of dummy variables*

In the data set the time period will be the columns going across and the sales pairs will be the rows going down (each sales pair is in its own row). Then a matrix is populated where the purchase prices of the sales pairs are replaced with '-1', '1' and '0' according to which sale it is.

Target

As the second step, a target or dependent variable is created for each row. Since the repeat sales model is based on pairs of transactions of the same property, the target is equal to the natural logarithm of the second transaction's price, divided by the first transaction's price (see formula 3.4). Using formulas 3.4 and 3.5, the growth index parameters for each month is then derived (see Figure 4.3).

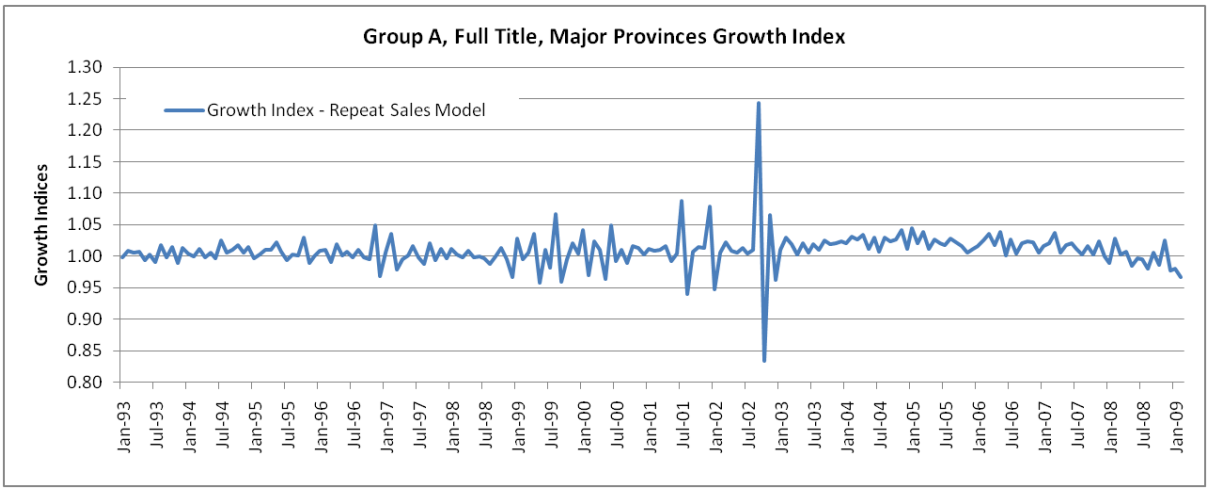


Figure 4.3: *The growth index using the repeat sales model*

4.3 Testing the accuracy of the three approaches using statistical techniques

In this section the three models will be compared using statistical tests. By reviewing the results the most accurate, robust model will be identified. All the tests were performed on the same dataset which contains only properties that sold more than once, to accommodate the repeat sales model, thus enabling a comparison of the predicted prices from the different models to the actual prices the property was sold for. The properties in the dataset are properties that were registered between January 1993 and April 2009. According to Leblond (2004) and Cook (2006), there are nine different tests that can be performed to determine which model is the most accurate when predicting house prices. All these tests were based on a random sample of 100 000 properties as a form of back-testing (predicting the second purchase price of the sales pair using the first purchase price of the sales pair and then comparing it to the actual value of the second purchase price).

The nine test statistics are the following:

1. Closest prediction to actual value;
2. Distribution of model errors;
3. Theil's U-statistic;
4. Mean error (ME);

5. Mean squared error (MSE);
6. Root mean squared error (RMSE);
7. Mean absolute error (MAE);
8. Mean prediction error (MPE); and
9. Mean absolute prediction error (MAPE).

4.3.1 Statistic 1: Closest prediction to actual value

The first statistic to measure which of the models is the most accurate in predicting the price of the property sold, is the closest prediction to actual value. This is also known as the model residual or model error. This statistic measures how close each model is in predicting the actual value of the property sold. The most accurate model, according to this statistic, is determined by counting the number of properties for which the particular model is the best performer, and then declaring the model that is the top performer in most of the cases as the 'winner' of the first statistic.

This statistic is measured using the following formula:

$$Closest\ \ prediction\ to\ actual\ value = \frac{(\hat{y} - y)}{y} \qquad (4.1)$$

where

$\hat{y}$ = Predicted value,

y = Actual value.

The results of applying the first statistic are presented in Figure 4.4. The repeat sales model was the best performer for 100 000 properties tested (73%), making it by far the best model according to the statistic.

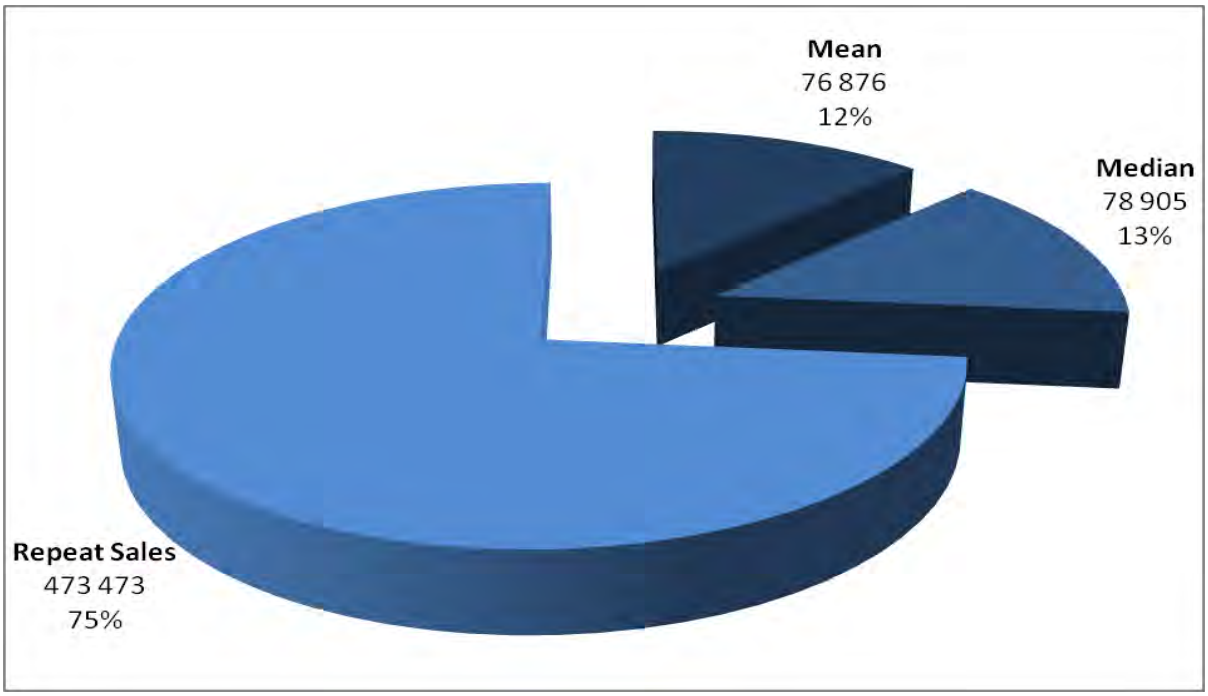


Figure 4.4: *The performance of the three basic models for Statistic 1: Closest prediction to physical assessment*

4.3.2 Statistic 2: Distribution of model errors

The model residual determined in the first statistic is used in the second statistic to determine the number (and percentage) of properties falling within 20% of the actual price (below or above) for each model. The results of the second statistic are presented in Figure 4.5. The model using the mean price to predict the value of properties, predicted 26.60% of the properties within 20% of their actual price, while the model using the median price predicted an average of 28.67% of the properties within 20% of their actual price. The repeat sales model predicted 51.84% of the properties within 20% of their actual price, once again making it the best model according to the statistic.

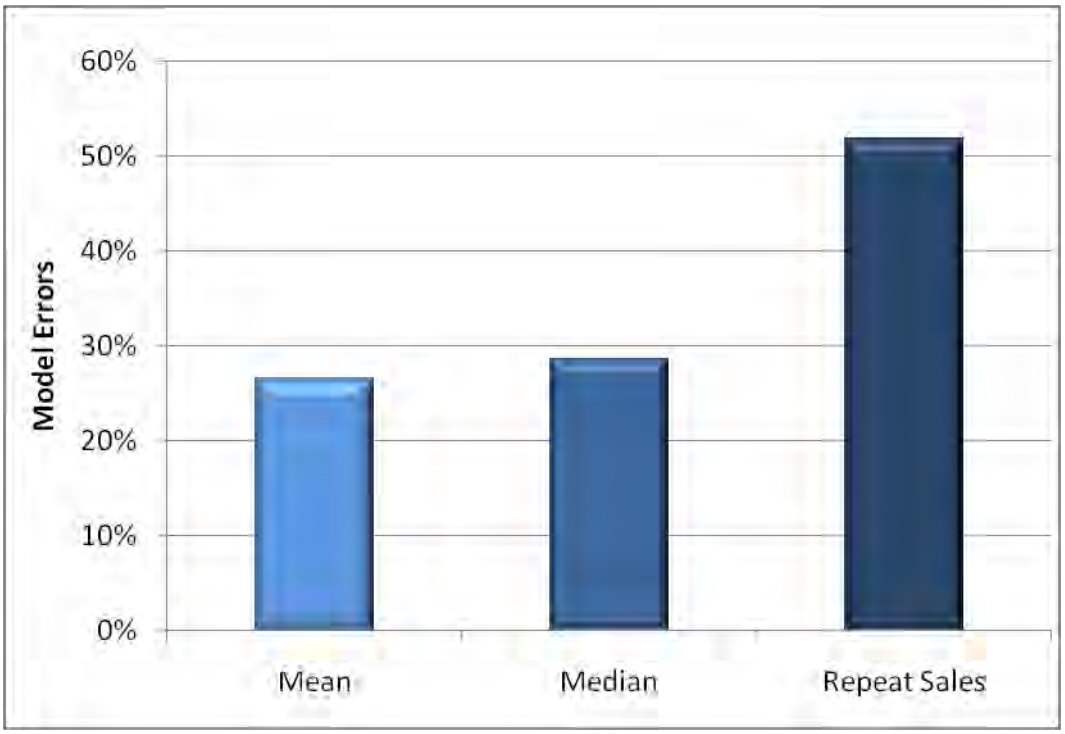


Figure 4.5: *The performance of the three basic models for Statistic 2: Distribution of model errors*

The results from this test can also be depicted as the cumulative residual distribution of properties by model (Figure 4.6). On the y-axis, the percentages of predictions within the error brackets (on the x-axis) are depicted. To illustrate, the number of predictions within -15% to +15% for the repeat sales, equals that of the mean model for a -30% to +30% error bracket. From the graph, it is clear that the repeat sales outperform the other two by far.

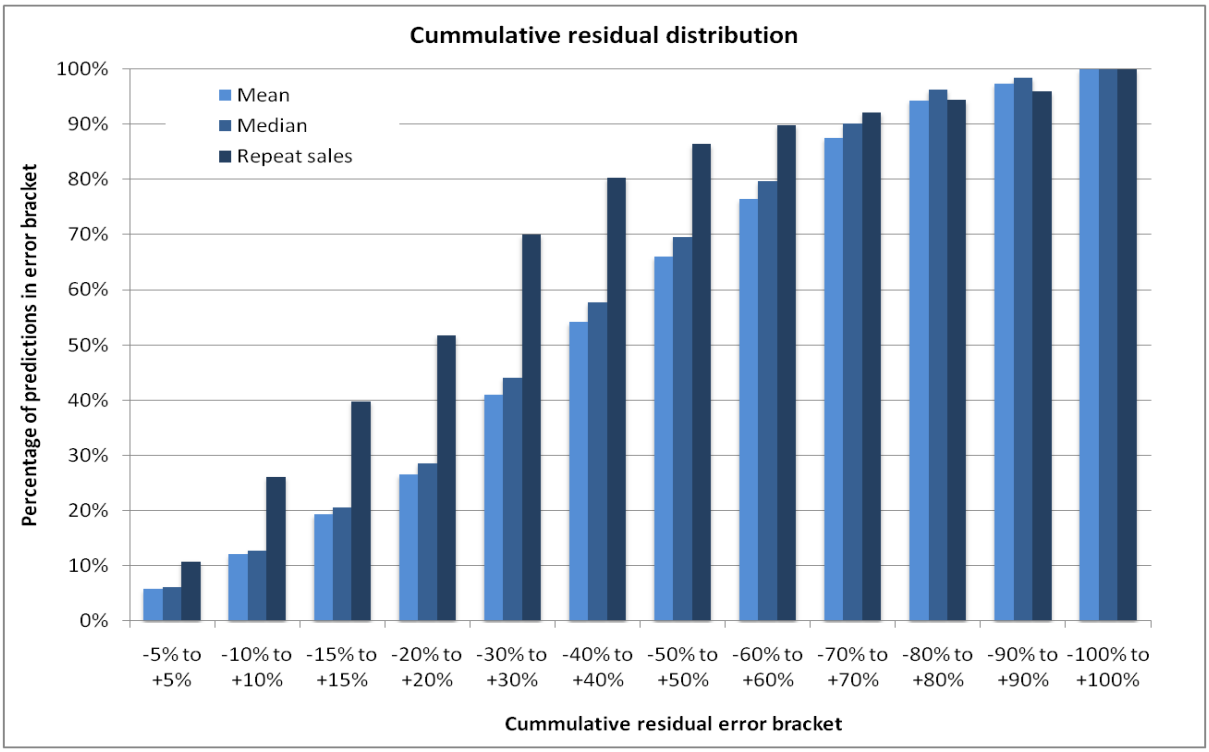


Figure 4.6: *Residual distribution for all properties*

4.3.3 Statistic 3: Theil's U-statistic

Theil's U-statistic is used to measure the quality of fit between actual and predicted property prices. The more accurate the prediction of a property, the lower the value of Theil's U-statistic. Theil's U-statistic is bounded between 0 and 1, with values closer to 0 indicating greater prediction accuracy. The formula used to measure Theil's U-statistic is the following:

$$Theil's\ U-statistic = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^{\,2}} + \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i^{\,2}} \qquad (4.2)$$

where

$\hat{y}$ = Predicted value,
y = Actual value,
n = Population size,
i = Jan 1993 ... Apr 2009.

The evaluation results of the third statistic are presented in Figure 4.7. The repeat sales model is the closest to 0, making it the best model according to the third statistic.

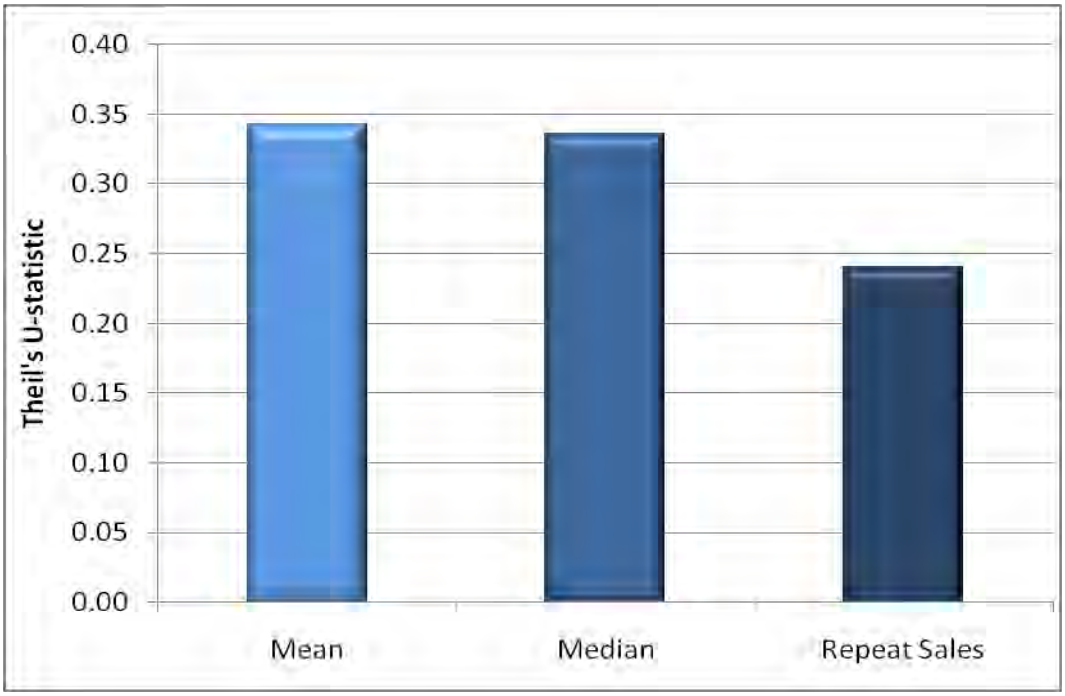


Figure 4.7: *The results of Statistic 3: Theil's U-statistic. The top performer for this statistic is the one closer to zero.*

4.3.4 Statistic 4: Mean error (ME)

In statistic 4, the mean error (ME) measures the mean residual of all predictions when compared to the actual values. The formula used to calculate the mean error is the following:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) \qquad (4.3)$$

where

$\hat{y}$ = Predicted value,

y = Actual value,

n = Population size,

i = Jan 1993 ... Apr 2009.

The results for the mean error are presented in Figure 4.8. Note that the lower the value, the lower the error, however a low value of the ME may conceal prediction inaccuracies due to the offsetting effect of large positive and negative forecast errors. However, despite the unbiasedness of the predictions, their inaccuracy becomes apparent from inspection of subsequent prediction evaluation statistics. The repeat sales model has by far the lowest ME.

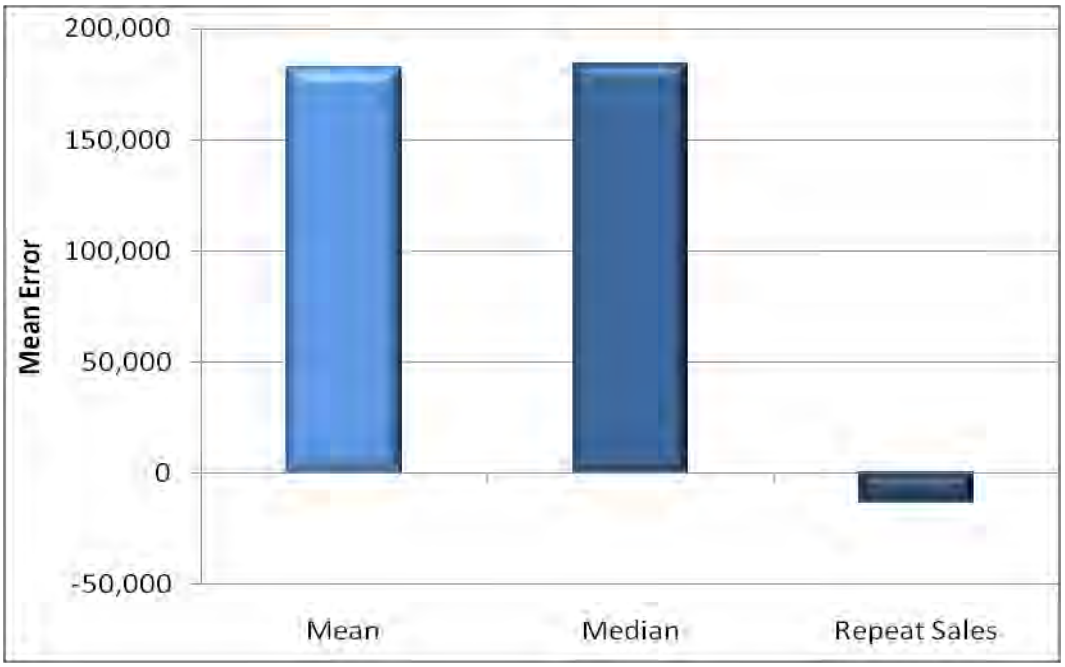


Figure 4.8: *The results of Statistic 4: Mean error (ME). The top performer is the one with the lowest ME.*

4.3.5 Statistic 5: Mean square error (MSE)

In statistic 5, the mean square error (MSE) corresponds to the squared error loss or quadratic loss. The formula used to calculate the mean square error is the following:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \qquad (4.4)$$

where

$\hat{y}$ = Predicted value,
y = Actual value,
n = Population size,
i = Jan 1993 ... Apr 2009.

The results for the mean square error are presented in Figure 4.9 (the lower the value, the lower the error). MSE may overcome the 'cancellation of positive and negative errors', but it fails to provide information on prediction accuracy relative to the scale of the series examined. The repeat sales model has the lowest MSE.

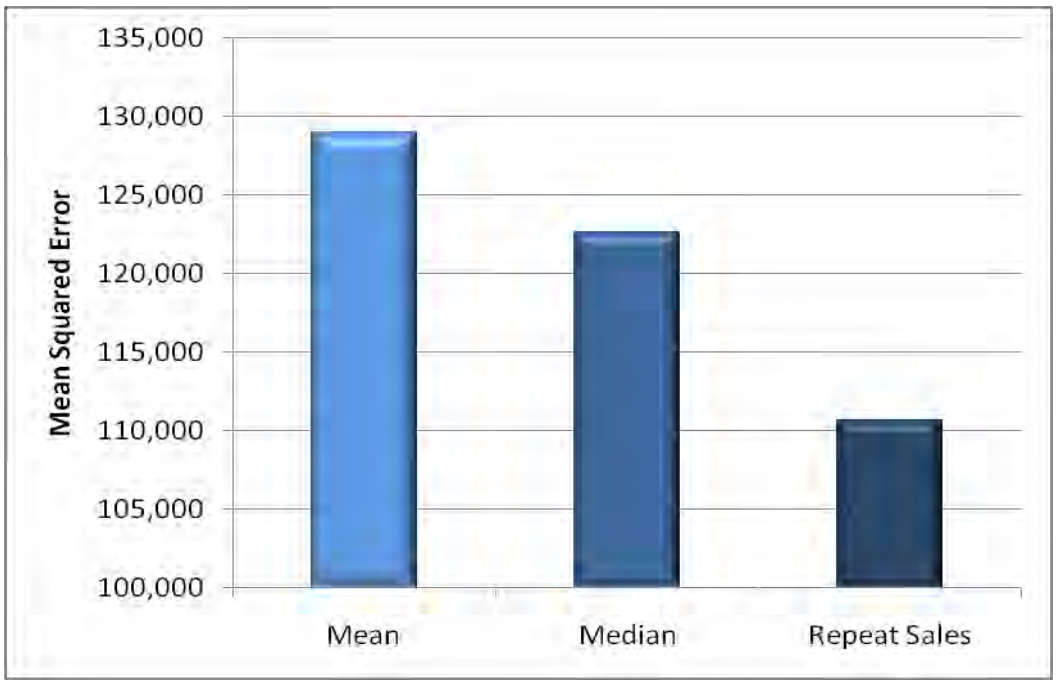


Figure 4.9: *The results of Statistic 5: Mean square error (MSE)*

4.3.6 Statistic 6: Root mean squared error (RMSE)

The sixth statistic, the root mean squared error, is a good measure of precision and is used to measure the differences between values predicted by a model and the observed values. The formula for the root mean squared error (RMSE) is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \qquad (4.5)$$

where

$\hat{y}$ = Predicted value,

y = Actual value,

n = Population size,

i = Jan 1993 ... Apr 2009.

The results for the RMSE are presented in Figure 4.10. The repeat sales model has the lowest RMSE making it the best performing model for statistic 6, although not by a large margin.

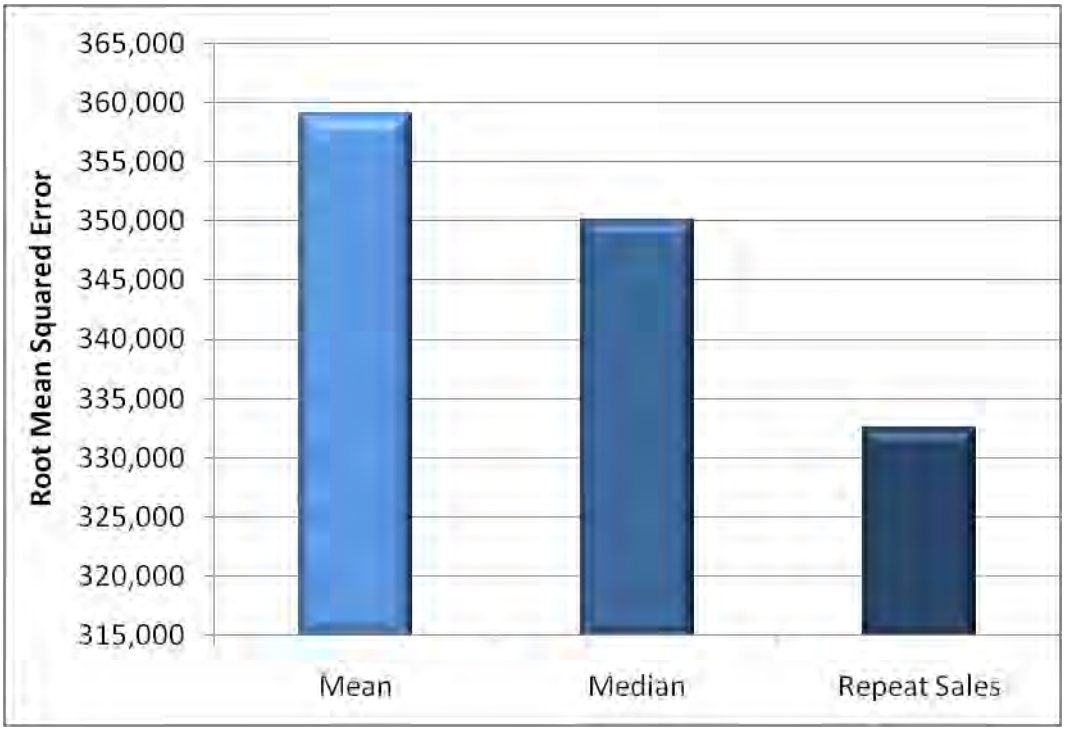


Figure 4.10: *Evaluation results for Statistic 6: Root mean squared error*

4.3.7 Statistic 7: Mean absolute error (MAE)

The seventh statistic, the mean absolute error, is, as the name suggests, the average of the absolute error. The formula for the mean absolute error (MAE) is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.6}$$

where

$\hat{y}$ = Predicted value,

y = Actual value,

n = Population size,

i = Jan 1993 ... Apr 2009.

The results for the MAE are presented in Figure 4.11 (the lower the value, the lower the error). MAE may overcome the 'cancellation of positive and negative errors', but they fail to provide information on prediction accuracy relative to the scale of the series examined like the MSE. MSE places a greater penalty on large prediction errors than the MAE. The repeat sales model has the lowest MAE making it the best performing model according to statistic seven.

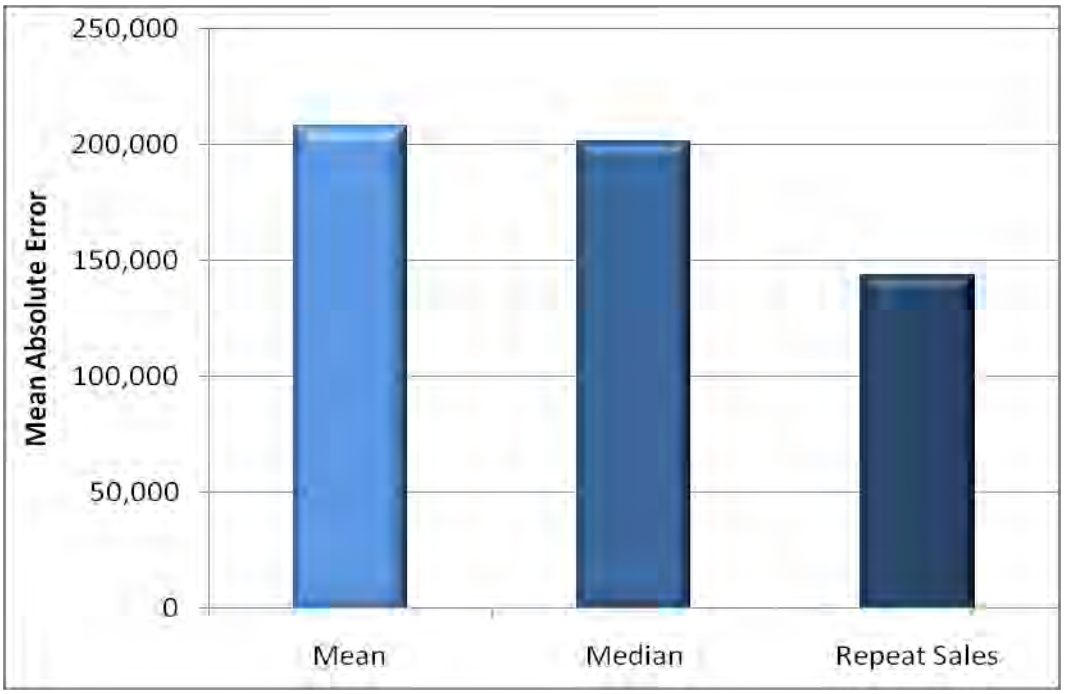


Figure 4.11: *Evaluation results for Statistic 7: Mean absolute error*

4.3.8 Statistic 8: Mean prediction error (MPE)

The eighth statistic, the mean prediction error, is the same as the fist statistic, but here the focus is on the actual error percentage, not the amount of times a specific model ‘won’. The formula for the mean prediction error (MPE) is:

$$MPE = \frac{1}{n}\sum_{i=1}^{n} 100 \times \frac{y_i - \hat{y}_i}{y_i} \qquad (4.7)$$

where

$\hat{y}$ = Predicted value,

y = Actual value,

n = Population size,

i = Jan 1993 ... Apr 2009.

