A PROPERTY VALUATION MODEL FOR RURAL VICTORIA

Kelly Nicole Hayles

Associate Diploma of Engineering (Surveying & Mapping)
Bachelor of Applied Science (Land Information)

Thesis submitted for the fulfilment of the degree
Doctor of Philosophy

School of Mathematical and Geospatial Science
RMIT University
GPO Box 2476V
Melbourne, Victoria 3001
Australia

Date: 29 June 2006

Declaration

This thesis contains no material which has been accepted for the award of any other higher degree or graduate diploma in any tertiary institution. To the best of my knowledge and belief this thesis contains no material previously published or written by another person except where due reference has been made in the text of this thesis. Furthermore, the work presented has been carried out since the official commencement of the program.

Kelly Hayles

Dedication

This thesis is dedicated to the memory of Dr Rodney Allan, who always had great beliefs in my abilities and achievements, but sadly is not here to see the submission of this thesis.

Thank you Rod.

Acknowledgements

Acknowledgments must be given to Spatial Vision P/L, Department of Infrastructure, Natural Resources and Environment (NRE), Land Victoria, West Gippsland and Yarriambiack Shires for the supply of the various data used throughout this study.

Many thanks must also go to the staff and post graduates at RMIT – School of Mathematical and Geospatial Science for their support and assistance throughout the many years of this research.

To Dr Ron Grenfell, I appreciate your expertise and your willingness to help overcome the many data integration problems encountered during this research, and your support during this time. Thank you for your persistence.

To Professor Tony Norton, thank you for giving me the determination to keep going and not letting me give up, even though I was half a world away. You have such a wonderful professionalism in all that you do and such a caring attitude to those that come by you.

Thanks must go to Dr David Silcock for the emotional support nearing the end, I can't wait to send you that long awaited email to say 'it's submitted'.

To Dr Daniel Kildea, thanks for your patience, determination and enthusiasm for my project and 'real estate data'. It was wonderful to work with you and to have someone so vibrant and enthused about my work. Thank you for the time you put aside for me whilst in France, it was a wonderful help and gave me the determination to going, especially the French cuisine.

To Peter Woodgate, I appreciate your time and willingness to provide different views and aspects to consider within this project.

To Dr Connie Spinoso, I am greatly appreciative of your valuation and GIS experience and for being there when needed, even after long lapses between consultations.

To Kathryn and Tim Hailes, thanks for your good humour, support and offers of assistance to read this thesis or help in any way, they were all greatly appreciated.

I wish to thank my husband, for his beliefs in what I can do both professionally and personally and for his endurance throughout my studies. Thank you Cameron for travelling halfway around the world to follow my dreams and aspirations and for putting up with those cold English winters for me. Lastly I wish to thank Jordan for providing a welcomed distraction. I wish that I could of spent less time on the computer working and more time playing with you. With this chapter of my life over, I can at least focus on giving you the time you have needed and deserved over the last 4 years.

Abstract

Licensed valuers in the State of Victoria, Australia currently appraise rural land using manual techniques. Manual techniques typically involve site visits to the property, liaison with property owners through interview or the use of questionnaires, and require a valuer experienced in agricultural properties to determine a value. The use of manual techniques typically takes longer to determine a property value than for valuations performed using automated techniques, providing appropriate data are available. Furthermore, manual methods of property valuation can be subjective and lead to bias in valuation estimates, especially where valuers have varying levels of experience and knowledge about a specific regional area. The use of more automated techniques may help reduce subjectivity, reduce the time spent valuing property and minimise the need for site visits. Automation of the property valuation process may lend itself to more accurate valuation estimates by providing greater consistency between valuations. Automated techniques presently in use for property valuation were reviewed and include artificial neural networks, expert systems, case based reasoning and multiple regression analysis. The latter technique appears most widely used for valuation.

The aim of my research is to develop a conceptual rural property valuation model, and to develop and evaluate quantitative models for rural property valuation based on the variables identified in the conceptual model that appear most likely to influence price. The conceptual model was developed by examining peer research, Valuation Best Practice Standards, a standard in use throughout Victoria for rating valuations, and rural property valuation texts that consider the more significant property characteristics commonly used for rural valuation. The research uses this conceptualisation to

examine the extent to which numerical models can be developed within a rural Victorian setting. Using data that are only available digitally and publicly, the research assessed this conceptualisation using properties from four LGAs in the Wellington and Wimmera Catchment Management Authority (CMAs) areas in Victoria. Cluster analysis was undertaken to assess if the use of sub-markets, that are determined statistically, can lead to models that are more accurate than sub-markets that have been determined using geographically defined areas. I discuss the impact of these statistically determined sub-market groupings and whether future hedonic research should determine sub-markets using this technique.

The research is divided into two phases; the 'available data phase' and the 'restricted data phase'. The 'available data phase' used publicly available digital data to build quantitative models to estimate the value of rural properties. The 'restricted data phase' used data that became available from Land Victoria near the completion of the research to develop additional models of rural property values.

The research examined the effect of using statistically derived sub-markets as opposed to geographically derived ones to estimate property values. Cluster analysis was used during both phases of the model development. During the 'available data phase' one of the clusters developed was superior in its model prediction compared to the models produced using geographically derived regions.

My research indicated a number of limitations with the digital property data available for Victoria. Although GIS analysis can enable more property characteristics to be derived and measured from existing data, it is reliant on having access to suitable data in digital

form. My research also identified limitations with the metadata elements in use in Victoria (ANZMETA DTD version 1).

It is hypothesised that to further refine the models and achieve greater levels of price estimation, additional properties would need to be sourced and added to the current property database. It is also suggested that additional research needs to address the issues with sub-market identification and the development of a more effective technique to segregate properties such that they are segregated both geographically and also using property characteristics. If results of additional modelling indicated significantly different levels of price estimation, then these models could be used with manual techniques to evaluate manually derived valuation estimates. However, it is envisioned that automated techniques to model rural property values would not be used as the sole valuation technique even if the accuracy of the models were to be improved dramatically.

Table of Contents

List of Appendices

List of Tables

List of Figures

List of Equations

Glossary and List of Acronyms

ABS	Australian Bureau of Statistics
AGD66/84	Australian Geodetic Datum 1966 or 1984, a datum derived from a least squares adjustment to obtain latitude and longitude coordinates
AMG	Australian Map Grid is a grid based coordinate system derived from a Universal Transverse Mercator projection on either the AGD66 or AGD84
ANN	Artificial Neural Network – a form of artificial intelligence which trains data by studying past relationships and patterns between data
ANZLIC	Australian and New Zealand Land Information Council
ASDTS	Australian Spatial Data Transfer Standard
AVM	Automated Valuation Model
CAMA	Computer Assisted Mass Appraisal
CBR	Case Based Reasoning
Centroid (of a polygon)	Is the notional middle of a polygon
CMA	Catchment Management Area
COD	Coefficient of Dispersion
COV	Coefficient of Variation

CPI	Consumer Price Index
DSS	Decision Support System
Euclidean Distance	The straight line distance between two points
ESRI	Environmental Systems Research Institute
ESRI ArcInfo	A command line based GIS software package developed by ESRI
ESRI ArcView	A Windows based GIS software package developed by ESRI
FME Universal Translator	Feature Manipulation Engine - A Windows based software package which translates between a variety of GIS software formats
GIS	Geographical Information System – a tool used for the analysis, management and display of spatial information
Highest and best use value	'the highest and best use to which the land might reasonably be expected to be put at the relevant time and to any potential use' (*Valuation of Land Act 1960*)
HTML	Hyper Text Markup Language
LGA	Local Government Area – a system used throughout Australia to divide each state into a number of areas with each managed as a local council
LVR	Loan to Value Ratio – used by mortgage lenders to ascertain the amount of borrowing compared to the value

	of the borrowings
MapInfo	A Windows based GIS software package developed by MapInfo Corporation
MCDM	Multi-Criteria Decision Making
Minitab	A statistical and graphical analysis package
MRA	Multiple Regression Analysis
Parcel	The smallest piece of land which is defined by cadastral boundaries that is able to be transferred within Victoria. A parcel can contain multiple properties.
PRISM	Property Information and Sale Data – a digital data set which details sale prices and property characteristics for Victoria
Property	Can consists of multiple land parcels, a single parcel or part of a parcel
Property Valuation Estimate	A property valuation that is obtained utilising a database of past sale prices to arrive at a valuation figure at a current point in time
Property Valuation Forecast	A property valuation that is obtained utilising past sale prices and valuation estimates to produce a series of valuation estimates into the future
P value	Indicates the level of significance of the independent variable
R^2	A measure of the fit of a regression equation

RPVM	Rural Property Valuation Model
SDRN	State Digital Road Network - a digital data set comprising the road network for Victoria
SDTS	Spatial Data Transfer Standard
SE coefficient	The Standard Error in the coefficients
SLTVF	State Land Tax Valuation File – a digital data set which details council rating valuations for individual properties
T value	Is a ratio of the coefficient and the SE coefficient
Unix	A command line based computer operating system
Value Drivers	Data elements of variables used to determine a valuation
VBP	Valuation Best Practice
Windows	A graphical based operating system developed by Microsoft Corporation

Chapter 1 General Introduction

1.1 Introduction

Licensed valuers in the State of Victoria, Australia currently appraise rural land using manual techniques. Manual techniques typically involve site visits to the property, liaison with property owners through interview or the use of questionnaires, and require a valuer experienced in agricultural properties to determine a value. The use of manual techniques typically takes longer to determine a property value than for valuations performed using automated techniques, providing appropriate data are available (Waller, 1999). Furthermore, manual methods of property valuation can be subjective and lead to bias in valuation estimates, especially where valuers have varying levels of experience and knowledge about a specific regional area (Bonissone & Cheetham, 1997). The use of more automated techniques may help reduce subjectivity, reduce the time spent valuing property and minimise the need for site visits. Automation of the property valuation process may lend itself to more accurate valuation estimates (Nattagh & Ross, 2000) by providing greater consistency between valuations.

Multiple Regression Analysis (MRA) and expert systems are employed in residential valuation, and are increasingly used for rural valuation. Some residential automated models are obtaining valuation results close to or superior to those produced manually (Gardner & Barrows, 1984; Xu *et al.*, 1993; Faux & Perry, 1999). Even so, automated methods have their limitations (Isakson, 2001, Detweiler & Radigan 1999). The quality and availability of digital data are key issues surrounding the ability of an automated model to perform valuations accurately. Another consideration is the selection of which variables to include within a model. Like manual techniques, automated

techniques may be subject to some form of error as they are effectively trying to mimic the complex processes undertaken by a human valuer.

Automated techniques are increasingly based on the theory of hedonic pricing said to be pioneered by Rosen (1974) in which implicit prices for particular variables are determined. Hedonic pricing theory has been combined with regression analysis in rural areas to study the affects that wetland areas have on rural values (Reynolds & Regalado, 2002), the impact of soil conservation (Gardner & Barrows, 1984) and the value of irrigated water (Faux & Perry, 1999). Each of these valuation studies have had varying results in terms of their accuracy levels.

The automation of the valuation process can take a variety of forms. A more simplistic approach can use GIS for mapping and visualisation. The creation of valuation models using Multiple Regression Analysis (MRA) or 'expert system applications' such as Case Based Reasoning (CBR) (Gonzalez & Laureano-Ortiz, 1992; McSherry, 1993) and Artificial Neural Networks (ANN) (Lenk *et al.*, 1997; Connellan & James, 1998; McGreal *et al.*, 1998) are more advanced applications of the automation process.

The use of GIS for valuation has increased over time and led to wider, more varied applications within the valuation industry. Some of these applications include the analysis of past sales information by using visual displays of past results, checking of anomalies in current valuation results, and supporting a greater understanding of the land being valued through visual analysis of the characteristics of neighbouring properties (Valuer-General Victoria, 2002). GIS has been used in valuation research for developing 3-dimensional surfaces representing property values (Ward *et al.*, 1999)

and for determining additional spatially-derived characteristics such as the geographic proximity of a property from town services (Rosiers *et al*., 2000).

Presently, the valuations for municipal rating and land taxation are performed every two years in Victoria and are undertaken in five key stages. The final stage is required to be submitted one year after the commencement of the valuation (Valuation Best Practice, 2005). The valuation process is tendered to Licensed Valuers who are required to produce municipal valuations following Valuation Best Practice (VBP) standards (Valuation Best Practice, 2005).

The VBP standards in Victoria specify the procedure and outcomes of the valuation process and the housing and property characteristics to be used when valuing property (Valuation Best Practice, 2005). Although the specifications outline the mandatory variables to be determined during the valuation process, not all variables are necessarily used to determine a final value. These variables depict property owner information that is used during the rating process along with additional variables such as reference numbers, lot, plan and section numbers and standard parcel identifiers. Whilst the characteristics included as part of the VBP guidelines relate to individual property, data pertaining to the variables are restricted in that they are not available to the general public for purchase. The information contained within VBP is relatively comprehensive and includes road frontage, pasture condition codes, vegetation type and soil types along with water access and water rights information. Past values for site value, capital improved value and net annual value are also included within the VBP property characteristic data elements.

The Property Sales and Information (PRISM) data set is the only available data in Victoria that contains sale price information at an individual property level. PRISM is not available publicly. Although PRISM is a restricted data set, it was made available for this research. The PRISM data set contains information regarding sale price, sale date, street number, street name, suburb, postcode, land use, crown allotment details, property area and a Melway map reference (a page numbering and grid based referencing system used in the Melway Street Directory) (Ausway, 2004). Typically, rural properties are valued using a combination of housing attributes (Valuer-General Victoria, 2002), although at present many of these are not contained within PRISM. Some of the variables for property valuation can be derived by obtaining additional data sets held by the State Government and performing spatial overlays with GIS tools. This technique can assist in enhancing limited data sets.

Within the VBP framework, GIS can be used to show changes between valuations of previous years, locate properties that have recently sold and to map property values. Although statistical analysis of previous valuations is performed to determine the reliability of results and detect outliers in valuations, the municipal rating process is still largely manual and involves site inspections within municipalities (Valuation Best Practice, 2005).

Currently four types of computer assisted valuation software are used throughout Victorian Local Government Areas (LGAs) for residential valuation, with only one of these utilising regression (Brett Reed 2003, pers. comm., 18 September). Three of these software packages use a series of look up tables to define each value driver for residential valuation and were developed through consultations with valuers (Brett Reed 2003, pers. comm., 18 September). To date, value drivers have not been

derived for rural markets due to the diversity of the market, nor have regression analyses been applied in a rural setting in Victoria. Thus, the automation of techniques to estimate the value of rural property in Victoria can be considered to be in its infancy.

1.2 Research Aims

The aim of my research is to develop a conceptual rural property valuation model, and to develop and evaluate quantitative models for rural property valuation based on the variables identified in the conceptual model that appear most likely to influence price. The research uses this conceptualisation to examine the extent to which numerical models can be developed within a rural Victorian setting. Using data that are only available digitally and publicly, the research tests this conceptualisation using properties from four LGAs in the Wellington and Wimmera Catchment Management Authority (CMAs) areas in Victoria. Cluster analysis was undertaken to determine if the use of sub-markets, that are determined statistically, can lead to models that are more accurate than sub-markets that have been determined using geographically defined areas. I discuss the impact of these statistically determined sub-market groupings and whether future hedonic research should determine sub-markets using this technique.

The research aims to enhance existing rural valuation methods by helping to automate the valuation process to develop numerical models using statistical regression analysis. The research aims to provide a greater understanding into the limitations of automation for rural valuation in Victoria and the degree of dependency of the process on available data. In addition, I consider current problems associated with data integration techniques and provide a framework for use during data integration and database development. The property valuation estimate determined, is based on the 'highest

and best use' to which the land might reasonably be expected to be put at the relevant time, and to any 'potential use' (*Valuation of Land Act 1960*).

Based on the above aims, the following hypotheses were formulated:

- that statistical regression techniques can be used to accurately estimate property values in rural Victoria
- that using statistical measures to re-group properties into sub-market areas will lead to higher modelling accuracy than numerical models which primarily segregate property into sub-markets using geographical techniques.

In order to test the above hypotheses, the following objectives were identified:

a) Examine the issues associated with the integration of numerous and disparate data sets for determination of a rural property valuation estimate based on the highest value and best use of land.

b) Develop and discuss an integration framework that can be used in the development of a model to determine a rural property valuation estimate based on the highest value and best use of land.

c) Examine the techniques used for both the manual and automated determination of property valuation estimates based on the highest value and best use of land.

d) Examine the value drivers and their degree of significance for the determination of rural property valuation estimates based on highest value and best use of land.

e) Create a database containing individual property variables within the four LGAs and use GIS analysis to perform spatial overlays to derive additional characteristics.

f) Test and evaluate the effectiveness of regression analysis for rural property valuation estimation utilising the database created in e).

g) Use cluster analysis to create statistically derived regions (or groupings). Re-test the initial regression models developed for each of the clustered regions.

h) Evaluate the effectiveness of regression modelling using 1) a-priori geographically defined regions and 2) statistically derived regions (or groupings) using cluster analysis to ascertain whether unconstrained statistical techniques can provide more accurate property valuation estimates than a-priori techniques for automated rural valuation in Victoria.

1.3 Research Scope and Approach

The research is applicable to rural property valuation in Victoria. Due to time constraints, it was not possible to consider the application of this modelling approach in other jurisdictions, although this is considered fertile ground for additional research within this field. Due to the research requirement to use digital data, the number of property characteristics implemented within the numerical models was somewhat less than that represented in the conceptual model. It should be noted that 'restricted-use' digital data became available during the latter stages of the research and as such additional processing was undertaken to ascertain the effects of modelling with the inclusion of these variables.

The development of numerical models and subsequent testing was undertaken for selected properties in the four Victorian LGAs: Horsham, Yarriambiack, Northern Grampians and West Gippsland. An internal validation process was used to verify and determine the accuracy of the regression results obtained from the models.

Lastly, numerical valuation models were developed to evaluate the manual valuation techniques currently used by licensed valuers.

1.4 Research Methods

The research involved five main stages; selection of the study areas, conceptualising the value of rural property in terms of measurable land characteristics, database development, development of the numeric rural property valuation models using 'a-priori' LGA sub-markets, and cluster analysis to determine sub-markets using statistical methods.

Stage one involved the *selection of the two study areas* in Victoria. This involved selecting regions where agricultural productivity was high and where there was some form of land degradation problems apparent within the regions. Stage two, the *conceptual model* stage involved reviewing existing research to determine suitable land characteristics that were applicable to rural valuation in Victoria. This review encompassed research on rural valuation and examined the frequency of use of specific property characteristics and their significance for use in rural valuation regression modelling. Stage three, the *database development* stage involved the creation of a database using the variables conceptualised to model rural property prices from stage two. Stage four involved the *development and testing of the numeric*

rural property valuation models using LGAs to assign properties to particular geographic areas. Stage five involved the use of *cluster analysis* to re-define properties into more homogenous groupings using statistical measures. These five stages are outlined below.

1.4.1 Stage 1: Study Area Selection

Stage one involved analysis of the agricultural regions in Victoria, Australia. The Wimmera and West Gippsland Catchment Management Authorities (

Figure 1.1) support relatively high agricultural productivity compared to some other regions in Victoria. The occurrence of land degradation and salinity is increasing in both CMAs and has a significant detrimental affect on agricultural productivity (Wimmera Catchment Management Authority, 2000; West Gippsland Catchment Management Authority, 2001). Each CMA is comprised of a number of LGAs. In this study the LGAs of Horsham, Yarriambiack, Northern Grampians and Wellington were the smaller geographical areas from which the majority of data was obtained for valuation modelling.

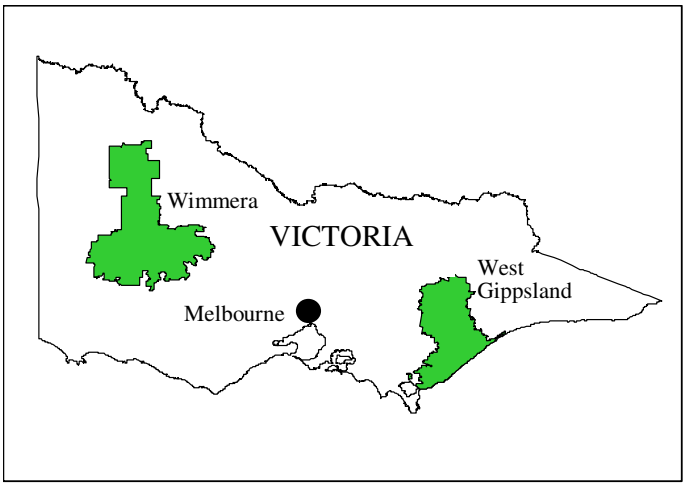

Figure 1.1 Wimmera and West Gippsland CMAs indicating the LGA extent of the two study areas

1.4.2 Stage 2: Conceptualising the Value of Rural Property

This stage involved examination of a wide array of property characteristics that have been used for rural property valuation. This involved analysis of various research papers, standards presently in operation for rural valuation (eg. Valuation Best Practice, 2005), rural land valuation manuals (Division of Property Assessments, 1992) and relevant texts (Baxter & Cohen, 1997). In addition, a number of research articles that have applied hedonic regression analysis to rural valuation were evaluated. These research papers provided a more substantial measure of the frequency with which specific property characteristics have been used, and more importantly, the level of significance of each characteristic with respect to property value, in each study.

The number of variables that have been used amongst the numerous studies for rural valuation is large. With each study using different combinations of property characteristics, the conceptual model I develop later in this thesis aims to specify only the most significant property characteristics that appear most likely to influence property values in Victoria.

1.4.3 Stage 3: Database Development

The aim of this stage was to obtain the required data sets and integrate these into a single GIS database. Due to the disparate nature of the data required for rural valuation, data sets were obtained from a combination of agencies. As a result, data formats, projections and the representation of variables differed amongst each data set. After integration of all data sets, spatial analyses were performed to enhance the final database by providing additional measurement variables for use in the study. Due to the integration issues that arose during this stage, a framework was developed to assist others during the integration of disparate data sets.

1.4.4 Stage 4: The Numeric Rural Property Valuation Models

Multiple Regression Analysis (MRA) was first used in the 1960's in the USA to generate models for residential property valuation (Adair & McGreal, 1988). This technique was used to develop a model or equation that attempts to explain the relationship between a dependent variable (sale price) and the independent variables that are characteristic of the property (Gallimore *et al.*, 1996). Once a model has been developed the technique enables the value of a property to be estimated in situations where there is no sale price (and where the characteristics of the property are known).

The use of MRA for both rural and residential valuation is widespread. MRA has been used for rating valuations in the United Kingdom and the USA (Reid Schott & White, 1977; Palmquist & Danielson, 1989; Elad *et al.*, 1994; Adair *et al.*, 1996; Kettani *et al.*, 1998). Automated Valuation Models (AVM's) and Computer Assisted Mass Appraisal (CAMA) are increasingly used in the valuation industry and the mortgage origination process (Poor's, 2004). The two major AVM's used in the USA for the mortgage origination process are based on MRA and hedonic theory. These models are used in approximately 10% of all commercial mortgage originations (Poor's, 2004).

Stage 4 involved using hedonic pricing theory and MRA to develop an implementation of the conceptual model that best represents property valuation within rural Victoria. A number of techniques including rank regression, best subsets and standard regression were used. Variations were made to the dependent variables by implementing the model with a 'sale price' as well as a 'sale price per hectare' dependent variable. Logarithms to the base 10 were also taken of the 'sale price' and 'sale price per

hectare' dependent variables. Each of the developed models were tested using various methods to determine their level of price estimation.

1.4.5 Stage 5: Cluster Analysis

This stage used cluster analysis to ascertain if properties could be classified into more homogenous regions and to statistically analyse if these sub-markets could lead to more accurate rural models. Cluster analysis is a statistical technique that is used in numerous disciplines to classify data into similar type groupings which may not often be obvious through visual analysis (Everitt *et al.*, 2001). The LGA boundaries used in Stage 4 were disregarded for this stage. Instead, variables were specified within a cluster analysis software program called SPSS13 using a two step clustering algorithm. Each set of derived clusters were then tested on the 8 models derived within Stage 4.

1.5 Thesis Structure

This thesis is presented as nine chapters. Chapter 1 provides an introduction to the research project highlighting the research aims, objectives, the research scope and research methods. Chapter 2 (Rural Property Value Determination) incorporates a literature review on the different manual and automated techniques commonly used to value property. The Chapter discusses the issues arising from present manual valuation techniques and the common problems associated with the growing use of automated techniques for rural valuation. A conceptual model is developed based on the analysis of property characteristics that have been presented in rural valuation manuals, texts and current valuation standards. Research papers using hedonic pricing were reviewed and an analysis was made of the different characteristics used within these studies and the significance obtained when using these variables to estimate property values.

Chapter 3 is entitled GIS Modelling and Sub-market Identification for Rural Valuation. This Chapter incorporates a discussion on the issues associated with integration of numerous data sets and presents various techniques that can be used to minimise problems during integration. This Chapter examines the use of GIS technology to derive additional property characteristics through the spatial overlay of different data layers. In addition, the Chapter discusses the use of cluster analysis within hedonic valuation and the need for sub-markets when using regression models for valuation. The Chapter incorporates a review of the cluster analysis technique and the various ways in which segregation of data into sub-markets is presently performed using both geographical and statistical measures. Finally, the Chapter discusses the ways in which cluster analysis has been used in property valuation.

Chapter 4 considers database development for the study. This Chapter provides background information on the selected study areas, the data sets obtained, the software and hardware used and the development of the property database. The development of the property database details the techniques used to integrate the acquired data sets into the one GIS database. This Chapter also examines the data integration issues that were encountered during the research and proposes a framework that can be applied to streamline future data integration tasks. Chapter 5 is titled Development of the Numerical Rural Property Valuation Models and describes the implementation of the conceptual model through the development of the numerical property valuation models. The Chapter presents the regression equations/models developed and the degree of significance of the property characteristics. Chapter 6 describes the implementation of the conceptual model through the use of cluster analysis. A number of clusters were developed using a two step clustering algorithm

with the total number of clusters constrained for the study. Each cluster developed was then tested on 8 regression models developed during Stage 4 of the research.

Chapter 7 is entitled Numerical Model Development with Restricted Digital Data and incorporates restricted digital property data (that were not available at the start of this research) into model development. These data relate to housing and farm buildings, and their improvements. The aim to this aspect of my research was to evaluate the models developed earlier in the thesis, and report on the affects of developing rural property models without data on house and building improvements. The processes undertaken within Chapter 5 and Chapter 6 are replicated in this chapter so that direct comparisons can be made of model performance. Chapter 8 discusses Numerical Rural Property Valuation Models. It outlines the development of numerical valuation models and considers the results obtained after applying the regression equations to each property to determine price estimates. This Chapter provides a measure of the ability of each model to determine property valuation estimates by examining the number of estimates that fall within 10% and within 20% of the actual sale price. Additional statistical measures were computed for each of the developed models and include the COV (Coefficient of Variation), COD (Coefficient of Dispersion) and PRD (Price Related Differential). This Chapter then discusses the use of cluster analysis in the determination of sub-markets. A comparative analysis is made of the use of geographical (a-priori) based sub-markets (Stage 4) to those defined using numerical techniques, such as cluster analysis. The discussion aims to provide a basis for utilising cluster analysis over a-priori techniques for sub-market identification.

The results of this research are compared to other rural valuation studies and a discussion is presented on the possible reasons behind the levels of price estimation

achieved from the models. Limitations associated with the availability and quality of digital data in Victoria for rural valuation are discussed, and some solutions are proposed.

To close the thesis, Chapter 9 provides a general discussion. I discuss some of the limitations of the research, and outline recommendations for further research.

Chapter 2 Rural Property Value Determination

2.1 Introduction

Property valuation can be performed using a number of different automated and manual techniques, each having various levels of success. The amount and type of data available can influence the techniques used and may result in many data integration issues depending on where the data are sourced and in what format and level of accuracy they are supplied. The use of automated techniques for rural valuation generally focuses on the monetary impact of a particular variable and how changes to the variable will affect the property valuation estimates obtained (Gardner & Barrows, 1984; Faux & Perry, 1999; Reynolds & Regalado, 2002). The price estimates found for each characteristic can then be used to determine a value for a property with differing characteristics.

The aim of this Chapter is to examine the various techniques available for valuation of property and discuss their advantages and limitations. Manual and automated valuation techniques will be considered with a more extensive examination made of the computer-assisted techniques presently in use for rural property valuation. A conceptual model is presented that is based on the property characteristics that appear most likely to influence property values in rural Victoria.

2.2 Land Valuation

Prior to a discussion on the approaches to valuation, a few definitions are necessary.

Property: A property is land that is capable of being owned and does not include leasehold land (Wooton, 1989). A property can consist of multiple land parcels, a single parcel or part of a parcel.

Parcel: A parcel is the smallest piece of land defined by cadastral boundaries that is able to be separately transferred within Victoria (Land Victoria, Victorian Department of Sustainability and Environment 2006; see http://www.land.vic.gov.au).

For rating valuation in Victoria, three main methods are used; Site Value, Capital Improved Value and Assessed or Net Annual Value (Nind, 2002). Site Value is the value of land after improvements have been made and includes any work completed on the property or any materials used to benefit the land and its value. It does not include buildings, fencing or planting (Nind, 2002). Capital Improved Value is the sum of the buildings and capital value that a property might realise if offered for sale at the date of valuation (Nind, 2002). Assessed Annual Value is the total rental that the land might be expected to realise on the basis of rent from year to year (Nind, 2002). During the 2004 rating valuations within Victoria, 71 municipalities used Capital Improved Value, 6 used Net Annual Value and 2 used Site Value (Land Victoria, Victorian Department of Sustainability and Environment, 2006; see http://www.land.vic.gov.au).

2.2.1 Approaches to Land Valuation

The three most common approaches to valuation are the Cost, Sales Comparison and Income Capitalisation approach.

2.2.1.1 The Cost Approach

The Cost Approach involves estimating replacement costs of farm buildings and machinery on the property to determine an estimated value. This approach requires the creation of a detailed inventory of all machinery and buildings within a property.

This approach is not as applicable to the rural market as some other techniques discussed below as it places a greater importance on the buildings of a rural property and not on the type and quality of the land (American Institute of Real Estate Appraisers, 1983). The Cost Approach, however, is used when there have been major improvements in the buildings that are deemed to make up a large part of the property's value. In some cases, the total of all farm buildings and developments is greater than the value of the property. This method can be used to provide a check of valuations derived using other techniques (Suter, 1974). One drawback of this technique for rural valuation is that the information it requires is not contained in publicly available data in Victoria.

2.2.1.2 The Sales Comparison Approach

The Sales Comparison Approach analyses sale prices determined by vendors and purchasers from previous sales of properties. A number of properties are selected so that they have similar characteristics to the subject property and can then be compared

more easily. Based on differences between the comparable and subject properties, the valuer makes adjustments to the sale prices to arrive at a valuation for the subject property (Suter, 1974).

This method is dependent on the availability of sales data and having adequate numbers of properties with similar characteristics. Rural environments are often characterised by few sales and in many cases, a significant period of time can lapse between sales. There can also be greater diversity amongst properties, such that obtaining enough sales over a similar time frame can be difficult. This can affect the degree of comparison that can be made between properties, (Walker, 1994) making the use of the technique within a rural setting problematic. For manual valuation the approach may be somewhat biased as comparable selection and any adjustments made are based on the expertise of the valuer (Adair & McGreal, 1988).

In rural environments, the quantity of sale's data available for modelling property values can be a serious constraint. One way to overcome this constraint may be to employ data spanning a longer time frame of perhaps 6-8 years. These data can be used to determine the change in average values per year such that the previous year's data can be used with the current year of data to increase the database of land sales. Creation of an index to compensate for varying sales prices over a period of time can extend the use of the previous year's data by providing more comparative properties (Suter, 1974).

2.2.1.3 The Income Capitalisation Approach

The Income Capitalisation Approach is based on the assumption that those wishing to purchase a property are mostly concerned with its current and future income producing

capacity (Suter, 1974; American Institute of Real Estate Appraisers, 1983). The valuer must determine the rate of productivity and expenditure of the business to present an estimated value of its worth based on the possible income and cash flow that can be generated. The estimation is an indication of the price an individual will pay for the income the farm produces and as such, other factors may affect the value. The purchaser may be willing to pay more for the property if there are other intrinsic characteristics that makes the property more special to a purchaser (Suter, 1974).

The Income Capitalisation Approach is primarily used for commercial properties and has had a somewhat limited use for rural valuation. This is mainly due to the technique requiring more detailed data regarding property income. Income over one year may also fluctuate. Thus, to predict future income, production rates from previous years are required to compensate for any market driven influences. Although the Australian Bureau of Agricultural and Resource Economics (ABARE) conducts farm surveys throughout Australia (known as 'farm gate' receipts) to collect more detailed income and economic information, the farm surveys are only available as an aggregate data set. Some of the information required to successfully use the Cost Approach is available through ABARE, but not on an individual property basis. This, therefore, limits the use of these data for this type of rural valuation (ABARE 2006, see http://www.abareconomics.com).

2.2.2 Techniques for Rural Valuation

Although similar in its principles to that of residential valuation, the process of valuing of rural properties is often more complex than urban properties due to the diverse value drivers involved (Australian Institute of Valuers and Land Economists, 1997). Rural valuations, in particular of income earning properties, are influenced not only by the

physical characteristics of the property, but also by broader economic factors, perceptions within the market place and the production capabilities of the property. Whilst characteristics can be obtained from sale data regarding land size, land use, crown allotment and location, additional information is necessary to differentiate the price influences between properties (Australian Institute of Valuers and Land Economists, 1997). This additional property information can be obtained from existing digital data sets such as those containing information on road or rail networks, hydrography, planning schemes or vegetation where available. In some cases not all of this information is available to the public. Where suitable data have not been collected, site visits and interviews with property owners may assist in providing some of this information. Clearly, these manual methods of data collection can minimise the effectiveness of automated techniques due to the additional time spent collecting and collating data.

The Valuation Best Practice Specifications (2006) incorporate a number of statistical testing procedures and GIS mapping components to assist in rating valuations for various property types within Victoria. The statistical procedures provide a series of techniques to evaluate valuation results, whilst the GIS mapping component of the specifications allows for cross validation of valuations between differing dates. The main impact of these specifications is that they aim to ensure consistency in valuations, reporting procedures and property data storage across Victoria. In addition, the specifications outline the standards to be adhered to, and the deliverables to be submitted during the rating valuation process.

To date, whilst most Victorian LGAs utilise value drivers (devised by valuers) to determine their residential valuations, this is not the case for rural valuations that are still performed manually (Brett Reed 2003, pers. comm., 18 September).

2.2.3 Manual Valuation Methods and Valuation Best Practice Standards

The methods used for the manual valuation of rural property are similar to that of other property types in terms of the principles being applied (Baxter & Cohen, 1997). The difficulties that arise with valuation of rural property is that there are often fewer sales, greater time between sales and rural property can be influenced by more factors than other types of valuation (Baxter & Cohen, 1997).

In addition to utilising a database of information regarding each property, manual valuation can also require site visits to obtain additional property characteristics. The accuracy of a property value is influenced by the data held by the valuer and the number of similar properties that the valuer can use in their comparable sale analysis. Automated techniques are also influenced by the available data and the rules that have been applied to develop the model.

Property tax legislation for Victoria requires that all property be valued using either Capital Improved Value, Site Value of Net Annual Value (*Local Government Act 1989 (Act No. 11/1989)*). The legislation outlines the procedures to be undertaken with regard to collecting property rates and administering changes to the rating system. The legislation documents the methods of collecting rates from property owners, exclusions for certain properties and available concessions to relevant individuals. The significance of this legislation is that it provides a model to base the Valuation Best Practice standards upon. Each municipality required to adhere to the *Local*

Government Act 1989 (Act No. 11/1989) for its administration of property taxes. This requirement leads to higher standards for property valuation and the Valuation Best Practice guidelines provide a basis to achieve this.

The requirement for higher standards led to the State of Victoria undergoing a substantial valuation review in the late 1990's. The development of Valuation Best Practice in 1998 was intended to coincide with the implementation of the first biennial valuation to commence in 2000. These standards detail the timelines, processes to arrive at a valuation and the deliverables required as part of the valuation process.

A key component of Valuation Best Practice is not only its determination of a standard and accurate valuation for each rateable property, but to ensure that an accurate property database containing the rate payers' details and details of each property is held state-wide. Having consistency across all municipalities of Victoria with regard to the way that this information is stored enhances the application of these data.

A brief review of the key components of the Valuation Best Practice will now be made. For a more extensive review regarding the outputs, the current 2006 specifications can be viewed and/or downloaded from Land Victoria, Victorian Department of Sustainability and Environment; see http://www.land.vic.gov.au.

One of the emphases of Valuation Best Practice is the project planning and tendering process to ensure all required property are valued by the dates set out in the specifications. In addition, it aims to ensure that consistent and quality data are held state-wide within Victoria and that the valuation derived from the process, have

undergone statistical evaluation of the values obtained. Within the preliminary stage, statistical analysis is undertaken through sales ratio testing and examination of sub-market groups using the previous valuation. For rural property, sub-market groups are defined through separation of townships from rural residential lands and through the use of the land classification code (Valuation Best Practice, 2005).

The second stage of the valuation process involves preliminary analysis of current sales information prior to field analysis. This involves performing GIS mapping and producing documentation regarding the valuations that have been determined after field visits. Stage three involves specialist and commercial property and is not addressed here. Stage four involves producing final valuations for rural and residential property and using statistical techniques to evaluate the results. Stage five involves submission to the Valuer General of the valuations along with the key outputs (Valuation Best Practice, 2005).

The major statistical tests used within Valuation Best Practice for the evaluation of results are the median sales ratio, Coefficient of Dispersion (COD), Coefficient of Variation (COV) and Price Related Differential (PRD). These statistical measures adhere to that of the Standard on Automated Valuation Models (AVM's) (IAAO, 2003) and the Standard on Ratio Studies (IAAO, 1999). Mapping outputs produced are primarily used to show location of sales and value shift maps which depict the change in values between dates. They can also show the percentage change between site value for previous valuations and that of the current valuation (Valuation Best Practice, 2005).

Overall, a large part of the valuation process for rating valuations in Victoria is optimised through more computerised means such as using spreadsheets, statistical software and GIS mapping for reporting and analysis of results. The procedure is still largely manual and involves site visits, manual generation of price estimates, and manual interpretation of results. The implementation of Valuation Best Practice provides a standard approach to the valuation process and for the outputs generated for all valuations state wide. The next possible step for valuation reform within Victoria could be the implementation of a more automated means to arrive at a valuation estimate. A discussion of present automated techniques in use is presented in the next section.

2.3 Computer Assisted and Automated Valuation

Greater development in computer technology and processing ability has lead to the use of Automated Valuation Models (AVM's) to speed up the appraisal process, to provide more accurate estimations and to minimise the cost of appraisals (Ross & Nattagh, 1996). Automated valuation, also known as computer-assisted valuation was first used during the 1960's in the form of regression analysis (Adair & McGreal, 1988). Other early automated valuation techniques utilised sale price data and database technology to find a property that had similar characteristics to the subject property. Adjustments were then made to account for differences between the subject and the comparable properties to generate a valuation for the subject property (O'Rourke, 1998). Regression analysis and comparable selection still feature prominently in current research with some residential research achieving a similar accuracy to manual techniques (Adair *et al*., 1996). Utilising computerised techniques enables data to be retrieved more effectively and at more rapid speeds, and through the application of regression analysis, allows large volumes of data to be analysed in ways that are not feasible using manual techniques (Sauter, 1985).