The results for the MPE are presented in Figure 4.12. Note that the lower the percentage, the lower the error. The repeat sales model has by far the lowest MPE.

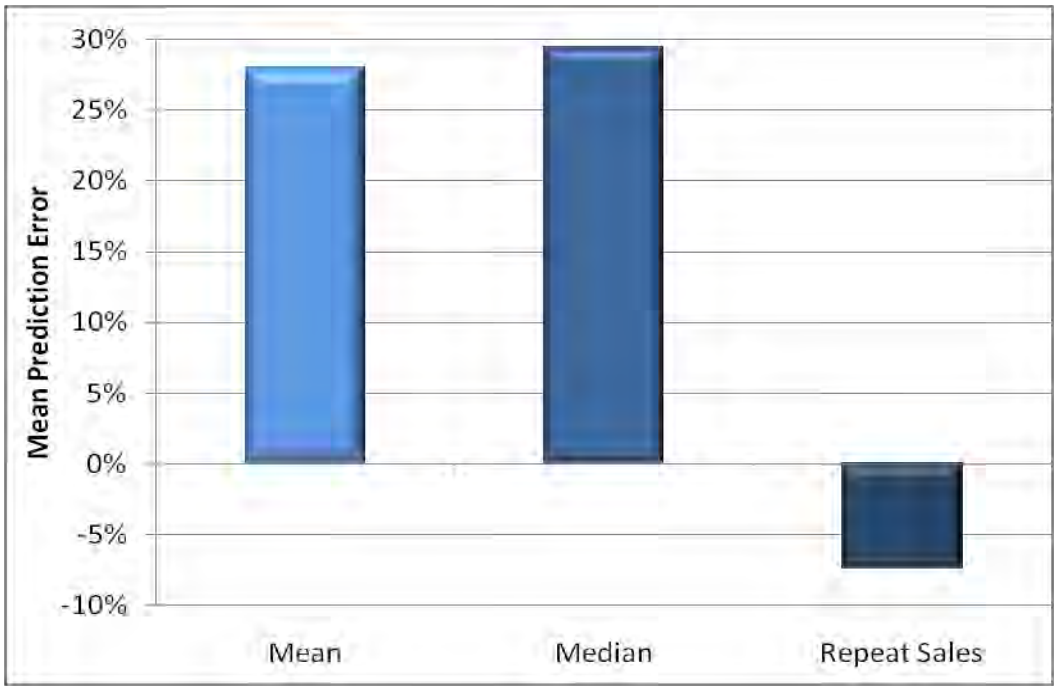


Figure 4.12: *Evaluation results for Statistic 8: Mean prediction error*

4.3.9 Statistic 9: Mean absolute prediction error (MAPE)

In statistic 9, the mean absolute prediction error (MAPE) measures the mean absolute error of all predictions when compared to the observed values (assessments). The formula used to calculate the mean absolute prediction error is the following:

$$MAPE = \frac{1}{n}\sum_{i=1}^{t} 100 \times \left| \frac{y_i - \hat{y}_i}{y_i} \right| \qquad (4.8)$$

where

$\hat{y}$ = Predicted value

y = Actual value

n = Population size

i = Jan 1993 ... Apr 2009

The results for the mean absolute prediction error statistic are presented in Figure 4.13. Note that the lower the percentage, the lower the error. The repeat sales model again features the lowest MAPE.

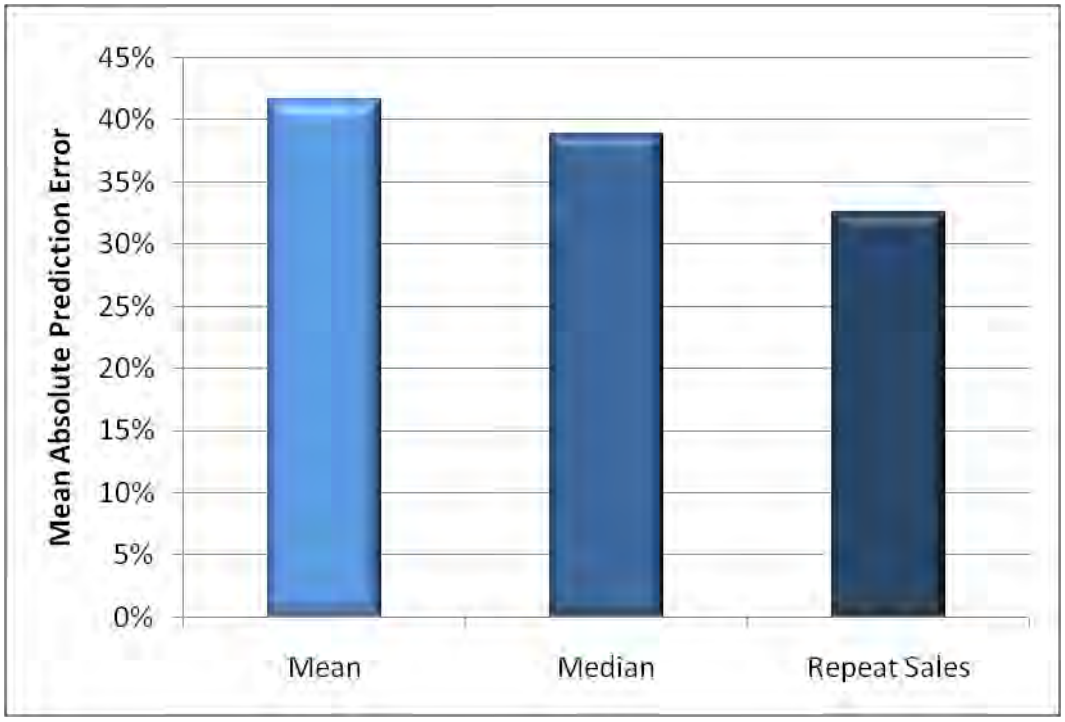


Figure 4.13: *The results of Statistic 9: Mean absolute prediction error (MAPE). Please note that a lower percentage indicates a lower prediction error.*

4.4 Conclusion

In conclusion, based on the data used for the evaluation exercise and the nine statistics discussed in this chapter, the repeat sales model is the most robust model to use when predicting residential property prices in South Africa.

The coverage of a model is also an important test, because it is of little use if a model is extremely accurate, but only for a small sample of properties. Since all three the models can be used to predict the property prices for all the data in the entire dataset for which the purchase price and registration date are available, all

three models automatically have a coverage of 88%[14] and therefore perform equally well. The topic of coverage will be discussed again in the next chapter.

In the next chapter, a range of potential improvements to the repeat sales model will be assessed. More comprehensive testing will be done and those improvements that do improve the repeat sales model beyond the basic one assessed in this chapter, will become part of the model used to determine the final property index.

[14] The percentage of property transactions with no price and/or date amounts to 12%.

Chapter 5: Development of a property price index

5.1 Introduction

This chapter explains how the performance of the basic repeat sales model was further improved with a series of modifications. Section 5.2 provides an overview of each of the model improvements that were explored and in Section 5.3 the outcomes of the various evaluation tests and the improvements it contributed to the model are discussed.

With the final model defined, further enhancements to improve the operational usability of the model are introduced in Section 5.4. Lastly, in Section 5.5, the end goal of the dissertation, the development of a robust repeat sales property price index is presented.

5.2 Further improvements to the repeat sales model

It was demonstrated in Section 4.3 that the repeat sales model is the most robust model to use when predicting property prices in South Africa. Although the model was much better than the models using the measures of central tendency, there was still scope for improvement. Various remaining shortcomings of the repeat sales model were investigated by implementing a number of changes to the basic model from the previous chapter. The following versions of the improved model incorporate these changes:

1. Model farms separately;
2. Model townships separately;
3. Segmentation into significant groups;
4. Improving the model error by using weights; and
5. Reducing the effect of volatility through smoothing.

It was decided to follow an approach where these changes are implemented in a sequential manner. This implies that, for example, version 3 would add segmentation to a model for which farms and townships are already modelled separately. Since the changes consequently escalates up to the final model, only the incremental impact of each change, compared to the previous version and the basic repeat sales version is presented.

Because of resource constraints, it was necessary to limit the size of the data sets used in studying the model improvements in this chapter to a random sample of 700,000 twice-traded properties. Versions 1 to 5 of the model were tested using test criteria 2 to 9 as presented in Section 4.2 and each version was then compared to the 'original' repeat sales model of Section 4.1. The ClusterPlus groups referred to in this chapter were briefly discussed in Chapter 3 and further detail on these groups can be obtained in Appendix B.

5.2.1 Version 1: Model farms separately

Transaction data for farms can be easily identified, because it belongs to the ClusterPlus group S (special cases) with the cluster code 1S (agricultural) or 16S (small holdings)[15], or by a "~" as the third character in the erf key. There are 98,329 transactions involving farms (sold and reported in the deeds data during the period of interest) in South Africa according to the definitions above.

Farms demonstrate very different growth behaviour than full title or sectional title properties. Firstly they are not traded as often as regular residential properties and their selling prices may vary considerably from one farm to another. This is invariably due to factors other than simply just the intrinsic value of the property (e.g. the extent to which the farm had been developed, stock levels and recent crop yields) which, in turn, might influence the modelled growth of residential property prices in an area in an unrepresentative way.

By removing the data for farms from the repeat sales models, potentially biased information will be removed, allowing regular residential properties to be valued more

[15] See section 3.2.4 Cleaning the data

accurately. The farm data can be used in a specific farm-model to model farms separately, following the same approach as for the previous repeat sales models.

5.2.2 Version 2: Model townships separately

The ClusterPlus groups that have the largest volumes are groups A, B and C and the groups with the smallest volumes are groups G, I and J. There are 27,692 transactions involving groups G, I and J (see Figure 5.1 below). Groups G (township living), I (dire straits) and J (below the breadline) represent groups that most likely form part of a township. Compared to groups A, B and C, the volumes in groups G, I and J is so small that it should not make a great difference in the total behaviour of the index if groups G, I and J are excluded. Figures 5.1 and 5.2 show clear differences in volume and behaviour between the different groups.

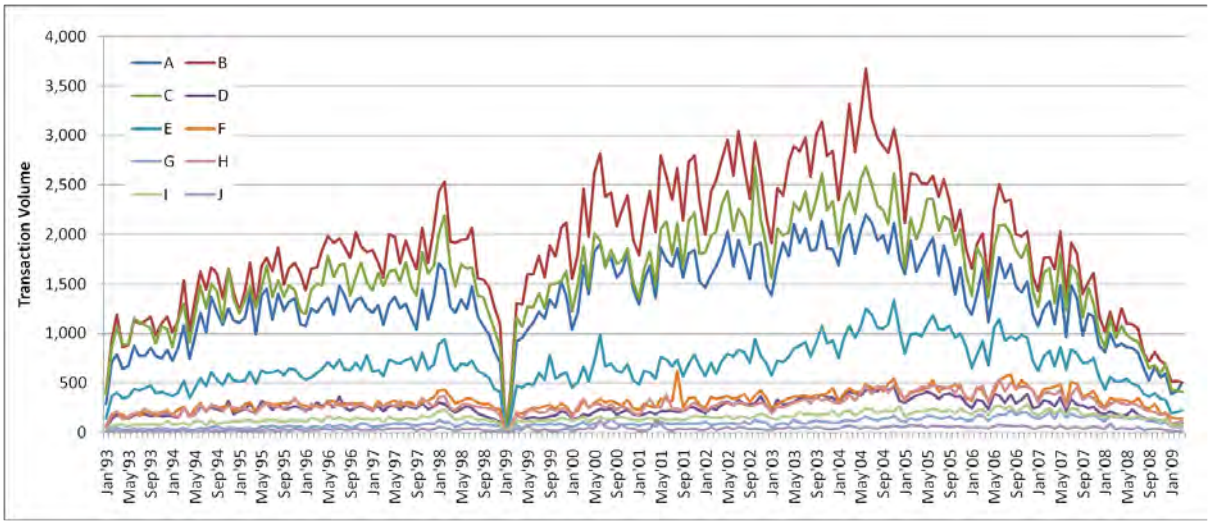


Figure 5.1: *Transaction volume by group code*

By using the median purchase prices in each month instead of the mean purchase price, the smaller and larger values will not influence the trend as much. The trend is then normalized with the natural logarithm to have a standard base with which to measure the trend. Because the volumes in groups G, I and J are so small, the model output is very erratic which can predict the rest of the results when included in a national index. Properties in townships often have a low basis purchase price due to subsidisation, but grow in price quite substantially as demand increases or once

the township have, for example, received electricity. An example of this rapid growth is illustrated in Figure 5.2.

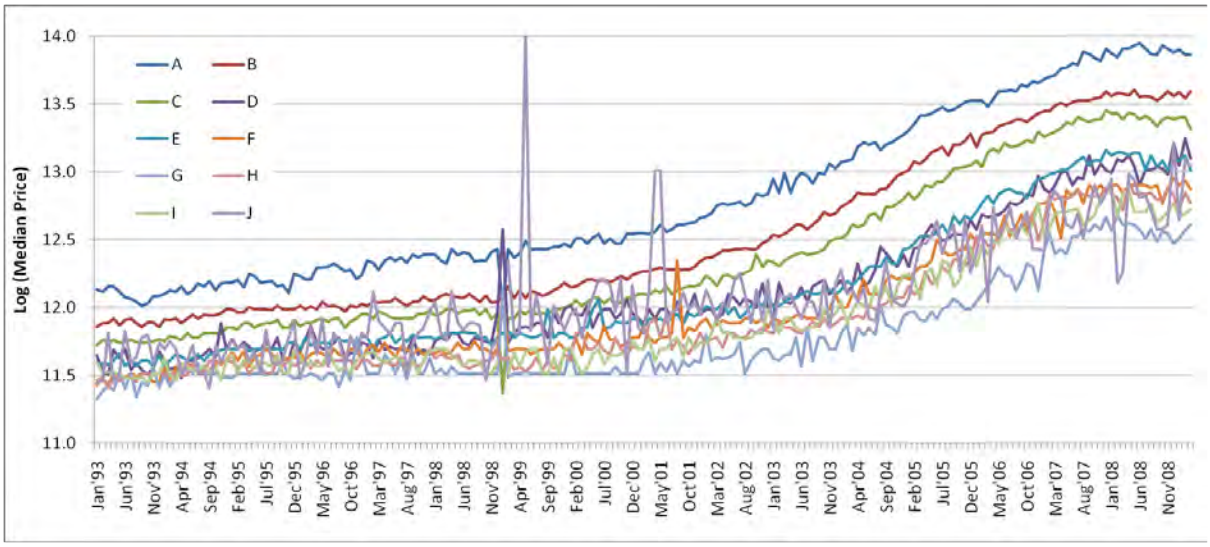


Figure 5.2: *The logarithm of the median price per group code, illustrating the rapid growth of groups G, I and J*

Since townships have different growth patterns from the majority of properties in South Africa, it was decided to exclude them from the data used to develop a national repeat sales index.

Township data can be identified with confidence and can therefore be used in its own specific township model, modelled separately using the repeat sales model[16].

5.2.3 Version 3: Segmentation into significant groups

The next improvement that was identified involved segmenting the data into statistically significant groups, based on a trade-off between the following two basic considerations:

1. Models built on homogenous segments of a population generally perform better.
2. The repeat sales model requires the occurrence of a sufficiently large number of repeat sales per market segment.

[16] Note that the detail of a township specific model is beyond the scope of this study.

As an example, segmentation by province and types of properties may improve the model due to the greater level of homogeneity of the data, but will result in a reduction in the number of repeat sales transactions upon which the index can be based. It was therefore decided to examine these trade-offs as a possible improvement. To decide how many models to build and the level of data at which the models need to be built, the data had to be analysed first.

An obvious initial split between different properties is that between *full title properties* and *sectional title properties*. The full title and sectional title properties were divided and analysed separately because of anticipated differences in their correlation behaviour. As mentioned before, the main variables used in building a property valuation model are purchase price and registration date. Thus, analysing the trends in the purchase prices each month (registration dates) for both full title and sectional title properties at different levels helped to inform a decision as to which trends display similar behaviour at the various levels and therefore could be grouped together.

The next dimension analysed as a possible segmentation dimension was provinces, because the prices for similar properties may differ significantly between provinces. Building models at a national level implies the basic assumption that properties in different provinces all experience the same growth, which is certainly not the case for South African properties. This implied that, to potentially segment by geographical area, was one of the dimensions that could be explored. The next task was to determine the level at which the estimates for the province and types of properties could still be regarded as reliable due to the reduced data volumes per model.

The last dimension that was considered for possible segmentation was socio-economic grouping, which represents the clustering of properties owned by people with similar income and density throughout South Africa, indicating the type of property in an area.

For the purposes of this study, it was decided not to utilise the next lower level of geo-demographic segmentation - the cluster level - or the suburb level even though

the data was available. By including these levels, the average number of records per unique data bin[17] would have been reduced by a factor of approximately four[18].

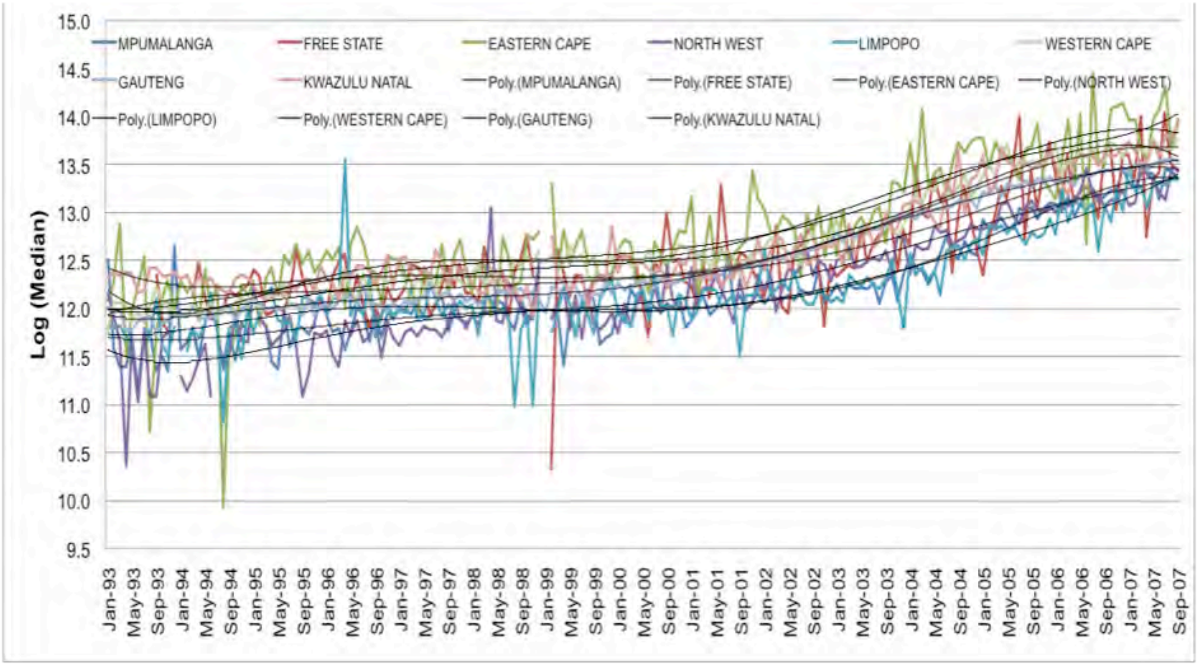


Figure 5.3: *The logarithm of median prices per province for sectional title, group A (for all provinces)*

Figure 5.3 shows the logarithm of the median purchase price per month per socio-economic group per province for group A sectional title properties. The normalised median purchase price per grouping is depicted on a single graph to enable (i) an analysis of the trends per socio-economic group in the different provinces and (ii) to determine statistically, according to the trends and the volumes in each grouping, which properties should logically be grouped together. If an unlimited amount of data had been available, properties in each socio-economic group and province would have its own model, but unfortunately the volume of data is limited when broken down to lower levels, especially when time intervals at monthly levels are used. In compiling groups, the basic assumption was made that there had to be at least thirty properties per month per province per bin to assume normality.

[17] A unique data bin is defined as a subset of transactional data from the same province, with the same title status (sectional title or full title), from the same socio-economic group and with the same median date and the same time interval between the two sales considered in the repeat sales model.

[18] The socio-economic group S (special cases) represents properties that are not generally residential properties and was therefore not considered for modelling.

It was decided to empirically explore the viability of constructing super-groups of the nine provinces by grouping the provinces as major provinces (Gauteng, Kwa-Zulu Natal and Western Cape) and minor provinces (Eastern Cape, Limpopo, Mpumalanga, North West, Northern Cape and Free State).

To illustrate a case where the volume of data is enough to model major and minor provinces separately[19], the volume of transactions for these two groups for full title transactions in socio-economic group A is presented in Figure 5.4.

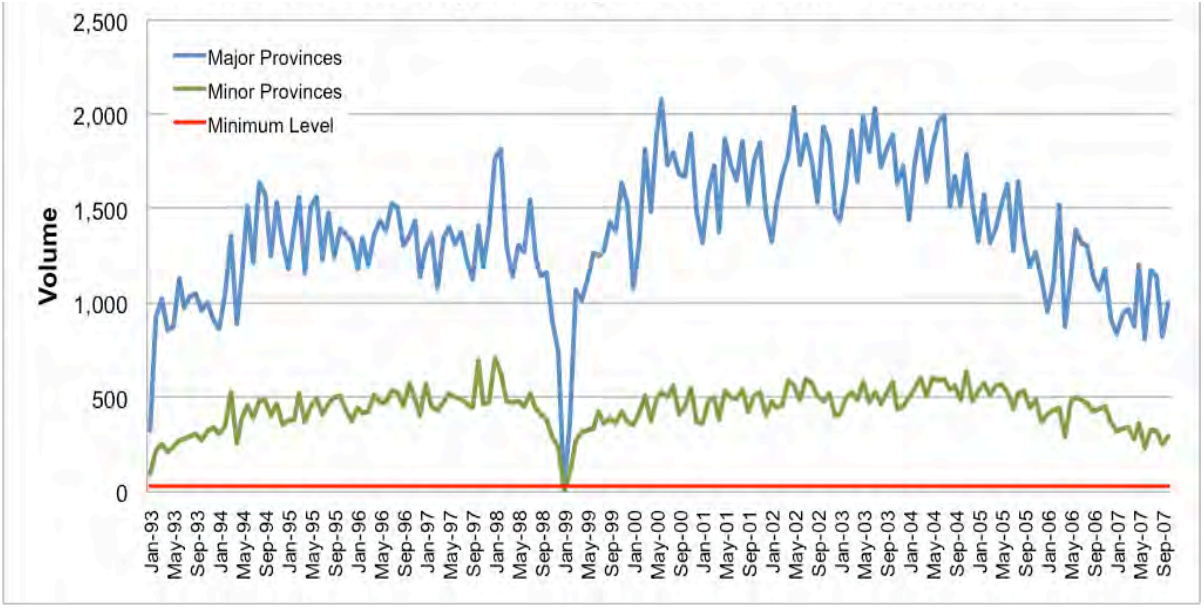


Figure 5.4: *The volume of transactions per province for full title group A*

Due to insufficient data per unique data bin and resource constraints in running the model, segmentation down to major provinces and minor provinces could not be implemented for some of the socio-economic groups. For these segments, the data for all the provinces was grouped together. There are a few cases where even the segments containing the data of all the provinces had too low volumes. For these cases, a number of the closely-related socio-economic groups were grouped together.

In Figure 5.5, the number of data sets per bin is presented for sectional title data, with group A, split into major and minor provinces. In this case the volumes for the

[19] The assumption that each unique data bin must include at least 30 data sets for statistical purposes is indicated on the following few figures as a horizontal red line.

minor provinces were not sufficient for these groups to be segmented separately, resulting in major and minor provinces being grouped together.

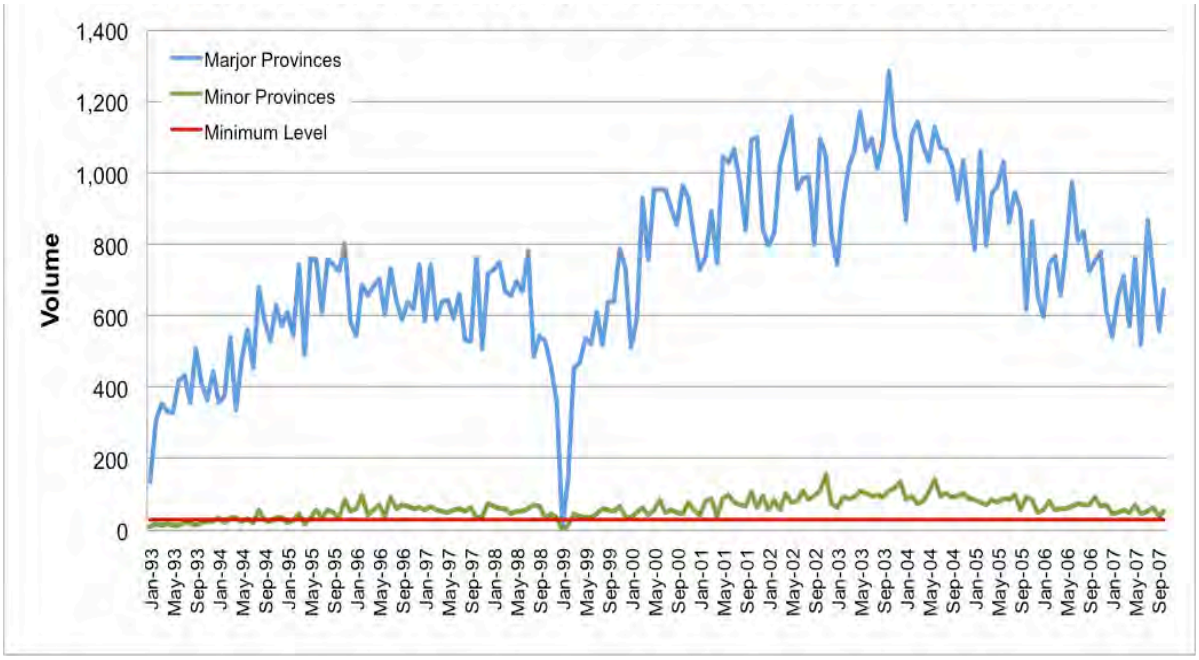


Figure 5.5: *Volume for full title group A for major and minor provinces*

In Figure 5.6 below, even though group E technically had sufficient transaction volumes to be in a group of its own, it was grouped it with group F which had insufficient volumes on its own.

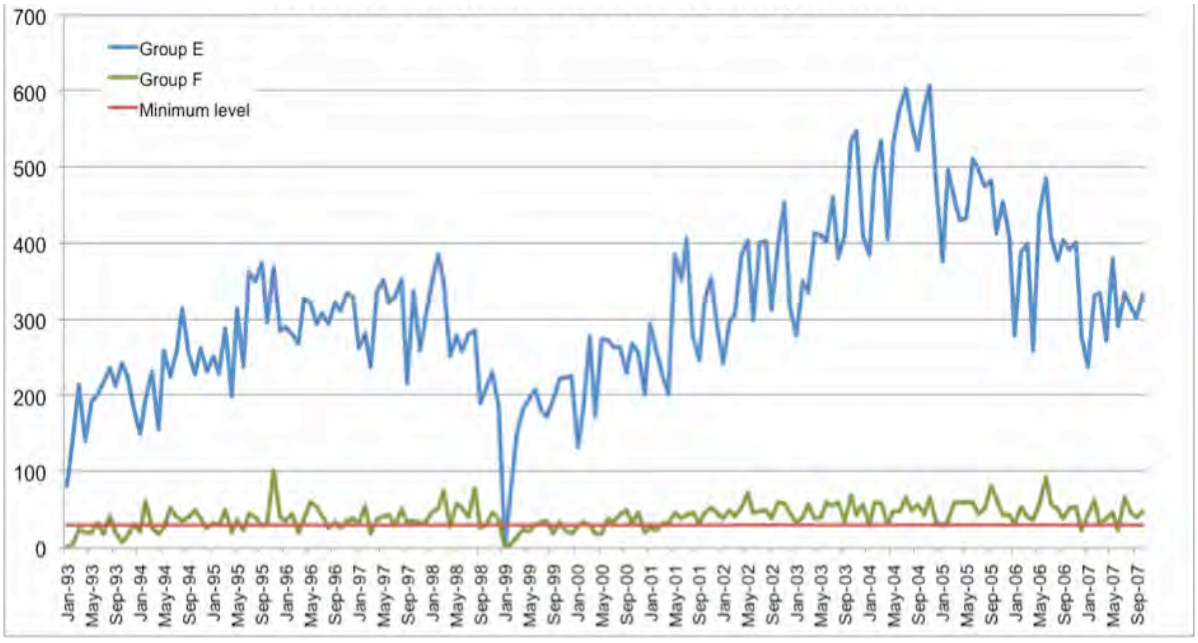


Figure 5.6: *Volume of sectional title groups E and F for all provinces*

After the volume in each segment had been reviewed, the trends of the resultant segments were analysed to assess the extent to which the data in the segments correspond (Appendix C). Figure 5.7 below illustrates the outcome of this assessment which was performed for all four super-groups discussed.