Initially, automated valuation was seen as a means to automate some of the processes involved in appraising properties. However, more recently it has been used as a technique to enable a property valuation estimate to be artificially generated (Rayburn & Tosh, 1995). For residential valuation, automated modelling has been extensively used within the USA in the mortgage origination process. The major drawcard of the development and extensive models already in use is the ability of models to deliver more accurate results than manual techniques (Adair & McGreal, 1988). This can reduce the costs involved in appraisal of property (Valentine, 1999; Waller, 1999).

Achieving greater accuracy in an automated model compared to manual methods poses issues relating to the future of valuers (Waller, 1999). However, to date, most models are still only used as an enhancement and not a replacement for manual techniques and the valuers, themselves. This is especially the case where there is a greater uncertainty involved in the valuation of a property or where the property is more specialised and requires a more detailed analysis of its price influences. Whilst much success has been achieved in some residential areas with automated techniques, they are considered to be in their infancy in Australia (Evans *et al.*, 1992).

Many studies have applied regression analysis in residential, commercial and rural markets though less focus has been in the latter due to the potential complexity of the rural marketplace and the lack of available sale price data in some regions. Whilst automated techniques may be less subjective and biased than manual ones, the accuracy of any automated model is in the most part determined by the quality and reliability of the data and the valuation methods used (O'Rourke, 1998). In comparing a manual human generated valuation as opposed to an automated one, data are a

primary influence in both of the methods (McCluskey *et al*., 1997). Data availability is a key issue to address in the determination of valuations for any type of property and in the generation of automated models.

AVM's have typically used regression, either linear or multiple; expert systems or neural networks in their modelling for the determination of property estimates (Waller, 1999). The three most commonly used models for automated valuation are repeat sales, tax assessment and hedonic models (Nattagh & Ross, 2000).

Repeat sales models use sale prices of properties to develop indices based on historical sale price data which can then be used to estimate property values. Tax assessment models use assessed values that are based on tax values (Nattagh & Ross, 2000). Hedonic models are based on the theory that the price of a property is a function of the property's various characteristics and that the purchaser is only willing to pay a set amount for these characteristics (Powe *et al*., 1997, Rosen, 1974).

The following sub-sections discuss specific automated techniques (neural networks, case based reasoning and regression analysis) and relate the use of hedonic models to regression analysis. A number of rural models is discussed. I review the variables used in each model, the number of sales that the model is generated from and the accuracy of the model's estimates. An attempt to determine the reason behind the more accurate models and their use of specific property characteristics will be made in the following sections.

2.3.1 Decision Support for Property Valuation

A Decision Support System (DSS) is a computer-based system that enables greater utilisation of data by enhancing and supporting solutions to complex and ill-structured problems (Guariso & Werthner, 1989; Klein & Methlie, 1990). Ill-structured problems can indicate that the problem is not repetitive or is ambiguous due to its complexity (Klein & Methlie, 1990). A DSS has the ability to provide an enhancement to decision making rather than replace manual techniques (Guariso & Werthner, 1989).

Teaming a DSS with a GIS enables greater searching and identification of similar characteristics which can then provide the user with a list of sites or properties that have the relevant criteria. The data from the DSS can then be used to assist in solving problems (Laaribi *et al.*, 1996). Whilst the DSS has the capacity to pass information from the GIS to assist in solving a decision problem, it is also able to utilise personal preferences, qualitative information and rank important data (Guariso & Werthner, 1989). Thus, the DSS teamed with a GIS can increase the development of solutions to problems whilst utilising the analytical functionality of a DSS.

The manual approach to valuation involves human based appraisal which can lead to subjective and biased results in the determination of property values (Adair & McGreal, 1988). Expertise within rural areas, especially within more specialised areas such as vineyards or sugar cane, requires significant knowledge regarding the industry being valued. Several days can be required to perform manual appraisals of properties (Bonissone & Cheetham, 1997) and site visits may be necessary. Site visits are frequently conducted without viewing the inside of any buildings on a property. This may influence the accuracy and quality of the appraisal (Longley *et al.*, 1994).

The decision making process, as discussed by Simon (1977), comprises three main stages - intelligence, design and choice. Intelligence is looking for decisions which need to be made, design incorporates the development of possible actions to be taken, whilst choice is the selection of a particular action to be taken (Simon, 1977; Malczewski, 1999a). Each of these phases can occur in any order during the decision making process and in terms of spatial applications, GIS can be used to enhance decision making during any or all of these phases (Malczewski, 1999b).

Decision support tools for residential valuation can use a variety of techniques ranging from regression analysis, expert systems, knowledge-based systems or case-based reasoning. Each technique uses different methods to determine a property valuation estimate. This can include statistical analysis of relationships between property value and area, construction of a knowledge base, and weighting specific characteristics depending on their relevance to property value (Waller, 1999). The application of decision support enables adjustments between comparable properties to be made by giving a higher rating to those property characteristics which have a greater influence over the price (Wyatt, 1997). Rule sets can be defined that enable deviation values to be set for each property characteristic. These deviations give maximum ranges with which each attribute from each comparable property must be within (Bonissone & Cheetham, 1997). After selection of appropriate comparable properties, the DSS can be used to write any new knowledge derived from specific analyses back into the database for reasoning (Zhu *et al.*, 1996).

A DSS can be implemented with GIS to enhance the complex issues and decisions that are required to be made by a valuer during the appraisal process. While a DSS is reliant on a human developing a system based on set rules, it may reduce some of the

bias and subjectiveness present in manual techniques by providing a more analytical approach to property valuation. A DSS may not be capable of making a judgement about a set of property characteristics to include in an appraisal, or which comparable property to use. However, this technique can assist by providing guidelines to enable an appraiser to make these judgements (Mathieson & Dreyer, 1993).

The method applied with the DSS such as case-based reasoning, multicriteria decision making or artificial neural networks, will determine the extent of automation that is possible. Multicriteria decision making will tend to lend itself to working side by side with an appraiser who is prompted for responses based on choices. Artificial Neural Networks (ANN) on the other hand can be considered as more of a 'black box' system where information is fed into the network and after processing, a valuation is determined. These systems tend to minimise valuer input and analysis as the pre-defined 'network' performs all the calculations to determine a valuation. The disadvantage of this technique is that the valuer does not know the nature of the price influences of the model. Multiple Regression Analysis (MRA) is also more open to valuer input in that the valuer will select the property characteristics to input into the model and after the analysis can ascertain which variables are significant through analysis of the regression result outputs (P-value).

2.3.1.1 Criteria Weighting and Decision Rules

Weighting of criteria is important in property valuation due to the differences in the influence that each criterion may have on the valuation price. Weights are determined by assigning a ranking, with a larger weight implying a greater dependency on the values of the criteria. This provides a basis to indicate the degree of variation between the different criterions (Malczewski, 1999b).

Techniques to determine weights for criteria include ranking, rating, pair-wise comparison and trade-off analysis as methods (Malczewski 1999b). Ranking and rating, though they are easy to use and can be calculated in a spreadsheet and then imported into a GIS environment, are not considered to be precise or suitably accurate (Malczewski *et al*., 1997). Trade-off analysis is considered precise, but relatively difficult to use. Pair-wise comparison is considered most effective for spatial decisions (Malczewski *et al*., 1997). Given a requirement of high accuracy in determining criteria weights for property valuation, the pair-wise method might be considered more appropriate due to its underlying statistical theory (Malczewski *et al*., 1997).

After identification of criteria that are influential to sale prices, figures representing the affects of each criterion, enable adjustments to be made to reflect a market value for each property (Kettani *et al*., 1998). Assigning decision rules for the initial retrieval of the comparables and the deviations for each characteristic is critical in estimating a market value. Bonissone & Cheetham, (1997) have determined rules outlining the maximum percentage value a single adjustment should have with respect to sale price. The net adjustment and gross adjustment are deemed to not exceed 15% and 25% of the sale price, respectively. This gives some assurance to the degree of accuracy of the resulting value as any large adjustments will indicate that the comparables selected are dissimilar and the criteria have been over adjusted (Gonzalez & Laureano-Ortiz, 1992).

To date, criteria weighting has not been applied on its own for rural property value determination. However, it has been used with multi-criteria decision analysis and case-based reasoning techniques for residential valuation (McSherry, 1993). The

underlying concepts of the approach are not dissimilar to that of multi-criteria decision making and case-based reasoning in that prior rules and reasoning need to be determined by one who is skilled in rural valuation. Once rules are set, the technique could be applied in a number of ways. It can be used in a fairly manual based approach to rank the closeness of comparable sales to the subject property and, thus, provide a means to determine which comparable sales are more appropriate to include in an analysis.

The use of criteria ranking can be used in a preliminary stage for the ranking of the importance of comparable properties. The technique is not as reliant on a substantial database of sales as other techniques, and is more likely to be used as a precursor to the employment of another technique. As such, the technique will always be reliant on initial user interaction to set the rules and ranking of criteria. After this initial stage the technique of criteria ranking can be used in a more automated approach to help generate valuation estimates.

2.3.1.2 Multi-Criteria Decision-Making

Multi-Criteria Decision-Making (MCDM) involves using decision support tools to assist in evaluating a set of choices or alternatives which often have conflicting criteria (Carver, 1991; Malczewski, 1999a). The basis of MCDM is not in solving a problem, but ranking the alternatives or reducing the number of alternatives to enable the decision-maker to find a preferable alternative (Jankowski, 1995) to a problem. MCDM is used for a variety of spatial problems ranging from site allocation of recreational or retail sites, as well as route selection for powerlines and other similar infrastructure (Jankowski, 1995).

Key elements of MCDM include defining the alternative and criteria sets and analysis of the impact of the alternatives on each criterion (Jankowski, 1995). The criteria can be qualitative or quantitative and may be conflicting. The decision-maker is able to weight or rank the criteria based on their expertise and experience in the area (Bui, 1987). In general, there are only small numbers of criteria and alternatives used in this method. The use of a larger number of criteria requires the modification of the MCDM techniques (Carver, 1991). Bui (1987) reported that after assigning weights to criteria and evaluating the alternatives, the analysis identified some preferences that required re-structuring.

Property values are affected by numerous factors. Some of these include the age, size and quality of the buildings, special features, and proximity to schools and shops (Adair & McGreal, 1988). These factors can be deemed as 'multiple criteria' that may influence property values to varying degrees. The factors that influence value can be thought of as a function of a property's value (Kettani *et al.*, 1998).

The valuation methods proposed by Kettani *et al.* (1998) utilised a data set consisting of 108 properties from a one year time period. The selected study area was Alberta, Canada and 'residential bungalows' were the sole property type under analysis. After consultation with a team of real estate agents, 11 criteria were identified that were deemed to have a significant influence on sale prices in this area. These criteria included house age, house size, number of garages, ease of access to the garage, presence of a basement, presence of a fireplace within the region. Kettani *et al.*, (1998) assumed that the sale prices of these bungalows represent a multi-criteria

approach taken by the buyers and that each criteria contribute to the sale price realised through the property sale.

The technique outlined by Kettani *et al.* (1998) required a valuer or real estate agent to determine the relevant ranking of importance of each property characteristic prior to further analysis. The technique is subject to human judgement in determining which property characteristics are important. The statistical allocation of features into ranges to determine an adjustment of price can limit the reliability of the results due to some pricing factors only having a few instances of property within each range. This could give an unrealistic view of the monetary effect of each value driver as the sample is not representative of a wider range of property. The method used is static in that new results cannot be updated after new prices are estimated for properties thus new prices cannot be added to the database. The process is limited as it uses statistical results to segregate the data into ranges for each criteria and to develop the relationship of its effect on price. In this sense it does not search for comparable properties first and then make adjustments based on the differences in features between the subject and the comparable properties. This could lead to less reliable results due to properties being selected from different locations or prices based on statistical relationships on the whole data set instead of on a selection of similar properties.

The Analytical Hierarchy Process (AHP) is another form of multi-criteria decision support used by Ameson *et al.* (1996). This technique is used to determine the ranking of the three approaches to valuation (Cost Approach, Income Approach and Sales Comparison Approach). When a property is valued using the above three techniques, it is rare that the determined values with each method is the same. Ameson *et al. (*1996) used the approach to determine which of the estimates determined using the

three approaches has more weight or importance. The four criteria used for determining the ranking are the appropriateness of the approach used for the valuation purchase, the appropriateness of the approach for the type of property, the adequacy of the data available and the quality of this data. The technique places its importance on data quality and availability as well as the suitability of the valuation approaches to the property being appraised.

Whilst the above approach can theoretically be used to rate which approach is more suited for each valuation purpose and the data that are available, it does not sufficiently deal with the underlying issues of the data quality of each data set, nor any inconsistencies with data used between valuation approaches. Likewise, it does not address the issues which may arise regarding any subjectiveness that the valuer may have in valuing property using any of these techniques. The AHP approach can assist by providing a measure of ranking of a suitable valuation after a valuation has been determined, however it may not be ideal for the automation or generation of actual valuation estimates.

2.3.1.3 Case Based Reasoning

Case Based Reasoning (CBR) uses a decision making approach to select examples from a database which are similar to a subject property (Holt & Benwell, 1996). The process of CBR involves retrieval of a similar case from the case library, adaption of the case to apply it to a new problem, then whilst 'learning' from the existing case and developing solutions, update the case library with results of the current problem. Adaption between the differing cases is necessary to account for any differences between the parameters of the old problem and that of the new problem being solved. As reasoning occurs, so to does learning and thus when a result of the solution or

problem is found, it can then be added to the case library to be recalled at a later date (Kolodner, 1993). This is especially valuable in the sales comparison approach where not only are properties selected based on specific attributes, but once an estimate in value has been made it could be used in further valuation estimates by writing the results back into the database with a specified confidence interval (Gonzalez & Laureano-Ortiz 1992).

CBR has been utilised by Gonzalez & Laureano-Ortiz (1992), McSherry (1993) and Bonissone & Cheetham (1997) for residential valuation. These studies use the comparable sales approach and utilise CBR to retrieve similar properties to use for comparable analysis. The procedure only utilises a few properties from the database and, thus, is not reliant on the strength of the whole data set to determine differences between the data. However, the technique does require a large database in order for appropriate selection of similar property to occur. The fewer the number of properties, the less chance of finding comparable properties that are the most similar to the subject property. The strength of the CBR technique is that it is constantly learning from and utilising the information regarding past experiences. The technique also requires domain specific rules to be set up prior to case selection. In most instances, the aim of this technique is to find comparable properties that are the most similar to the subject property (McSherry, 1993).

A prototype program was developed by McSherry (1993) for residential valuation. The aim of this system was to develop another means for selection of cases from the case library. Existing CBR techniques require rules to be set up to adjust for differences between properties, thus the expertise of a valuer is required to assess the importance of the differences between property (McSherry, 1993). This specification of rules is

somewhat similar to a manual approach of comparable adjustment which is also reliant on a valuer to determine the level of adjustments required for each difference between the subject and the comparable properties selected.

McSherry (1993) aimed to develop a technique to find one case that represents an upper bound and another case which represents a lower bound and not to find the most similar case during case selection. McSherry (1993) used a ranking based on similarities between the subject and the comparables so that interpolation between the upper and lower bounds could be undertaken to enable an estimate to be determined. The prototype used by McSherry (1993) appears rather simplistic in its selection of the property characteristics. Only a few characteristics are used with these being the number of bedrooms, style of the property and its location. If no match occurs between the subject and the comparable properties then linear interpolation is undertaken based on pre-determined rules.

Gonzalez & Laureano-Ortiz (1992) applied CBR to residential property using the comparable sales approach. After consultation with valuers, a set of property characteristics were selected to develop the system. Selection and retrieval of appropriate comparables was undertaken and adjustments are made to account for differences between the comparable properties selected. Ten comparable properties were selected and an adjustment process was applied through a set of pre-determined rules. Another set of rules are then used to select the best three comparables from the ten initially selected. A weighted average was calculated to determine a valuation based on the differences in the adjustments. The rule based system provides a dollar value for the amount to adjust the subject property for each difference in a property characteristic between the subject and the selected comparables. In this case, the test

system consisted of 70 properties. Thirty seven per cent of the properties had an appraised value of 5% or less of the listed price. In 68% of the properties, the appraised value was within 10% or less and in 90% of the properties, appraised values were 20% or less than the listed price (Gonzalez & Laureano-Ortiz 1992).

The system and associated rule set developed by these authors is dependent on an appraiser or multiple appraisers to determine which property characteristics to include into the model. Another set of rules is required to provide a means to assign a dollar value to compensate for differences in characteristics between the subject and comparable properties. Whilst the technique used by Gonzalez & Laureano-Ortiz (1992) achieved high levels of price estimation, like other CBR techniques it does require input from appraisers to determine levels of adjustments and rule sets, and as such is traditionally based on the technique used by appraisers to determine a valuation estimate. A high level of input from an appraiser is required prior to construction of the technique and over time this will require some adjustment to the rule set to deal with changes in the dynamics of the real estate market of the area. Although the technique aims to provide some automation to the valuation process it is still largely influenced by valuer input and expertise in setting up rules. A CBR uses prior property values to adjust for differences between the subject and the comparable using a simpler rule set, whilst a rule based system determines its valuation estimate from scratch using a number of rule sets (Gonzalez & Laureano-Ortiz 1992).

Bonissone & Cheetham (1997) used fuzzy logic with CBR to estimate residential property values in California. The selection of comparables was based on six attributes which are rated and ranked to how similar they are to the subject using weighted aggregation. Again, the sale prices of each property are adjusted to reflect additional

characteristics and differences between the subject and comparable properties. The selection of comparable properties was based on the sale date, distance of properties from one another, the lot size, living area, number of bedrooms and number of bathrooms. Each property characteristic was assigned a maximum allowable deviation between the subject and the comparable property with those falling within the deviation selected as a 'comparable'. These characteristics and the deviations selected were based on consultations with two appraisers. Bonissone & Cheetham (1997) found that between four and eight of the comparables were required to achieve a higher level of similarity between the subject and comparable. If too many comparables are selected then the technique introduces the possibility that the comparables will deviate further from the subject and thus require more adjustment than another comparable that is more similar in its property attribute levels.

The strength of CBR is that the case library evolves and is never complete. With new cases added over time it is constantly learning from and updating information from past experiences (Clayton & Waters, 1999) and thus provides great benefits to allow for changing knowledge. The technique aims to mimic that of a human appraiser and their approach to comparable selection and adjustment to account for differences between the subject and the comparable property. To mimic this approach a valuer or team of valuers are required to determine a rule base to apply for comparable selection and then another set of rules to enable adjustments to be determined. As the CBR technique uses existing adaptions from previous analyses, it is not as dependent as the rule system on the generation of rules for adjustments only that of how the selections of comparables are made (Gonzalez & Laureano-Ortiz 1992).

Although the approach is similar to that of manual techniques it does provide a higher degree of automation once the rules are pre-determined. The approach is also dependent on data and finding similar properties like that of manual valuation techniques. It also requires knowledge regarding the forces driving the market in which the approach is being used to generate the rules.

2.3.2 Artificial Intelligence

For valuation, two forms of artificial intelligence exist. These are Expert Systems and Artificial Neural Networks (ANN's) (Rayburn & Tosh, 1995). The way in which these forms or artificial intelligence have been applied to property valuation will be discussed within the following sections.

2.3.2.1 Expert Systems

Although an expert system can be classified as a form of artificial intelligence, they are in fact an extension of a rule based system and used to provide a more automated approach to property valuation (Rayburn & Tosh, 1995). They are called 'expert' as they are based on rules specified by an 'expert' or a valuer with these rules trying to mimic the reasoning that an expert uses for problem solving within property valuation. The expert system is based on human thought processes used for the determination of property value (Bryant, 1991). The development of an expert system can be defined as the "elicitation of the knowledge from an expert or experts, and representation and validation of the knowledge in a computer program" (Nawawi *et al*., 1996).

The expert system, like that of the CBR technique and other rule-based techniques require valuers or experts to collaborate and determine the core rules for developing

the system. For rating taxation purposes, an expert system was developed in Malaysia for commercial and industrial property (Nawawi *et al.*, 1996). The elicitation of knowledge was obtained initially through a panel of valuers. However, to minimise gaps in knowledge, academics, property managers and other experts were used to broaden this process.

During the knowledge elicitation process, Nawawi *et al.* (1996) experienced conflicts between the different experts which were used to develop the key components of the system. Selection of appropriate comparable property for the sales comparison approach was developed by the 'experts' to mimic the way in which identification of comparable property is undertaken. Multiple regression analysis was incorporated into the model to determine weightings for location and buildings so that the valuation process was simplified (Nawawi *et al.*, 1996). Site inspections of the 92 office buildings and 14 shopping complexes was undertaken to derive additional property characteristics.

Twenty properties were tested and results show that the values estimated were within 10% of the value determined by a human valuer (Nawawi *et al.*, 1996). Issues with this technique were that the use of regression analysis led to the incorporation of weightings for property attributes which did not strictly mimic the process used by a human valuer. The model required a large amount of rental information to be incorporated whereas a valuer generally only uses a small subset of this information. In addition, as the model had pre-determined weightings of property characteristics, any new knowledge could not be included in the current model (Nawawi *et al.*, 1996). Any new influences in such a dynamic market would not be able to be incorporated unless new rules were developed to include these changing dynamics.

The use of expert systems evolved due to the so-called rigidity of regression analysis methods which are claimed to be data driven and based on linearity between characteristics (Czernkowski, 1989). Expert systems have been favoured by those seeking to devise a technique which more closely resembles the manual valuation process. The development of an expert system is reliant on obtaining or having access to adequate sale price or valuation data as well as property characteristics, as is the case for most other techniques in use.

The technique and development of the system requires a series of rules to be determined which resembles the thought processes of the human valuer. Whilst expert systems have a benefit in that they can more closely resemble the processes undertaken during manual valuations by valuers, this can lead to the rules developed containing some of the bias that can be found in manual methods. The process relies on a valuer or other experts to generate rules for comparable selection and then another set of rules for adjustment of these differences between the subject and comparable property. During this knowledge elicitation from the valuers, differences between influences may occur when these rules are defined.

The valuation of property for rating purposes typically is aimed at high accuracy such that there are minimal appeals against the values by rate payers (Czernkowski, 1989). Valuations are also performed on a regular basis by those that have been trained to achieve consistency amongst other valuers (Czernkowski, 1989). In Victoria, the rating valuation process is tendered and may result in valuers working within geographical areas with which they have little or no experience. The use of VBP does aim to ensure consistency between valuation years and between valuers by incorporating statistical

testing to minimise anomalies and outliers in valuations. Thus, for valuation for rating purposes using manual techniques and using VBP guidelines, a set of rules regarding the process is ingrained into the process which could be used within an expert system for rating valuation. The drawback of this is that for rural property, there are greater influences and price drivers over residential valuation (Walker, 1994). Rural property can be more difficult to value due to lack of data detailing property characteristics and as a result of less frequent sales of property (Walker, 1994).

2.3.2.2 Artificial Neural Networks

Artificial Neural Networks (ANN's) involve training to learn relationships and patterns from the data to mimic the learning that a human appraiser performs during appraisal (Rayburn & Tosh, 1995). The technique can also be limited by the quality and size of the data sets used as is the case with manual techniques and rule-based systems. ANN's have been used in residential and commercial valuation since the early 1990's (Tay & Ho, 1991; Do & Grudnitski, 1992; Evans *et al.*, 1992; Worzala *et al.*, 1995; McCluskey, 1996; Lenk *et al.*, 1997; Connellan & James, 1998; McGreal *et al.*, 1998), but appear to have had limited application in rural markets (Kwon & Kirby, 1997). ANN's have been used quite extensively in the UK (Tay & Ho, 1991; Evans *et al.*, 1992; Worzala *et al.*, 1995; McCluskey, 1996) and to a lesser degree within Australia (Rossini, 1998; Rossini, 1999). Results of these studies vary with some research obtaining far superior results for valuation using ANN's whilst others report poor results (Worzala *et al.*, 1995).

ANN's are not based upon rules determined by a valuer; rather they 'learn' by example (Paris *et al.*, 2001). ANN's aim to determine often complex relationships which can be difficult to discern through traditional means (Paris *et al.*, 2001). The ANN has the

ability, with adequate quality and quantity of data, to analyse relationships between prices, establish how these alter over time and perform generalisations to estimate a value (Rayburn & Tosh, 1995). The ANN is dynamic as it adapts to new knowledge and information when it is input into the system (Waller, 1999). A more comprehensive description on ANN's can be found in McCluskey (1996), Tay & Ho (1991) and Lenk *et al.* (1997).

Connellan & James (1998) applied ANN's to commercial property using a relatively large database of sale information and property characteristics. These data covered a time span of over eight years with the properties held in the database containing valuations taken monthly in some areas, leading to numerous properties having multiple valuations undertaken during this time span. Having an extensive array of properties to use whilst training the ANN led to a reasonable accuracy in the prediction of values for the proceeding five months.

Tay & Ho (1991) applied an ANN and a Multiple Regression Analysis model (MRA) to the appraisal of residential apartments using 1,055 properties within Singapore. They found that the ANN outperformed the MRA in this geographical area. The authors acknowledged that ANN's are a 'black box' technique that requires minimal input from a valuer. In this case, a valuer need only input a text file containing the sales information. The number of inputs (property characteristics) and outputs (sale price) information is specified and the network is then created and trained. The valuer is required to enter in the new property characteristics prior to a valuation estimate being determined by the ANN (Tay & Ho, 1991).

Lenk *et al.* (1997) applied an ANN to 288 residential properties in Fort Collins, Colorado, USA over a three month period. The sale price database obtained incorporated 12 property characteristics for each property, which were atypical of the characteristics used during residential valuation. Results indicated that the ANN did not outperform the model created using hedonic regression modelling. Lenk *et al.* (1997) found that 18% of the estimations devised by the ANN were greater than 15% of the actual sale price, thus indicating a considerable variation in it model's predictive capability.

The use of ANN's by McCluskey (1996) showed that initial results were still below that considered acceptable through a manual appraisal. The network was trained using 375 sales over a two year period with the network then tested on 41 properties which were held out from the sample. After some reassessment and re-training, the modelled results increased marginally. Analysis of the Coefficient Of Dispersion (COD) and Coefficient Of Variation (COV), standards set by the IAAO (IAAO, 2003), indicated that the ANN achieved an acceptable level of price prediction.

Evans *et al.* (1992) used ANN in a study of property values in the Midlands region of the United Kingdom. The residential sale data spanned a six month time period and 34 properties were used to train the network and 13 properties for testing. In a technique that is generally reliant on large data sets, Evans *et al.* (1992) reported that the ANN was able to estimate to a high level of accuracy (average errors were within 5-7%). However, the authors concluded that the ANN is best used as a support tool for valuation or as a preliminary analysis tool (Evans *et al.*, 1992). The higher level of accuracy achieved using a relatively small number of properties could be attributed to a fairly homogenous study area, with all properties located within a few streets of each

other or could be indicative of the small number of property characteristics actually included in the ANN. The short time span over which the properties were sold may have attributed to the higher accuracy achieved within this study.

In a residential environment with 288 sales within Fort Collins in Colorado, USA, Worzala *et al.* (1995), found that results with ANN's were inconsistent and did not achieve a high level of accuracy. From the total sale database, 217 properties were selected for training whilst 71 were used for testing of the network. Worzala *et al.* (1995) aimed to replicate the sample used by Do & Grudnitski (1992) to the extent of selecting similar price brackets and time spans for sales that occurred. A third training set was undertaken using more homogenous property types to determine if the modelled results could be improved. During this research two different ANN software packages were used and compared to results obtained using multiple regression. Worzala *et al.* (1995) found that the ANN performed slightly better in estimation than MRA however not noticeably. Whilst using two different software packages for ANN development, each package performed differently, but not always did both ANN's outperform the MRA model. The ANN also alternated between packages on which technique performed better in the different cases used. When using the same sample for training as Do & Grudnitski (1992), they found that the results were inferior. Overall, they concluded that the use of different software packages for the ANN development, may have undermined the reliability of the results.

Within an Australian context, Rossini (1998) and Kershaw & Rossini (1999) used ANN's to value residential property within three LGAs within South Australia. Rossini (1998) conducted surveys within Adelaide and additional regional centres throughout the State to obtain a property data set of 1940 sales for residential valuation. The

survey data were merged with the property data obtained from the Department of Environment and Natural Resources (DENR). Using the variables recorded as part of the DENR data set, MRA and ANN techniques were applied. During this research they found that the ANN tended to perform better with smaller data sets than larger ones.

Kershaw & Rossini (1999) used 18 years of property data from South Australia to determine if ANN's were more superior than multiple regression for estimation of price indices to account for transactions of property over time. The study showed that the ANN was not significantly advanced to that of multiple regression. In each study area, both ANN and MRA produced similar results.

Most of the research undertaken using ANN's has been applied to residential valuation. Kwon & Kirby (1997) is an exception since they used an ANN for agricultural valuation. A data set of 155 sales was used for properties located within Illinois, USA between 1990 to 1996 with 10% of the sample data used as a test data set. Kwon & Kirby (1997) achieved a high level of price estimation with their model achieving 90% accuracy. This research reported achieving reliable results in a market that had previously not has ANN applied to rural valuation (Kwon & Kirby, 1997). The authors stated that the ANN 'predicted the price of farmland averaging 90% of actual selling price'. This might be reasonably assumed to imply that the ANN achieved results within ± 10% of the sale price.

Although some studies have utilised ANN's for their ability to identify non linear patterns (Connellan & James, 1998) and show some promise as an effective valuation tool, they have been limited by others given the reluctance of valuers to rely on a method whereby the actual process taken by the neural network is not known or

understood (Worzala *et al*., 1995; Waller, 1999). With ANN's less emphasis is placed on how the model was developed, and more on how reliable and accurate the model developed is (McCluskey, 1996). The nature of ANN's are that they do not require any a-priori knowledge as they are 'self learning' or 'self adaptive'. They can be used to characterise a novel problem and make generalisations about it based on rules developed from problems considered earlier (Paris *et al*., 2001). The non-linearity of the ANN has made it the preferred technique by those who believe that MRA does not wholly represent the market appropriately due to the complexities and drivers in force throughout the rural market place (Tay & Ho, 1991).

To some extent ANN's are seen as a tool to complement the practising appraiser and not replace them (Do & Grudnitski, 1992; Evans *et al*., 1992). In some cases there has been advice to 'treat with caution' any result that has been generated by an ANN (Worzala *et al*., 1995). The accuracy of the neural network is reliant on the choice of data to be trained and having the quantity of data available for training and testing. Like that of CBR and MRA techniques, ANN's are data intensive (McCluskey & Anand, 1999) and in the case of rural property, it can be problematic to obtain adequate sale price and property characteristic information. Large variations in the quantity of data used for residential property ANN development ranges from 34 (Evans *et al*., 1992; Kwon & Kirby, 1997) to approximately 1940 (Rossini, 1998). Some of these studies have obtained accurate results. However, given the relatively small number of properties considered by some authors, (eg. Kwon & Kirby, 1997; Worzala *et al.* 1995) the training stage of the process must be quite lengthy for the ANN to be able to detect such patterns.

As highlighted by Worzala *et al.* (1995), ANN's were found to predict different results when using the same data with different software packages. This irregularity and inconsistency in property estimates can lead to uncertainty in the technique and its appropriate usage within the valuation industry. Many studies have compared MRA with ANN and found that the ANN's tended to report similar or higher accuracy than the MRA (Do & Grudnitski, 1992; Kershaw & Rossini, 1999) indicating that with appropriate data, they can be used to determine accurate property estimates which can then be used to verify manual techniques.

The time to run an ANN is somewhat lengthier than that of MRA which can take seconds (Rossini, 1998). ANN's can run for hours (Rossini, 1998) and still not reach a suitable learning stage after this time. Thus, although an ANN can be seen as a valid technique for valuation of rural property, there are still issues that need to be addressed with regard to their stability and the time taken to train the network compared to other automated techniques.

2.3.3 Regression Analysis and Hedonic Pricing Theory

2.3.3.1 Multiple Regression Analysis (MRA)

Regression analysis has been utilised in many disciplines as a tool for explaining relationships between variables. In its simplest sense, it is concerned with fitting a straight line to data so that the sum of the squared residuals is minimised through the technique of least squares. In many cases it can explain relationships between variables and often transformation of variables can increase the explanation between the dependent and the independent variables (MathSoft, 1997).

In applying regression analysis to property valuation, the objective is to determine the relationship, if any, between the dependent variable (property sale price) and a variety of property characteristics (eg. property size, land use, soil types). The significant coefficients can then be used with the developed regression equation for prediction purposes to estimate a value for another property which has different levels of property characteristics. Regression analysis has been used for property valuation as an enhancement to manual approaches as it is less subjective and more cost effective than manual methods (Adair & McGreal, 1988). The approach can help overcome the data limitations of the Income Approach with sale price data being more readily available than property income data (Reid Schott & White, 1977).

In the USA in the 1960's, Multiple Regression Analysis (MRA) was used as a statistical technique for computer assisted property valuation in residential areas (Adair & McGreal, 1988). MRA was used to determine the effect that size, age and building quality have on the valuation of a property and led to further computerised techniques being developed to assist with the selection and analysis of comparables. Regression techniques have been used to analyse data sets and arrive at a value for a set change in a variable and have been more widely used in residential valuation. The following sections outline the use of hedonic regression analysis to property valuation and discuss the number of properties used in each study and their results.

2.3.3.2 Hedonic Price Theory applied to Rural Regression Analysis

Regression analysis has been used quite successfully in residential valuation in combination with hedonic modelling or hedonic price theory. Hedonic pricing is based on the theory that the price of non-market goods and services (ie: the characteristics of a property) can be modelled as a function of the characteristics of those goods and

services. In terms of valuation, the price of a property can be modelled as a function of the structural, neighbourhood and environmental variables of that property. Given that there is equilibrium in the housing market that assumes individuals base their selection of properties on alternative locations (Mahan *et al.*, 2000) and that there is market clearing, then the following is described as a hedonic price function:

$$\mathbf{P_h = P_h (}S\mathbf{,}\ N\mathbf{,}\ Q\mathbf{)}$$

Equation 1 Generalised Hedonic Price Function (Mahan *et al.*, 2000)

Where

P_h is the price of a property

S is a vector of structural characteristics,

N is a vector of neighbourhood characteristics and

Q is a vector of environmental characteristics

The hedonic price function, through the use of partial derivatives can be used to determine the implicit price for a characteristic given a one unit increase in the change of a good. If, for instance, the distance to wetlands was the measure of value chosen by buyers, this implicit price indicates the additional amount that would need to be paid to be located an additional unit closer to a wetland area (Mahan *et al.*, 2000). Many studies have utilised hedonic pricing theory to determine implicit prices of a variety of goods in residential areas. The demand for floodplain areas (Donnelly, 1988), urban wetlands (Mahan *et al.*, 2000), square feet of living space (Palmquist, 1984), access to woodlands (Powe *et al.*, 1997), road development impacts (Lake *et al.*, 1998) and

clean air (Harrison & Rubinfield, 1978) were all studied using the hedonic pricing technique.

Rural areas have been studied in terms of the effect that wetland areas (Reynolds & Regalado, 2002), irrigation water (Faux & Perry, 1999), soil conservation (Gardner & Barrows, 1984) and erosion control (Miranowski & Hammes, 1984; Palmquist & Danielson, 1989) have on the value of rural farmland prices. These studies have used data spanning varying time frames and have considerable differences in the number and type of property characteristics used in their models.

Elad *et al.* (1994) applied hedonic regression analysis to rural properties in Georgia, USA. The sale price data used in this study were derived from records of land sales from unpublished Farm-Rural Land Market Surveys which were conducted by the University of Georgia between 1986-1989. Additional data were obtained from the U.S. Census of Agriculture and the Georgia Statistical Abstract. The study region comprised five geographic areas with varying amounts of homogeneity within each region. Within each sub-market there were between 201 to 386 properties.

The aim of the research by Elad *et al.* (1994) was to find the best functional form for modelling of rural land within the study areas. It was found that within each geographic region, the price influences of the properties varied. Elad *et al.* (1994) demonstrated that the linear and log linear models tested did not capture the pricing relationship and thus results of the R^2 values ranged from 29% to 60% for these models.

Reynolds & Regalado (2002) incorporated wetland area variables into their models to determine their affect on rural land values. The aim was to assess if wetland areas affect rural property values as they limit the use of the land and thus may not be perceived as an asset in landscapes used for agricultural purposes. Four counties were studied within Florida, USA and incorporated 212 sales between 1988 and 1993. Wetland information was obtained from the National Wetland Inventory maps on a parcel by parcel basis. The dependent variable 'price per acre' was used in the modelling. The study used a large number of indicator variables to represent the county location as well as the presence of an irrigation well and road frontage. In addition, in the two stage modelling process there were six variables that were included which were indicator variables of different wetland types. Overall, the models performed well which could be attributed to the more detailed information regarding wetland area types which were hypothesised to influence value. The research was significant in identifying that high levels of price prediction can be achieved when using only a small portion of housing/property specific variables. The use of data spanning a five year time frame could also contribute to price fluctuations and a time variable was included initially in the model.

Xu *et al.* (1993) applied regression analysis for agricultural land within six regional areas of Washington State, USA. The study period was between 1980 and 1987 and data were obtained from sales books from the Farm Credit Service. A total of 928 properties were used (initially 1806 however responses from surveys yielded only 928). A further 23 properties were excluded due to data inconsistencies. In the six regional areas from which models were developed, the R^2 values range from 80% to 94% - indicating models which fitted the regional areas well. The study area comprised 24 states within Washington State and these were segmented into six sub-markets to help provide more homogeneity within each region. The sub-market regions were

specified using location rather than a statistical means. However, there was no justification as to why specific regions were classed into each sub-market (Xu *et al.* 1993). The developed models indicated that across each sub-market, price influences are varied and the nature of this variation and its potential affect on property values needs careful consideration.

Gardner & Barrows (1984) achieved high levels of price estimation with their two models developed within Wisconsin, USA using 158 properties. The study used 'sale price per hectare' as a dependent variable and achieved R^2 values over 90%. This study included a large number of variables within each of the models and these were primarily comprised of land classification codes, slope and erosion information.

From the results of Gardner & Barrows (1984), when erosion is visible, it may have a significant effect on the price a buyer will pay for the land. This result can indicate that erosion does not affect prices if it is minimal, which is contrary to most beliefs (Gardner & Barrows, 1984; Baxter & Cohen, 1997), or could simply indicate that a buyer finds it difficult to gauge how much erosion has taken place and, thus, how it may affect the land. When the erosion is more severe or noticeable, then the buyer can more readily see the affect of the erosion on the land (Gardner & Barrows, 1984). The strength of the models developed in terms of their accuracy could be attributed to a study region that appears particularly susceptible to erosion.

Faux & Perry (1999) achieved high levels of accuracy in their property valuation models for the Malheur County in Oregon, USA. Sale price information was obtained from 1991 to 1995 for 225 properties. Faux & Perry (1999) used a time variable to account for sales spanning multiple years which is in contrast to Miranowski &

Hammes (1984) who adjusted to a common year whilst Reynolds & Regalado, (2002) did not include any variable to indicate time differences between sales. Market segmentation was also tested within this research to ascertain if the data could be segmented into different sub-markets. No market segmentation was found to exist in the study region for rural land. The high levels of accuracy achieved by Faux & Perry (1999) are possibly indicative of a homogenous study region. All properties used were zoned as 'exclusive farm use'. This approach may reduce the risk of using non rural farm land in an analysis.