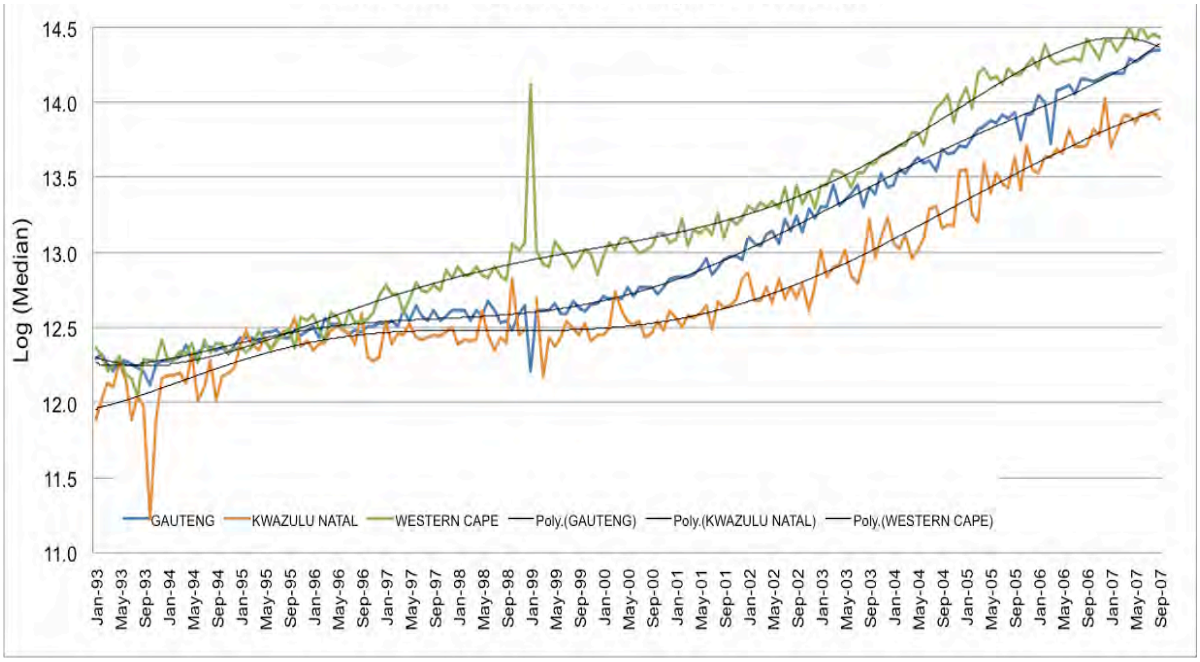


Figure 5.7: *Log (Median price) of full title, group A for major and minor provinces*

After applying the considerations illustrated above and detailed analyses, the data was finally grouped into 18 statistically chosen groups (Table 5.1), of which 12 groups are full title groups and 6 are sectional title groups. The repeat sales model parameters were derived for each of the 18 segments. Each segment was given a weight based on the data volume in each segment for each period, divided by the total volume for each period. The parameters were then weighted back using one set of model parameters.

FULL TITLE		
1	Group A	Major Provinces
2	Group A	Minor Provinces
3	Group B	Major Provinces
4	Group B	Minor Provinces
5	Group C	Major Provinces
6	Group C	Minor Provinces
7	Group D	All Provinces
8	Group E	Major Provinces
9	Group E	Minor Provinces
10	Group F	Major Provinces
11	Group F	Minor Provinces
12	Group H	All Provinces

SECTIONAL TITLE		
1	Group A	All Provinces
2	Group B	All Provinces
3	Group C	All Provinces
4	Group D	All Provinces
5	Group EF	All Provinces
6	Group H	All Provinces

Table 5.1: Statistically chosen groups of data

5.2.4 Version 4: Improving the model error by using weights

The difference obtained by comparing the actual price a property was sold for with the predicted price for the property for the same period is called the residual or error. In a similar way, by predicting the model target, there can also be a model residual or error between the predicted target and the actual target. To minimise the effect of a model error, the residual can be weighted back into the model (Case and Shiller, 1989). Case and Shiller (1989) argued that the variance in house prices widens as the period between sales increases. This could be, for example, because some houses are very well maintained whereas others are not maintained at all. As a result, the variance in the residuals (i.e. the variance between predicted and observed house prices) will increase the longer the period between sales and is linearly correlated.

Abraham and Schauman (1991) argued that the variance of the error term associated with any repeat sales pair will not increase linearly with the time between sales indefinitely. Instead, they proposed a quadratic model which states that the increase in variance would decrease as the period between sales increases. In the data used in this study, the difference between the linear curve and the quadratic

curve is insignificantly small when applied to the segmented model against time between sales.

As discussed in Section 3.2 on the SAS procedure used to model the repeat sales, the general linear model (GLM) procedure uses the method of least squares to fit a linear regression curve. For the method of least squares, it is assumed is made that the errors for different observations are assumed to be uncorrelated with identical variances. It was decided to pursue two approaches towards estimating the parameters $(\alpha_i, i = 0, 1, \ldots, t)$ in this version of the repeat sales model:

1. Ordinary least squares, which assumes that the error terms $\varepsilon_i = 1, \ldots, n$ all have the same variance, and
2. Weighted least squares, with weights equal to: $\frac{1}{t_{i+1} - t_i}, i = 1, \ldots, n$

This approach is based on the assumption that the errors follow a normal diffusion process (i.e. the errors in the unlogged model follow a lognormal diffusion process). This in turn implies that the error variance is proportional to the period between the two sales, i.e. sale prices that are four years apart will have half the weight of sale prices that are two years apart in the application of the model.

The result of this analysis proved that the second approach yields the most accurate and reliable results, since there was some evidence of the error variance increasing with time.

The residual for time between sales (in years) vary for different years. If a property takes between 0 and 2 years to sell, the price growth trend will be very different from that of a property that took 4 or 6 years to sell, as seen by the fitted trends in Figure 5.8. Also, the period between sales differ at different times, e.g. if a property took 4 years to sell in a growing property market, the trend will be very different to that of a property that took 4 years to sell in a generally depressed economic environment. Time between sales was therefore identified as suitable criteria for another level of segmentation. This was, however, not pursued, since it would be beyond the scope of this dissertation.

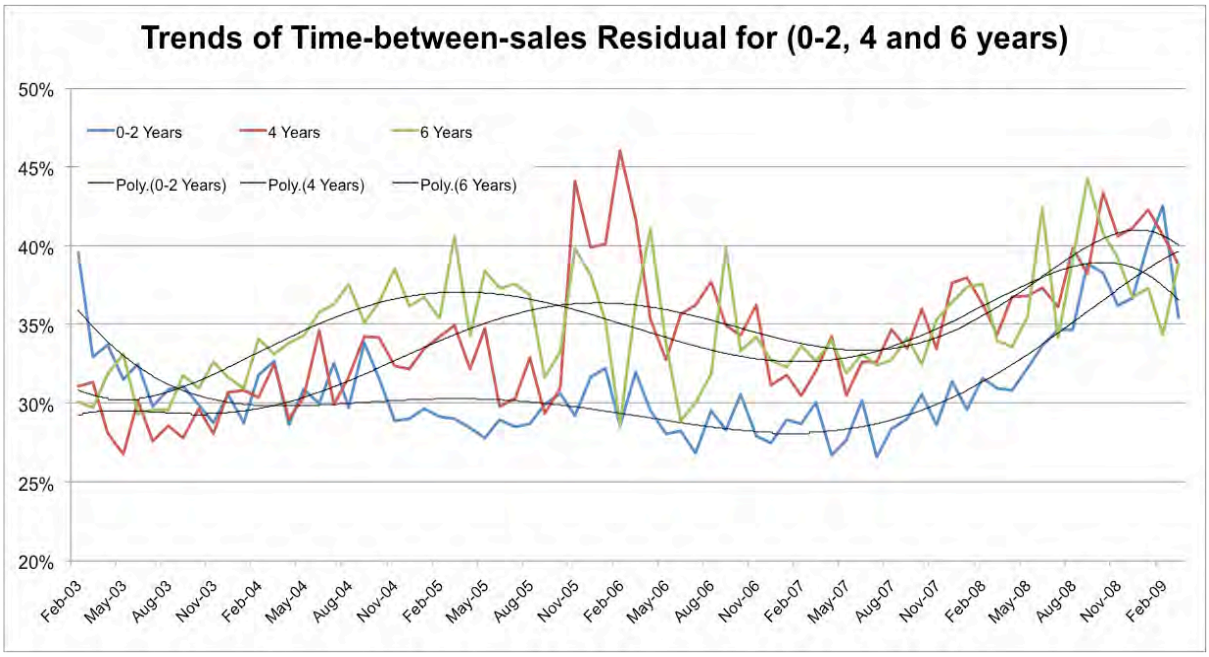


Figure 5.8: *Trends for different time-between-sales*

To add a weight in the repeat sales model, the GLM procedure in SAS includes a 'weight' statement with which to minimize a weighted residual sum of squares:

$$\sum_{i=1}^{n} w_i \left(y_i - \hat{y}_i\right)^2 \qquad (5.1)$$

where

w_i = Value of the variable specified in the 'weight' statement,

$\hat{y}_i$ = Predicted value of the response variable,

y_i = Actual value (observed value or response variable).

If the weights for the observations are proportional to the reciprocals of the error variances, then the weighted least-squares estimates are the best linear unbiased estimators (SAS, 1976).

5.2.5 Version 5: Reducing the effect of volatility through smoothing

The growth factor time series for the models above for each segment can be very volatile due to varying monthly growth. In practice, a price index should not exhibit such erratic behaviour.

A pattern of pairs of significant periodic spikes was identified in the data, occurring around December / January each year. The spikes occur in pairs of generally the same magnitude, but in opposite directions. It was postulated that this could be due to irregularities regarding the Deeds Office's operations when transactions are registered around the end of the year and during the Christmas vacation. This could however not be confirmed.

An easy way to smooth these spikes is to take the average of the two months with outliers. However, removing these pairs of spikes does not smooth all the data and further smoothing is required. One of the most common methods recommended for smoothing the time series is the Hodrick-Prescott filter (Hodrick and Prescott, 1997), see Figure 5.9.

The Hodrick-Prescott filter is a mathematical tool used in macro-economics to obtain a non-linear, smoothed representation of a time series of data. The tool is generally more sensitive to long-term than to short-term fluctuations. The extent of this bias can be adjusted by a value selected for a multiplier λ. For the purpose of this study we have chosen a value of 50 as the multiplier, based on empirical analyses. The mathematical expression of the Hodrick-Prescott filter follows:

$$\min \sum_{i=1}^{t} \left(y_i - y_i^*\right)^2 + \lambda \sum_{i=2}^{t-1} \left[\left(y_{i+1}^* - y_i^*\right) - \left(y_i^* - y_{i-1}^*\right)\right]^2 \qquad (5.2)$$

where

y_i = Actual value (observed value or response variable) in period *i,*

y_i^* = Trend output in the same period *I,*

λ = Determines the degree of smoothing of the trend, quantified on a discretionary basis. A high λ value results in trend growth moving toward a linear trend and $\lambda = 0$ means that the trend output is identical to the actual value,

i = Jan 1993 ... Apr 2009.

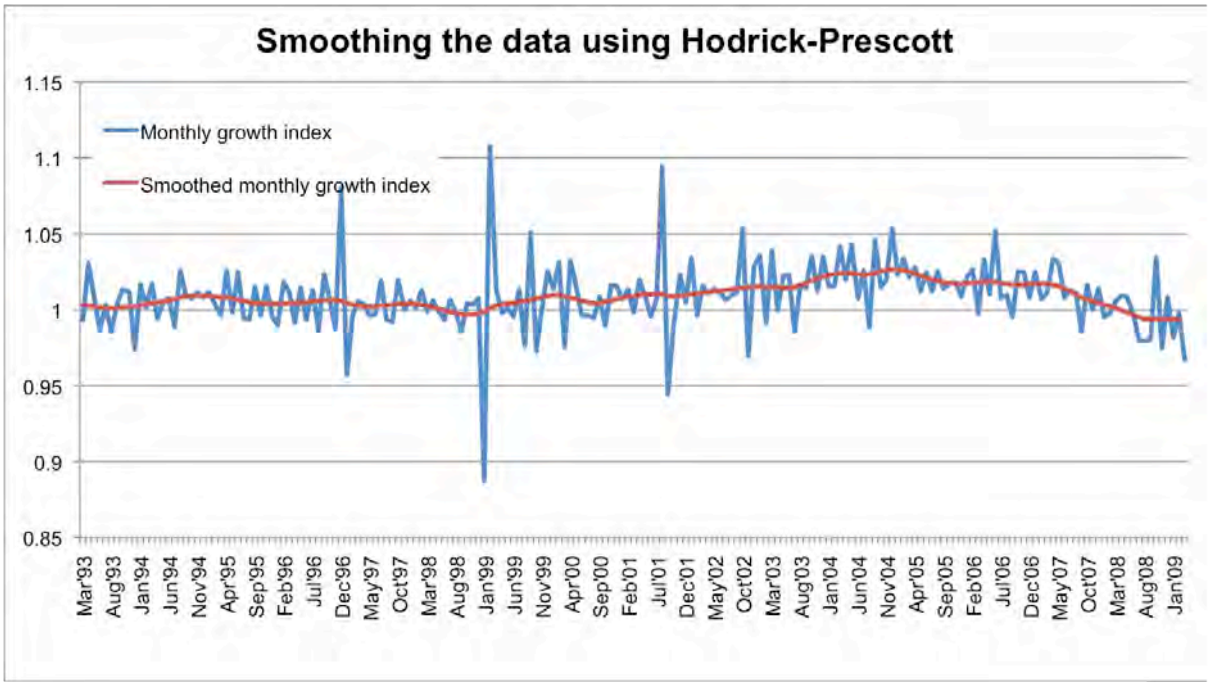


Figure 5.9: *Example of a smoothed property price index using the Hodrick-Prescott filter.*

In order to perform smoothing in SAS, the *Expand* procedure is used. The *Expand* procedure automatically converts time series data from one sampling interval or frequency to another and interpolates missing values in the time series.

5.3 Comparison of results of the improved method vs. previous methods

In this section each version of the improved model is compared to the original 'basic' model and to each other using the same test statistics as in Chapter 4. All the test statistics were applied again apart from the first one, as this statistic makes more sense when comparing independent models, not models derived sequentially as in the case of the versions of improved model changes. Some tests may carry a higher priority than others, depending on the intended function of the model, e.g. in some cases the accuracy of the model might be of greater importance than the size of the population, while in other cases the accuracy may be important, but not as important as the ability to predict the price of as many properties as possible. In this dissertation, the tests were not prioritised. This is a topic that could be investigated in

future. The same random sample of 100 000 properties was used as the one in Chapter 4 to derive the results for the test statistics by means of back-testing. In the graphs that represent these statistics, the basic repeat sales model and the final model incorporating all model improvements are represented in a darker colour than all the intermediate versions.

The eight test statistics are as follows:

1. Distribution of model errors;
2. Theil's U-statistic;
3. Mean error (ME);
4. Mean squared error (MSE);
5. Root mean squared error (RMSE);
6. Mean absolute error (MAE);
7. Mean prediction error (MPE); and
8. Mean absolute prediction error (MAPE).

5.3.1 Statistic 1: Distribution of model errors

The model residual was determined by using the formula 4.1 and then statistic 1 was calculated to determine the number (and percentage) of properties falling within 20% of the actual price (below or above) for each model. The 'basic' repeat sales model predicted 51.84% of the properties to be within 20% of their actual price. The model with farms excluded predicted 51.70%, where townships and farms were excluded the prediction was 51.92%, while the segmented model predicted 53.24%, the weighted model predicted 57.00% and the final smoothed model predicted 59.90% of the properties within 20% of their actual price. Applying all these improvements therefore improved the basic repeat sales model by 15.5%. The results for the first statistic are presented in Figure 5.10.

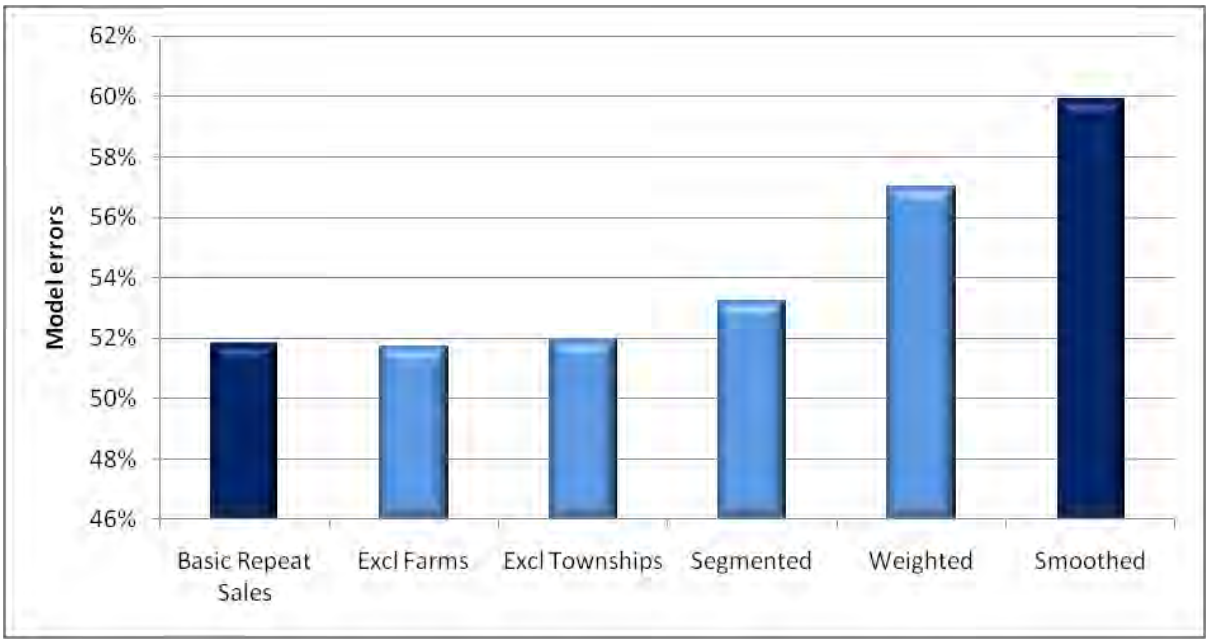


Figure 5.10: *Improvements in the performance of the model according to Statistic 1: Distribution of model errors*

5.3.2 Statistic 2: Theil's U-statistic

As mentioned before (Section 4.3), Theil's U-statistic is used to measure the quality of fit between actual and predicted property prices. Theil's U-statistic is bounded between 0 and 1, with values closer to 0 indicating greater prediction accuracy - the closer Theil's U-statistic is to 1, the closer the prediction error's variance is to the variance of the actual prices. Formula 4.2 was used to calculate Theil's U-statistic for versions of the improved model.

Theil's U-statistic for the 'basic' repeat sales model, the model change where farms were excluded, where townships and farms were excluded and the segmented model was 0.24. For the weighted model, Theil's U-statistic was 0.23 and for the final smoothed model it was 0.22. The model change that was the closest to 0 using Theil's U-statistic is the final smoothed model as can be seen in Figure 5.11.

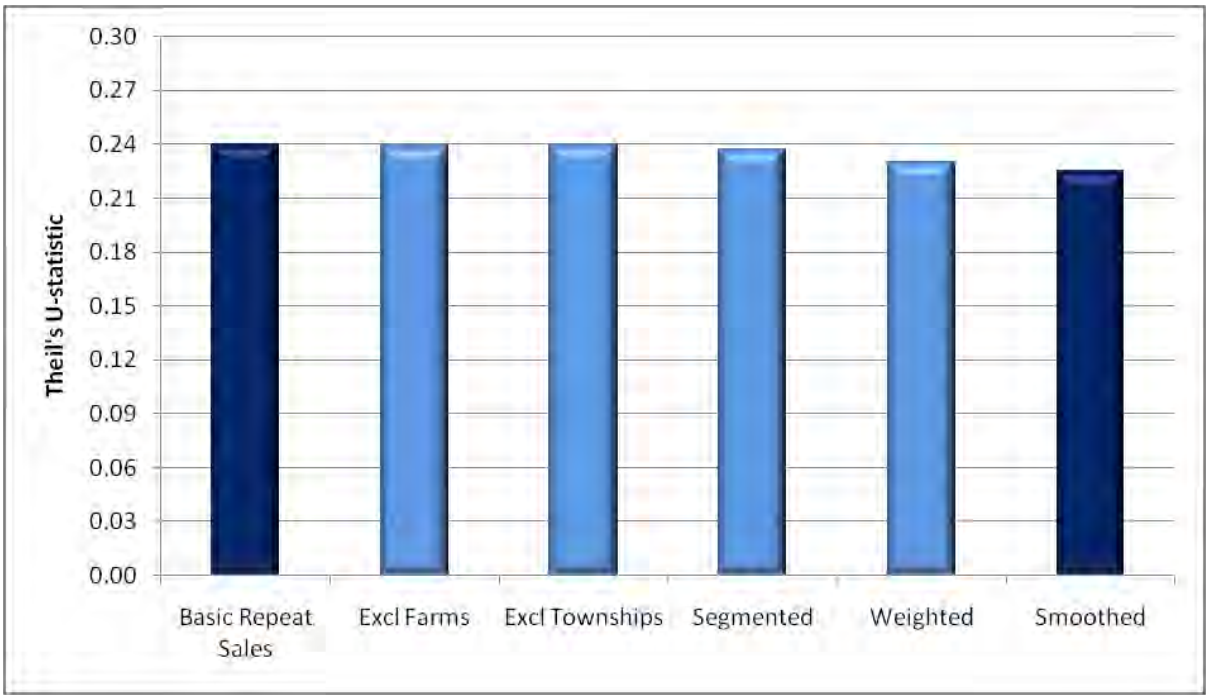


Figure 5.11: *Theil's U-statistic*

5.3.4 Statistic 3: Mean error (ME)

In statistic 3, the mean error (ME) measures the mean residual of all predictions when compared to the actual values. The results for the ME statistic is obtained using formula 4.3. The lower the value, the lower the error. However a low ME value may conceal prediction inaccuracy due to the offsetting effect of large positive and negative forecast errors. However, despite the unbiasedness of the predictions, their inaccuracies become apparent when inspecting subsequent prediction evaluation statistics.

The ME for the 'basic' repeat sales model was -R13 339. For the model change where farms were excluded the ME was -R11 336. Where townships were excluded the ME was -R15 400. For the segmented and weighted models the ME was -R6 066; and the final smoothed model's ME was -R4 685. The results for the mean error are presented in Figure 5.12.

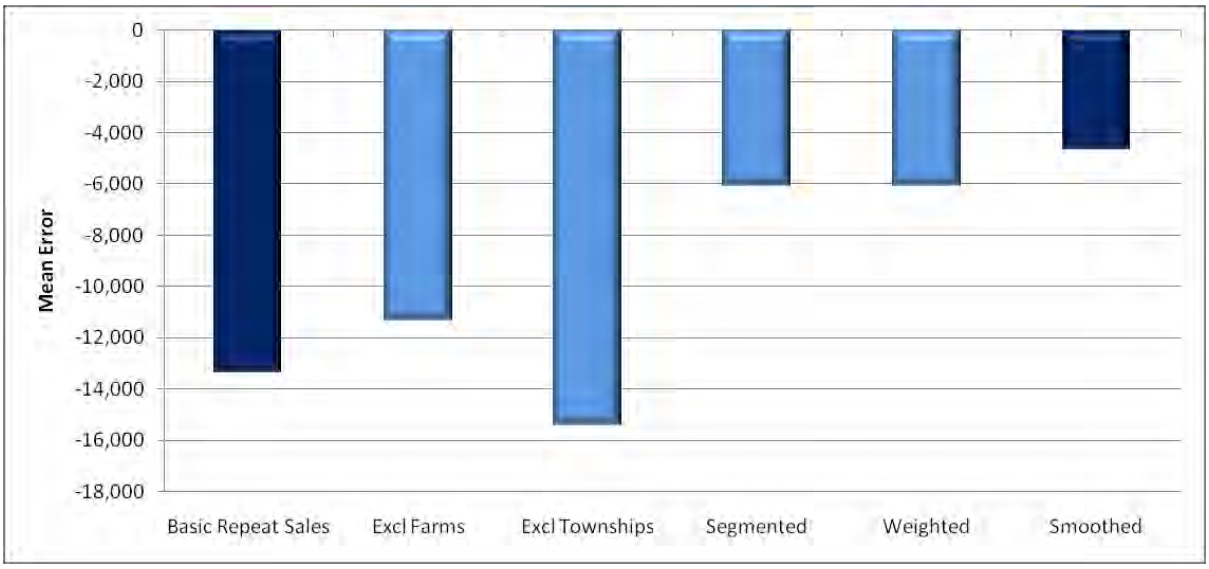


Figure 5.12: *The results of Statistic 3: Mean error (ME)*

5.3.5 Statistic 4: Mean square error (MSE)

In statistic 4, the mean square error (MSE) was calculated using formula 4.4. The results for the mean square error are presented in Figure 5.13 (the lower the value, the lower the error). The final smoothed model has the lowest MSE.

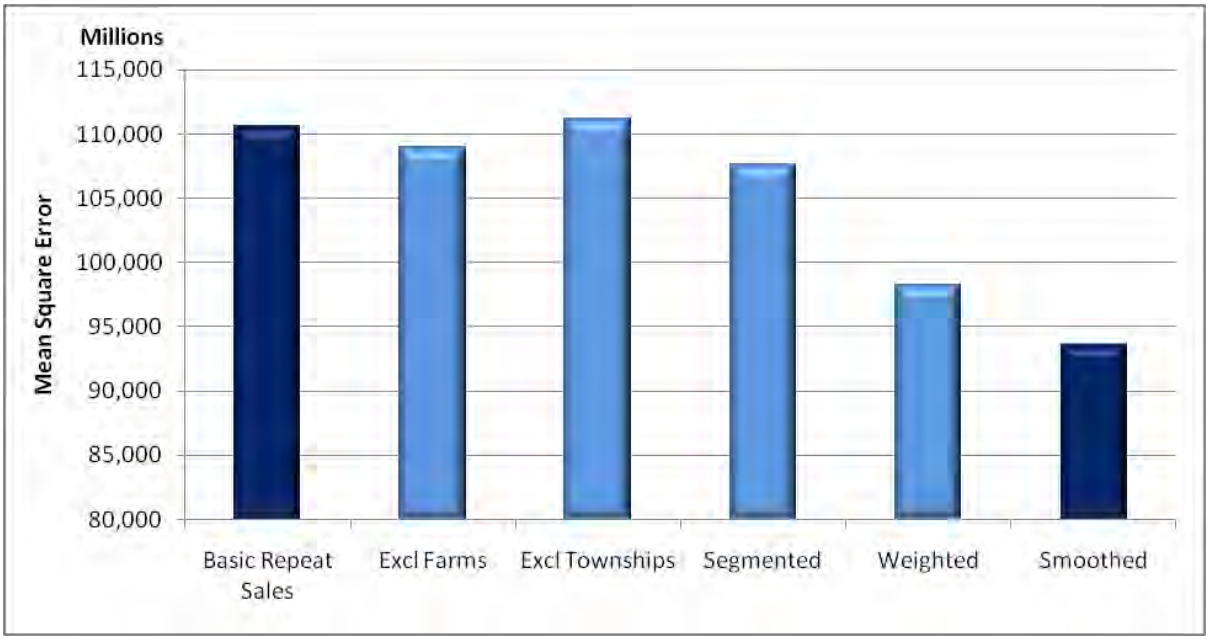


Figure 5.13: *The results of Statistic 4: Mean square error (MSE)*

The MSE (depicted in millions) for the ‘basic’ repeat sales model was R110 602. For the model change where farms were excluded the MSE was R109 010 and where townships and farms were excluded the MSE was R111 212. For the segmented model the MSE was R107 646. For the weighted models the MSE was R98 220 and the final smoothed model’s MSE was R93 681, or 84.7% that of the basic model.

5.3.5 Statistic 5: Root mean squared error (RMSE)

The fifth statistic, the root mean squared error (RMSE), is a good measure of precision and is used to measure the differences between values predicted by a model and the observed values. Formula 4.5 is used to derive the results for RMSE and the results are presented in Figure 5.14. The final smoothed model has the lowest RMSE of all the model changes, namely R306 073, an average improvement of 7.97% over the original ‘basic’ repeat sales model.

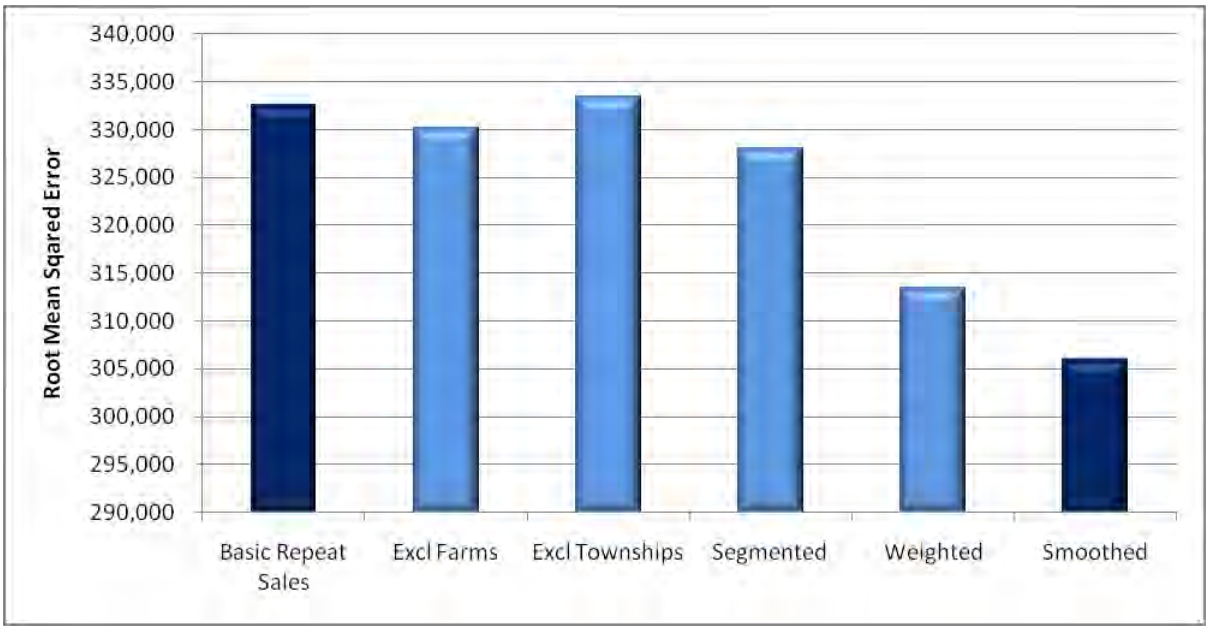


Figure 5.14: *Evaluation results for Statistic 5: Root mean squared error*

5.3.6 Statistic 6: Mean absolute error (MAE)

The sixth statistic, the mean absolute error, is, as mentioned in Section 4.3, the average of the absolute error. The formula used to calculate the MAE is formula 4.6. The results for the MAE are presented in Figure 5.15 (the lower the value, the lower the error). The final smoothed model has by far the lowest MAE (R123 034) of all the model changes.

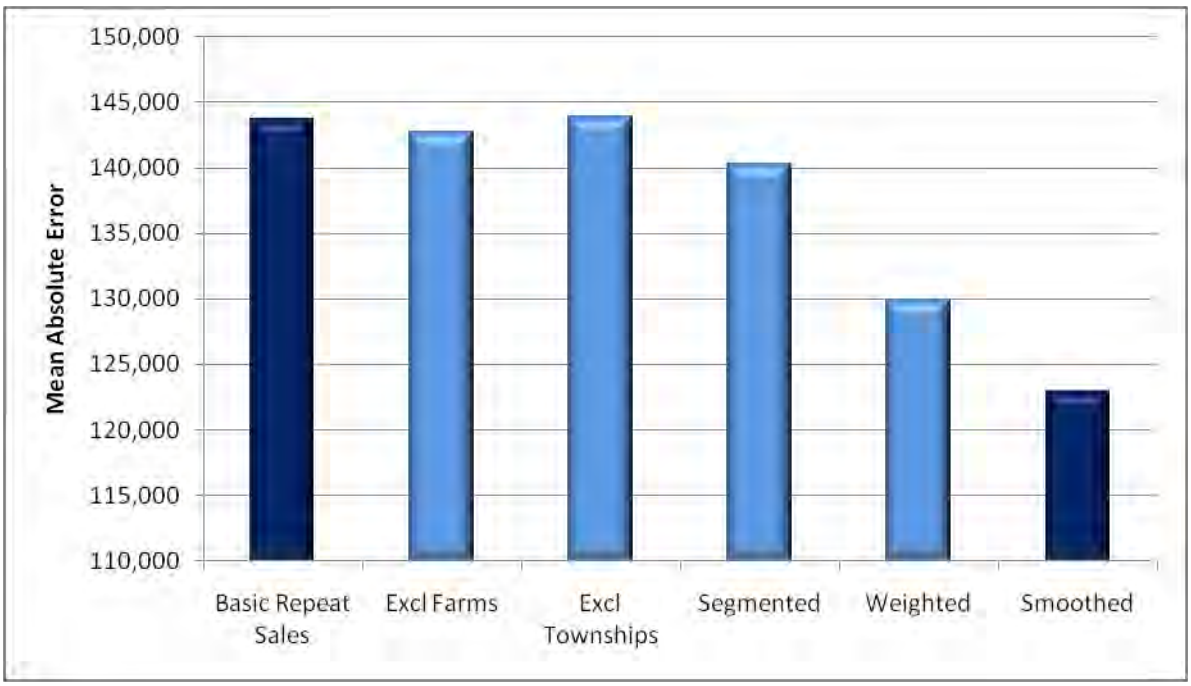


Figure 5.15: *Evaluation results for Statistic 6: Mean absolute error*

5.3.7 Statistic 7: Mean prediction error (MPE)

The seventh statistic, the mean prediction error, is calculated using formula 4.7. The results for the MPE are presented in Figure 5.16. Note that the closer the percentage is to 0, the lower the error. The final smoothed model's MPE is the closest to 0 of all the model changes, with a value of -6.21%.

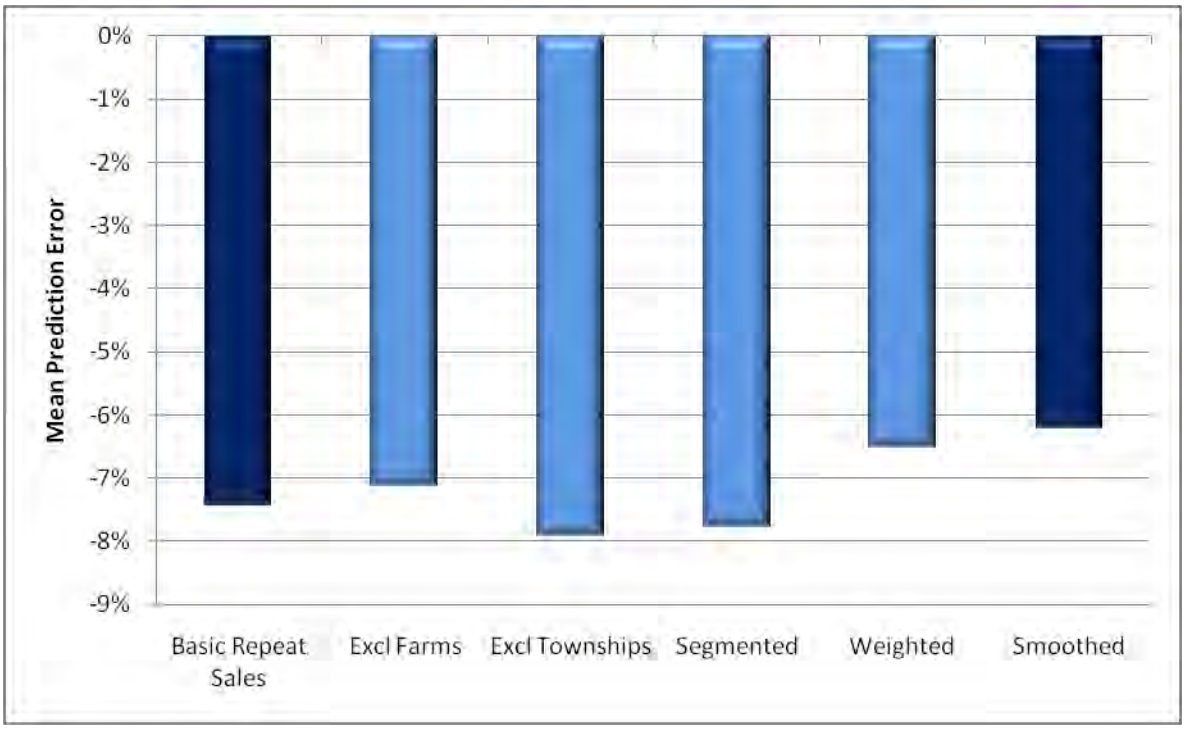


Figure 5.16: *Evaluation results for Statistic 7: Mean prediction error*

5.3.8 Statistic 8: Mean absolute prediction error (MAPE)

In statistic eight, the mean absolute prediction error (MAPE) measures the mean absolute error of all predictions when compared to the observed values (assessments). The formula used to calculate the mean absolute prediction error is formula 4.8. The results for the mean absolute prediction error statistic are presented in Figure 5.17. Note that the lower the percentage, the lower the error. There is a clear downward trend towards 0 as the model changes are consecutively applied which results in the lowest percentage, or error, for the smoothed repeat sales model with a value of 28.45%.

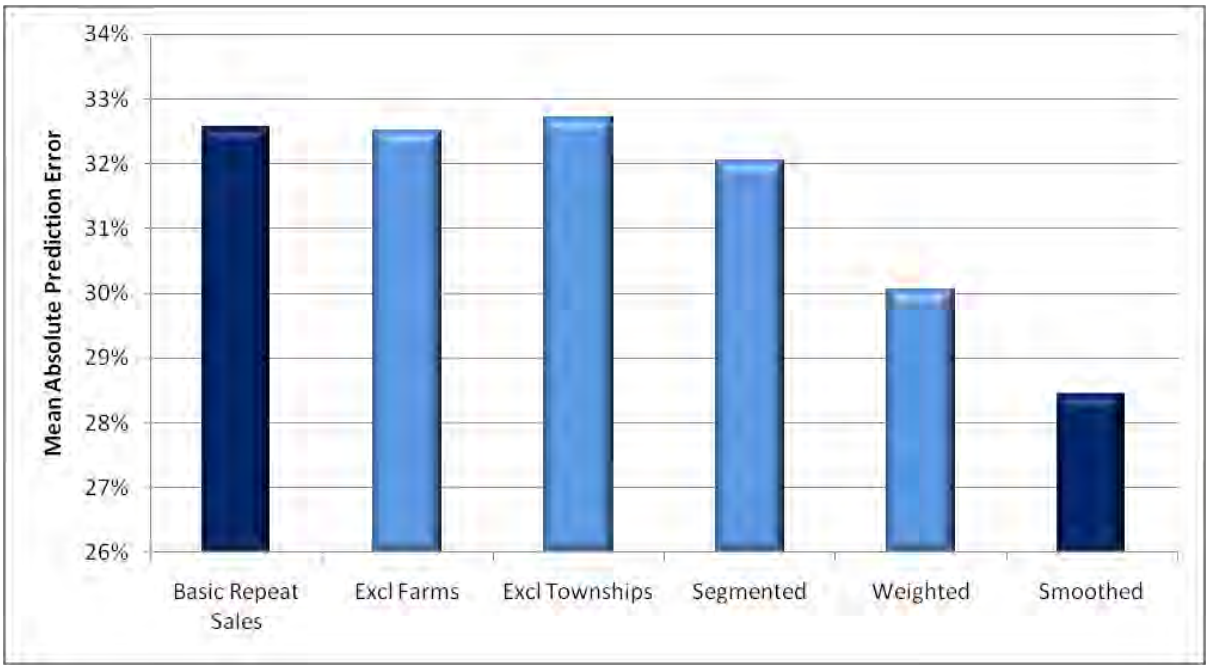


Figure 5.17: *The results of Statistic 8: Mean absolute prediction error (MAPE)*

Summary

In summary, the variance in the test results between the 'basic' repeat sales model, introduced in Chapter 4, and the 5 variants, was significant especially with that of the final smoothed repeat sales model. All the graphs follow a trend that indicates improved accuracy as the model changes were implemented sequentially. The smoothed model has shown better results than all the versions for each of the test statistics, making it the best model to use for deriving the property price index.

5.4 Deriving the property price index

In this section the derivation of a property price index is discussed, based on the proven best model from the previous section. The derived property price index will then be compared to other financial institutions' indices by correlating each index to prime interest rate. The repeat sales model has been proven to be the best model with which to valuate properties and several model changes have been added to it to improve the model even further.

The final model to be used for developing the property price index is based on the repeat sales methodology, but with farms, special cases and townships excluded from the data. Thus with the model have been segmented and weighted back using reciprocals of the volume in each segment and, finally, been weighted against the time-between-sales and smoothed.

To derive a property price index from a valuation model, the following steps need to be followed (see the illustration of the procedure in Figure 5.18):

- *Step 1: Cumulative parameters* $(\alpha_i, i = 1,\ldots,t)$- by fitting the ordinary least squares regression to the data, cumulative parameters (growth factors) are obtained.

- *Step 2: Smoothing* - one of the steps for the improved model changes was to smooth these parameters.

- *Step 3: Monthly growth factors* $(\alpha_{i+1} - \alpha_i, i = 1,\ldots,t)$ - these growth factors are calculated from the smoothed cumulative parameters:

- *Step 4: Exponent of monthly growth factors* $(e^{\alpha_{i+1}-\alpha_i}, i = 1,\ldots,t)$ - the exponent of the monthly growth factor is taken to correct the logarithm used for the target.

- *Step 5: Choose a base period* - A base period needs to be chosen for the index which is set to 100. Most financial institutions choose a base period of Jan 2000, thus Jan 2000 was used for comparison purposes.

- *Step 6: Calculate the property price* index in terms of *basis points* - If each month's basis points are known as M_i for the i-th period and the monthly growth factor is G_i, then the following month's basis point is calculated as $M_i = M_{i-1}e^{G_i}$.

- *Step 7: Calculate the property price index in terms of year-on-year growth (for the purpose of interpretation the abbreviation used for property price index in this study will be HPI (House Price Index), not PPI, as this may be confused*

with other existing indices in the market) – the basis point for the i-th period is divided by the basis point for the $(i-12)$-th period: $HPI = \frac{M_i}{M_{i-12}}$

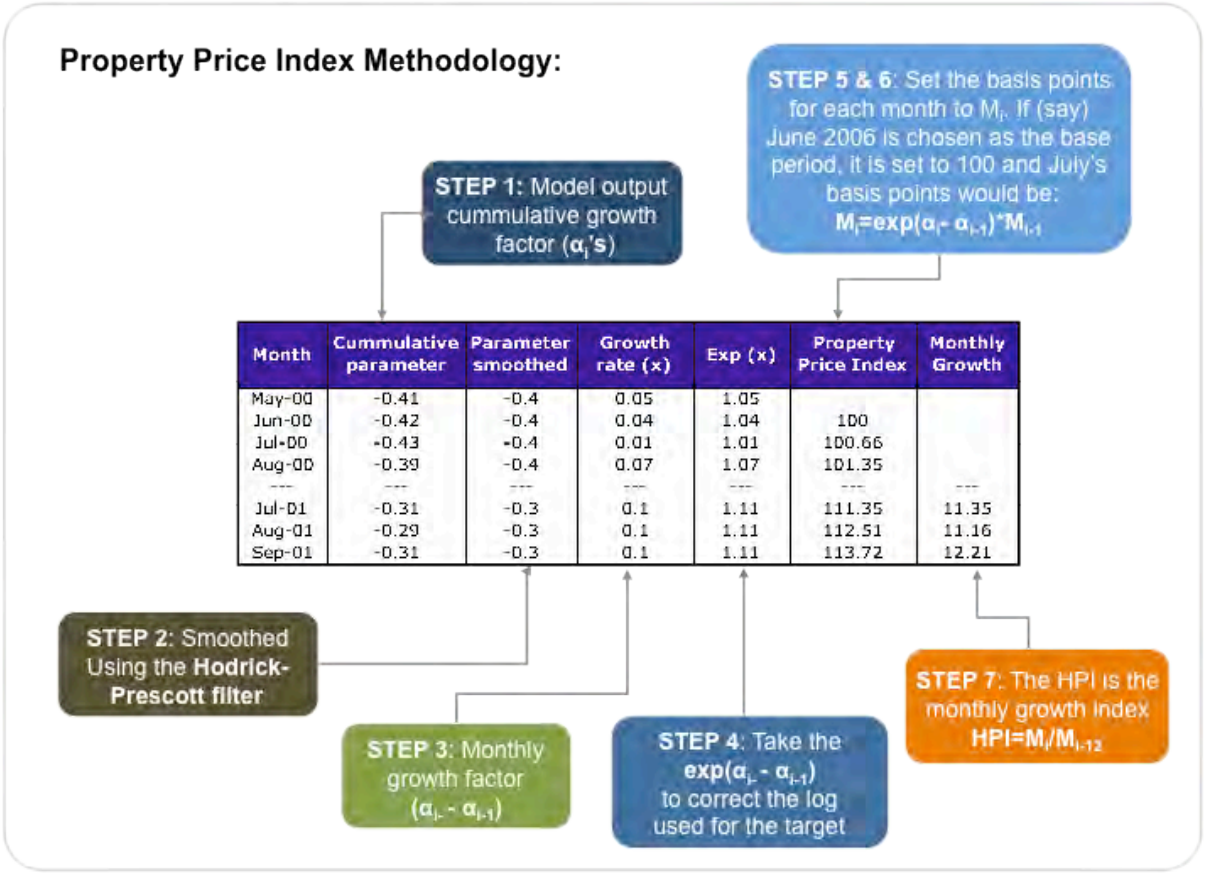


Figure 5.18: *Calculation of the Property Price Index*

In summary, the formula to derive a property price index from the smoothed cumulative model parameters is:

HPI – Basis points: $M_i = M_{i-1} e^{\alpha_{i+1}-\alpha_i}$ (5.3)

HPI – Year-on-year growth: $HPI = \frac{M_i e^{\alpha_{i+1}-\alpha_i}}{M_{i-12}}$ (5.4)

where

α_i = Smoothed cumulative parameters for period i,

M_i = Monthly basis points for period i,

i = Jan 1993 ... Apr 2009.

The property price index based on the basis points is presented in Figure 5.19 and the property price index based on the year-on-year growth is presented in Figure 5.20. From the year-on-year growth in Figure 5.20 and from the basis points in Figure 5.19 it can be seen that before the economic crises there was a strong increase in growth of property prices which turned to negative growth by the end of 2008 when the global economic crises started.

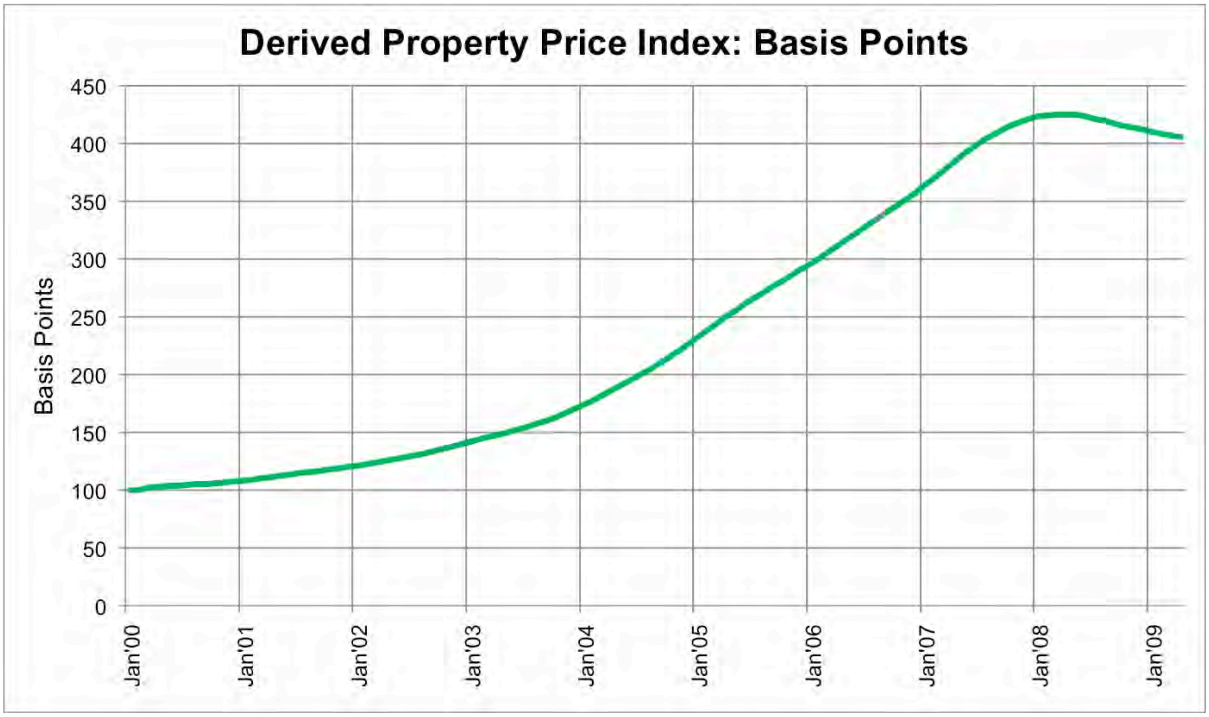


Figure 5.19: *Derived property price index – basis points*

Figure 5.20: *Derived property price index – year-on-year growth*

As a final validation, the new property price index resulting from this study was compared one-on-one with the indices of three major financial institutions in South Africa, as well as that of Lightstone. All four entities publish property price indices on a regular basis. The new property price index derived from this study (indicated by "HPI") and the property price index from Lightstone both use the repeat sales methodology for calculating their indices (Lightstone), implying that the resultant graphs should follow a similar pattern. The two indices do indeed follow one another over the whole period (see Figure 5.21, green line: HPI, yellow line: Lightstone).

ABSA uses average house prices to derive their property price index and it is based on only the transacted properties in the ABSA book. The FNB index is similarly constructed, using the average value of housing transactions financed by FNB (FNB). Standard Bank's index is based on the median house prices of the transacted properties in their loan book. Because of their significant market share of 27.7% and the lag in the Deeds data, which limits other indices based on Deeds data, Standard Bank argues that their index is a good proxy for the national market (Standard Bank). The following figure depicts the published year-on-year growth curves of the major financial institutions in South Africa, together with that of the model developed in this study, in green (see Figure 5.21).

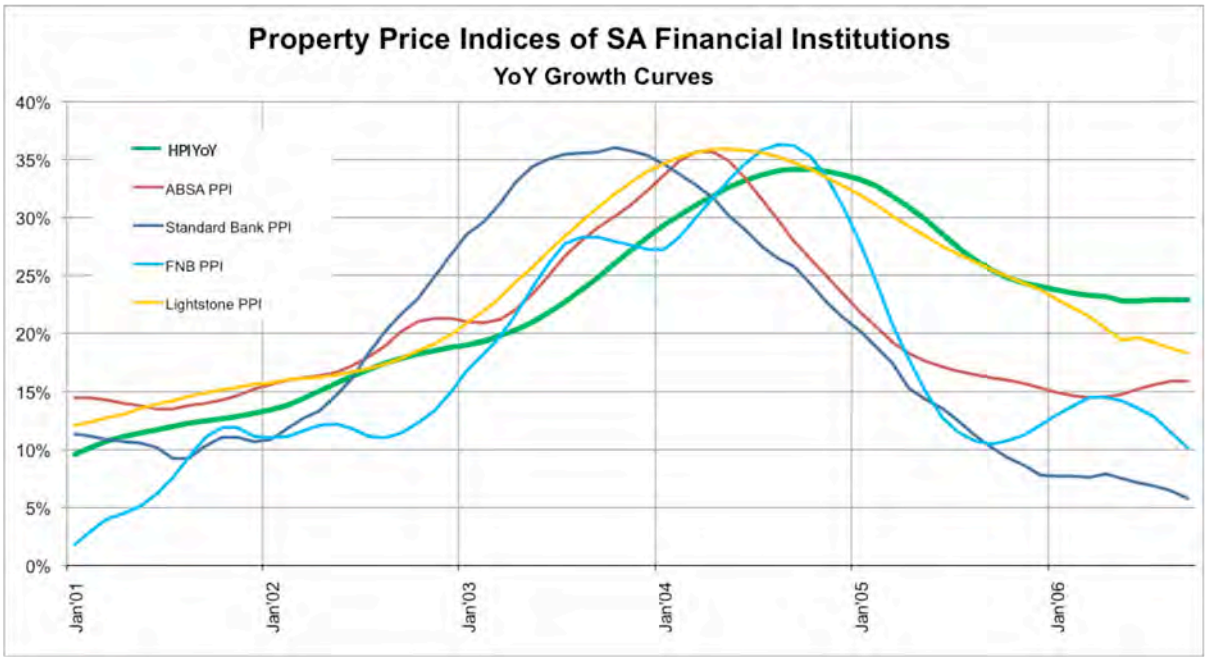


Figure 5.21: *Year-on-year growth of property price indices of SA Banks*

Property price indices normally have a negative correlation with the prime interest rate. If the prime interest rate increases, people's debt-to-income ratio increases and the demand for properties generally decrease, resulting in a slowdown in the growth of the price of properties, whereas if the prime interest rate decreases the opposite is true.

In Table 5.2 below, the correlations of the indices of the same four entities, as well as the HPI index with the prime interest rate, are compared. All indices are published monthly. Lightstone's index and the HPI has a data lag of at least four months, as both are derived from the Deeds data and the three banks' indices have a lag of probably less than one week, as their indices are derived from their own internal data. ABSA's index has the most history, dating back to January 1980; the HPI is next dating back to January 1993; and Lightstone, FNB and Standard Bank only publish indices dating back to January 2000.

Comparison Indices	Correlations to Prime Interest Rate	Frequency	Data Lag
Standard Bank	-0.1516	Monthly	< 1 week
ABSA	-0.3551	Monthly	< 1 week
FNB	-0.5071	Monthly	< 1 week
Lightstone	-0.7254	Monthly	4 months
HPI	**-0.7458**	**Monthly**	**4 months**

Table 5.2: *Comparison of South African property price indices*

The smoothed repeat sales model, with its improvements developed in this study, has the best (negative) correlation with prime interest rate, followed by the Lightstone repeat sales model. The correlations of the other three models, used by three major South African financial institutions, are significantly less.

5.5 Conclusion

In this chapter the repeat sales model was further improved by implementing five different model variants. All the consecutive changes improved the 'basic' repeat sales model to a final smoothed repeat sales model in which farms, special cases and townships were excluded, weights were applied, the models were segmented and weighted back and finally the outliers and volatility in the parameters were smoothed, resulting in the best model with which to predict property prices.

The derivation of the property price index was discussed and a final validation was done by comparing the property price index from this study with other property price indices in the market.

Current and future applications for the repeat sales model are discussed in the next chapter and further operational enhancements are presented.

Chapter 6: Current and future applications and enhancements

6.1 Introduction

A range of property valuation components and/or tools are already implemented by some of the major financial institutions in South Africa. This is evident from the fact that three of the four major banks, as well as a number of other financial institutions regularly publish their own property indices.

This chapter proposes a number of potential business benefits that the financial institutions could derive from having a quality property valuation model. The benefits are in most cases relatively obvious and will only require integration with each institution's business architecture and approach before implementation, while the implementation of the last example presented in this chapter, will require a bolder change in the institution's business strategy, but with the potential reward of becoming the market leader in this field. This chapter provides insight into the various contributions and applications that this valuation model can contribute to a financial institution above and beyond the derivation of the property price index.

6.2 Applications of valuation models in the property domain

Having a robust property valuation model could offer a number of opportunities to financial institutions which might provide them with a leading edge in the property market. For example:

- **Cost-effective, real-time desktop valuations** - Provided that the methods used to estimate the current market price of a property are statistically accurate and that the property valuation model (PVM) is validated, the ability to value any individual property for the purpose of home loans approval provides a financial institution with a useful starting point in their credit approval process;

- **Accurate revaluations of an institution's home loans book** - Having a suitable PVM could enable financial institutions to (i) perform accurate revaluations of their home loans book, to (ii) analyse the current and historical Loan-to-Value[20] (LTV) of home loans lending and to (iii) forecast future property prices in order to assess the effect of interest rate movements on property prices;

- **Property investor tool** - Investors who invest in the property market require reliable and current estimates of the value of properties which they may consider purchasing. The use of a PVM could serve as a value-added service that a financial institution could offer to selected clients;

- **Provider of private client information -** By interrogating Deeds Office data and utilising the output of a valuation model, the tool can enable financial institutions to ascertain the current market value of property portfolios owned by high-value individuals; and

- **Lead generation -** Once a financial institution has identified clients with large equity values in their property portfolios, the institution has an opportunity to offer the client products with which to release this equity for their use to investment elsewhere as a sell-on opportunity.

Several further uses of a robust property valuation model within the domain of a financial institution with interest in the property sector, are depicted in Figure 6.1 below and discussed in the following sections.

[20] LTV = Outstanding balance on a loan divided by the valuation of the property

Figure 6.1: *Further uses of a robust property valuation model*

6.2.1 Property portal

A property portal can serve as a value-add to affluent clients who are granted access to the data through a web interface (see Figure 6.2 for an example of a conceptual property portal). The information that is presented on a portal may be selected to match the needs of the audience. The audience of a property portal is expected to be investors, people considering purchasing or selling property, property developers and builders.