Miranowski & Hammes (1984) used 94 sales in Iowa, USA between 1974-1979 to model property values. All sale prices were converted to a common year value, 1979 in this instance. Their research obtained data from a non-random sample and it is suggested that this may be the cause of the poor model results (R^2 values ranging between 33% to 51% for the three resulting models). The study used a non-random sample which meant that all properties were within the one study region (central Iowa) and thus there could be bias in the selection of properties. Random sampling of the data would have reduced the number of properties available for modelling. Miranowski & Hammes (1984) also did not use any distance variables (distance to nearby city or town), land use or parcel size variables in their study. Converting all the sale prices into a common year format may have contributed to the poor results depending on the method undertaken to perform these adjustments.

The above studies undertaken for rural valuation using multiple regression analysis all report different levels of accuracy in their models. No two studies use the same number of property characteristics or the same sample size. Xu *et al.* (1993) used a much larger data set than others (n>900) and achieved accurate results. Although it is

suggested that large data sets of property sales are required for any automated valuation technique to be successful (Poor's, 2004), accurate results have been obtained for rural studies that have used a smaller number of properties (Gardner & Barrows, 1984; Reynolds & Regalado, 2002). This may indicate that using a smaller data set does not necessarily yield less accurate results.

The variation in the property valuation models reported in the literature may suggest that each study area is unique in its geographic location and a model developed for a particular regional area may not necessarily be applicable for another nearby area (Xu *et al.*, 1993).

2.4 A Rural Property Valuation Model for Victoria

Rural property values are influenced by numerous factors with each market area subjected to different price influences or value drivers (Fletcher *et al.*, 2000). Baxter & Cohen (1997) discuss the various factors that influence agricultural property values. These range from the climate of the region, presence of irrigation, soil type and capability, economic factors, machinery cost and presence and age of buildings.

Within a number of rural hedonic pricing studies, property area, land use classification and improvements have been used as price influences (Gardner & Barrows, 1984; Xu *et al.*, 1993). In addition, more economic based studies have identified that inflation, interest rates and the tax rate on the land sale are important factors affecting rural values in some regions within the United States of America (Just & Miranowski, 1993).

Valuation Best Practice specifications (Valuation Best Practice Specifications, 2005) document characteristics to include that are specific to rural valuation. These include arable and non-arable areas of the land, access, water supply, fencing condition, water rights, pasture condition, vegetation type, soil type, and unused roads and water frontages. Whilst these characteristics are classed as elements required for rural property within Victoria, many regression analysis studies for rural valuation in other jurisdictions do not incorporate these elements (Miranowski & Hammes, 1984; Faux & Perry, 1999). Typically, each region has specific individual influences that may not always be encapsulated into one global model or set of variables. This could be the reason behind why such variability exists between research in different study areas and why there is no apparent definitive list or model of significant characteristics.

Within the various research reported in the literature, property characteristics have been categorised as belonging to structural, economic, neighbourhood, environmental and accessibility classes (eg. Chicoine, 1981, Miranowski & Hammes, 1984, Elad *et al*., 1994, Lake *et al*., 2000). Not all studies have employed each of these classes in their models (Elad *et al*., 1994). Some studies concentrate on economic variables (Just & Miranowski, 1993).

The basis for the selection of property characteristics for the conceptual models developed in this thesis is based on the rural valuation literature using regression modelling. The following sub-sections consider the property characteristics used in a number of rural hedonic regression analyses, and are used to set the scene for the development of my conceptual approach.

2.4.1 The Development of the Model

Whilst residential property values are primarily based on location and structural improvements, rural properties differ in the role that these and other additional characteristics play. The number of bedrooms, bathrooms and living areas, and proximity to transport, shops and schools may influence the price a purchaser will pay for a residential property (Wyatt, 1997). Rural properties are influenced by the size of the land, its quality, the production rate of crops, irrigation, distance to transport services and the income producing capabilities of the property (Chicoine, 1981; Elad *et al.*, 1994; Reynolds & Regalado, 2002). Other variables of influence include road frontage and building value (Reynolds & Regalado, 2002), land slope and erosion (Gardner & Barrows, 1984), the value of improvements and reason for land purchase (Vandeveer *et al.*, 2000). The Consumer Price Index (CPI) along with season, month, and year indices have been utilised to account for time differences between sale data (Elad *et al.*, 1994).

A number of studies have used regression analysis for the determination of rural property values (Gardner & Barrows, 1984; Palmquist & Danielson, 1989; Elad *et al.*, 1994; Roka & Palmquist, 1997; Bastian *et al.*, 2001). Each study has been applied within different countries, have variations in the number of property characteristics used in their models, have used different numbers of properties and have used different dependent variables.

Gardner & Barrows (1984) used a large data set of property characteristics in their two models which achieved levels of R^2 over 91%. The inclusion of property size, improvements, building quality, building value, house presence, house value, and

distance to towns and regional centres are all used within most of the research in this area (Chicoine, 1981; Xu *et al.*, 1993; Faux & Perry, 1999; Vandeveer *et al.*, 2000).

Additional characteristics more commonly used are some form of time variable that is either monthly or yearly; land use type, irrigation, soil types, soil quality, soil wetness and erosion (Chicoine, 1981; Xu *et al.*, 1993; Elad *et al.*, 1994). All of these characteristics are used in some form, whether they are depicted as indicator variables (showing the presence of absence of the characteristic) or through a quantifiable characteristic such as the percentage of land or number of acres on a property with a specific land use.

Other characteristics that have been used are indicator variables for the reason for purchase (ie: agricultural or industrial) (Elad *et al.*, 1994; Vandeveer *et al.*, 2000) and the inclusion of paved access roads or types of road frontages for a property (Chicoine, 1981; Vandeveer *et al.*, 2000). The addition of erosion, slope and soil characteristics have only been included in a minority of studies. Likewise, more regional characteristics which can often be derived from Census data include population density and the presence of community housing and have only been used in a few studies (Palmquist & Danielson, 1989). Loan details and interest rates have also been used (Gardner & Barrows, 1984).

Xu *et al.* (1993) included more specific property characteristics that included age and number of stalls in milking parlours, percentage of total land acres irrigated by three different types of irrigation along with information regarding barn size and age, house size and age and a variable to include machinery value. The use of more environmental based variables has also been used by Bastian *et al.* (2001), Chicoine

(1981), and Xu *et al.* (1993) who included some form of water body or stream variables along with land use types.

2.4.1.1 Structural Characteristics

Property size, parcel size and farm size are all variations of the same characteristic. This measure of the farm or property is almost universal in its use for rural valuation. Bastian *et al.* (2001), Chicoine (1981), Xu *et al.* (1993), Elad *et al.* (1994) and Vandeveer *et al.* (2000) all use property size and have found the variable to be significant.

Housing and farm building presence, age, condition and value are all variations to building quality used within a number of studies. Variations of these characteristics exist in the research, with Vandeveer *et al.*, (2000) and Reynolds & Regalado (2002) using a generalised 'improvements' and 'building value' variable to depict the monetary contribution of all improvements on a property. Xu *et al.* (1993) used more specific variables to depict house quality and the worth that a building may have to a property. These include the presence of a house, then a variable to depict the size of a house. The use of the conditions of buildings, building presence, building age and building value on a property has been found to be significant in all research studied (Vandeveer *et al.* 2000; Reynolds & Regalado, 2002; Xu *et al.*, 1993; Chicoine, 1981). The use of these type of housing and building variables is not atypical of all rural research and in some cases building information has been ignored. Gardner & Barrows (1984), for example, still obtained R^2 values over 91% despite the absence of building information in their model - indicating that other variables were more influential to price.

Variation in the use of water type and quality variables in the property valuation literature exists. Bastian *et al.* (2001) used an 'irrigation' type variable and found it to be significant, as did the study of Xu *et al.* (1993). However, Chicoine (1981) found the presence of a waterbody or stream to be insignificant. Within VBP (Valuation Best Practice Specifications, 2005) water rights information is retained during the valuation process. The "Smith vs Shire of Gannawarra Supreme Court Case in 2002" led to water rights information having to be included in the valuation of a property for rating and municipal valuations (VBP Fact Sheet, 2004). In an industry such as agriculture that places a large emphasis on water for stock maintenance and crop production, one would expect that some form of water access or water rights variable would be influential to price. As at 2005, half of the State of Victoria was drought declared (Department of Primary Industries, 2005) and the scarcity of water impacts farming operations and places a greater importance on water rights.

Recent legislative changes to water rights has led to changes to allow for the unbundling of water entitlements (*Water (Resource Management) Act* 2005). This unbundling enables the separate entities of water rights to be managed more effectively than when they are one. The implications of this for valuation in Victoria are that for the 2004/5 revaluation, supplementary valuations will need to account for those properties affected by the amendments to the *Water Act 2005*. In addition, the Valuer General of Victoria and the Australian Property Institute are developing strategies to deal with the lower values which may result from unbundling of water entitlements to irrigated property (VBP Fact Sheet 2004). Due to the emphasis of water rights on property values within Victoria and other states and territories of Australia, water rights information is included as a significant variable within the conceptual model even though this type of information has not been used in the regression modelling research reviewed in this thesis.

If water rights information was not available for use in a numerical model, it would be expected that other characteristics depicting the amount of water or proximity of possible water sources to a property may be influential to price. Dams and the length of a water course are other variables that may be used to depict information regarding irrigation on a property. However, these variables have not been used in the other rural research models reviewed here.

Soil type information has had minimal use in rural research with the exception of Palmquist & Danielson (1989) and Miranowski & Hammes (1984) who used a wider range of soil type and quality information. In the case of Miranowski & Hammes (1984), a higher level of price estimation was not achieved, possibly due to a limitation of the small study size and the few characteristics selected. Palmquist & Danielson (1989) incorporated erosion in addition to indicator variables of soil quality. Again, the poorer modelled results may be indicative of the lack of housing and building information, or the use of too many regionally-based characteristics such as population density. Another possible influence to the results may have been the inclusion of such a substantial number of indicator variables compared to the number of categorical variables. Pasture condition is a characteristic representative of the type of land/soil of the property and used within VBP (Valuation Best Practice Specifications, 2005). To date, this variable has had only limited use for this purpose.

Fencing type or presence would be expected to be significant to rural property whose farming type is classed as a stock land use. For grazing of stock, one would assume that some form of fencing and its quality would influence a purchaser (Baxter & Cohen, 1997), especially if large proportion of a property fenceline is in disrepair and requires

replacement. The inclusion of this information in VBP (Valuation Best Practice Specifications, 2005) warrants its use in the model to ascertain the significance considering VBP places it as an important variable.

Windbreak presence was used by Xu *et al.* (1993) as a variable in estimating property values and found to be significant. This variable has not been used in other research reviewed here. However, since it was a variable that could be readily derived from existing topographic data, and considered potentially important, it was included in the conceptual model developed in this thesis.

2.4.1.2 Environmental Characteristics

The use of environmental characteristics varies considerably in rural hedonic research and is dependent on the environmental issue being investigated. This type of variable differs in its definition in that some variables may be seen as belonging to the class of 'structural' characteristics. For my research, a number of additional environmental characteristics were considered to be potentially important. Typically, the susceptibility of an area to fire and flood can be seen as an environmental issue as it also has follow on affects to the land and surrounding soils, flora and fauna. Although a flood prone area is a more regional consideration, a property subject to inundation is likely to be more specific to certain property or properties.

Dryland salinity is a threat to regional Victorian property as it can have a detrimental affect on the land and production capabilities (Wimmera Catchment Management Authority, 2000; West Gippsland Catchment Management Authority, 2001). Dryland salinity was specified in the conceptual model developed in this thesis for this reason. Likewise, pest species may have a damaging effect on rural property and were

specified as a variable in the model (Wimmera Catchment Management Authority, 2000; West Gippsland Catchment Management Authority, 2001).

2.4.1.3 Accessibility Characteristics

Typically, accessibility characteristics are likely to influence the value of residential and rural properties. Distance to towns and major cities has influenced housing prices with higher prices being paid for property closer to regional centres and towns. For rural property, the influence of a regional centre may not be as apparent as there may be multiple regional centres influencing to valuation. Nevertheless, distance to a town is a characteristic widely used in the rural hedonic research literature (Vandeveer *et al.*, 2000; Chicoine, 1981; Elad *et al.*, 1994). Some studies use only one major town from which to derive a distance measure (Vandeveer *et al.*, 2000) whilst Chicoine, (1981) used multiple variables to depict the distance from a variety of regional centres and major cities. The inclusion of multiple variables depicting multiple town distance measures can have the effect of differentiating which town distance is significant and thus which distance measures are not influential to price in that region.

Another variable employed less frequently in the literature is road frontage. Indeed, road presence (Reynolds & Regalado, 2002), road frontage type (Chicoine, 1981) and paved access road (Vandeveer *et al.*, 2000) have been used as variables in modelling, and found to be significant. Road frontage is generally specified as an indicator variable and unless there are multiple properties within a study region where there is no access road or road frontage, then there is often little or no variation amongst the properties with respect to this variable.

2.4.1.4 Neighbourhood Characteristics

Climate plays an important role in rural production and valuation of land (Baxter & Cohen, 1997), especially on a dry continent such as Australia. Set amounts of rainfall are required at specific times in the growing season and these vary for different crops. Excessive rainfall and flooding can have devastating affects on the land, not to mention the loss of livestock in severe cases of flooding. Land subject to inundation may also directly influence the capacity and likely production of a property.

The property valuation literature reviewed here do not incorporate any form of rainfall, temperature, flooding or inundation characteristics into their models. Perhaps the experienced farmer has ascertained the influences of the regional climate and has selected the crop type based on this knowledge. In most cases this is possibly a fair assumption in that crops that require a more temperate climate are not likely to be found in regional Victoria. The purchaser of a property may not be directly able to identify this information as affecting the value of a property if their substitute property is within the same area and most likely subject to the same climate, flooding and temperature conditions. Climate variability may only have a minor affect on price and, thus, may not be directly measurable against how much a buyer is willing to pay for a property in another slightly different geographical region. The use of climate as a variable may be been discounted if researchers believed that its influence on property values would be captured (indirectly) within another property variable such as soil quality or pasture condition.

Planning zones are an indication of the potential use of a property and exactly what prescribed uses can be made of a property. The significance of the planning scheme of a property is reflected in many hedonic regression studies, either as a indicator

variable to depict the land zone a property falls within, or as a purchase reason indicating the property was purchased for commercial or recreational uses (Vandeveer *et al.*, 2000). Vandeveer *et al.* (2000) found that purchase for commercial use positively affected price whilst recreational use negatively affected the price. Chicoine (1981) found that commercial/industrial and also mining and quarrying land use types had a significant effect on price whilst land zoning classes were not found to be significant. A zoned agricultural property has the potential for rural land uses whilst one that has a rural/residential has multiple. The location of a property in a rural/residential zone may lead to the property being valued at a higher rate, or being sold for a higher amount due to buyers wanting a 'sea-change'.

The county or LGA to which a property belongs has also been used in hedonic research. Whilst some studies use the county as a means to segregate property into different sub-markets and run separate regression models for each county, others have incorporated indicator variables to depict county location and develop models for the whole region (Reynolds & Regalado, 2002). Xu *et al.* (1993) amalgamated multiple counties (between two to six counties) to form six sub-markets and subsequently developed one model for each sub-market. This may have been a more intuitive means to deal with the location of a property within a specific region, or allowed for more homogenous regions to be developed using these amalgamations. The research, however, fails to indicate the relationships behind these amalgamations. It is difficult to ascertain why these county groupings into sub-markets appeared to work more effectively as reflected by the more accurate models. The sub-market grouping technique used by Xu *et al.* (1993) is geographical and it asks the question as to how sub-markets may be developed using more statistical techniques.

2.4.1.5 Economic Characteristics

Economic influences to rural valuation are significant in Australia as buyers perceive market place fluctuations to gauge a property's future earning potential (Baxter & Cohen, 1997). Seasonal trends may operate in an area and natural influences may heighten any economic fluctuations. An example of this is the location of a property in a flood, drought or fire prone area may decrease the worth of a property given these natural risks. The use of 'site valuation' to gauge or determine trends in an area has not been used in the academic research analysed. Site valuation is monitored over time to determine variations between municipal site valuations during the Valuation Best Practice process. Although having not been used in academic research it was included within the model due to its possible use for trend analysis.

The use of the Consumer Price Index has not directly been used in the research studied. Indirectly it is incorporated into a time based variable for this research to provide a means to align sale prices over time to a common year representation.

Production of a property can help measure the economic characteristics of a property. Suter (1974) described a tabular 'field book' for rural appraisers using various property details to assist in valuing rural property manually. Whilst the reference is quite dated, it does emphasise the need to detail the production capacity of each property, specifying the acreage of crops and their yield, the amount and type of livestock, required feed for stock etc to then enable the earnings of each property to be generated. Productivity of a farm has been less utilised in hedonic research. A soil productivity index was devised by Chicoine (1981), however, the variable was not found to be significant at either the 0.05% or 0.10% level. Bastian *et al.* (2001) used

fish productivity and carrying capacity of the property as variables to assess properties in an agricultural region primarily influenced by fishing and elk habitats.

The lack of specification of production characteristics in rural models may be an indication of the complexity of the rural market or the difficulty in obtaining more specific data. Production may not greatly influence price, especially where properties are being used for non-agricultural purposes. If this is the case, then the buyers may simply not incorporate income earning information when trying to find similar properties with the same characteristics. Another possibility is that production information was not easily able to be obtained for rural valuation and, thus, many studies have simply not used the characteristic for hedonic modelling. It is more likely that income and production of a property does affect price, but may not directly be capitalised into the sale price.

2.4.2 Restatement of the Model

As highlighted from the review of the research undertaken for rural valuation, property characteristics used and reported as significant to the price of a rural property differ. Some studies have more of an environmental focus (Bastian *et al*., 2001, Chicoine, 1981), whilst others use more locational-based property characteristics (Vandeveer *et al*., 2000).

A number of property characteristics are used more frequently (land use, distance to town, soil type, monthly time index), however, in the research studied it was strongly evident that three to four particular variables are the best indicators to use to predict property prices for rural land. In residential valuation, a more definitive list of characteristics is more evident and widely used in most regression studies ie. (house

size, location, property area, number of bedrooms, number of bathrooms). For rural valuation, my literature review indicates that some combination of structural characteristics (buildings, building age), environmental features (land use, soil types, erosion), and regional features (climate, population, flood prone regions) is important for undertaking reliable valuations (Baxter & Cohen, 1997). It should be noted that the extent that each of these may affect price can differ in each regional area (Xu *et al.* 1993).

Table 2.1 provides a summary of the property characteristics and their significance based on a review of the rural property valuation literature. The characteristics house size, house age, building size and building age and the building value per acre are most frequently reported in the literature to be significant factors affecting rural property values. In addition, the characters distance to city, distance to freeway, abuts a town and soil quality were found to be significant in some studies reported in the literature.

	Xu *et al.* (1993)	Vandeveer *et al.* (2000)	Reynolds & Regalado,(	Gardner & Barrows (1984)	Palmquist & Danielson (	Miranowski & Hammes (	Elad *et al.* (1994)	Chicoine (1981)	Faux & Perry (1999)	Boisvert *et al.* (1997	VBP (2005)
Property Size				U						X	U
Time / year of sale/ time owned	U										
House size											
House age											
Barn size											
Barn age											
Building value or building value p/acre											
Building quality – average/good					X						
Machinery value											
Value of improvements				U						U	
Distance to town/ nearest town								X			
Distance to city											
Distance to freeway											
Abuts a town											
County location	U										
Percentage of land in pasture/crop	U										U
Non-arable area											U
Land capability class/code	U										
Soil quality											
Soil type											U
Erosion/ potential erosivity				U							
Population density										U	
Pasture condition code											U
Water supply code											U
Fencing condition code											U
Water rights											U

Table 2.1 Summary of the variables most frequently used to estimate property values based on a review of the property valuation literature (- significant, X - not significant, U – unclear)

2.4.3 Description of the Conceptual Model

A conceptual model was developed to provide a basis for the automation of rural property valuation in Victoria (Figure 2.1). The model was developed to represent the influential variables hypothesised to be potentially most significant for rural valuation. Implementation of the conceptual model to develop various numerical models allowed their performance (defined by the level of accuracy of valuations) to be examined.