The example below contains a few categories of information that could be of interest to the audience above:

- *Hot suburbs this month* - Showcasing suburbs with the highest growth indices;

- *Economic information* – Information that could potentially influence the institution's Property Price Index or the growth of the property market in general;
- *Property news* – New developments in an area or crime statistics, etc. This category could also link to other relevant websites such as a local news website;
- *Latest property developments* – This category could be tailored to specifically interest developers, builders and investors; and
- *General* - The property portal could also include trends, such as interest rate information, tools for comparing asset performance, a bond calculator and a watch list.

Finally, the property valuation model is expected to play an important role in the property portal's property search function, enabling clients to search for a property's predicted current or future prices by using the PVM as a tool in the calculations.

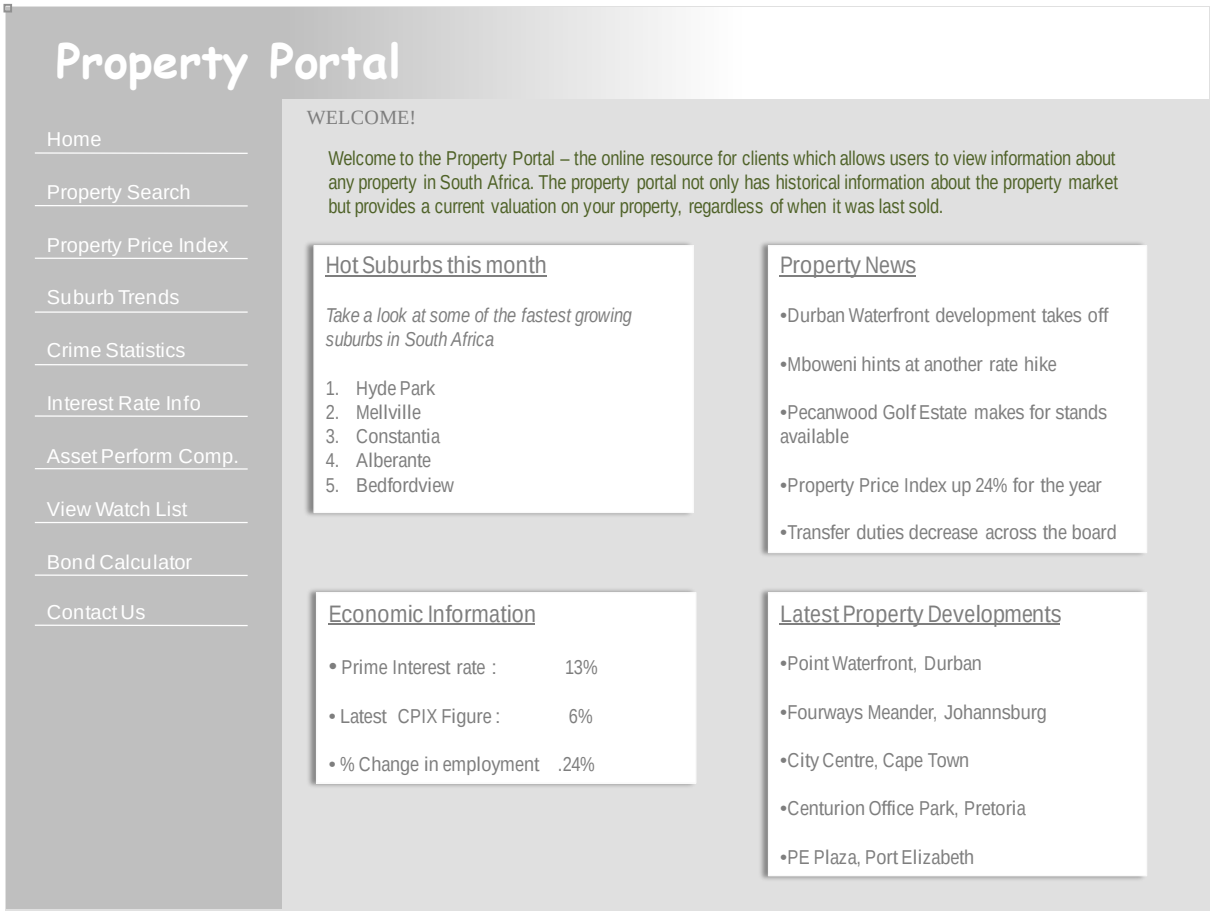


Figure 6.2: *Example of a conceptual property portal*

6.2.2 Bond pay down estimation model

A bond pay down estimation model can be developed, based on a property valuation model to evaluate the pay down curve of home loans in the South African home loans market, in order to determine the available equity.

The model can be constructed by first calculating the pay down curve of the financial institution's home loan portfolio and then calculating the equity available as the difference between the predicted balance of a specific home loan and the predicted value of the property to which the home loan applies.

6.2.3 Switch propensity model

The switch propensity model is a tool that helps financial institutions functioning in the property sector to predict the switching probability of their client base and to use this information as an input into focused marketing campaigns.

From the Deeds data one can determine with which financial institution specific bonds are registered. By looking at the switching patterns of properties between different financial institutions, regression models can be developed with the valuation amount of a property and/or the social grouping of the property as key variables to determine the switch rate curve, based on bond tenure. Being able to predict the probability of a bondholder switching his or her bond to another institution is very useful in focused marketing campaigns. See Figure 6.3 below for an example of a switch rate curve. Note that the curve follows a Weibull.

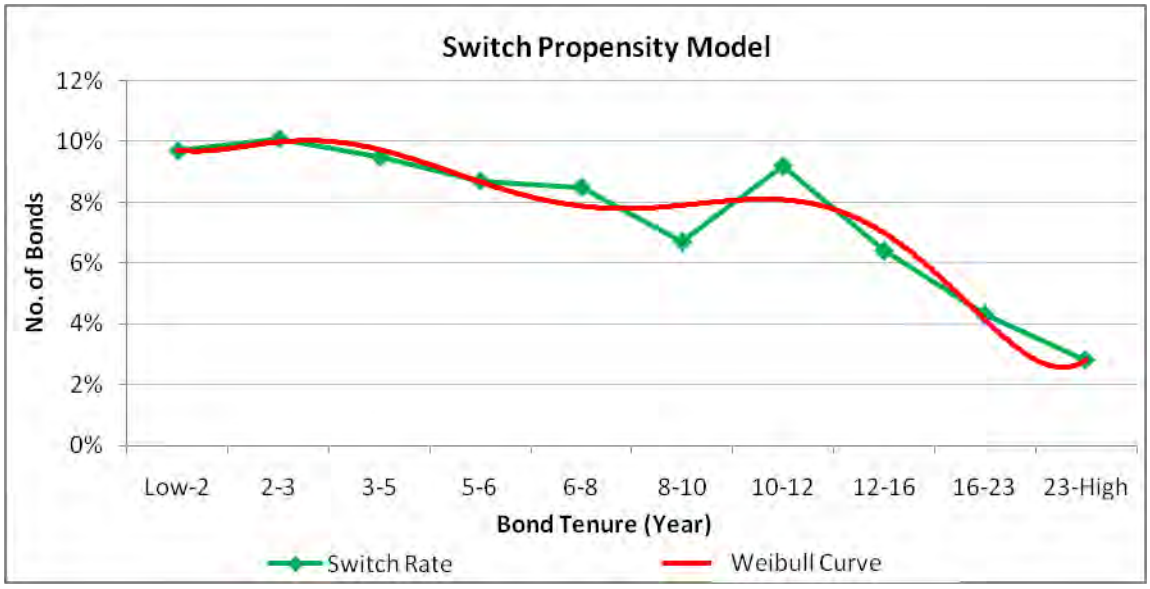


Figure 6.3: *Example of a typical switch rate curve*

6.2.4 Home loan equity release model

A property valuation model also enables a financial institution to manage the equity in a home loan. By deriving a bond pay down model and a switching propensity model based on a robust PVM, a strategy can be developed to manage the equity in

a home loan. A schematic diagram depicting a home loan equity release model is depicted in Figure 6.4 below.

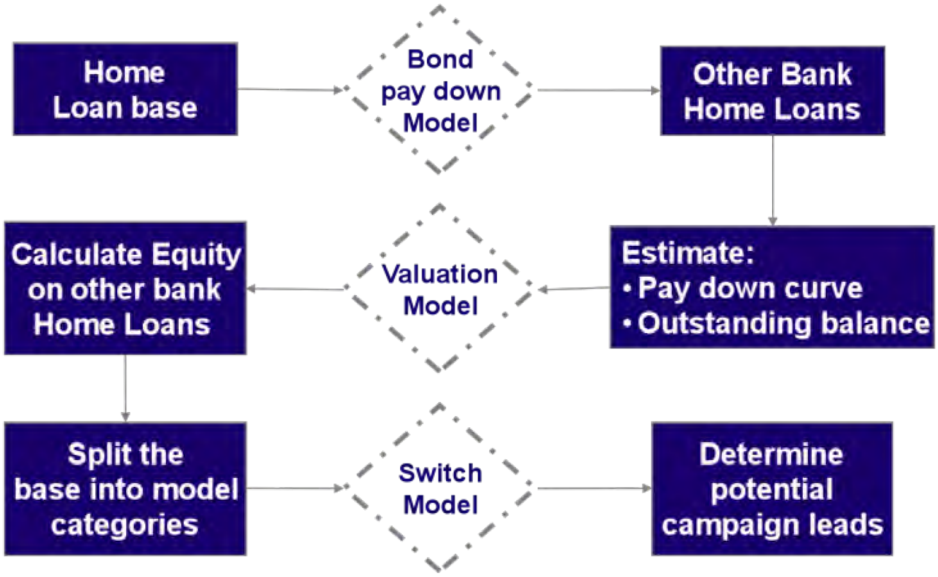


Figure 6.4: *Overview of home loan equity release model*

The bond pay down curve is estimated on the financial institution's home loan portfolio by using the property valuation model. By using the financial institutions data in the Deeds database, the bond pay down curve is applied to loans financed at other financial institutions to estimate their outstanding balances. The equity for each loan is calculated by subtracting the outstanding balance from the valuation amount of the property. A switching propensity model is run after calculating the equity. It is therefore possible to derive the equity for each of a client's properties, even at other financial institutions. Combined with the outputs of the switching propensity model, it provides very valuable input for focused marketing campaigns.

As an example for a potential marketing campaign, it is potentially possible to switch a client's home loan with or without cross-selling to the current financial institution and a further loan with or without cross-sell and with or without registration. The value propositions for clients for an equity release strategy include:

- A home loan with preferential rates;
- The bundling of buy-to-let clients and property investors; and
- A re-advance or a further loan with registration.

6.2.5 Wealth index

A wealth index can be developed, based on the valuation of an individual's property, to identify "super affluent" individuals in order to cross-sell capital related products. South African property owners can be sorted according to their total property portfolios and available equity estimated as discussed in the previous section, as well as in the case of the pay down curve.

6.2.6 Home loan attrition model

A home loan attrition model can be developed, using regression, which predicts when a client will close their home loan, using the valuation amount of the property as a key variable in the model. Clients who are likely to close their accounts can be targeted in focused marketing campaigns to address the issue.

6.2.7 Monthly revaluation of home loan book

An important advantage of having a robust property valuation model is the ability to re-evaluate the home loans portfolio of a financial institution. This exercise helps to determine how much capital should be held towards a client's home loan, taking into account the probability of a default for the client or the home loan.

In the event that a client defaults on their home loan and the property is undervalued, too little capital would have been held and the financial institution may lose a considerable amount of money. On the other hand, in the event that a client defaults on their home loan and the property is overvalued, too much capital would have been held which the financial institution could have utilized elsewhere. Knowing the value of a financial institution's home loan portfolio can help to better manage the risk in its property portfolio as well as to identify opportunities to borrow more or less and to manage strategies around the equity in the portfolio.

6.2.8 Property price index

The flagship output of a property valuation model is a property price index, representing an average property price growth for residential properties in South Africa which was done in this study. Such an index could be released on a regular basis and become a major publicity tool for the financial institution to promote their prowess, knowledge and understanding of the South African property market.

Three of the four major South African banks have created their own property price indices which are quoted regularly in the press. Property price indices can be created at any of the following levels:

- National level;
- Provincial level;
- Major city;
- Residential properties only;
- Commercial properties only;
- Coastal properties only;
- Inland properties only;
- High value properties only;
- Medium value properties only;
- Low value properties only;
- Full title properties only; or
- Sectional title properties only.

6.3 Model enhancements

The operational usability of the final improved model (the smoothed repeat sales model) can be further enhanced by adding certain features. These are enhancements that were identified in the business environment and were implemented in this study.

- **Forecasting -** To compensate for the lag in the Deeds data and the inaccuracy in the most recent months' data, forecasting can be applied to the parameters of the smoothed repeat sales model;

- **Comparable sales** - To enhance the model's coverage (the ability to accurately predict the prices of as many properties as possible), a technique known as the comparable sales method, is proposed. It does not require the sales price or date of transactions, but uses alternative information instead;
- **Upper and lower bounds** - Another enhancement to the improved model is the introduction of a confidence level and upper and lower bounds to predictions, which may help to improve the accuracy of certain predictions;
- **Reliability score** - A reliability score can also be added as a feature specifically for desktop valuations to assist in determining the accuracy of the index; and
- **Area quality grade** - Finally, an area quality grade can be developed to understand the qualities of predictions in certain geographical areas. Most of these enhancements are key features incorporated by various users in their property price indices.

The abovementioned enhancements will be briefly explained in the subsequent sections.

6.3.1 Forecasting

To register a property at the South African Deeds office usually takes at least three to four months. Therefore a property transaction will on average only appear in the Deeds data four or more months after it was registered. It also takes time for Knowledge Factory (KF) to clean the data to a useable form for their clients. Therefore, when looking at the most recent Deeds data, the number of transactions could still change and the predictions might not be accurate. It is important to use the smoothed data in forecasting for these months in order to prevent erratic results. Forecasting uses the last two years' data to predict the last four months of data as illustrated in Figure 6.5. Forecasting can also be used to predict the future behaviour of property prices, but this is beyond the scope of this dissertation.

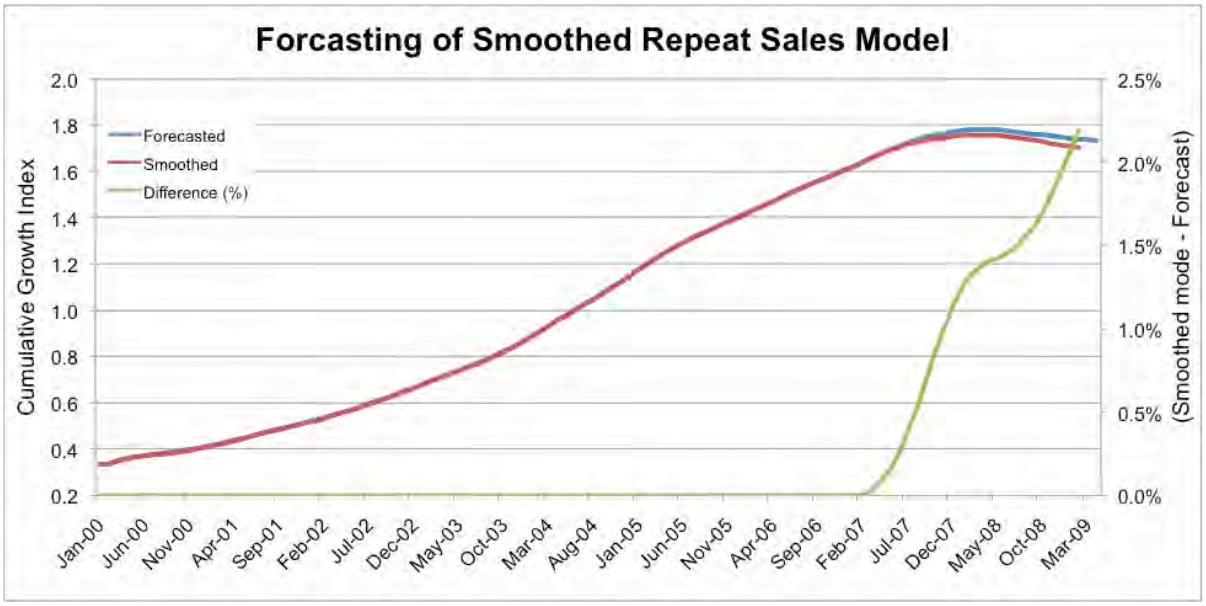


Figure 6.5: *Forecasting of the last 4 months for the smoothed repeat sales model*

To forecast in SAS, the unobserved components models (UCM) procedure is used, which analyzes and forecasts equally-spaced univariate time series data. A UCM decomposes the response series into components such as trend, seasonal cycles, and the regression effects of to predictor series. The components in the model are programmed to capture the salient features of the series that are useful in explaining and predicting its behaviour. A good exposition of the technical details of the UCM procedure can be found in the SAS manual (SAS, 1976).

6.3.2 Comparable sales

There are a number of factors that can make property transactions in suburbs more or less predictable. Some suburbs may have more property transactions than others while other suburbs may have similar price ranges for properties in that suburb.

Some models may also be more predictive in certain suburbs but less predictive in other suburbs depending on the suburb's homogeneity. If there is, for example, a new development in an area with building restrictions, the price range of properties could be very similar which makes models using the measures of central tendency (mean and median) very predictive in this case. Alternatively, in some suburbs there may be farms, smallholdings and full title properties with a large range of prices, making these models less predictive.

As mentioned before, the Deeds data is notoriously 'dirty' and there are quite a few properties for which the purchase price and/or date data is missing. These properties may be considered as outliers compared to, the rest of the suburb/sectional scheme or enumerator area (EA) group. A different method known as the comparable sales method[21] (or inferred analysis method) of property valuation can be used for these cases to improve the coverage of a model.

This method estimates the value of a property by comparing the prices of properties with similar specifications, in similar locations, that transacted within a recent timeframe.

One of the advantages of this method is that comparatives are the most straightforward method of valuing property and it closely reflects the instinctive market value of a property. A disadvantage of using this method is the difficulty to locate enough similar, recently-sold properties.

The basic idea behind the comparable sales model is to compare properties with different parameters. Depending on the quantity and quality of data available, the comparable sales model can be split into four different categories ranging from more to less predictive:

Category 1: Compare price per m² to other properties within the EA *(full title)* or sectional scheme *(sectional title).*

This category applies to properties where the purchase price and date data are missing, but the size of the erf is available. For sectional title properties, the average price per square meter of the properties in the sectional title (normally a complex) is calculated and multiplied by the size of the property with the missing information. For full title properties, the average price per square meter of all similar properties in the EA is calculated and multiplied by the area of the property with the missing information.

[21] "UK's leading resource for house prices and property valuation data" (Mouseprice)

Category 2: Compare price to other properties within the EA *(full title)* or sectional scheme *(sectional title)*

This category applies to properties where the data for the purchase price, date and size of the erf is missing. An assumption is made that the predicted price of a sectional title property (with missing information) is equal to the average price of all properties in the sectional title, while the predicted price of a full title property (with missing information) is equal to the average price of all properties in the EA.

Category 3: Compare price per m² to other properties within the suburb

This category applies to properties where the purchase price, date, sectional scheme and EA are missing, but where the data for the size of the erf and the suburb exists. The assumption is made that for both full and sectional title properties, the predicted price is equal to the average price per square meter of all the properties in the suburb, multiplied by the size of the particular property.

Category 4: Compare price to other properties within the suburb

This category applies to properties where the data for the purchase price, date, sectional scheme, EA and size of the erf is missing, but the suburb data exists. The assumption is made for both full and sectional title properties that the predicted price is equal to the average price of all the properties in the suburb.

6.3.3 Confidence level

The accuracy of an index value is defined as the normalised length of the 95%-confidence interval around an estimated index value. The closer the lower and upper values of the confidence interval are to one another, the more accurate the estimated index value is considered to be. No references to measures of accuracy could be located following a focused literature study. Similarly, no references could be found regarding how narrow a confidence interval had to be in order to be described as ‘accurate’, or what constitutes the minimum required accuracy for a sample.

In Figure 6.6, the final smoothed repeat sales model and its upper and lower bounds are presented. The upper and lower bounds can be used to determine the accuracy of an index at different points in time by calculating the difference between the upper and lower bounds, the smaller the differences, the higher the accuracy and vice versa. This technique was not further evaluated.

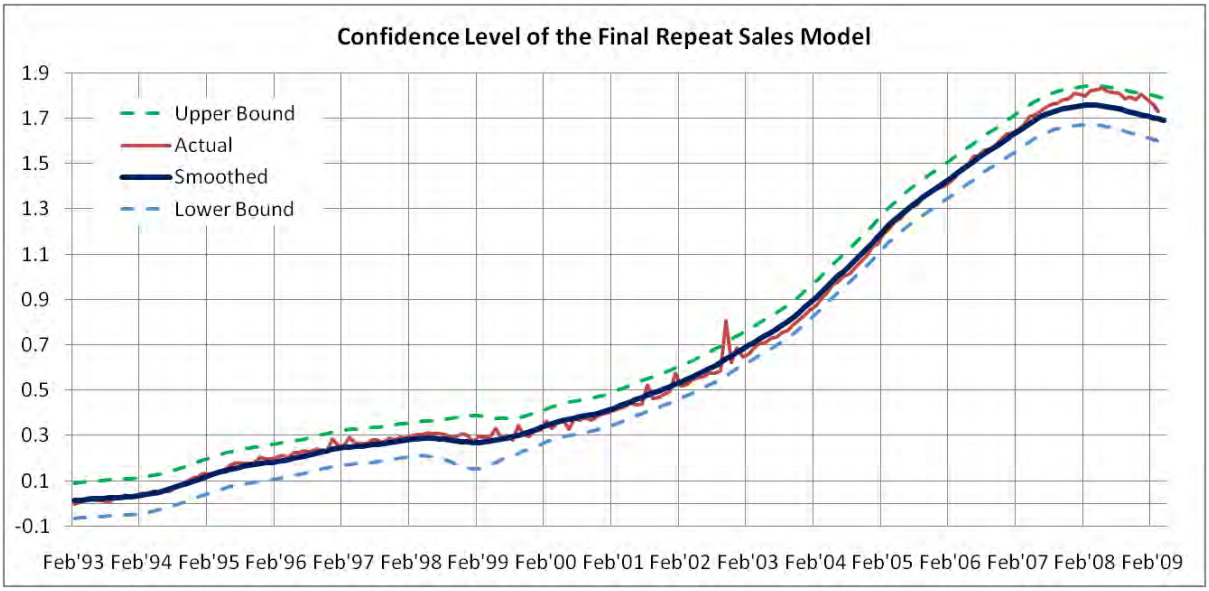


Figure 6.6*: Confidence level of the final repeat sales model*

6.3.4 Reliability score

A reliability score is an enhancement that could assist a financial institution with desktop valuations. In practice, a number of assumptions have to be made to eventually determine whether the predicted price for a property is accurate enough to use, or whether a property assessor should visit the property to assess the price in person.

The reliability score is split into three components or sections namely policy rules, comparable sales reliability (assuming that the comparable sales is applied) and repeat sales reliability. In Table 6.1 various policy rules are depicted that could influence the reliability of the repeat sales model, together with the percentages that would be assigned to properties as reliability scores if it were to fall in any of these categories. In the comparable sales section, the reliability of the pricing is based on

the specific method used to evaluate the property, bearing in mind that some methods are more accurate than others.

An adjustment can be made (policy rule nr. 2 in Table 6.1) based on the number of properties that the specific property was compared with to obtain the comparable sales price. When determining the repeat sales reliability, accuracy is determined for each individual model, based on back-testing and the upper and lower bounds. It is then adjusted for the period between sales and the month it traded in. If the price falls below one standard deviation of the suburb mean, the reliability is adjusted downwards. The minimum reliability for the repeat sales model is 63% and the maximum is 83%.

Policy Rules		Comparable Sales Reliability		Repeat Sales Reliability	
1. Predicted value detected as an outlier within suburb, but no replacement valuation by comparable sales possible	20%	1. Determine accuracy based on difference in valuation for repeat sales and comparable sales for:		1. Model accuracy determined for each individual model then adjusted for hold period and trade month	
2. Less than 3 properties per suburb	20%	a. EA per m²	65%		
		b. EA average	49%		
		c. Suburb per m²	31%	2. Below 1 ∂ from suburb μ(suspected land transactions) reliability adjusted downwards	
		d. Suburb average	22%		
3. Predicted value in top 5% of predicted values in suburb	60%	2. Score is adjusted for the amount of properties it was compared to	-10%		
4. Entire suburb where standard error (∂/√n) is in the top 5% of all standard errors	60%	a. 0-5	-7%	Maximum reliability	83%
		b. 5-10	-5%	Minimum reliability	63%
5. Farm predictions less reliable	50%	c. 10-15	-2%		
		d. 15-20	0%		
		e. 20+			

Table 6.1: *Reliability score rules*

Figure 6.7 illustrates typical results that can be obtained when applying the reliability score to the predicted prices. The blue bar indicates the properties that, based on the reliability score rules in Figure 6.7, would have been approved and the green bar indicates the properties that would have been referred to a property assessor. The red line indicates the number of properties considered.

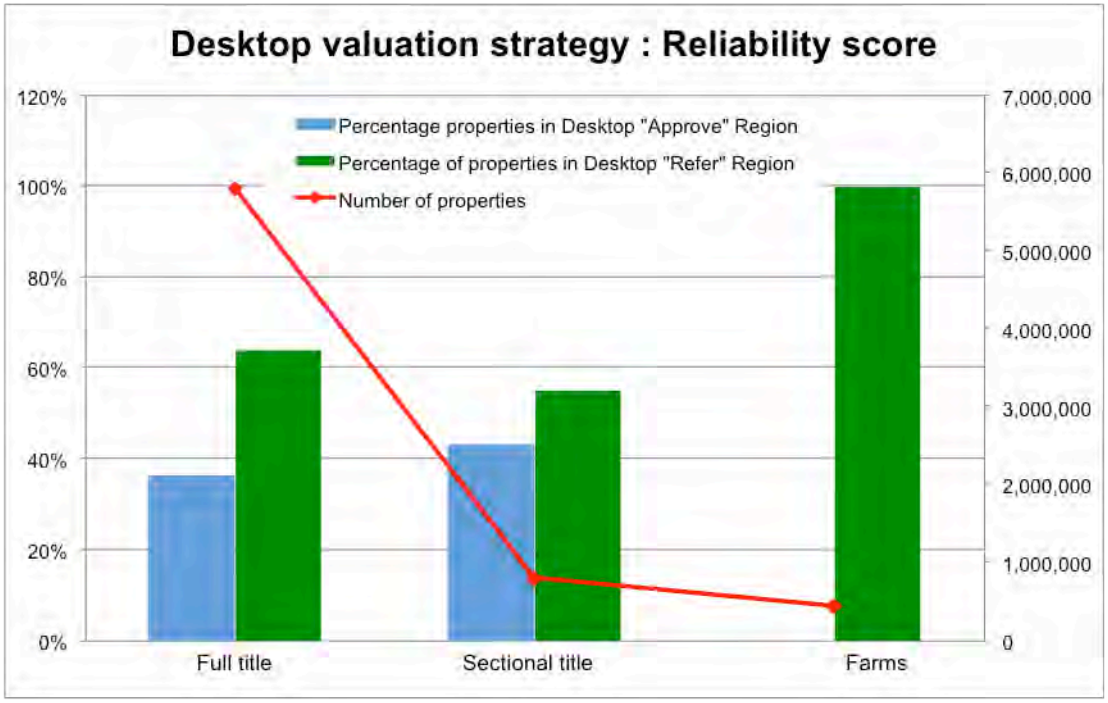


Figure 6.7: *Results applied to actual properties*

6.3.5 Area quality grade

The area quality grade is a grading that can be put in place to help determine the quality of a specific geographical area, from suburb level up to provincial level. The area quality grade is made up of two main components, the price level and the market liquidity which are weighted (40% price level and 60% market liquidity according to a business need together with senior management input and experience) in calculating the area quality grade (see Figure 6.8).

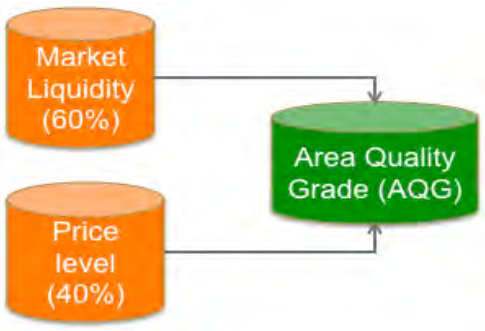


Figure 6.8: *The main components of area quality grade*

Market liquidity is determined by calculating the number of property transactions in a specific area during the previous 18 months, while price level is calculated by taking the absolute value of the difference between the average price of all properties sold during the previous 18 months and the average price of properties in the specific area for the same period.

The area quality grade is a possible enhancement to a desktop valuation tool, as well as an indicator to investors of the quality of the area they are interested in. Many factors such as crime statistics or census data can also be built into an area quality grade, but this is viewed as being beyond the scope of this study.

In summary, by implementing the enhancements described in this section, a desktop valuation tool can be developed for a scenario where 40% of valuations are accepted and 60% referred to a property assessor. Figure 6.9 illustrates what a desktop valuation tool might typically look like.