The dependent variable (sale price) along with the independent variables of the conceptual model are shown in Figure 2.1. The 'Price' of a rural property is made up of a vector of Structural (S), a vector of Environmental (EN), a vector of Accessibility (A), a vector of Neighbourhood (N) and a vector of Economic (EC) variables. Under each of these category headings is the list of independent variables to which each property characteristic belongs.

The 'Structural' characteristics are concerned with the individual characteristics applicable to each property and not a regional area. These include the size of the property, the presence of a house, its age and quality and the presence of farm buildings and their condition. More specific to each property is the presence of irrigation, presence of dams, length of watercourses and water rights information, each providing information regarding water access and availability. The type of fencing along with more agricultural specific variables such as land use type, proportion of each land use on the property, soil type, pasture condition and road frontage provide further information about the agricultural productivity and influences of the property.

The 'Environmental' variables depict any pest species on the property, the type of dryland salinity if applicable, the extent to which the property may be affected by being

located in a bush fire prone area or in a flood prone area. These variables are categorised as 'Environmental' due to their more regional exposure in the community and influences by the environment. They could, however, be classed as 'Structural' (as they do depict variables specific to each property) or 'Neighbourhood' in that pests and bush fire generally affect a wider locational area and may not just apply to one single property.

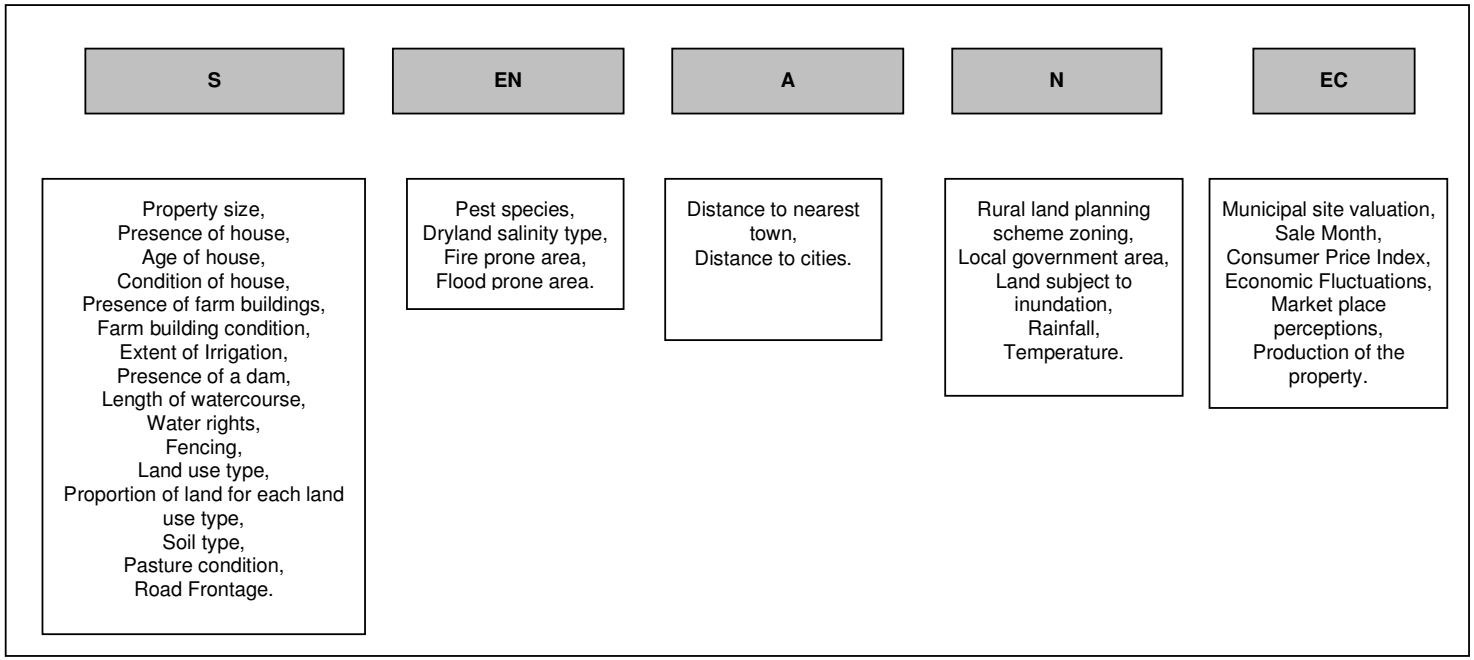

Figure 2.1 Diagram of Conceptual Rural Property Valuation Model

The ‘Accessibility’ variables included were the distance to the nearest town and distance to cities providing a measurement of the closeness of each property to local centres and major regional centres. These variables are important when transporting produce and stock to the marketplace. GIS analyses could measure the time component of each of these instead of the distance, and thus provide a more accurate measure of proximity.

The ‘Neighbourhood’ variables include the planning scheme code to which each property belongs, the LGA, whether the land is subject to inundation and the amount of rainfall and temperature of the regional area.

The ‘Economic’ factors include the municipal site valuation of each property, the month of sale, the consumer price index, economic fluctuations, perceptions within the marketplace and production of the property. Some of these could fall into the ‘Structural’ characteristics category as they are applicable to each individual property and not a regional area (ie: municipal site valuation and production of the property).

The acquisition of digital data on the conceptualised variables and their integration into a property database are discussed in Chapter 4.

2.5 Summary

This Chapter examined the various reasons why property is valued and the different methods used for rating taxation valuations (Site Value, Capital Improved Value, Assessed Annual Value). The Cost Approach, Income Capitalisation Approach and the Sales Comparison Approach are the three main approaches to land valuation. The Cost Approach uses replacement costs of the property, in particular inventories of machinery and the type and quality of buildings located on the property. The Income Capitalisation Approach assumes that buyers are concerned with the income stream of a property that they are purchasing and thus information regarding the income sources is required for this type of valuation. The Sales Comparison Approach is reliant on searching for similar properties and adjusting the sale prices to account for differences between the subject and the selected comparable properties.

Manual techniques for valuation are discussed - in particular the present Valuation Best Practice standards in use in Victoria for rating valuations (Valuation Best Practice, 2006). Although computer based techniques are incorporated into the process and mapping and statistical analysis is performed to validate results, the rating valuation process for Victoria is still largely a manual process.

A discussion is presented on the various automated techniques presently in use worldwide for both residential and rural valuation. The use of rule based techniques (Case Based Reasoning, Criteria Weighting) is highlighted along with their limitations which include the requirement for elicitation of knowledge from an 'expert' prior to development of the models. Expert systems and Artificial Neural Networks (ANN) are examined. ANN's are deemed more suitable in a complex market as they are not linear based and are able to 'self learn' from previous modelled examples and

determine relationships between data. Limitations in the consistency of results may arise when different software packages are employed (Worzala *et al.*, 1995). ANN's can take considerably longer to arrive at a valuation estimate than manual methods or techniques which use Multiple Regression Analysis (Rossini, 1998).

Multiple Regression Analysis and its use within the theoretical aspects of hedonic pricing has ensured that it has become the most widely used technique for valuation (McCluskey, 1997; Rossini, 1997; Poor's, 2004). Although this technique is widely used, it is not without its limitations as it is based on assumptions of linearity between variables this may be invalid in many situations when undertaking property valuations. The technique involves a degree of statistical knowledge in producing models and requires some expertise for the selection of appropriate characteristics to include within a model so as to avoid issues associated with multi-collinearity.

The Chapter examines the various models that have been developed for rural valuation. In particular, it seeks to provide some insight into the accuracy of the rural models developed and the property characteristics used within these models.

Chapter 3 GIS modelling and Sub-market identification for Rural Valuation

3.1 Introduction

A GIS is ideal to store property sale information and to display and query data. GIS can facilitate a greater understanding of each property by deriving additional variables during the spatial overlay of multiple data layers, perform measurements between features and often show relationships with greater ease (Lake *et al.*, 2000). In many cases, the wide array of characteristics required for rural property valuation are generally not contained in sale price data thus additional data is required from a combination of sources. Obtaining data from various sources can lead to data issues and integration problems. Once data is integrated, additional variables can be derived through GIS analyses.

Watkins (1999) argues that a regression equation determined without the segmentation of property will result in aggregation bias in the models developed and that property needs to be segmented for hedonic valuation. Cluster analysis is a numerical technique used to segment or classify data into natural groupings (Everitt *et al.*, 2001). Cluster analysis is a technique which has been used to segment markets for property valuation. This Chapter aims to discuss the present use of cluster analysis in the valuation industry as a tool for statistically segmenting property into homogenous regions. As Watkins (1999) suggests, this will possibly enable more accurate estimations to be determined than models produced without segmentation or through models developed using geographical 'a-priori' techniques. This section details the use of cluster analysis for determination of classes, the basic principals of the technique

and the different variations that have been applied within housing and valuation research.

3.2 GIS Data Integration Techniques and Issues

3.2.1 Considerations affecting Database Design

Database design should take into account any limitations of the data, the future application of the database, and should in turn drive the data required for the project. Adequately documenting and analysing the possible future applications and ensuring growth is catered for, will enable the database to be designed with the current use and likely future changes taken into consideration.

The knowledge of the end user is another factor affecting the design of the database in terms of how data are stored and represented, and the type of processing and queries that are likely to be performed with the data. Considerations to be addressed during the database design stage are whether the data need to be updated by multiple users at the same time and whether multiple users can query the data simultaneously.

For property valuation, the number of attributes required to describe each property can be quite large and in most cases not all found within the one data set (Wyatt, 1997). In designing the various layers in use for a property data set, consideration should be made as to what attributes are necessary and the representations of the data to be stored, along with the time and cost involved in obtaining and applying regular updates to the database.

Thomas (2000) discussed a GIS database designed for storage of property information at a local county level. Considerations highlighted are address determination of parcels, utilising the same line work for a number of layers where lines overlap, and not having duplicate lines in multiple layers. Other important aspects of property data are that whilst some features are representative only of each individual property, other data may be common to multiple properties and this aspect needs to be factored into the design from early on.

The way in which attributes are presented within a database is an important consideration. The use of data in statistical software for regression analysis will reflect the way that attributes require representation in the database. Whilst only the table of information relating to the data is required to be exported and used in statistical software, it is necessary to specify any categorical variables as indicator variables prior to statistical analysis. An example of this is where the 'susceptibility to erosion' may be classed as either high, medium or low in a GIS database. This information is required to be represented through the use of 3 indicator variables in a statistical software package (Table 3.1).

Indicator Variable Name	Indicator Variable Description
Erosion_high	(1,0) 1 if erosion susceptibility is high, 0 otherwise
Erosion_medium	(1,0) 1 if erosion susceptibility is medium, 0 otherwise
Erosion_low	(1,0) 1 if erosion susceptibility is low, 0 otherwise

Table 3.1 Indicator Variable Example Definition

The issues with a new attribute specification is that if the original attribute is retained in the data then it can be used for visual display to show concentrations of high, medium and low erosion susceptibility. If it is removed then GIS mapping requires the use of

three attributes to show the same thing. An easier solution would be to retain both so that the indicator variables can be used for statistical purposes whilst the initial categorical variable can be used more readily in GIS analyses by not having to perform queries on three attributes as in the above example.

3.2.2 Technical Issues Associated with Data Integration

The technical issues associated with integration of multiple data sets, range from scale and format variations, age of data, data inconsistencies, different representation of features, data quality and data standardisation problems (Longhorn, 1998). Often data distribution agencies store the data in multiple formats and can offer a conversion facility for the purchaser thus minimising the integration the purchaser must perform on the data. However, a major concern with this is that the purchaser is not aware of the techniques that have been used to convert the data set and whether there have been inaccuracies introduced.

Incompatibilities with software storage formats are a common problem since there is a greater range of software packages on the market with each creating unique data formats. To overcome these problems, a further method of processing is often required on the data (Shepherd, 1991). There are a number of conversion translators available that will convert from one data format to another either as a stand-alone translator package or within an existing GIS such as the Universal Translator within MapInfo Professional. Common methods of conversion are either a one-to-one conversion translating directly from one format to the new format, or translation using an intermediate format. To assist in future integration, creation of in-house formats to store data should be analysed as to whether these will benefit the process or hinder it in future endeavours. Non-standard formats can further escalate the problem for

standardisation and for future integration problems as there are more format discrepancies to overcome.

Although the transfer of digital data is improving and the different format types in use by various GIS software packages is managed through the use of translators, the issues of data incompatibilities, incomplete data sets, accuracy, scale and format differences can still be an impediment to the effective use of the digital data.

Scale is a common problem in the data integration process with different organisations collecting and representing data at different scales (John, 1993). Whilst the collected data may be accurate for a particular data set, often the integration of multiple data sets leads to problems with errors arising in the merged data set. This can be in the form of resolution problems or small differences in lines (termed sliver errors or overlaps) being formed, with the quality of the data being compromised upon integration.

The extent of the data set and any edge matching considerations can affect the quality of the data. Data stored in different map sheet files may not have been edge joined to abutting sheets. Where similar type data is obtained from different agencies there can be discrepancies in the representation of the same feature. For example, a lake may be classified as perennial on one map sheet and intermittent on another even though it is the same body of water. Some areas of data may even be incomplete and may have to be added through additional digitising or there may even be overlap of features.

The entity representations and attributes of real world features will differ between organisations depending on the purpose of their information and, therefore, some features will be described using different attributes (Shepherd, 1991). Kuhn (1994) discussed the semantics of spatial data and perceives data sharing and transfer as a communication problem. For information sharing to be successful, a common language or standard needs to be developed to represent these features with Kuhn (1994) stating that if the discipline areas are too distant between the users' data, then a common representation and classification is very difficult to achieve. The possible applications of spatial data are somewhat reduced if restricted to a specialist discipline area. On the other hand, increasing the range of users will lessen the chance of success of integration and accuracy in the data.

3.2.3 Data Quality, Standardisation and Metadata Usage

Data quality issues are an integral part of integration for decision support and GIS based applications. John (1993) highlighted the need for accuracy and for recording the data set accuracy to obtain effective output from a Decision Support System (DSS). The origin of the data and its accuracy, determined from measurement errors associated with the data collection, will generally prevail throughout the entire data integration and processing (Goodchild, 1995) and can influence the quality of decisions being made.

The metadata, which provides detailed information concerning the data set it accompanies (Kim, 1999) can provide a user, or data purchaser with relevant information to decide if the data set will be useful and to what extent the data needs to be manipulated to serve the users' purpose. Metadata should provide information on the data sets purpose, who the original data was collected by, the custodian and the

geographic extent of the data set. Other details should include the currency, its status and whether updates and maintenance are continuous. There should be information regarding the access of the data including what format/s the data are stored in, what data formats are available for supply and if there are any restrictions on data access. Current Australian metadata standards (ANZMETA DTD version 1), detail the lineage, attribute accuracy, logical consistency and completeness of the data as outlined by the Australian Spatial Data Transfer Standard (ASDTS) (Wong & Wu, 1996; Kim, 1999).

Metadata standards are beneficial by providing relevant details to ensure a user has a thorough understanding of the accuracy and intended use of the data and the capacity in which it can be used, however, they fail to support variation in accuracy in the data (Wong & Wu, 1996). Current standards only allow for one description of the accuracy of the entire database so the confidence of the decision could fluctuate over a specific spatial area and uniformity issues could arise.

In perusing spatial data catalogues to find relevant data sets, it is often difficult to determine the type of information and attributes which are stored in the data set. Kim (1999) devised seven essential metadata elements for spatial information and apart from the standard elements already present in the ASDTS, new elements include the price of the data along with entity and attribute information. The additional entity and attribute elements proposed by Kim (1999) aimed to highlight the features that are used (ie: roads, elevation) along with the attributes of those features (ie: widths, heights). This additional information would be beneficial as often the data set description is not detailed enough to provide the user with sufficient information about the contents of the data set.

Current Australian metadata standards for the transfer of geospatial information do not provide information on the datum or coordinate system used. The translation of data to a specific datum or coordinate system can influence the accuracy and the processing time required for data translations and integration. In cases where a projection is new, often the projection parameters have not been updated in GIS software and thus to convert from one projection to another requires that the parameters of the new projection are known. In this research, coordinates were supplied as VICGRID however the version of ESRI ArcInfo that was used in this research did not have any parameters specified for this projection. The latest version of ESRI products does have VICGRID projection parameters specified so that it is now easier to translate. MapInfo Professional also allows for input parameters to be specified.

Although metadata standards are sufficient in providing information regarding the accuracy of each data set available for purchase, they seem to be lacking further information which specify the attributes and extents of each coverage. Attribute accuracy, positional accuracy, lineage, logical consistency and completeness are specified, however there is little additional information specifying the extent of the coverage. The 'bounding box' details the upper left and right and lower left and right coordinates of the data set. This gives an indication of the extent that the data encompasses, yet when there may be no features in specific sections of the data set, this can be misleading for users' that only need a small portion of the data. In essence the metadata does provide the relevant information regarding accuracy and data quality issues, but may fail to provide sufficient information concerning the contents of the data set in terms of attributes and quantity of information contained in each data set.

The metadata contains a description of each data set and may include additional data layers that accompany the data theme, yet it does not provide more detail regarding the fields within each data layer and the representation of these within the database (such as a data dictionary or index to fields). Although VICMAP, a vast series of different spatial data covering Victoria does have a 'data dictionary' detailing all fields and data layers within each data set, it is not connected to the metadata and is sometimes not available. Within the VICMAP data themes, Property, Hydrographic, Transport etc; a data dictionary is available to be downloaded within the product description of the data set which is separate to the metadata of the data set. For the remainder of the data sets available for public use and managed through the Corporate Spatial Data Library document; (a document outlining the data sets held by the State Government of Victoria), this information is not readily available. Providing an additional document or category within the metadata to provide this information would enable a more thorough analysis of each data set to be made prior to purchase.

Once a user has ascertained if the data are adequate for the required project, there needs to be some additional form of information to depict the actual features within each data set. Data sets can often be purchased to find that the features within the data did not overlay a users' study region. When obtaining data within a small study region, say 20km^2 then there may not actually be any features within that region from the acquired data set. This identifies a key issue in making spatial data available and accessible to others. There are a number of ways that this could be dealt with. The use of a descriptor within the metadata to identify the geographical extent of each clustering of features may be more suitable to provide the areal extent of each major groupings of features. However, this may not be suitable for when there are multiple clustering of features. Other techniques for point data would be to provide a count of the number of features within the data set. Another issue associated with this is that

often metadata is applicable to a data set which has state-wide coverage and thus for users who only require information for one or a few LGAs, the metadata may not be totally representative of that LGA. This issue may also arise for users specifying their data requirements based on bounding extents.. Further discussion on ways to specify information in metadata of point, line and polygon data is presented in Section 8.2.

3.3 GIS use for Automated Valuation

Geographical Information Systems (GIS) are used for the handling, storage and manipulation of large volumes of spatial information. A GIS allows data of a spatial nature to be stored, retrieved and displayed with an ability to transfer data into information, thus giving more value to existing data (Shepherd, 1991). GIS can be a powerful tool due to its visualisation and processing abilities.

GIS can be a useful tool for both manual and automated property valuation due to the spatial nature of the data used in valuations. The value of a property is influenced by location as well as structural and environmental influences. GIS may assist when specific data are not available or accessible, by utilising existing data sets to perform GIS overlays which will produce additional property variables. Although it may take considerable time over manual methods to incorporate various data sets into the one GIS database, once integrated, modification and measurement of variables is less time consuming than manual methods. The following sections provide a more detailed analysis of the ways in which rural valuation can be improved through the use of GIS.

3.3.1 Use of GIS within current Automated Valuation

Lake *et al.* (1998) used GIS as a tool to derive additional property characteristics. A number of distance variables were computed, including distance from shops, parks, rail stations and a travel time variable. This involved using network modelling within the GIS to determine time travelled and determining a ranking to specify the speed at which each road segment is travelled. In total 327 variables were derived using GIS for this research (although many were eliminated due to collinearity with other derived characteristics such as travel time and distance). Euclidean distance measures are frequently used to determine measurement characteristics from GIS analyses. Distances to nearest schools, colleges, highways and towns are common distance measures used for valuation (Rosiers *et al*., 2000). 'Distance to highway' and 'distance to town' being more common to rural valuation.