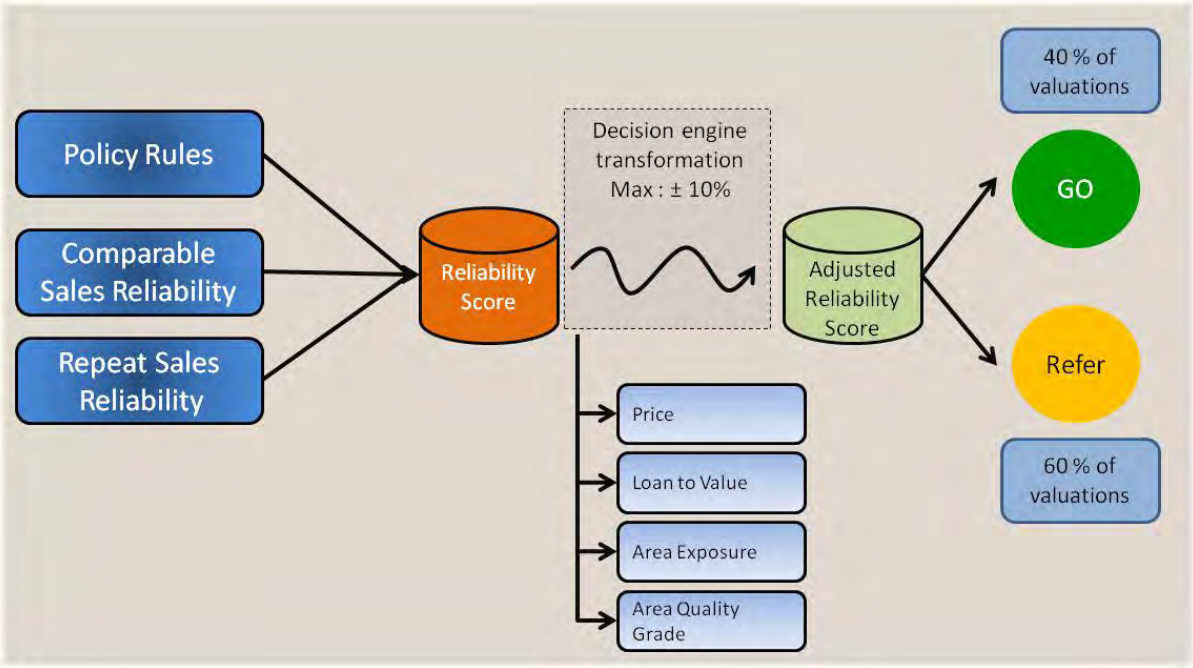


Figure 6.9: *Example of a process for a typical desktop valuation tool*

6.4 Conclusion

As illustrated in this chapter, there are many valuable applications which could utilise a robust and accurate property valuation model. As discussed, the property price

index could also provide those financial institutions that implement it, a leading edge in the property market. All these applications and enhancements depend on the quality of the underlying valuation model and property price index. This illustrates the importance and value of the work in Chapters 3 to 5 of this study.

Chapter 7: Conclusion

7.1 Introduction

In this chapter, the final comments and concluding remarks for the study are presented, together with a summary of the objectives of the study and how they were achieved. New challenges and opportunities for further study that presented itself during the research project will also be outlined.

7.2 Objectives of the study

In Chapter 1 it was stated that the primary objective of this research project is to determine the most robust approach available with which to valuate properties in South Africa and to develop a residential property price index. To accomplish this, five secondary research objectives that had to be achieved, were defined. These five objectives were:

- To gain a clear understanding of and present an introductory overview of property valuation modelling and property price indices and to understand the arguments and contributions made by various authors on the different approaches they studied;
- To understand the data sources used for modelling and the necessary clean-up processes, validations and exclusions that need to take place, as well as to clearly understand the model methodology of the three valuation models used in this study;
- To determine the most accurate approach to valuating properties by applying various statistical tests to compare the three valuation models;
- To improve the most promising model further to become the most robust solution with which to valuate residential properties in South Africa. Using this solution to further derive a national property price index; and

- To gain an understanding of various operational model enhancements that will add business value through current and future applications.

A summary of how these objectives were achieved follows:

To gain a clear understanding of and present an introductory overview of property valuation modelling and property price indices and to understand the arguments and contributions made by various authors on the different approaches they studied.

A literature research study was conducted (Chapter 2) to identify and understand the contributions of various authors and their different approaches to valuate property prices and develop property price indices. The main findings were:

- Measures of central tendency using the mean is one of the most basic approaches used to model property prices, but lacks reliability when the mean price does not represent the total population;

- Measures of central tendency using the median is also a very basic approach used to model property prices. Median prices can skew the results when more properties are sold at higher or lower prices;

- The hedonic regression valuation model can be applied to properties traded once or more but requires a large number of accurate property characteristics to enable the development of a valuation model that is sufficiently reliable;

- Repeat sales valuation models remove quality differences between pairs of properties sold at different periods and only requires information on the property price and the date of sale, but it can only be developed based on properties that sold more than once; and

- Hybrid valuation models involve using hedonic regression modelling for properties that only traded once, and repeat sales modelling for properties that traded more than once. Hybrid models also feature a combination of some of the advantages and disadvantages of both the hedonic regression model and the repeat sales model.

To understand the data sources used for modelling and the necessary clean-up processes, validations and exclusions that need to take place, as well as to clearly understand the model methodology of the three valuation models used in this study.

The data behind the model, the South African Deeds Office data, was investigated in Chapter 3. The data contains property transaction information dating back to 1993. Data prior to this date was deemed to be unreliable and not useable in the development of various valuation models.

The selected data was cleaned using the following clean-up processes:

- Removal of garages indicated as residential properties.
- The exclusion of multiple property purchases on the same date.
- The exclusion of invalid dates.

Certain exclusions were made from the cleaned data, including the following:

- Duplicate records were removed.
- Transactions before 1993 were excluded.
- Property prices below R70 000 and above R6 000 000 were excluded.
- Missing province and suburb information was excluded.
- All ‘special cases’ not including farms were excluded.

The data was then further validated and exclusions were made based on the following statistical assumptions:

- Outliers were removed based on two standard deviations.
- Properties that traded only once were removed.
- Where there were more than one transaction in the same month, the most recent transaction was kept in the data and the other transactions were removed.
- Where the price of the second transaction was less than 50% of the price of the first transaction for a sales pair, the sales pair was excluded.
- Where the price of the second transaction was more than 60% of the price of the first transaction for a sales pair, the sales pair was excluded.
- If a property traded more than 10 times from 1993 to 2009, the property was excluded.

- Extreme deviations in price (from the norm) were excluded.

Given that the South African Deeds data set is very large, but does not include property characteristics, the hedonic regression and hybrid valuation models could not be applied to the data. The theory underlying the development of the three remaining methods was explained, namely the measures of central tendency using the mean and median as well as and the repeat sales method.

To determine the most accurate approach to valuating properties by applying various statistical tests to compare the three valuation models.

Growth indices were derived for all three valuation models, based on the same data (Chapter 4). The valuation models were compared using the following nine different statistics in order to determine the most accurate approach to valuating property prices:

- Statistic 1: Closest prediction to actual value;
- Statistic 2: Distribution of model errors;
- Statistic 3: Theil's U-statistic;
- Statistic 4: Mean error (ME);
- Statistic 5: Mean squared error (MSE);
- Statistic 6: Root mean squared error (RMSE);
- Statistic 7: Mean absolute error (MAE);
- Statistic 8: Mean prediction error (MPE); and
- Statistic 9: Mean absolute prediction error (MAPE).

The repeat sales model outperformed the other two models in all cases and proved to be the most accurate approach with which to model property prices.

To improve the most promising model further to become the most robust solution with which to valuate residential properties in South Africa. Using this solution to further derive a national property price index.

Further potential improvements to the basic repeat sales model were identified and six additional versions of the basic repeat sales model were proposed to improve the model further (Chapter 5). The five versions were based on the following:

- Version 1: Model farms separately;
- Version 2: Model townships separately;
- Version 3: Segmentation into significant groups;
- Version 4: Improving the model error by using weights; and
- Version 5: Reducing the effect of volatility through smoothing.

These improvements were compared using the statistics 2 to 9, mentioned above, to determine whether it did indeed improve the model.

Some operational enhancements to the model were discussed, ? namely:

- Forecasting;
- Using four different categories of comparable sales to help predict outliers in the data and to improve the coverage of a model:
    - Category 1: Compare price per m² to other properties in the EA *(full title)* or sectional scheme *(sectional title)*;
    - Category 2: Compare price to other properties within the EA *(full title)* or sectional scheme *(sectional title)*;
    - Category 3: Compare price per m² to other properties within the suburb;
    - Category 4: Compare price to other properties within the suburb;
- Confidence level with an upper and lower bound;
- Reliability score; and
- Area quality grade.

Finally, a property price index was developed from the improved repeat sales model and compared to the published indices of various financial institutions in South Africa. By analysing the correlation to prime, the property price index developed in this dissertation has the best (negative) correlation when compared to four other property price indices in the South African market place.

To gain an understanding of various operational model enhancements that will add business value through current and future applications.

The property valuation models have many applications in financial institutions. A number of these applications (see below) were investigated, of which the property price index is but one:

- A property portal;
- Bond pay down estimation model;
- Switch propensity model;
- Home loan equity release model;
- Wealth index;
- Home loan attrition model;
- Monthly revaluation of the home loan book; and
- Property price index.

7.3 Problems experienced

A few problems were experienced with regards to the data on which the models were developed. The data does not contain any characteristic information on properties, which made it impossible to test all four methods on South African Deeds data. The data is also notoriously “dirty”, which can cause greater bias in the results and many clean-ups had to be performed on the data.

Property indices for financial institutions in South Africa are published in the media and can give an institution a competitive advantage. Therefore the data behind the indices is very sensitive and not freely available to enable in-depth comparisons between indices.

7.4 Possibilities for further studies

One of the future applications of an authoritative, independent, and stable house price index that covers all properties in South Africa, is property price index derivatives. To develop property derivatives a property price index is required that

covers all properties in South Africa. Although there are suitable property price indices that fit the profile and which could be used to support property price index derivatives, no financial institution currently offers property derivatives in South Africa, but there is a great opportunity to satisfy the demand for these derivatives in the property and investment market (e.g. developers, builders, holders of mortgage portfolios, mortgage insurers).

7.5 Conclusion

Chapter 7 concludes this study. In this chapter, a summary of the initial objectives and how they were achieved was presented, together with possible future research opportunities.

References

1. Abraham J.M. and Schauman W.S., 1991. “New evidence on home prices from Freddie Mac Repeat Sales”, *AREUEA Journal*, 19, 333-352.

2. AfriGIS, 2010. “AfriGIS Data Catalogue”, available at: http://maps.afrigis.co.za/DataCatalogue/Users/Default.aspx (Retrieved 2010).

3. Bailey M.J., Muth R.F. and Nourse H. O., 1963. “A regression method for real estate price index construction” *Journal of the American Statistical Association*, 58, 933-942.

4. Botha, J. 2010. “House price growth: marginal positive growth in March”, South Africa: Standard Bank. Available at: ws9.standardbank.co.za/sbrp/LatestResearch.do (Retrieved 2010).

5. Bourassa, S. C., Hoesli M.and Sun J., 2004. "A simple alternative house price index method.". Working paper.

6. Butler, J. S., 2005. “Revision Bias in Repeat-Sales Home Price Indices”. Philadelphia: Working paper.

7. Case, B. And Quigley, J., 1991. “The dynamics of real estate prices”, *Review of Economics*, 73 (3), 50-58.

8. Case, B., Pollakowski, H. And Wachter, S., 1991. “On choosing among house price index methodologies”, *AREUEA Journal*, 19 (3), 286-307.

9. Case K. E. and Shiller R. J., 1987. “Prices of single-family homes since 1970: New indexes for four cities”, *New England Economic Review*, 45-56.

10. Case, K. E. and Shiller R. J., 1989 “The effciency of the market for single-family homes”, *American Economic Review,* 79 (1), 125-137.

11. Clapp J.M. and Giacotto C., 1992. “Estimating price indices for residential property:a comparison of repeat sales and assessed value methods”, *Journal of the American*, 87, 300-306.

12. Cook, S., 2006. “Understanding the construction and interpretation of forecast evaluation statistics using computer-based tutorial exercises”. Available at: www.economicsnetwork.ac.uk/showcase/cook_forecast , from Swansea University. (Retrieved 2010).

13. Court, A., 1939. “Hedonic price indexes with automotive examples”, General Motors, New York: *The Dynamics of Automobile Demand*.

14. Crone, T. a., 1992. “Estimating house price appreciation: a comparison of methods”, *Journal of Housing Economics*, 2 (4), 324-338.

15. Deeds, 2001. “Frequently Asked Questions”. Available at: http://www.deeds.gov.za/ITSODeedsWebB/deedsweb/faqs.jsp?tagHeader=Frequently%20Asked%20Questions, from DeedsWeb. (Retrieved 2010).

16. Du Toit, J., 2010. “March ABSA House Price Indices”, South Africa: ABSA. Available at: www.absa.co.za (Retrieved 2010).

17. Ferri, M., 1977. "An application of hedonic indexing methods to monthly changes in housing prices: 1965-1975", *AREUEA Journal*, 5 (4), 455-462.

18. Footprint, 2009. “About Us”. Available at: http://www.footprintonline.co.za/about.aspx, from Footprint. (Retrieved 2010).

19. Gatzlaff, D. and Haurin, D., 1994 “Sample Selection and Biases In Local House Value Indices working paper”. The Ohio State University.

20. Gatzlaff, D. and Haurin, D., 1997. “Sample Selection Bias and Repeat-Sales Index Estimates Journal of Real Estate Finance and Economics”, 14 (1/2), 3-50.

21. Gatzlaff, D. a., 1994. "Measuring changes in local house prices: an empirical investigation of alternative methodologies", *Journal of Urban Economics*, 35, 221-244.

22. Goodman, A. a., 1995. “Heteroskedasticity in repeat sales house price equations". Wayne State University: working paper.

23. Griliches, E., 1971. Hedonic price indexes revisited, in Z. Griliches (ed), *Price Indexes and Quality Change.* Cambridge, Massachusetts: Harvard University Press.

24. Hodrick R. and Prescott E. C., 1997. “Postwar U.S. Business Cycles: An Empirical Investigation". Available at: http://en.wikipedia.org/wiki/Hodrick-Prescott_filter, from: Journal of Money, Credit, and Banking.

25. Hosios, A. a., 1991. "Measuring prices in resale housing markets in Canada: evidence and implications", *Journal of Housing Economics*, 1 (4), 303-317.

26. Investopedia, 2010. “Investopedia”. Available at: http://www.investopedia.com/terms/p/property-derivative.asp, from Dictionary.

27. Jansen, S. a., 2006. "Developing the House price Index in the Netherlands:A practical application of Weighted Repeat Sales". Delft University of Technology, The Netherlands: OTB Research Institute for Housing, Urban and Mobility Studies.

28. Knowledge Factory, 2008. “Deeds Data and predictions since 1993”. Available at: www.knowledgefactory.co.za, from Knowledge Factory. (Retrieved 2010).

29. Leblond S. P., 2004. “Comparing predictive accuracy of real estate pricing models: an applied study for the city of Montreal”. Universite de Montreal.

30. Lightstone, 2010. “Residential Property Indices: Lightstone April Repeat Sales Indices”, South Africa: Lightstone. Available at: www.lightstone.co.za/LSC/Content/NewsRoom/HousePriceIndex.aspx

31. Loos, J., 2010. “March FNB House Price Index”, South Africa: FNB. Available at: www.fnb.co.za/home-loans/house-price-index.html (Retrieved 2010).

32. Mark, J. and Goldberg, M., 1984. “Alternative housing price indices: an evaluation”, *AREUEA Journal*, 12 (1), 30-49.

33. Meese, R. and Wallace, N., 1991. “Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices” *Journal of Real Estate Finance and Economics*, 14 (1/2), 51-74.

34. Meese, R. and Wallace, N., 1997. “The Construction Of Residential Housing Price Indexes: A Comparison Of Repeat Sales, Hedonic Regression, And Hybrid Approaches”, *Journal of Real Estate Finance and Economics*, 14 (1/2), 51-74.

35. Mouseprice, 2007. “The Comparable Sales Method”. Available at: http://www.mouseprice.com/Articles/The-Comparable-Sales-Method.ashx?AspxAutoDetectCookieSupport=1 (Retrieved 2010).

36. Muller J., 2008. “Property values: What’s the truth?”, Available at: www.property24.com/articles/property-values-whats-the-truth/8470

37. Nielsen B., 2008. “Fannie Mae And Freddie Mac, Boon Or Boom?”. Available at: http://www.investopedia.com/articles/07/fannie-freddie.asp#12888433090933&close (Retrieved 2010).

38. Oobarometer, 2010. “Oobarometer”, South Africa: Oobarometer. Available at: www.ooba.co.za (Retrieved 2010).

39. Pendleton, W. C., 1965. "Statistical inference in appraisal and assesment procedures", *The Appraisal Journal*, 33 (1), 73-82.

40. Rosen, S., 1974. “Hedonic prices and implicit markets: product differentiation in pure competition”, *Journal of Political Economy*, 82, 34-55.

41. S&P/Case-Shiller., 2009. “Home price indices index methodology”. United States: Office of Federal Housing Enterprise Oversight (OFHEO).

42. SAS, 1976. “SAS”. Available at: www.sas.com, from: Products & Solutions / Analytics. (Retrieved 2010).

43. Shiller, R. J., 1991. “Arithmetic Repeat Sales Price Estimators”, *Journal of Housing Economics*, 1, 110-126.

44. Wang F. T. and Zorn P. M., 1997. “Estimating house price growth with repeat sales data: What’s the aim of the game?”, *Journal of Housing Economics*, 6, 93-118.

45. Zanola, R., 2007. “The dynamics of art prices: The selection corrected repeat-sales index”, Alessandria: Universita’ Del Piemonte Orientale “Amedeo Avogadro”

Appendices

Appendix A: Property identification (i.e. Erf Key)

To successfully use Title Deed data for developing a valuation model, it is important to define propertyies uniquely. The jargon often used when referring to a property, whether full title or sectional title, is an erf. Almost every unique property in South Africa has an erf key, which is a derived field using the metro-town name, the erf number (a 5-digit number) and the sub-division number of the erf or erf portion (also a 5-digit number). The 3 variables are separated with a '~'. For example, a typical erf key would be CAPE TOWN~64624~00001.

For each sectional scheme, there is a sectional scheme name and sectional scheme number (Sectional Scheme ID). Since an erf key refers to the erf registered at the Deeds Office, and erven (multiple of "erf") can range from a development with a few complexes to a single house, an erf key can have more than one sectional scheme, and each sectional scheme can contain several units. To distinguish between full title and sectional title erf keys, we refer to the Sectional Scheme ID and the unit number. For a full title erf, the Sectional Scheme ID and unit number fields will both be zero (0) or not populated (.). For a sectional title, the erf key, the Sectional Scheme ID and unit number will both be populated and not equal to zero.

In summary, there are two types of properties, a unit in a Sectional Scheme ID that belongs to an erf key, which is a sectional title property or and simply an erf key (where unit and Sectional Scheme ID are both zero or missing), which is a full title property. These erf keys exist in an EA, which can be in one suburb or overlap over more than one suburb. The suburb is defined as a socio-economic group and cluster and is part of a metro-town. The metro-town exists in a province in South Africa.

Each unique erf is assigned a title deed number by the South African Deeds Office. Together with a few other variables, this title deed number helps to define unique properties. For the man on the street, an erf key is defined by a property with a street name, street number and street type (avenue, lane, street, etc). Each erf has a land size and a property size, but in the Deeds Office data, only the land size of the erf is

supplied. This field is known as “extent sqm” and gives the size of the the erf in square meters.

Appendix B: ClusterPlus socio-economic groups by Knowledge Factory

Knowledge Factory's Clusterplus is a geo-demographic segmentation system which provides insight into the behaviours, characteristics, lifestyles and locations of the people of South Africa. It is developed at a suburb and sub-place level and modelled utilising primarily Deeds Office and Census information. Clusterplus is based on comprehensive datasets and provides coverage of the entire South African population. Distinguishing itself in terms of specificity, Clusterplus segments the South African population into 10 main groups (labelled A to J and S), which are further sub-divided into 38 clusters. These groups are statistically calculated for the whole of South Africa. The socio-economic groups are based mainly on income and density and are calculated using various clustering methods. The groups are ranked or classified, with group A (Silver Spoons) being the top end of the South African population and Group J (Below the Breadline) being the bottom end. Group S is classified as special cases such as golf courses, cemeteries, etc. These groups and clusters have been defined in terms of the following variables:

- Socio-economic rank – income, property value, education and occupation.
- Life stage – age, household and family structure.
- Dwelling type – size, type and age of structure.

More detail on the different groups:

Group A – Silver Spoons - "the most exclusive neighborhoods, inhabited by the elite of South Africa"

Group B – Upper Middle Class - "older, more stable suburban areas with several pricey new developments also included"

Group C – Middle Suburbia - "residents are reasonably educated (matric or secondary qualification) but the belt is tightening far more than in Group A or B "…children of Middle Suburbia can take nothing for granted other than that their parents will do everything to enable their children to achieve"

Group D – Community Nests - “Large blocks of flats and townhouse complexes, with single houses in between…typically close to the city centre and the majority of residents rent their accommodation”

Group E – Labour Pool - “mixture of dwellings, with houses in the majority… strong sense of community in which reliance on each other has ensured survival”

Group F – New Bonds - “young parents who are likely to be the first in their family to own property… stretched by financial demands and bond payments”

Group G – Township Living - “designed by the old regime as row upon row of low cost housing and has developed into strong communities with emphasis on group values”

Group H – Towering Density - “Crude tenement blocks, low cost semi-detached council houses and once-proud inner city blocks that fell victim to decay…teetering, but not yet falling”

Group I – Dire Straits - “represents townships bursting at the seams… every bit of open space filled with dilapidated shacks, either as free-standing home or extensions of the original matchbox houses”

Group J – Below the Breadline - “No formal planning, no infrastructure…informal settlements of shacks of varying size, in very rare cases replaced by formal structures”

Group S – Special Cases - Agricultural land, golf courses, cemeteries etc

The Family Tree

How to interpret the family tree

The family tree runs roughly from high to low income as one moves from left to right and high to low density from top to bottom. Cluster 4M can therefore be described as "high income, high density".The linkage indicated on the family tree shows which customers are closest in profile. Cluster 4M – high income, high density, (exclusive cluster homes and expensive but small homes) – is relatively close to cluster 7M – slightly lower income, high density, (cluster homes and townhouses).

Group	Cluster (M)etro & (T)own Range	Group	Cluster (M)etro & (T)own Range
SILVER SPOONS	1M Upper Crust 1T Big Fish 2M Pearl Strings 3M Cheese and Wine 4M Fashion Café Society	NEW BONDS	15M Bond Battalions 16M Developer's Dream 17M Strugglers Reward 7T Young Blues Town
UPPER MIDDLE CLASS	5M Suburban Bliss 3T Platteland Pearls 6M Dish and Decoder Set 7M Terracotta Terraces 2T Retreat	TOWNSHIP LIVING	15M Bond Battalions 16M Developer's Dream 17M Strugglers Reward 7T Young Blues Town
MIDDLE SUBURBIA	8M Pram Pushers 9M Settled Suburbia 4T Small Town Families	TOWERING DENSITY	21M City Strugglers 22M Modest Masala 23M Wilted Neon 24M Tenement Trenches
COMMUNITY NESTS	10M Silver Threads 11M Melting Pot 5T Modest Main Street	DIRE STRAITS	25M Chakalaka 26M Poor Neighbours 9T The Other Town
LABOUR POOL	12M Suburban Stagnation 13M Family Street 14M Family Strugglers 6T Rusty Blues Town	BELOW THE BREADLINE	27M Tin Town 28M eKaya 10T Forgotten People

Special Cases

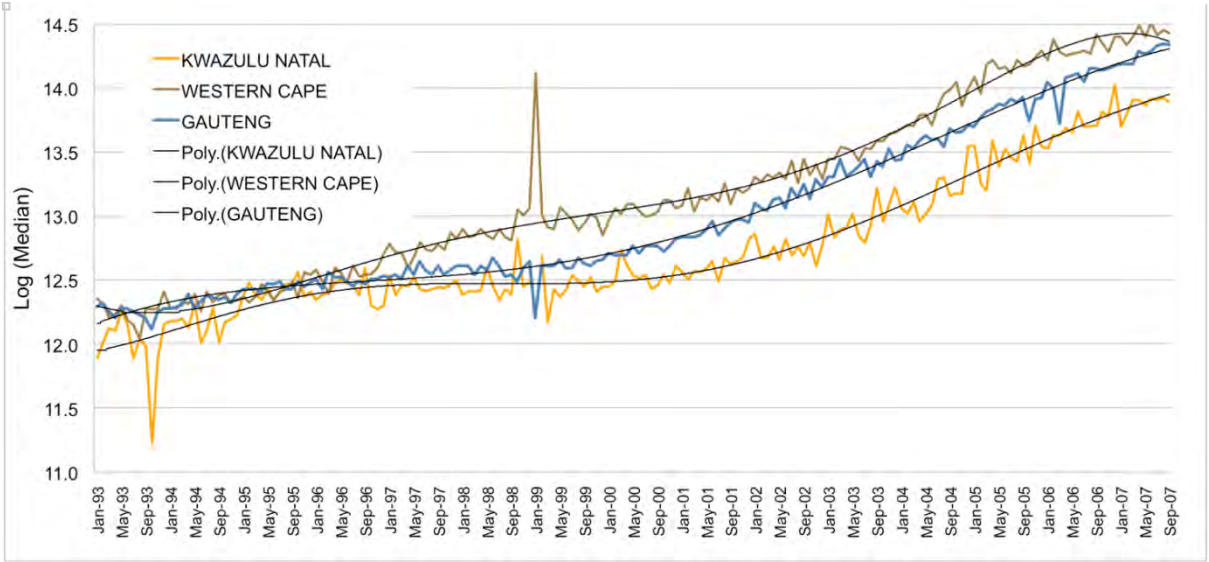
1S Agricultural	4S Commercial	7S Hospital	10S Institution	13S Air Field	16S Small Holdings
2S Sparse	5S Community	8S Hostel	11S Water	14S Recreational	17S Tribal
3S Industrial	6S Golf Course	9S Cemetery	12S Mine	15S Insufficient Data	

Appendix C: Grouping the data

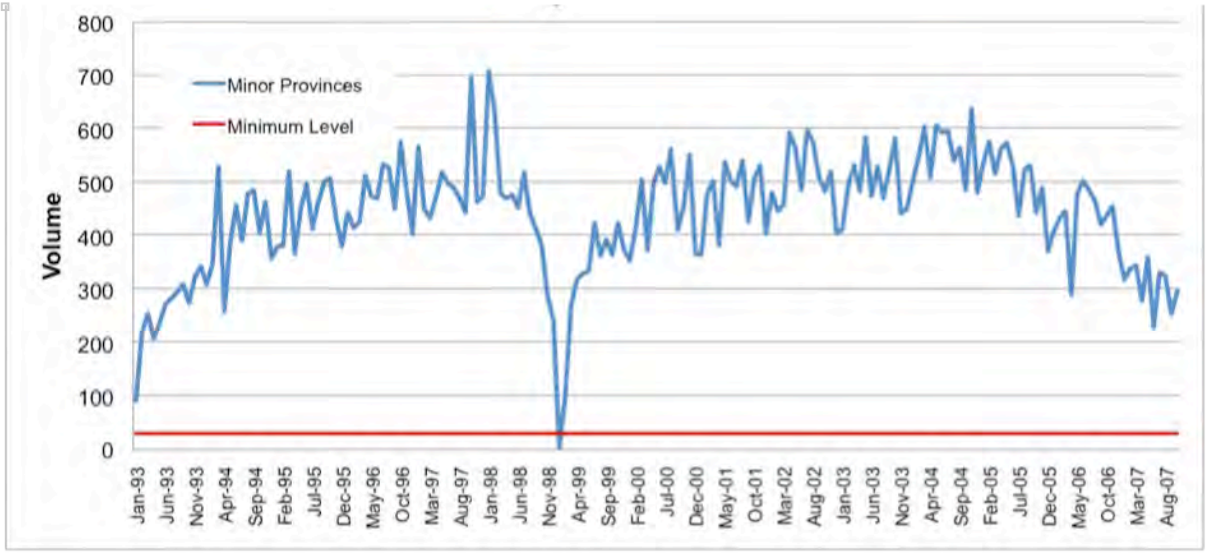
The outcome of the grouping of the data into 18 groups is presented in this appendix as the (i) data volume per bin and (ii) the logarithm median of the data with trend line for each of the 18 groups.