More detailed property characteristics were determined through GIS functionality by Bastian *et al.* (2001). These were mainly 'distance to town' and the area of 'Elk Habitat' on each parcel derived through the use of an intersection function between the parcel polygon layer and the Elk habitat layer. Determination of this through manual methods would have been time consuming and less accurate than a computer generated technique which more readily measures areas.

A number of applications have been developed which link the spatial information of a property within a GIS. Hardester (2002) used ESRI GIS software to develop a property database for comparable sale analysis. The developed application 'ProMap' facilitates display and query, and the mapping of property more readily (Hardester, 2002). The system does not actually generate values based on hedonic regression, but can generate a number of reports and has a comparable sales algorithm incorporated into

the software to verify assessed property values determined through more traditional means. The GIS functionality of the application enables multiple themes of data to be displayed and mapped and is Internet based and accessible to the general public so they can query property data online.

GIS can also be used to select similarly located property for comparable analysis. GIS enables an appraiser to determine similar property by enabling the query of all properties within a set distance from the subject property that have the same number of bedrooms (Castle, 1994) or other variations on similar characteristics. GIS can also be used for thematic mapping of property characteristics, for analysis of comparable property through a spreadsheet software program (Castle, 1994) and in the display of estimated values obtained from regression analyses (McCluskey *et al*., 1997). The GIS enables outliers in results or errors in data to be more readily determined through visual analysis. Through thematic mapping of land zoning, one can easily depict instances of incorrect codes being applied to properties (Castle, 1994). Discrepancies and errors in data can be detected and rectified.

Once data are integrated into a GIS, a wide array of spatial tools can be used. Display and query of data for use in valuation prior to developing models is often undertaken to remove outliers, and visualise discrepancies in data, however the latter could be performed more readily using a query on the data to identify irregularities within the database. GIS use to derive more meaningful property characteristics through measurement of distances, areas and the more wider use of spatial analyst functions, enables more variables to be defined within a data set or database.

3.3.2 GIS for Rural Valuation

Existing property information that contains sale price data on Victorian rural properties may hold limited information pertaining to each property. To use any automated techniques in Victoria the tabular property sale data (PRISM) requires geocoding within a spatial polygon data set so that the textual descriptors of the property can be spatially linked. With the lack of detailed information representing property within the PRISM data set, additional data needs to be acquired and property characteristics derived to develop a more comprehensive property database. GIS can be used for this purpose whereas, previously, much of the information required by valuers was represented in textual hardcopy data and was not easily integrated within a digital environment. Development of an appropriate database containing variables hypothesised to be significant for rural valuation involves:

1. Determination of the property characteristics likely to be influential,
2. Searching for appropriate data sets which contain information regarding property characteristics,
3. Integration of the data into a GIS to enable further information to be populated into a property database, and
4. GIS analyses on multiple GIS themes or layers to derive additional property characteristics.

Integration of data, once acquired, can often be relatively straightforward, or in cases where data are supplied in different projections and software formats, may take considerably longer to integrate. Where data are spatial in nature, but not represented by geographically located polygons, points or lines, it can be more troublesome to

integrate. Any data supplied only as a Microsoft Excel spreadsheet, which does not have any spatial linking attribute attached to the data, will require either automated geocoding (if property address details are sufficient) or a manual approach to be used. This influences the amount of data that can be integrated especially if a manual approach does not yield high matching rates between the textual data and the spatial GIS data.

The enhancement of a property database to enable adequate property characteristics to be populated can be assisted through the use of GIS technology. The research which has applied hedonic regression techniques to rural valuation was predominately completed between the early 1990's (Palmquist & Danielson, 1989; Xu *et al.*, 1993; Elad *et al.*, 1994) and the late 1990's (Boisvert *et al.*, 1997; Marano, 2000). The earlier research in rural valuation did not use GIS technology for display and query nor for visual analysis of results. The latter research for rural valuation is increasing in its use of GIS, however in comparison to residential valuation, its use for rural valuation is minimal.

Soto (2004) used GIS to derive some variables for rural land in Louisiana. The most notable were the distance variables which determined travel time through network analysis. Although a Euclidean distance measure provides a simplistic approach to compute distance and time from a property to a town, Soto (2004) used a network approach which measures a route along a set of roads, highways rather than a straight line. As indicated by Soto (2004), a further refinement could be undertaken to incorporate travel delays amongst the network model thus leading to a more accurate measure of time between two points. Whilst this could enhance estimates, travel delays may be seasonal or related to the time of day travelled and thus could not be

incorporated into a variable to account for these variations. In addition, as undertaken by Soto (2004), the use of distance as well as time variables in a regression model can lead to variables which are collinear, especially if the distance has been determined along the same path as that of a time travelled variable.

Bastian *et al.* (2001) used GIS to measure recreational and scenic variables associated with rural land. A 'distance to town' measure was performed to derive a variable indicating the 'distance from the centroid' of a parcel to a town with a population of over 2000. Area of elk habitat was defined as mentioned in Section 3.3.1. A Digital Elevation Model (DEM) was constructed to provide a theme depicting the visibility at the centroid of a property. This used vegetation height and a land cover data set to classify the DEM into average vegetation heights for each land cover. 'On parcel trout productivity' was also determined by using a river data set and intersecting with the property data set. A manual process was undertaken to classify each stream and allocate stream names to each line that was intersected in the stream data. A further database was obtained detailing trout productivity, but required manual interaction to populate the information into the resultant intersected stream data set. Within the research performed by Bastian *et al.* (2001), the R^2 values of the models ranged from 0.60 to 0.61 which didn't indicate a high level of model predictability. The use of GIS may have lead to a higher degree of accuracy. However, regression models were not developed using the variables that were not derived through GIS analyses, therefore a comparison between the two could not be made. It remains to be seen if the development of further measurement variables from GIS analysis will enhance these models.

Vandeveer *et al.* (2000) used GIS to a lesser degree and fewer variables in their models than Bastian *et al.* (2001). Initial visual analysis through a GIS software package showed that properties located closer to a commuting area had a higher sale price per acre. The article fails to indicate what variables were derived through GIS overlays, or if the distance measures used were based on Euclidean distance however it is likely that they were derived using GIS. It is possible that the presence of a 'paved access road' may have been generated through overlay of any GIS data layer, however, considering the data was obtained through mail surveys then it is possible that the population of each variable was performed manually. Although minimal variables, if any, were populated through the use of GIS techniques, Vandeveer *et al.* (2000) used GIS to create a contour map based on property estimates. The 'iso-price' map was generated using GIS software and depicts areas which have similar price estimates per acre. A contour line was drawn at $500 price intervals and was found to better depict areas where prices rise significantly between neighbouring properties. This approach provides another tool for appraisers to visualise pricing fluctuations amongst geographic regions.

Although an extensive array of techniques can be used to derive variables/property characteristics within a GIS, there still remains a need for suitable data with which to perform these analyses. As was the case for Bastian *et al.* (2001), some data and attributes required manual input to enable further analysis to be undertaken. Where information is available from a printed map, transfer of the information either manually or digitally through digitising can take some time depending on the number of features within the data.

3.3.3 GIS Techniques to enhance Rural Valuation

Taking the step towards a more automated approach with GIS, then allows for the spatial overlays between data layers to be performed to generate additional property characteristics for each property. This can involve intersecting data layers and populating the database with these new values to obtain more intelligent characteristics regarding each property. Where the data containing sale price information or previous valuations contains little information, this allows for characteristics for individual property to be derived. Spatial overlays, intersections and unions between data layers, creation of digital elevation models (DEM's), performing network modelling along paths and distance-allocation modelling using a raster data set are all examples of more advanced uses of GIS.

In rural valuation where data are sometimes obtained through questionnaires or where data is in a tabular non-spatial data format (ie: Microsoft Excel with no spatial link to a graphical data set), obtaining additional information is important to enhance the existing data. Once digital data is integrated into a common GIS software format and map projection, it can then be used to populate additional property characteristics through spatial overlays between data sets.

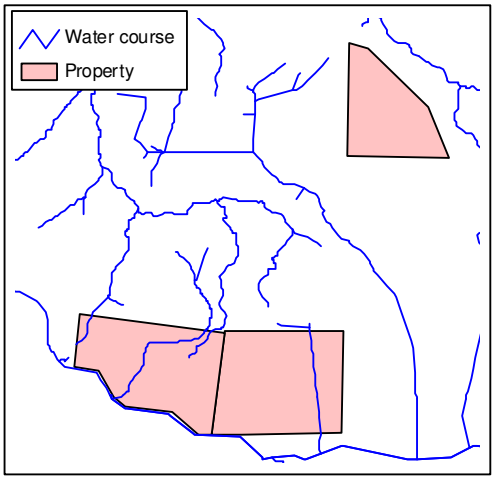

Figure 3.1 Unclipped data layers

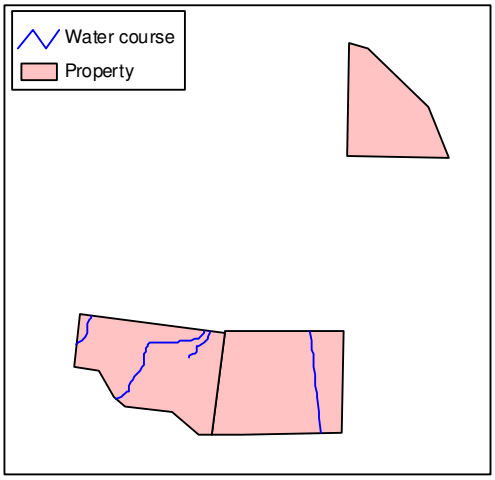

Figure 3.2 Clipped data layers

Figures 3.1 shows a selection from two themes, water courses and the property data containing polygons for a small portion of the Horsham LGA. A 'clip' function has been performed within Figure 3.2 that shows the polygons containing water courses. The resultant table depicts the length of each segment of water course within the clipped layer. This enables aggregation of data to be performed once all line segments of water course have a property identifier which relate each line segment to the property that it falls within. Data can be aggregated whereby a field can be selected from the table (such as the 'length') and the values of the 'length' field will be summed during this process.

The above procedure can be used for point, lines or polygon features. For a point feature, the number of points within the property polygon can be obtained, so too can the area of polygons (such as lakes) be determined through the above procedure. These additional characteristics for each property can be obtained for length of water courses, number of dams, area of lakes or larger waterbodies for each property or other similar property characteristics.

Calculations between fields can also be performed and allows computations such as the 'sale price per hectare' or acre to be computed by dividing the sale price by the number of hectares. In addition new variables can be created which represent the percentage area of the waterbodies on each property, or the percentage of a particular land use for each property providing the land use information is available. Likewise any areas which are presented in km^2 can also be readily computed to hectares or acres depending on the requirements of the study.

Any distance measures from the centroid of a property to a feature can be measured using two techniques. Distances to towns, cities, streams, highway, schools, railway stations etc can be determined using either a 'Euclidean' distance (Figure 3.3) or a route network analysis measure (Figure 3.4). Likewise, time travelled to these features can be determined from the distance, however computing it from the actual distance obtained will make the two variables correlated. An example of this is that if the distance to a town is 10km and the speed travelled to reach that town averages 60km p/hr then the time taken to reach the town will be 10 minutes. Computing the variable using this technique means that the resultant variable (time) is correlated with distance. Another technique to minimise correlation is to apply different speeds for different road segments based on the class of each road. A highway will demand more speed of the motorist, whilst a minor unsealed road will generally have a lower speed and therefore travel time will be greater. This procedure requires that a specification is made of what limit the speed will be on each class of road and assigning these to each road type. It also requires a network of routes to be determined which have some inbuilt knowledge on the best route for the motorist to take from each property. An example of this may include specifying that only roads that are sealed are to be travelled on or that you can not travel a set distance in the opposite direction of travel.

Distance measures using Euclidean distances are easier to determine as they are not reliant on determining a route network and performing route modelling or shortest path analysis. Euclidean measures are determined by creating a centroid for each polygon and then computing a distance between this point and a town/city point. The resultant measures then have to be populated back into the original data set from the temporary centroid data layer created during this process.

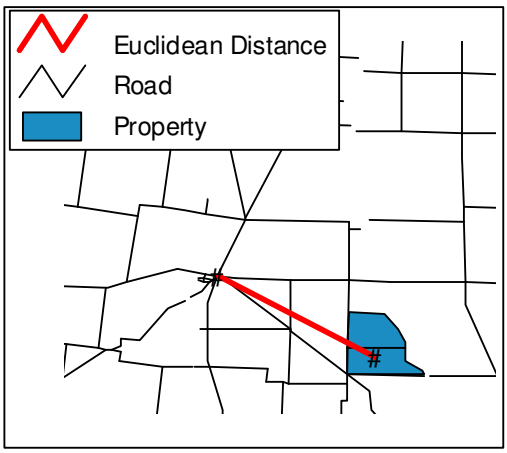

Figure 3.3 Euclidean Distance

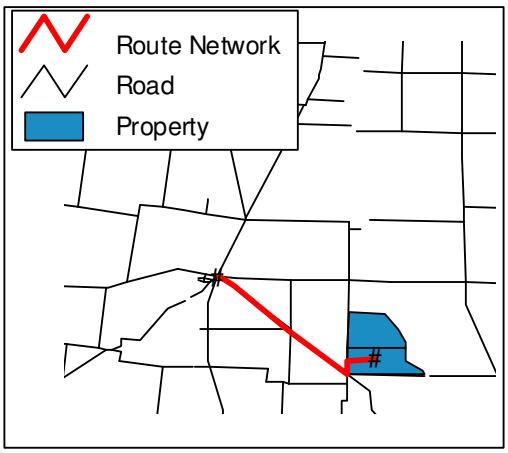

Figure 3.4 Route Network Analysis

Further analytical distance functions can be performed on the various data sets. There are 'distance' (described above), 'straight line', 'allocation', 'cost weighted' and 'shortest path'. The 'straight line' function measures the straight line distance from a feature. In the example of a town, it creates a raster data theme of the towns in the data set and provides a measure of each cell in the raster to the town point (Figure 3.5). This enables an analysis to be undertaken whereby each property can be assigned a distance based on the range which they fall within.

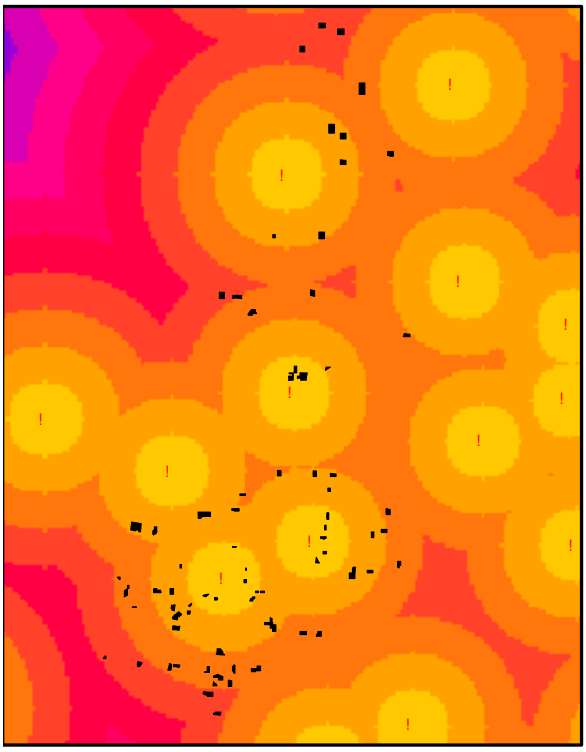

Figure 3.5 Straight line distance for towns (ESRI ArcGIS)

When using a linear based feature such as highways, a ‘straight line’ distance raster can also be generated. Figure 3.6 depicts the resultant straight line distance raster when computing a distance to highways. This allows each property to be assigned a distance from the highway within its attribute table.

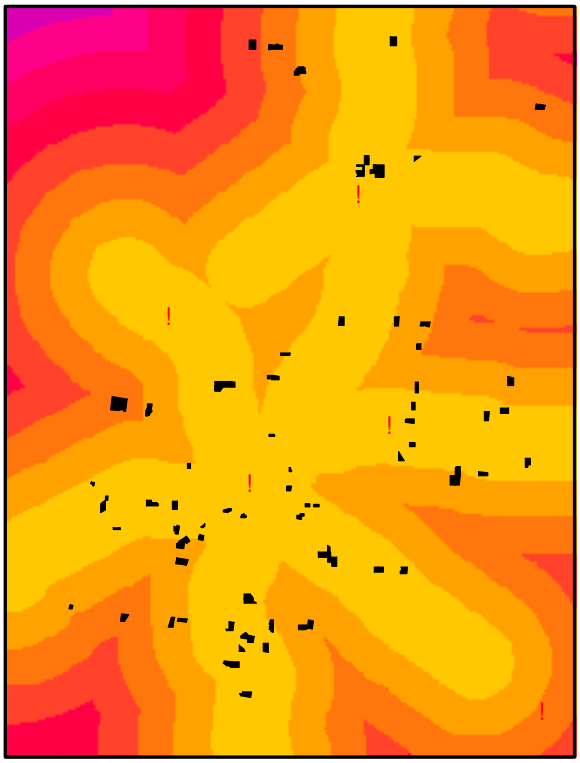

Figure 3.6 Straight line distance for Highways (ESRI ArcGIS)

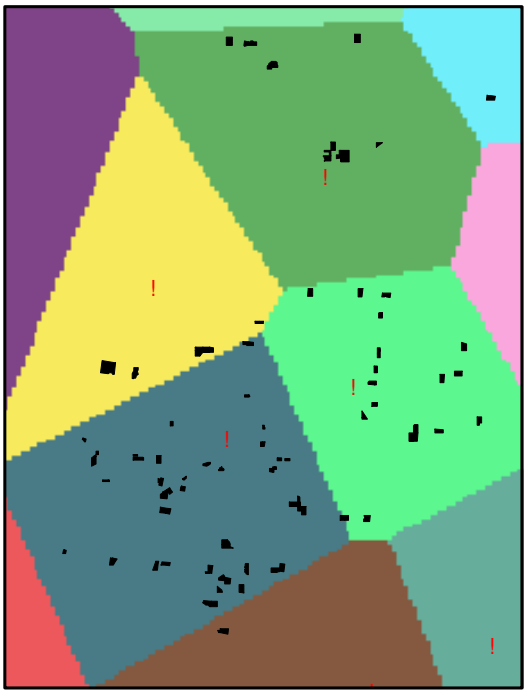

Figure 3.7 Allocation' function to towns (ESRI ArcGIS)

'Allocation' functions provide a means to determine which town is the closest to each property. These functions create an output raster which records for each cell in the raster the closet source cells based on either straight line distance or cost weighted distance. The function allows property to be segregated into allocation areas, due to their proximity to each town. This can be useful for finding out which town is the closest to each property or which properties are not served by a town. Figure 3.7 depicts the result of the allocation to towns based on straight line distance. The technique does not take into consideration the location and type of roads available from each property and in fact some properties may actually be closer to another town depending on the road network that is available.

Cost weighted distance relies on performing an allocation prior to using this function. The function creates an output raster which for each cell is the least accumulated travel cost from each raster cell to the source cell. This function is ideally used for movement or consumer behaviour regarding travel. It is reliant on setting up in the input data the preferences for travel, such as not travelling over steep slopes even though the route may be the shortest path. The distance model generated may then take longer to travel over a mountain range than through a longer route which by-passes this mountain range. This procedure involves re-classifying the raster data set such that it is representative of the preferences for determining least cost routes over the study region.