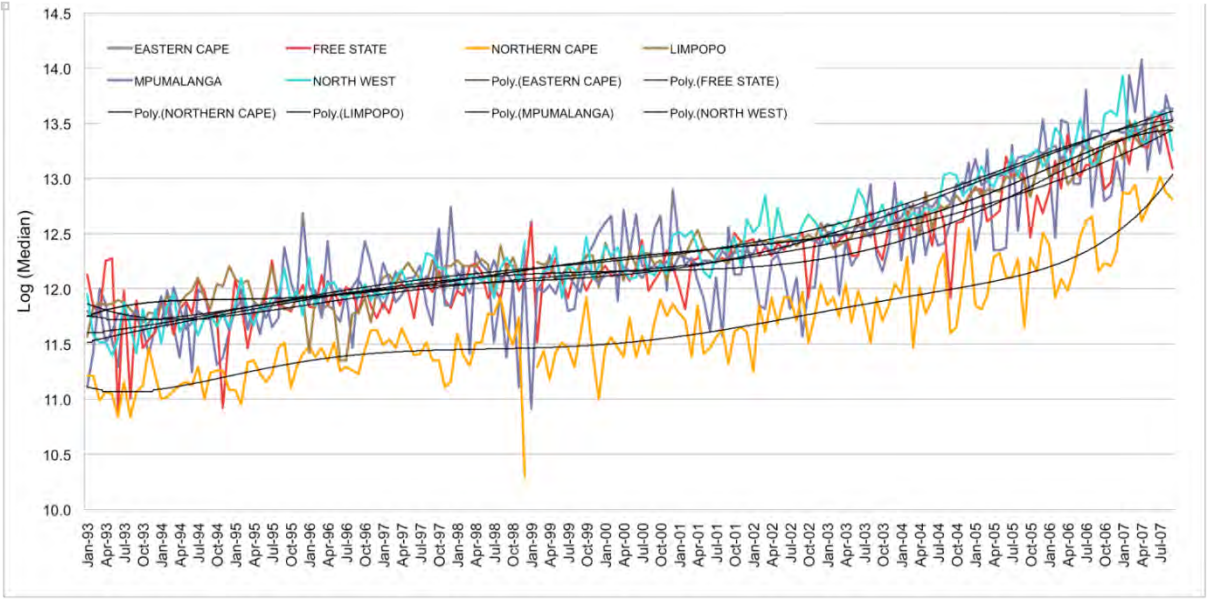
Full Title Properties

Model 1: Group A – Major Provinces

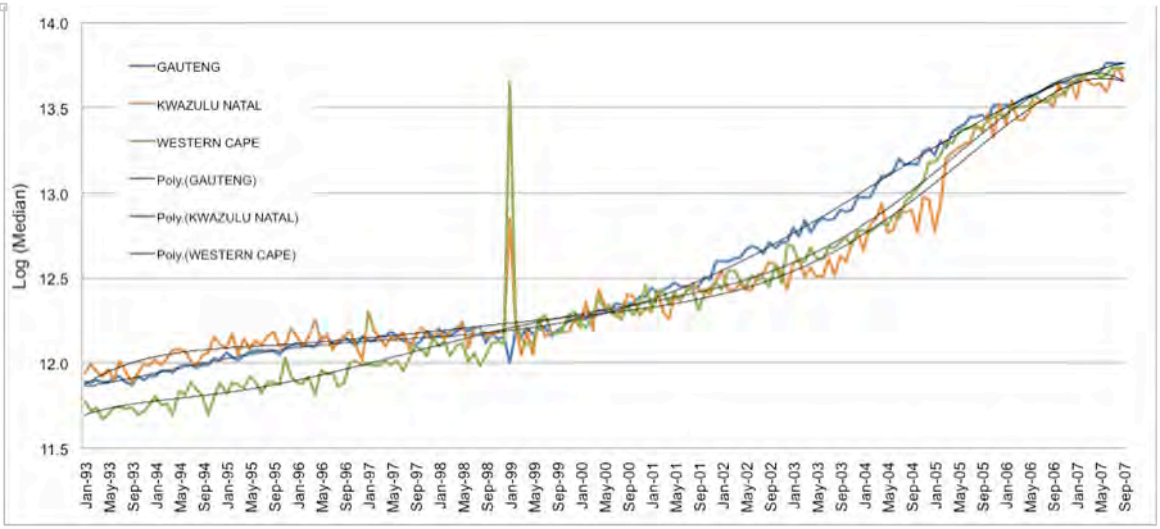


Model 2: Group A – Minor Provinces

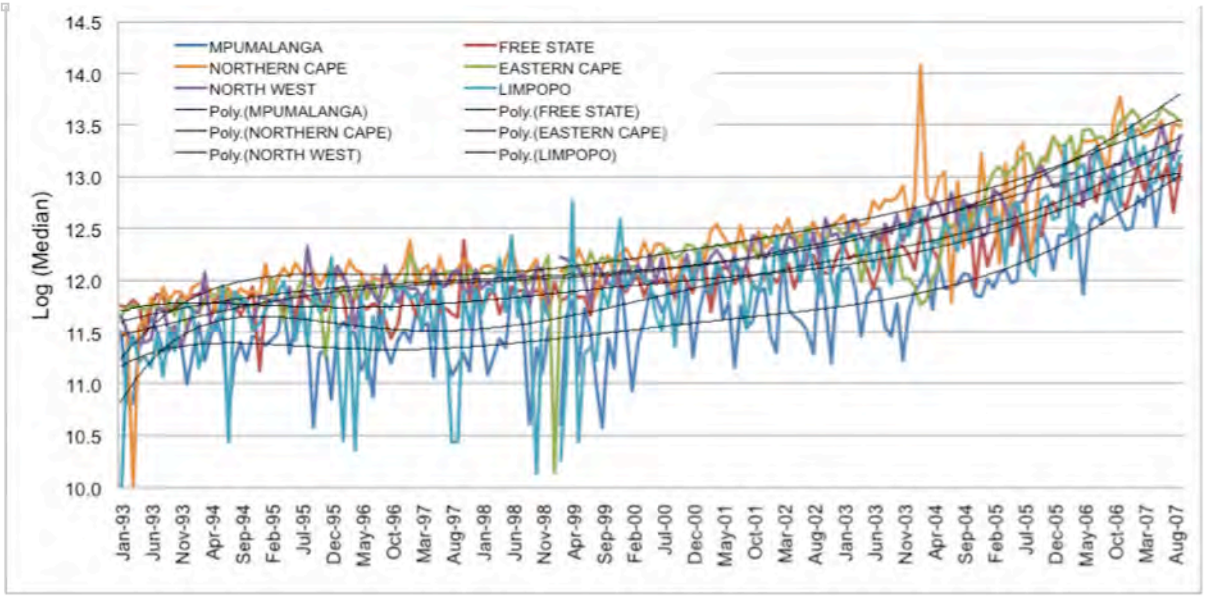




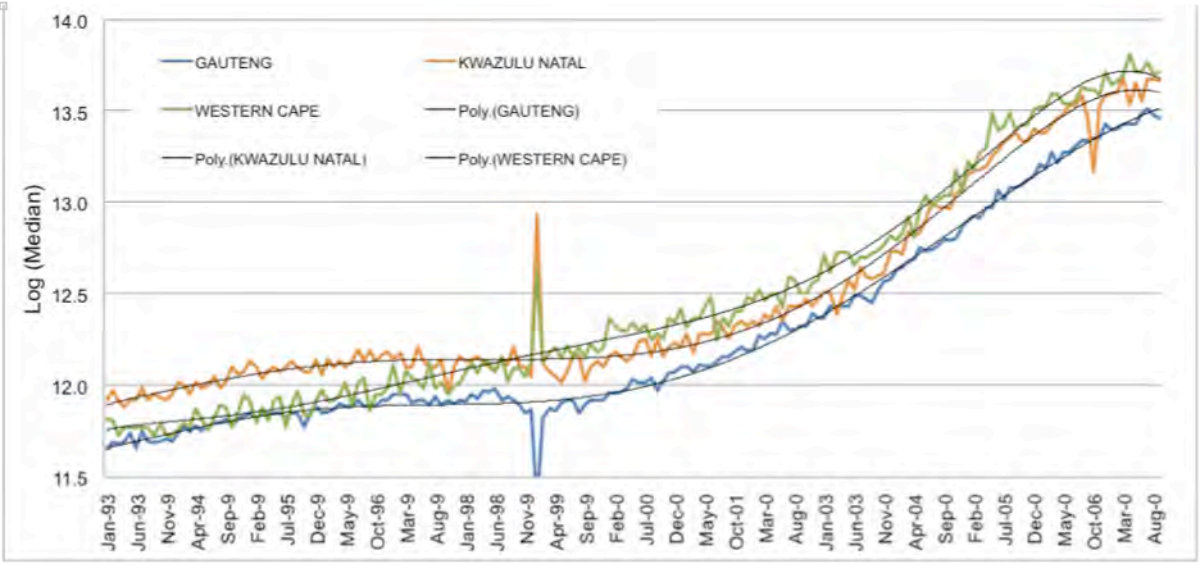
Model 3: Group B – Major Provinces



Model 4: Group B – Minor Provinces

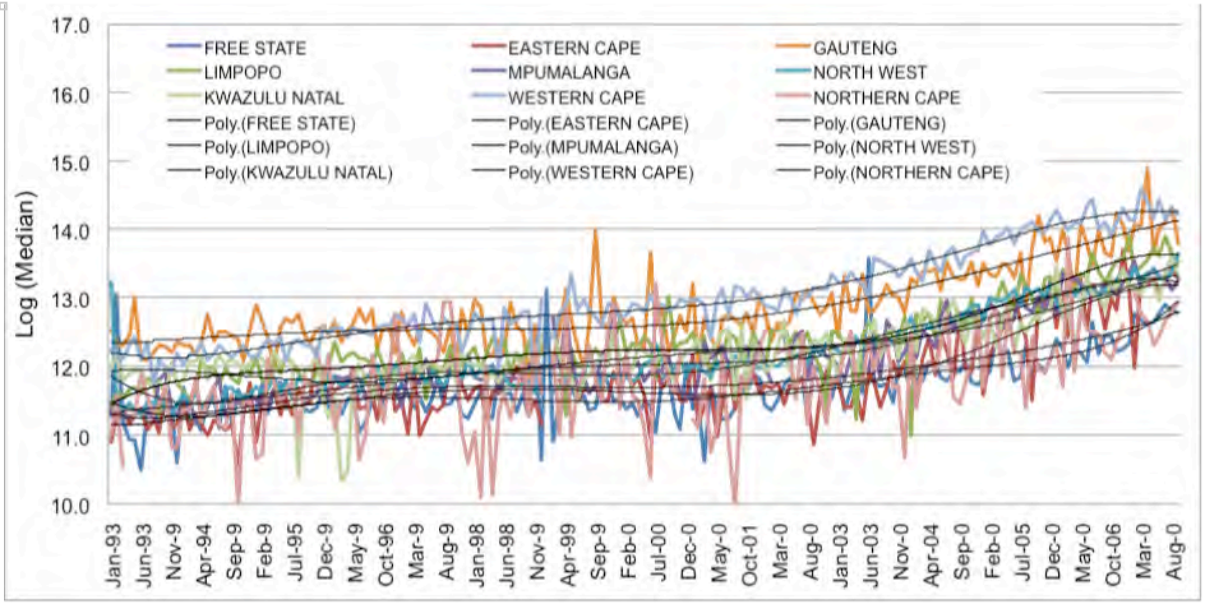


Model 5: Group C – Major Provinces

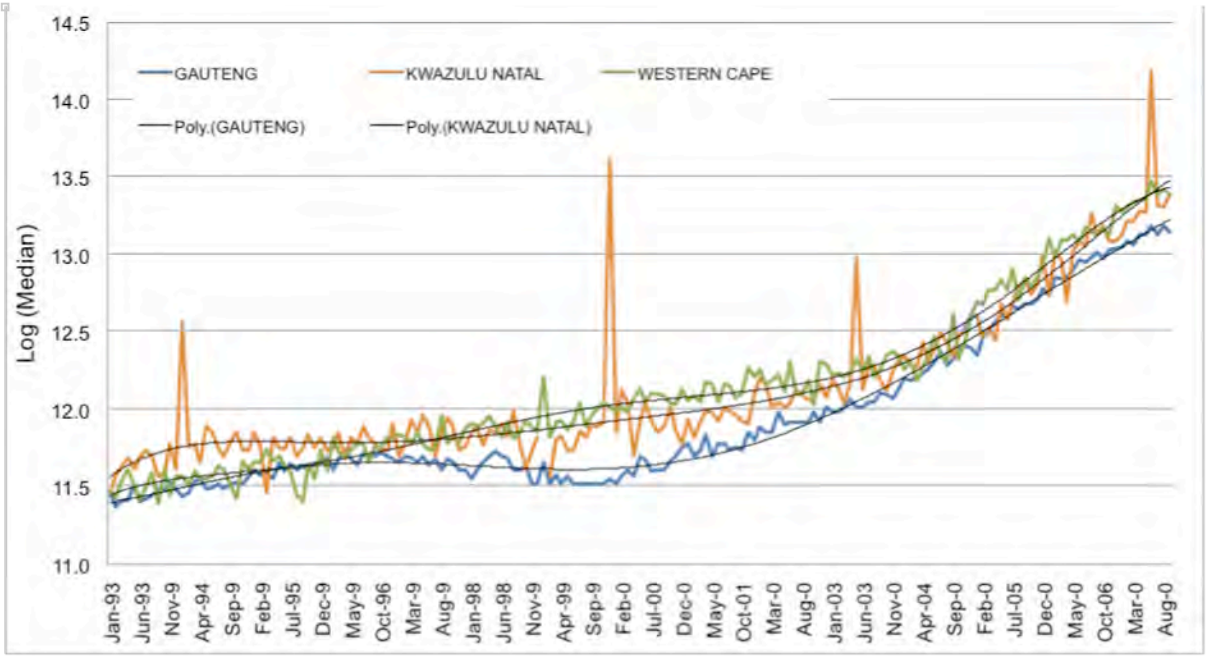


Model 6: Group C – Minor Provinces

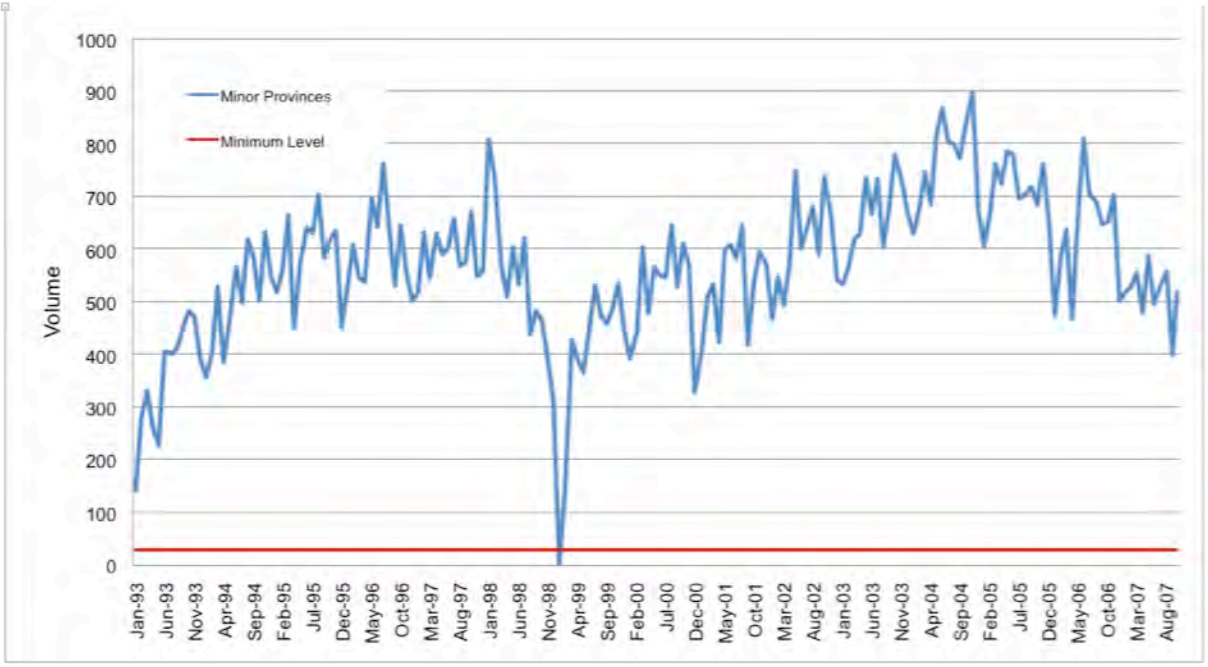
Model 7: Group D – All Provinces

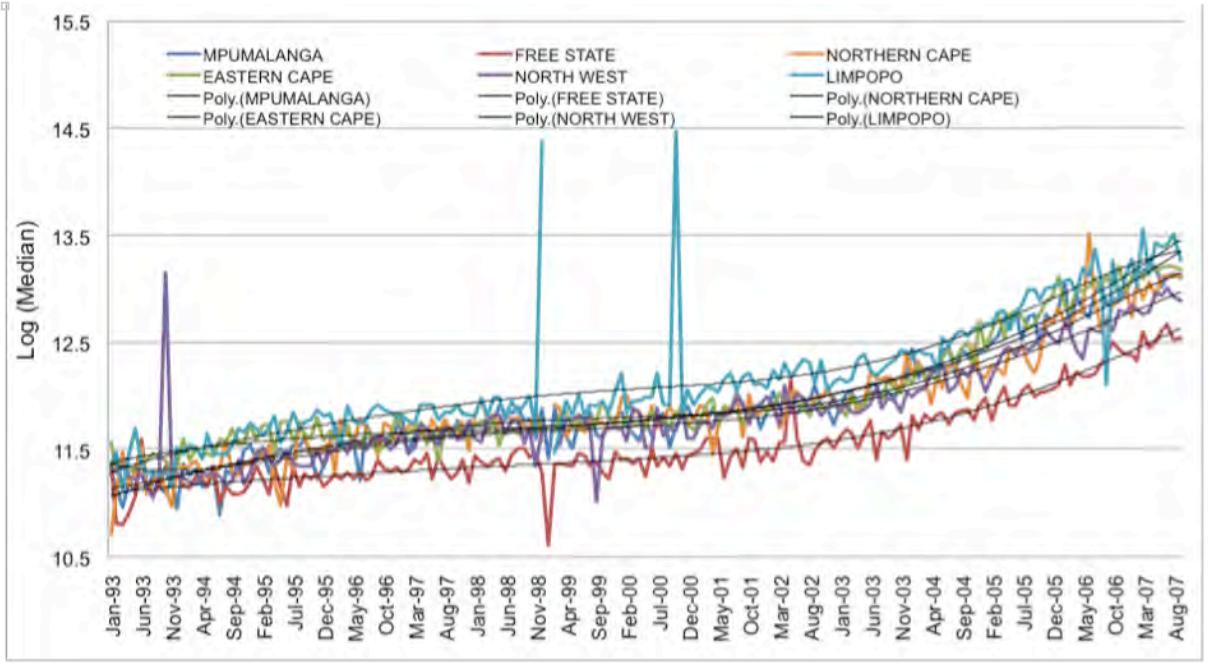


Model 8: Group E – Major Provinces

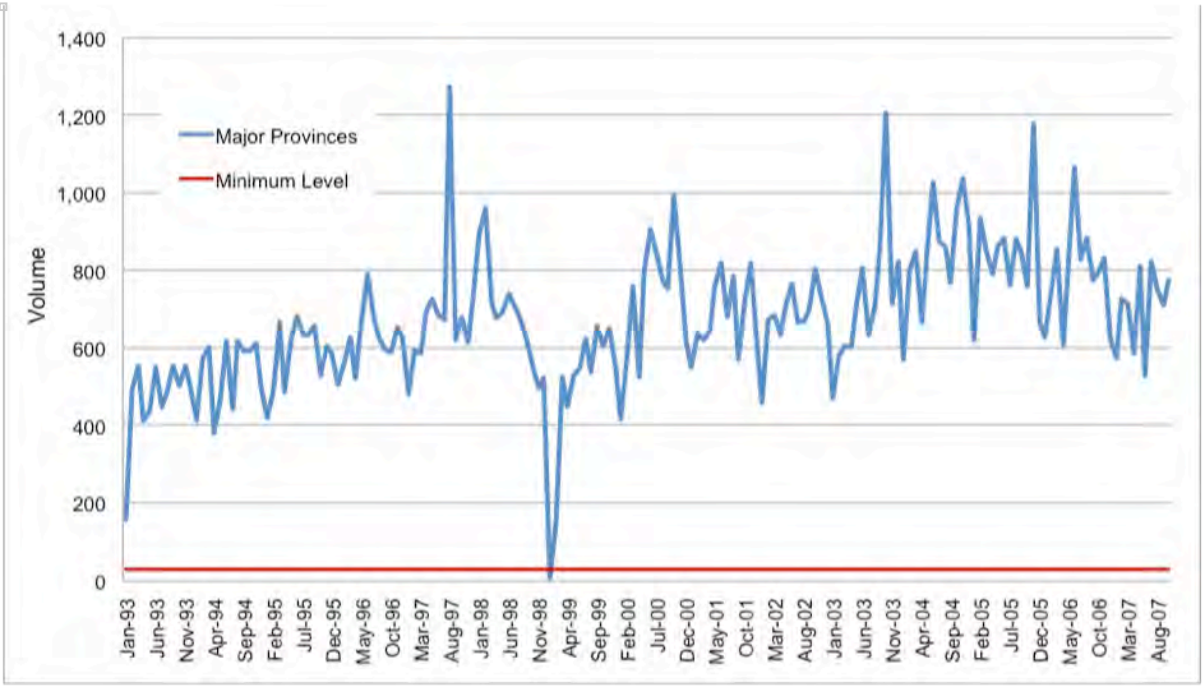


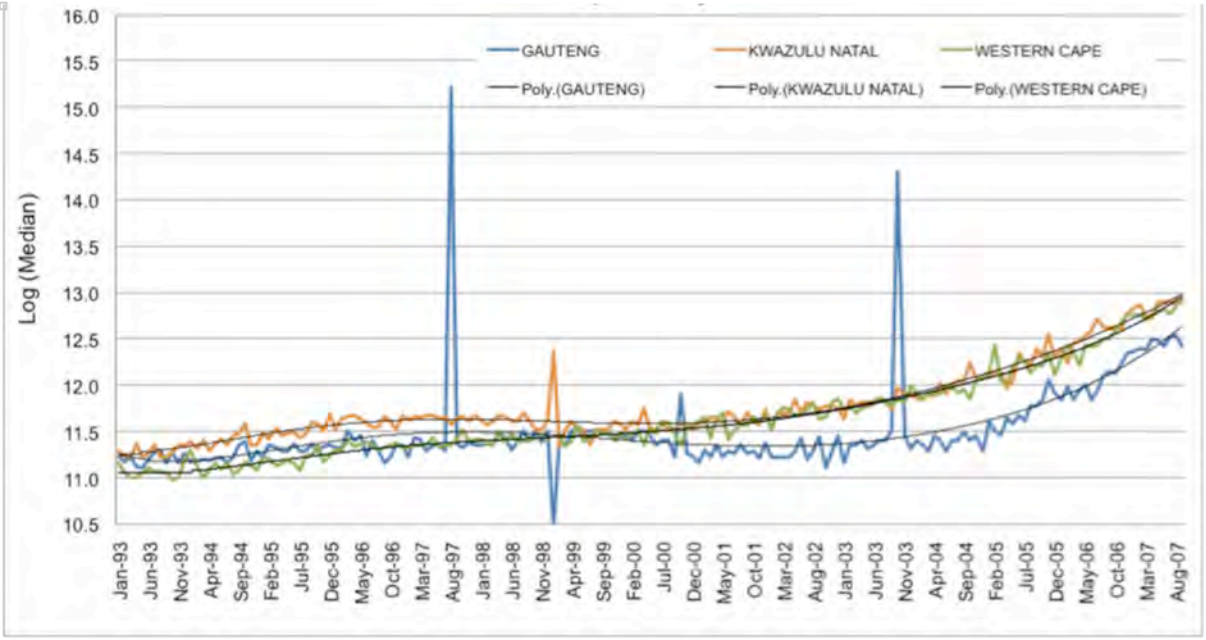
Model 9: Group E – Minor Provinces



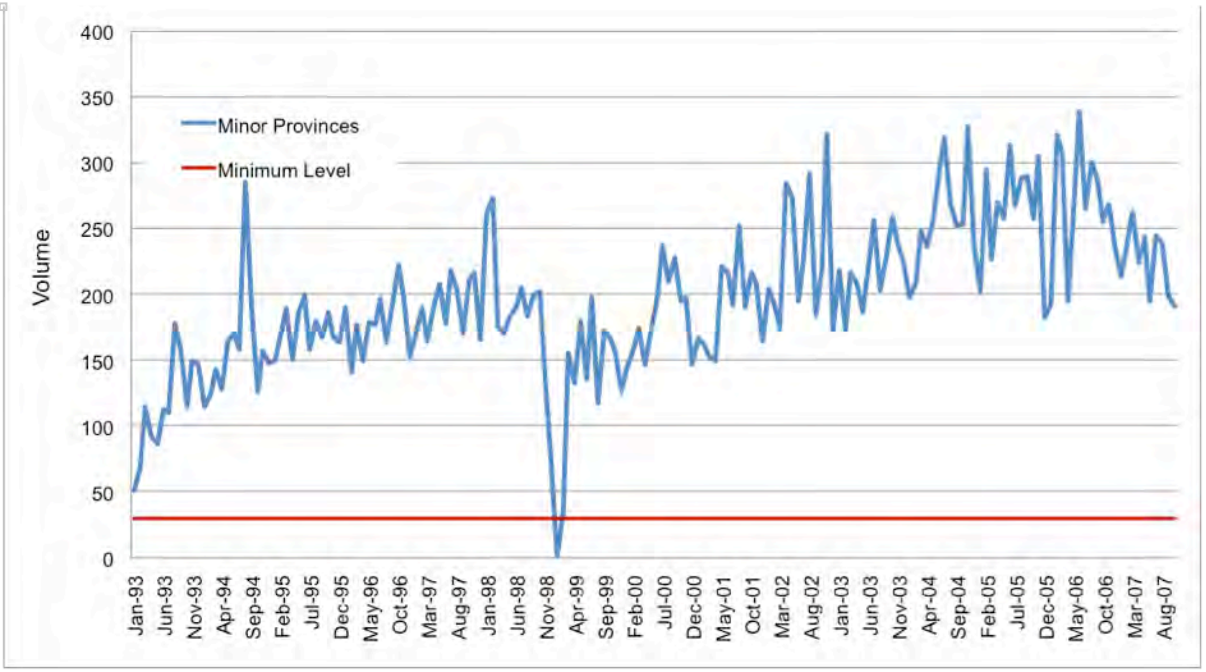


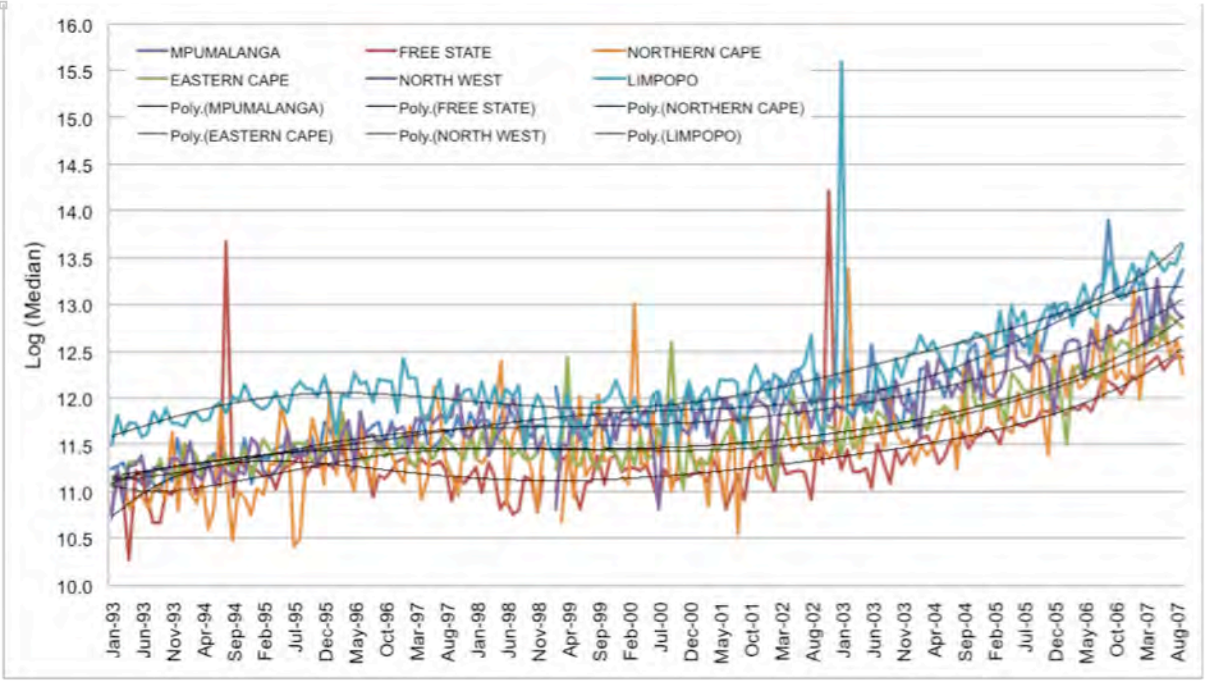
Model 10: Group F – Major Provinces



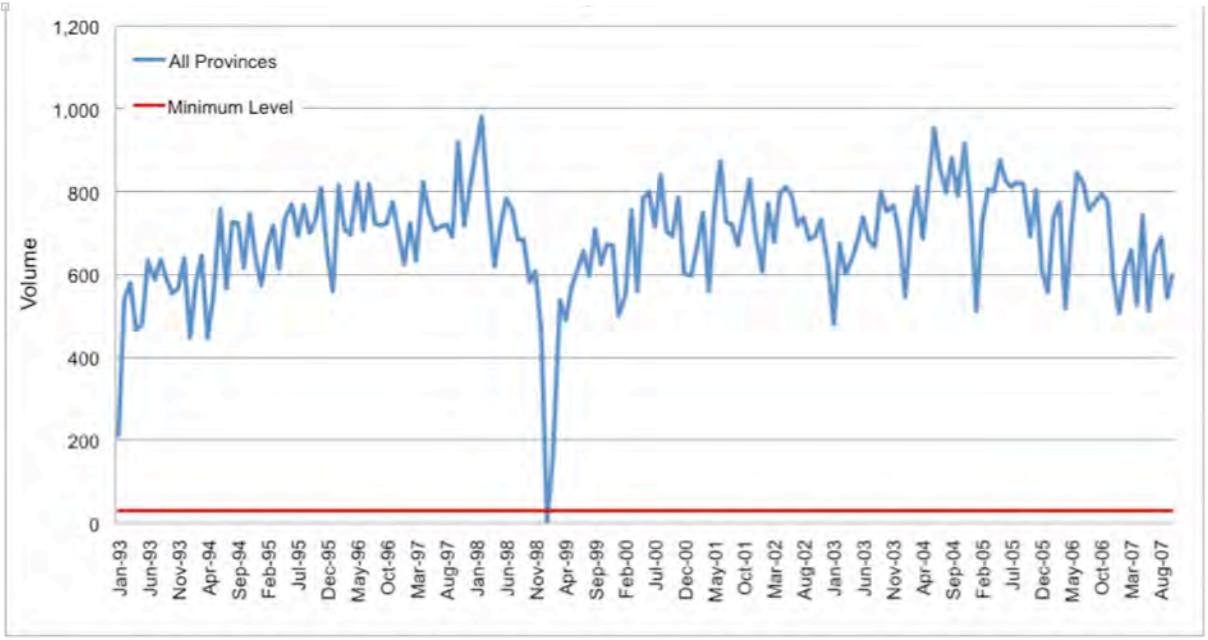


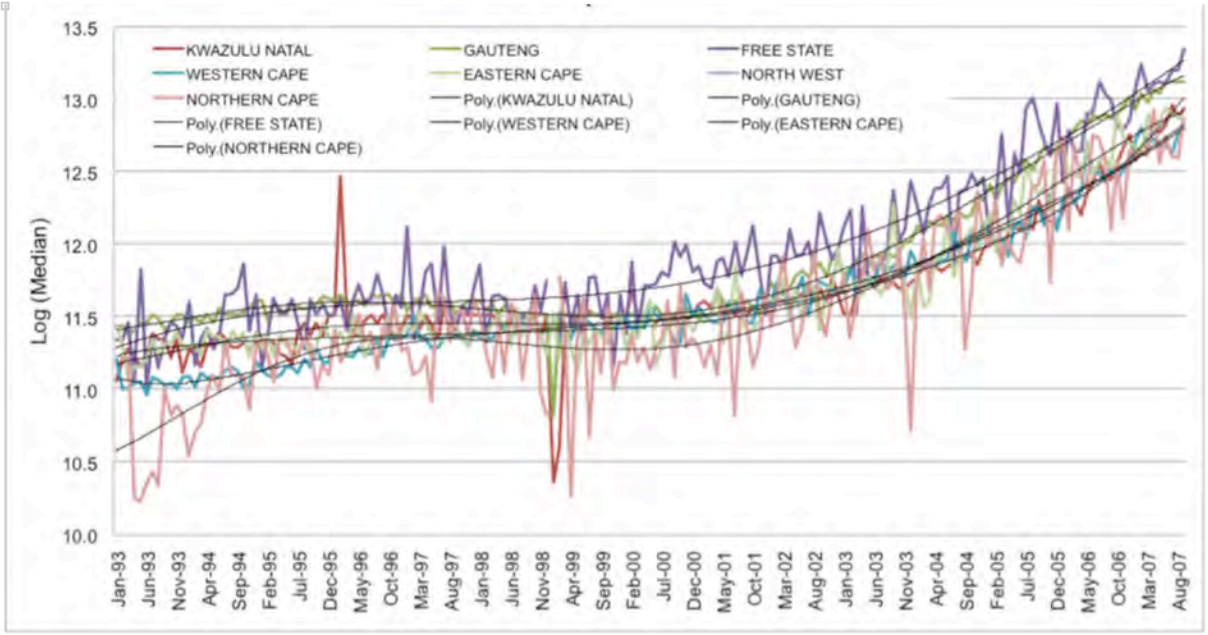
Model 11: Group F – Minor Provinces





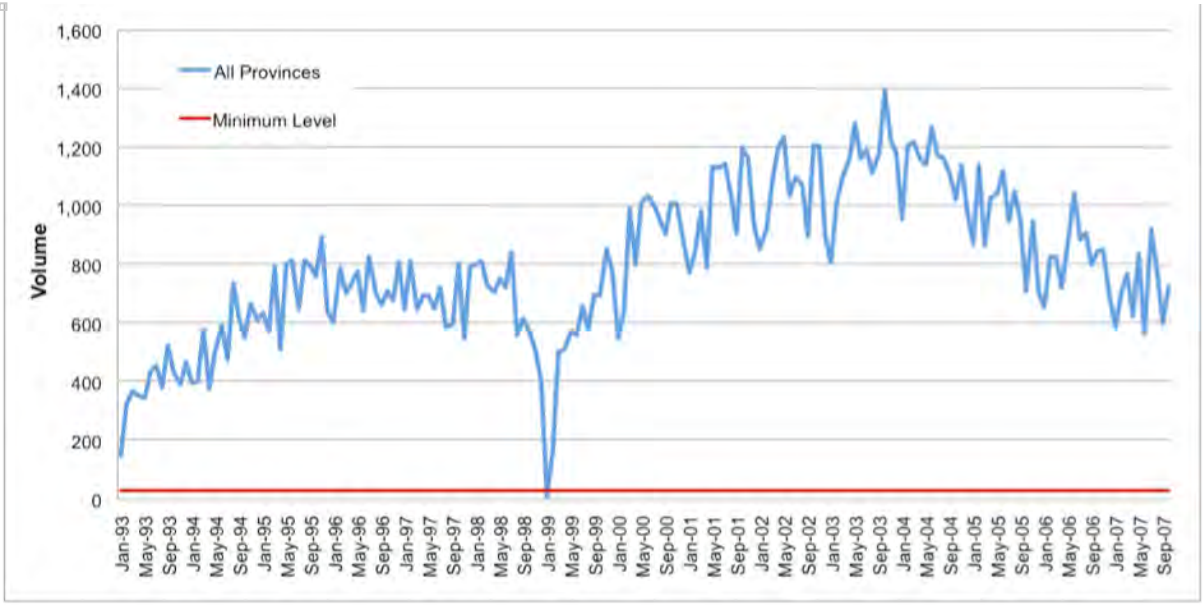
Model 12: Group H – All Provinces

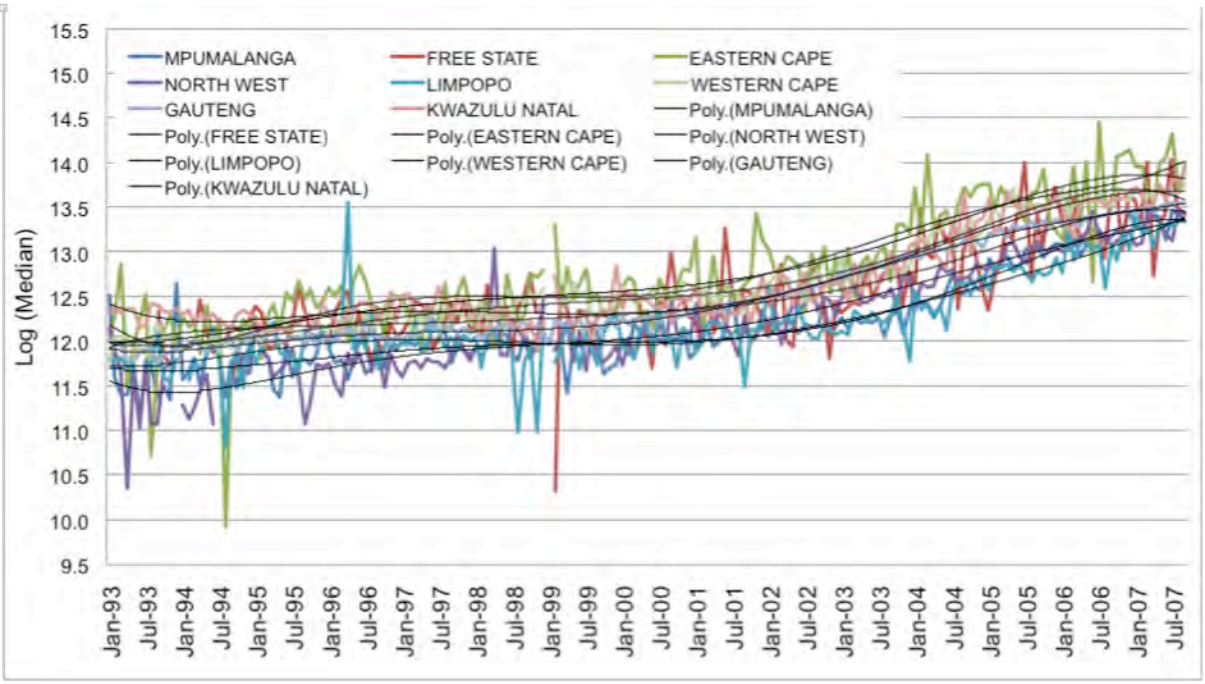




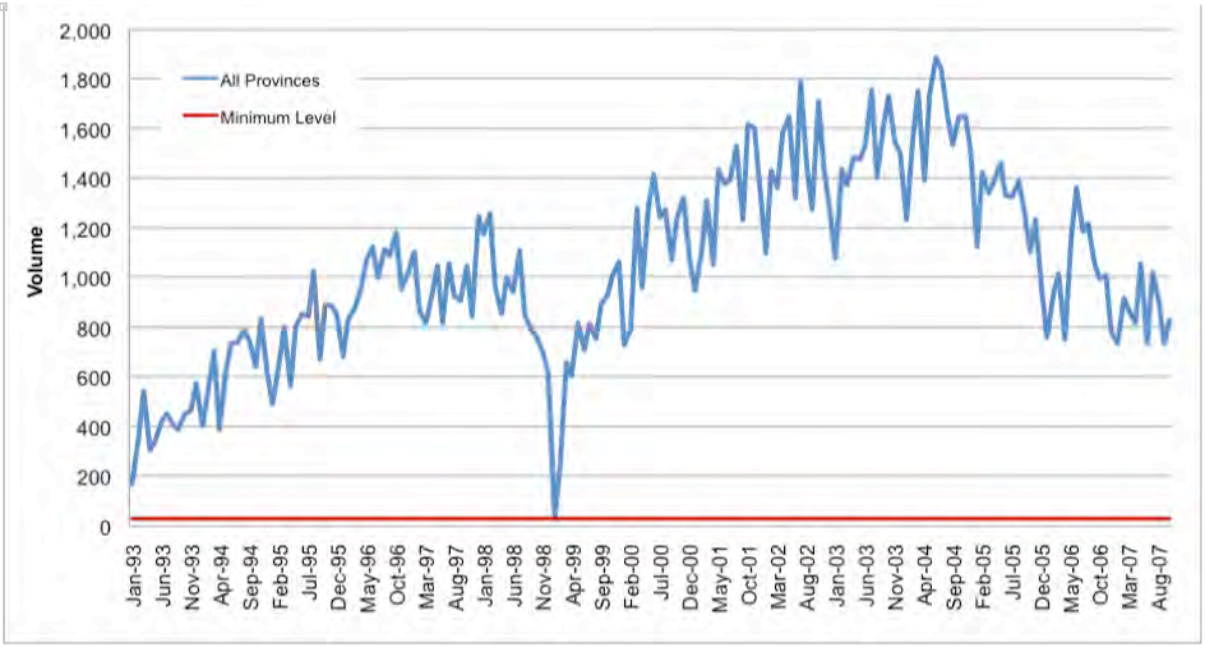
Sectional Title Properties

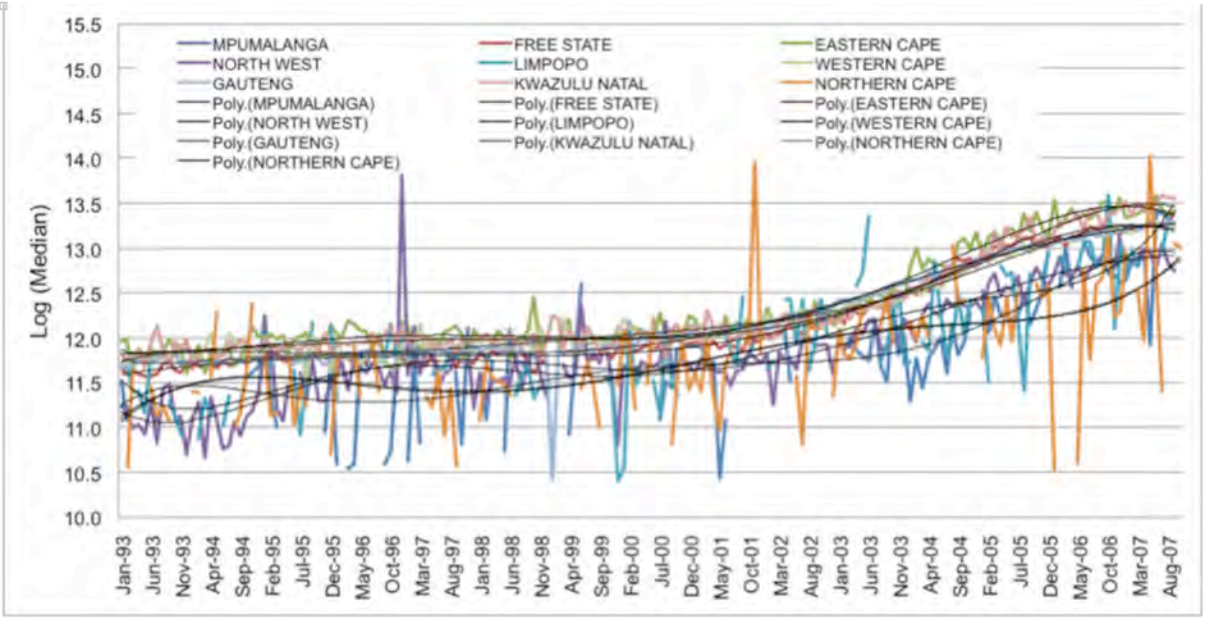
Model 1: Group A – All Provinces



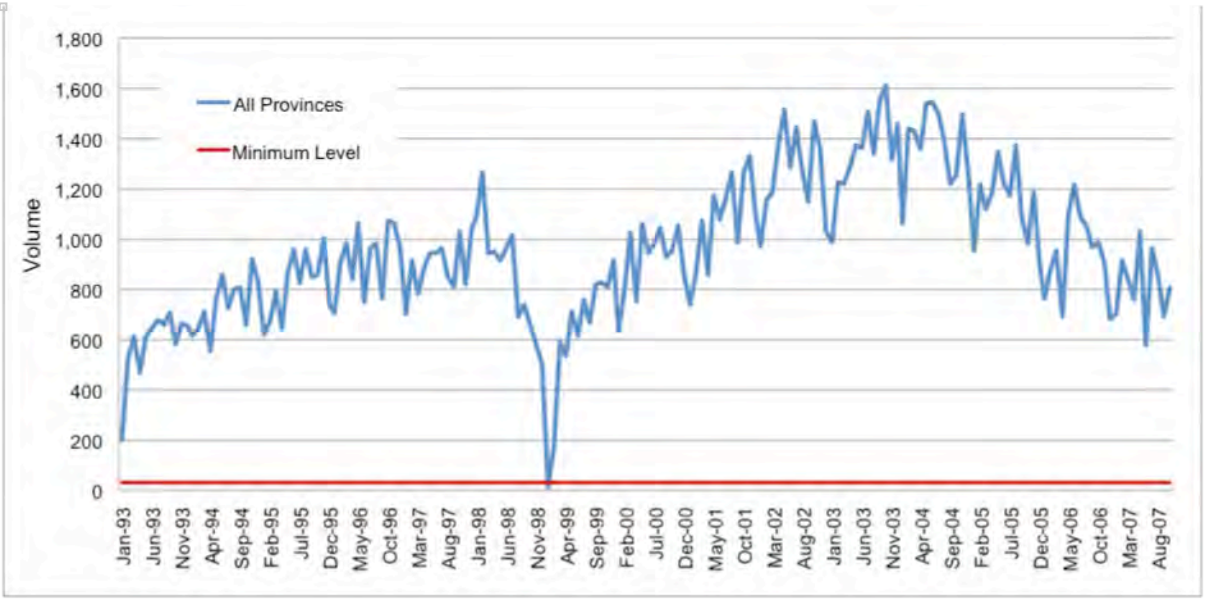


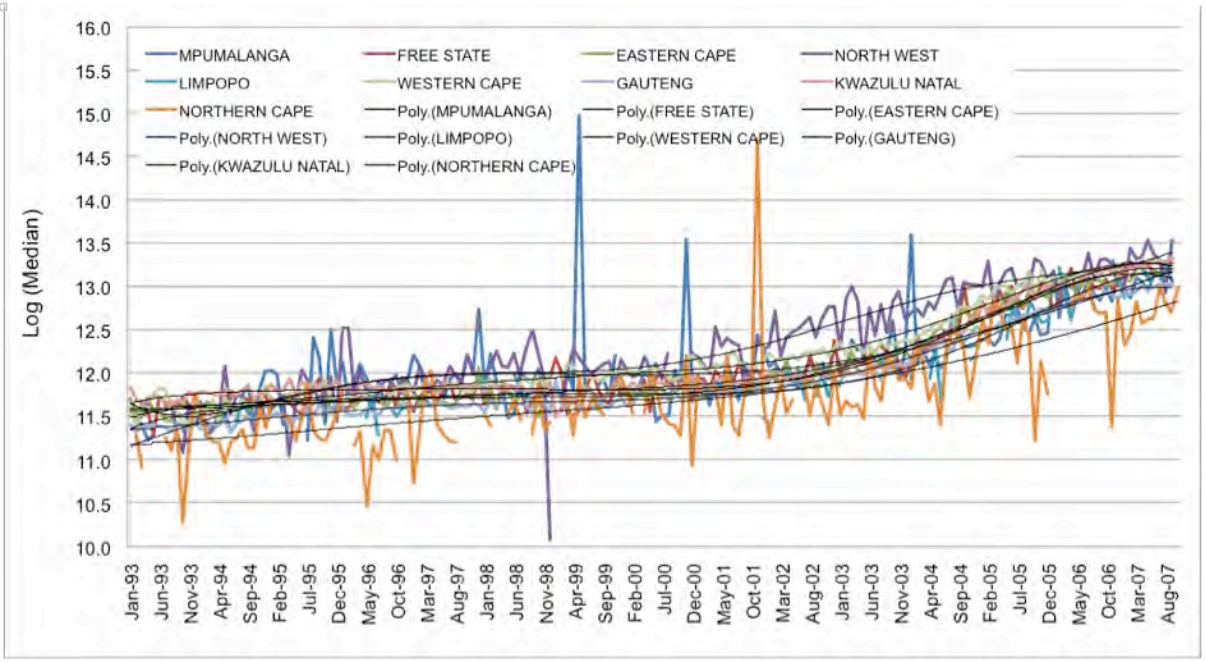
Model 2: Group B – All Provinces



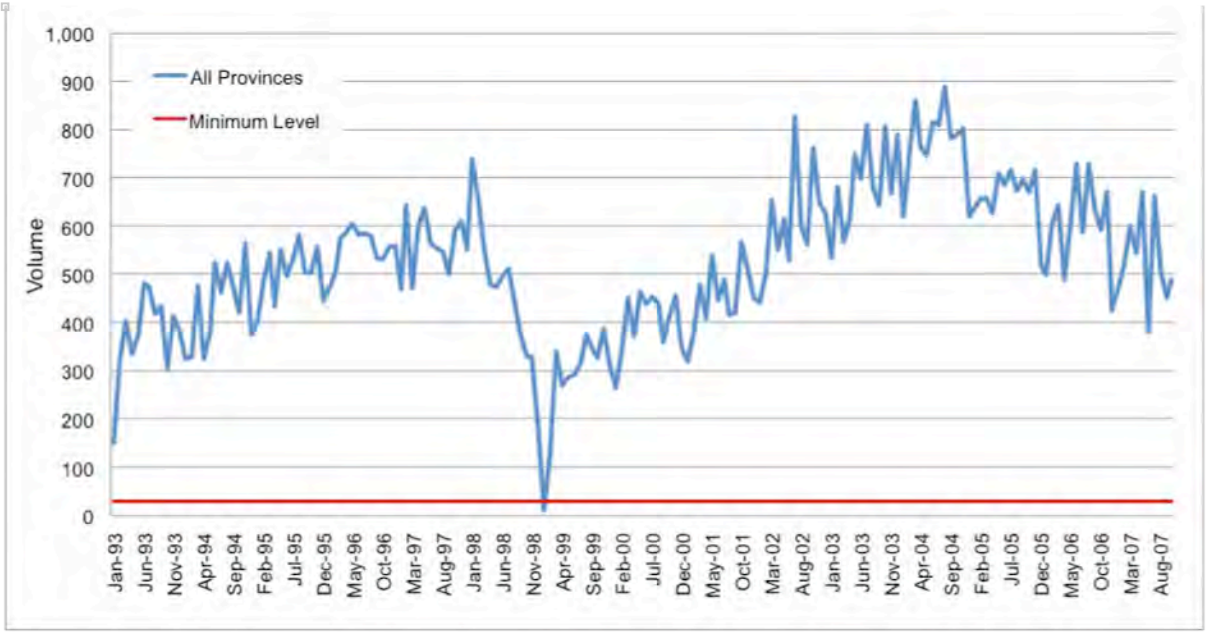


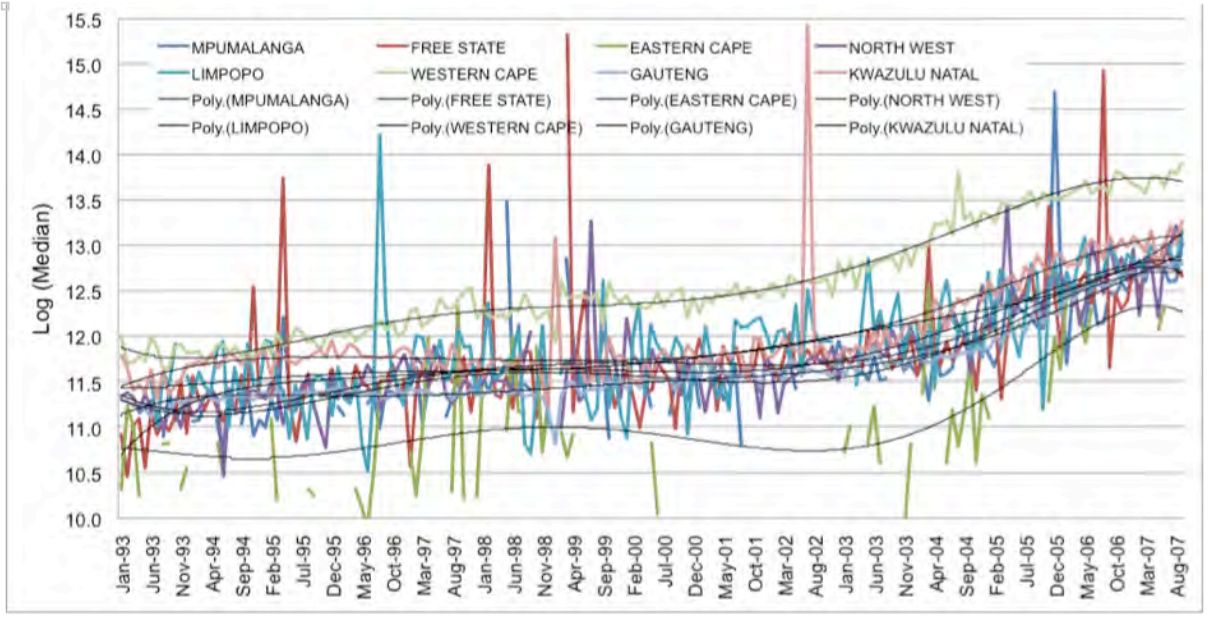
Model 3: Group C – All Provinces



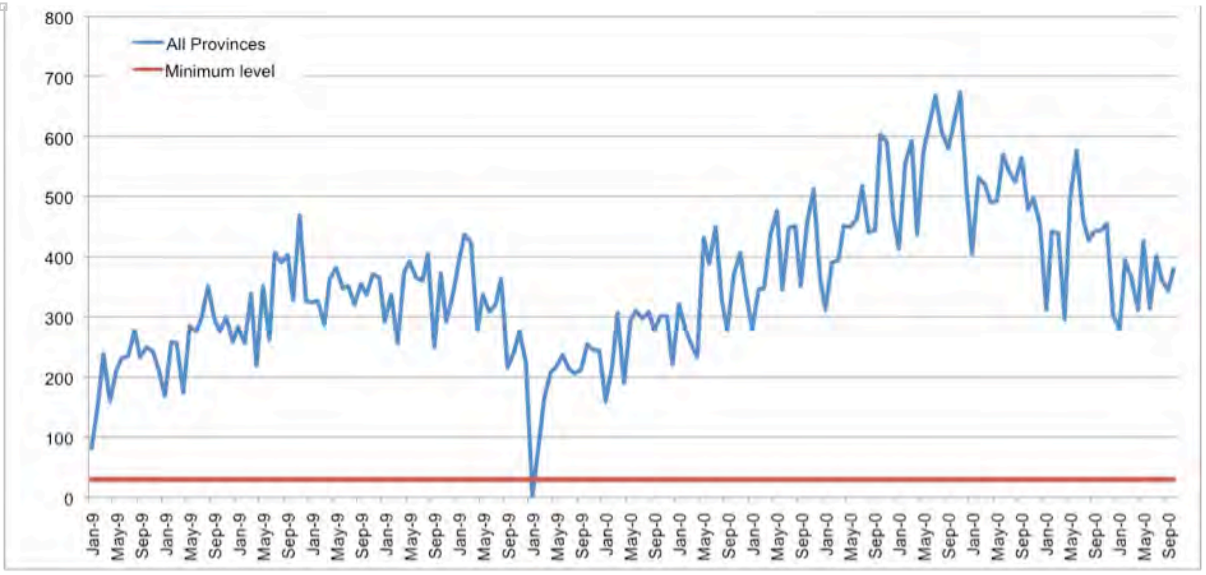


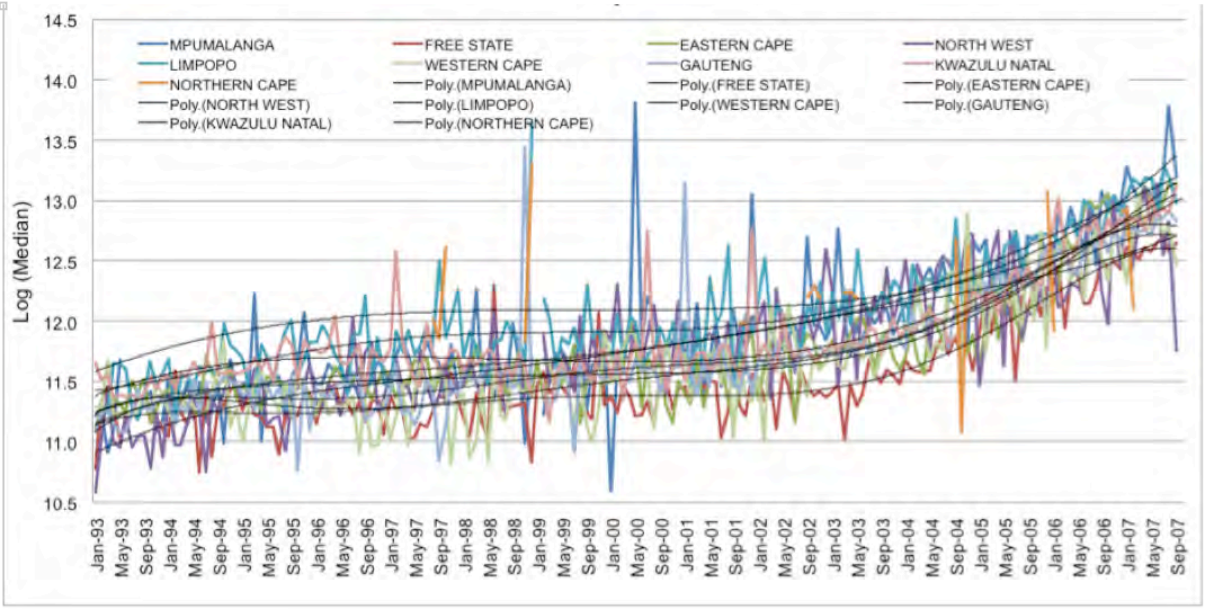
Model 4: Group D – All Provinces





Model 5: Group E and F – All Provinces





Model 6: Group H – All Provinces