A wide array of techniques can be performed. However, it is not my aim to detail all GIS software functions. Rather, the aim of this section is to provide a sample of the techniques which may help to improve data sets that have limited property characteristics and in essence show how using GIS may help to improve model estimates by deriving more distance related variables. Some functions provide a more theoretical approach to measurement and allocation in that they don't take into account the other variables which can influence distance or time travelled such as the road network, traffic in major towns and road conditions. They do however provide a way to calculate distances and time/cost allocations, a truer representation would involve using network analysis and route creation.

3.4 Sub-market identification for Automated Valuation

It has recently been acknowledged that regression models developed for valuation may be subject to aggregation bias if property are not segmented into distinct sub-markets (Watkins, 1999). Watkins (1999, p.159) defined a sub-market as "a dwelling that is a

close substitute for would be buyers". Thus, research has been undertaken to examine the effect that the use of sub-markets have on valuation models to determine if greater accuracy can be achieved.

Property can be segmented using property type, census tract, postcode area, LGA or numerous other structural characteristics. This type of segmentation is known as 'a-priori' based. A-priori segmentation is dependent on a human/valuer segmenting the property into groups based on what they feel are important contributions in terms of property pricing influences. Additional techniques have been used to segment housing data into homogenous regions. Cluster analysis "is a generic term for a wide range of numerical methods for examining multivariate data with a view to uncovering or discovering groups or clusters of homogenous observations" (Everitt *et al*., 2001, p.ix). Cluster analysis has been increasingly used in real estate and property analysis however has also been used in many other disciplines including zoology, biology and medicine (Everitt *et al*., 2001). Although its use is not widespread for valuation, it is gaining popularity and slowly being used as an alternative technique to a-priori based techniques for classification (Day, n.d. ; Smith & Kroll, 1989; Bourassa *et al*., 1997; O'Roarty, 1997; Watkins, 1998).

With emphasis on the need to segment housing data into sub-markets to enable more accurate regression models to be developed, cluster analysis may prove to be a beneficial technique due to its non-reliance on a human to determine sub-markets. The following sections detail the methods used to segment property and the use of cluster analysis as a segmentation tool within the valuation industry.

3.4.1 Segmentation of Property into Sub-Markets

3.4.1.1 Sub-Market Definition

Bourassa *et al.* (1997) defined a sub-market as a "set of dwellings that are reasonably close substitutes for each other". Typically a sub-market has been defined by either spatial or structural terms, or in some cases a combination of both (Bourassa & Hoesli, 1999) with Dunse *et al.* (2001) indicating that socio-economic factors also contribute to the definition. The definition still remains unclear as to whether a sub-market should be defined spatially, by property characteristics or by house price (McCluskey & Deddis, n.d.).

There appears to be little consistency in the techniques or methods used to define sub-markets. Some studies derive sub-markets using a-priori techniques such as using a geographical administrative boundary, others use structural characteristics of each property to segment the data by house type such as units, houses or apartments. Adair *et al.*, (1996) used a nested solution to segment the housing market by first segmenting using three geographical areas and then uses structural characteristics to further segment the market. Sub-markets defined using a-priori techniques are not defined in an optimal way (Bourassa *et al.*, 1997; Watkins, 1998) and can be biased if no clear method or reasoning is provided as to the choice of characteristics on which to base the segmentation of properties.

Some authors suggest that sub-markets should not be defined a-priori; rather it is argued that the data should be used to determine the sub-market and that sub-markets should not be constrained (Day, n.d.; Bourassa & Hoesli, 1999; Dunse *et al.*, 2001). Statistical or numerical techniques used to define sub-markets vary. Principal Components Analysis (PCA) has been used for variable reduction and help to exclude

the variables deemed to be collinear. Cluster analysis has also been used to create natural groupings of property into distinct markets (O'Roarty, 1997). Many studies have incorporated both a-priori and numerical techniques to determine sub-markets to ascertain which technique is more appropriate and yields more accurate models (Bourassa *et al*., 1997; Dunse *et al*., 2001).

3.4.1.2 The Need for Sub-Markets

The need for sub-markets has arisen due to the concern that hedonic regression analyses will be subject to bias if the market is not segmented (Watkins, 1999). Watkins (1999) stated that to enable better performance of the property market, segmentation is necessary prior to the determination of regression models. If regression models are determined without segmentation then the resultant models and the implicit prices which are defined, will not reflect the differences in pricing attributes that are in force across different markets (Watkins, 1999).

Similarly, Baum *et al*., (n.d.) and Day,(n.d.) noted that rural areas cannot be classed as homogenous and as such specific characteristics that are influential to property estimation will vary within and also across counties (Wilhelmsson, 2004). This is evident by the research of Elad *et al.* (1994) and Xu *et al.* (1993) in that not only do different quantities exist for each pricing influence, but some characteristics were found to be highly significant in some sub-markets whilst not in others. Thus, the consensus is that data needs to be drawn from distinct markets to compensate for varying levels of property characteristics and the varying significance of these characteristics within a study area.

A-priori type classifications are generally used to create sub-markets for valuation. Xu *et al.* (1993) identified six sub-markets which appear to be devised through the amalgamation of multiple counties within Washington State. Elad *et al.* (1994) developed geographic sub-markets based on an amalgamation of counties. Defining a sub-market through geographical administrative boundaries is unlikely to ensure that the sub-market is homogenous (Wilhelmsson, 2004) especially as its boundary was generated for another purpose, thus different markets will tend to be operational within and even outside of these boundaries. It is also unlikely that sub-markets will nest spatially within the geographical administrative boundaries, although this is likely to be dependent on the study area under investigation.

There is an increased requirement to use a more intuitive means to develop homogenous sub-markets that may enable more accurate regressions to be determined. Current research is now investigating the best techniques for segregating property into sub-markets to reflect the underlying characteristics of the market region.

3.4.1.3 The Creation of Sub-Markets

As discussed in the previous sections, sub-markets have been defined using a variety of methods. A-priori techniques have been developed using specific housing characteristics like Bourassa & Hoesli, (1999) who used property type, property value and a geographical area to construct sub-markets. As discussed in Wilhelmsson (2004), other studies used census tracts and postal regions. More recently, cluster analysis has been incorporated into valuation research (Day, n.d.; O'Roarty, 1997).

Some issues that arise with the segmentation of property into distinct sub-markets include the problem of the data set size, and the reduced variation amongst the

variables or property characteristics. Typically, any break down of the existing data set in segments will result in the data being split into smaller regions containing fewer numbers of property. In addition, due to the nature of segmentation, individual regression models will need to be developed for each sub-market. As Wilhelmsson (2004) noted, different hedonic price functions will be operational for each sub-market. As such, property characteristics will vary across these sub-markets as shown by Xu *et al.*, (1993). As discussed by McCluskey & Deddis (n.d.) although the sub-markets defined may more accurately reflect the pricing influences, a smaller sample size in each region is likely to introduce error, which may then compound the initial problems of model accuracy. The use of dummy or indicator variables to depict each sub-market may also contribute to issues in model bias and model accuracy. Running one regression model but depicting each sub-market using a dummy or indicator variable makes the assumption that the influential characteristics are uniform across the whole study region and that the characteristics do not alter for each region which is contrary to most of the research (McCluskey & Deddis, n.d.; Xu *et al.*, 1993; Wilhelmsson, 2004).

Another problem which may arise in sub-market determination is the degree of variation of the variables. Watkins (1999) stated that the use of factor analysis reduces the usefulness of the resultant regression models as the procedure tends to cluster the property into groups that have similar property characteristics. Within each cluster there will tend to be less variation in the characteristics seeing they were clustered based on set property characteristics which are likely to be homogenous. This was not deemed a problem by Watkins (1999) as the research was only undertaken to determine if sub-markets exist and not whether the models developed would lead to more accurate results.

When using Multiple Regression Analysis (MRA) for property valuation, location has been found to be an integral component of the price of a property (McCluskey & Deddis, n.d.) and can be incorporated into a model in a number of ways. Wilhelmsson (2004) constrained sub-markets by defining the market such that only property which are adjacent to each other are allowed in the one sub-market. This method could be used as an exploratory process however it was found that this technique did improve the prediction of the models over those sub-markets and models derived using pre-determined administrative boundaries. The disadvantage of constraining the sub-market based on adjacency is that a large database of properties is required which abut.

As highlighted by Day (n.d.), it is likely that sub-markets segregate spatially as well as structurally. Some researchers tend to use pre-determined geographical constraints as sub-markets (census tracts, postal zones, LGA boundaries) and then create a nested approach which uses structural characteristics for sub-market determination (Dunse *et al.*, 2001). This approach coincides with that of McCluskey & Deddis (n.d.) who recognised that location is an important function of price. Adair *et al.* (1996) used a nested approach to segment property spatially into three geographic areas, and then further segment using structural components. Again, this nested segmentation relies on having a large quantity of data and is constrained in that it places emphasis on either structural or location based characteristics, or a combination of the two.

As noted by Dale-Johnson (1982), the determination of sub-markets should be more rigorous given the ad hoc nature of previous techniques (cited by Wilhelmsson, 2004). The generation of sub-markets using cluster analysis can eliminate researcher bias and provides a basis for determination of markets which are not constrained.

3.4.2 Principles of Cluster Analysis

3.4.2.1 Introduction to Cluster Analysis

A cluster is a homogenous grouping of objects in which there is some form of internal cohesion and external isolation between the members of the group (Everitt *et al.*, 2001).

Cluster analysis has typically been used when there is a need to classify or re-group data using numerical of statistical techniques (Everitt *et al.*, 2001). It involves the placement of objects into groups that are not apparent visually (Aldenderfer & Blashfield, 1987), therefore, uncovering some form of structure in data (Everitt *et al.*, 2001).

Cluster analysis has evolved from different disciplines and, thus, may have inbuilt bias depending on the application use. Cluster analysis is a simplistic procedure that relies on heuristics; as such it is not surrounded by a large body of statistical knowledge (Aldenderfer & Blashfield, 1987). The procedure should be used with some caution as data may be segmented into different clusters depending on the different clustering algorithm that is used (Aldenderfer & Blashfield, 1987).

Cluster analysis has been used in demographic and marketing analysis, crop classifications, archaeology and botany to name a few application areas. For demographic and marketing applications, it is often used to find a niche market or demographic profile to enable specific products to be marketed. For crop classifications, the technique has been used to identify different crop types from

satellite imagery, whilst for archaeology it has been used to create taxonomies of tools, or taxonomies of plants in the field of botany (Everitt *et al.*, 2001).

For property valuation, cluster analysis has had minimal use. Primarily the technique has only more recently been used in the valuation industry due to the emergence of the significance of segmenting the property market to enhance property valuation estimates. The technique aims to detect similarities in the preferences of buyers and enable property that are considered to be close substitutes for each other to be amalgamated within the one cluster. Goetzmann *et al.* (1998) applied cluster analysis after constructing housing price indices of metropolitan postal code areas in California. Bourassa *et al.* (1997) used cluster analysis on metropolitan data from Sydney and Melbourne, whilst Smith & Kroll (1989) studied rental markets in suburban Texas. O'Roarty (1997) used clustering for retail store space requirements within the UK and Dunse *et al.* (2001) applied the technique for commercial property in Scotland. Wilhelmsson (2004) applied clustering to residential housing in Stockholm whilst Day (n.d.) studied the effects of sub-markets in Scotland using a combination of property types in Glasgow.

3.4.2.2 Clustering Algorithms

A variety of clustering algorithms have been developed which are based on hierarchical techniques (agglomerative and divisive clustering), iterative partitioning and factor analysis. Density search, clumping and graph theoretic algorithms are not as popular and thus a broader overview of these techniques can be found in Everitt *et al.*, (2001) and Aldenderfer & Blashfield (1987).

When clustering using hierarchical techniques, 'the data are not partitioned into a particular number of classes or clusters in a single step' rather, the algorithm attempts to 'fuse the data into sub-divisions to find the optimal number of clusters' (Everitt *et al.*, 2001). Agglomerative techniques arrive at a single cluster in the final stage and thus the analyst is required to observe the results to determine the number of clusters that should be used to arrive at an optimal solution. Divisive techniques work in the opposite way to agglomerative and thus start with one large cluster and segregate the data into multiple clusters. Issues with both of these hierarchical techniques are that as each stage sub-divides the data into a new cluster, the data from the previous stage cannot be re-created (Everitt *et al.*, 2001).

Iterative partitioning algorithms work on raw data and not on a matrix of similarities like that of the hierarchical techniques. Thus, iterative methods allow for more than one pass through the data and enable a previous poor partition of the data to be compensated where this not possible with hierarchical techniques. The major drawback of the iterative algorithm is that to obtain an optimal cluster or partition, the technique aims to observe all possible partitions on a data set which is not possible computationally (Aldenderfer & Blashfield, 1987). Thus, researchers have then applied heuristical rules to the technique to determine the most optimal data partition (Aldenderfer & Blashfield, 1987) without the need to examine all possible data permutations.

Factor analysis aims to reduce the number of variables to a smaller number of components or factors (Everitt *et al.*, 2001). The new components of variables are then used in other analyses such as cluster analysis or regression analyses. The technique

allows for interrelated variables to be identified and thus minimises the use of correlated variables.

The different clustering algorithms in use are varied and quite extensive with the algorithm used being dependent on the application and data being used (Aldenderfer & Blashfield, 1987).

3.4.2.3 Variable Selection and Proximity Measures for Cluster Analysis

The selection of variables for inclusion in cluster analysis has been undertaken using Principal Components Analysis (PCA) by Bourassa *et al.*, (1997) and Dunse *et al.*, (2001). Eight variables were obtained by Dunse *et al.*, (2001) and PCA was undertaken to produce "a limited set of uncorrelated factors which, together, retain most of the variance and information contained in the original variable and assigns a factor score to the property" (Dunse *et al.*, 2001, p.241). It is argued that PCA is able to extract out the underlying components that characterise sub-markets (Watkins, 1999). The factors are then used to determine sub-markets using cluster analysis.

Other techniques used to identify variables for use in clustering have been a-priori determined. For optimization of rent using rental markets, Smith & Kroll, (1989) used age and income to determine clusters based on demographic type, whilst unit size and monthly rent was used to cluster units. O'Roarty, (1997) used survey techniques to determine the prime influences of retail property. From this analysis, four main variables were selected and later used for cluster analysis.

Typically, variables within a data set requiring clustering can contain either continuous variables, categorical variables or a combination of both, although the latter have issues concerning the standardisation of data and the techniques best used for multi-mode data. Cluster analysis involves finding similarity or dissimilarity between observations in a data set to enable them to be grouped together (Everitt *et al.*, 2001) in a homogenous group, or not grouped together if they are dissimilar. The methods used to measure similarity are frequently concerned with distance and measuring how far apart observations are. As a consequence of this, variables need to be of a similar scale in order for them to be compared together and for similarity or distance measures to be calculated.

For data that are categorical and binary, the data are generally scaled to be either 0 or 1 (Everitt *et al.*, 2001). Where data have multiple categories such as land use type the data can be represented with a 0 or 1 for each level in the category. An example is for a land use of dairy, the value may be 1 or 0; for a land use of beef, the value can be 1 or 0. The problem with creating too many indicator or dummy variables as in the case of the land use variable is that there can become too many variables which hold the value of 0 affecting the construction of similarity between data observations (Everitt *et al.*, 2001).

For data that are continuous, a number of distance measures can be used to determine similarity between observations. As all data will be of the same nature in terms of distance, the scale of each variable is required to be the same. An example is distance to a railway and distance to a town. Both outcomes should be measured in the same unit to enable comparison (ie: all measured in metres, or all measured in kilometres).

A more detailed analysis and explanation of the variety of dissimilarity measures for this type of data can be found in Everitt *et al.* (2001).

When data are represented as a combination of both continuous and categorical variables, then different techniques are required to standardise the data to enable clustering analyses to be undertaken. To deal with this type of data a number of alternatives can be taken. One technique is to split the data and use appropriate proximity measures for each type of data and reassemble the data by then applying weights (Everitt *et al.*, 2001). Another method is to apply a proximity measure where each continuous variable is scaled such that it is based on its range within that variable. The categorical variable will still need to be kept the same. The approach taken in this research was to use a 'two step clustering' algorithm, (available in SPSS) as it enables the categorical variables and the standardised continuous variables to be specified and processed.

3.4.3 Cluster Analysis for Automated Valuation

Cluster analysis has been used for classification of residential sub-markets (Day, n.d.; Bourassa *et al.*, 1997; Goetzmann *et al.*, 1998; Watkins, 1998; Wilhelmsson, 2004) and for office sub-markets (Dunse *et al.*, 2001), storage space (O'Roarty, 1997) and rental markets (Smith & Kroll, 1989).

Most applications of cluster analysis for valuation utilise Principal Components Analysis (PCA) to determine factors which are then specified in the clustering algorithm. Traditional clustering methods incorporate specific variables into the clustering algorithm and determine clusters based on the variables rather than factors or factor scores. Watkins (1999) used factor analysis to determine three factor scores which

were derived from the 14 property variables. Cluster analysis is then used to develop sub-markets based on these factor scores and then regression analysis is applied to the clusters to determine if there are any significant price differences between the procedures (Watkins, 1999). Likewise, Bourassa *et al.*, (1997) and Bourassa & Hoesli (1999) used PCA to extract factors from the variables. The factors were then standardised and weighted and cluster analysis was then used on the weighted factors. The use of factor scores is quite widespread as it aims to reduce any collinearity between variables that may be apparent during the clustering process.

Day (n.d.) used a hybrid approach to clustering of properties in Glasgow, Scotland. The technique involves using a partitioning method, known as k-means clustering and is then followed by a hierarchical technique. The k-means clustering algorithm is used when the number of clusters is known and it was used by Day (n.d.) to develop 100 clusters in the initial stage of clustering. A hierarchical technique was then used on these 100 clusters. The procedure involved averaging the values of the characteristics of each cluster and using hierarchical clustering when eight clusters were identified. This approach reduces the computational requirements imposed by hierarchical techniques through the use of the k-means clustering technique which constrains the number of clusters during development. The hybrid approach then allows the analyst to have more control on the clustering groups defined by the hierarchical technique, thus providing a more intuitive approach.

Wilhelmsson (2004) used a slightly different approach. A regression analysis is performed to determine the characteristics to use in the cluster analysis by using the residuals of the regression. The residuals of the regression are divided into two sub-samples; those residuals which are positive form one cluster, those which have

negative residuals form another. Cluster analysis is performed for each of the two pre-determined clusters or groups based on the positive and negative residuals. This is undertaken using a hierarchical technique with the Ward approximation method (Wilhelmsson, 2004). Finally, regression analysis is undertaken using the determined clusters as indicator variables. The problem with this approach is that although sub-markets may exist and be found in the data, it assumes that all variables or property characteristics which influence price will be the same (McCluskey & Deddis, n.d.; Xu *et al.*, 1993; Wilhelmsson, 2004). An alternative approach is to develop regression models for each sub-market that is operational within the study regions.

3.5 Summary

GIS is an ideal medium with which to store property data due to location being a key factor affecting property (McCluskey & Deddis, n.d.). Once a variety of data is obtained, GIS can be used to derive additional property characteristics through a number of various techniques. Euclidean distance measures from each property to a local town, major city, school, highway etc can be performed, providing ancillary data such as schools, highways and the like are available. Spatial overlays or intersections of layers (data sets) can enable waterbody areas to be determined for each property in the data set or even stream length can be calculated once streams have been clipped to each polygon and aggregated.

For comparable sale analysis, a GIS can be used to query a large database and perform a search which selects comparable sales that are a predetermined distance from the subject property. Through visualisation, GIS can be used to enhance results, detect errors in prior results or find anomalies within data values.

With raster GIS analysis based on distance mapping; a straight line distance, allocation, cost weighting and shortest path analysis can all be undertaken but require varying degrees of input in the specification of cost weightings for this type of analysis. Although these functions allow for more detailed analysis to be undertaken with respect to determination of shortest path and cost weighted analysis, they are an additional technique which may enhance a property database and result in a more accurate regression model. The use of GIS for rural valuation is somewhat limited and is affected by the availability of key data sets to enable this form of analysis to be undertaken.

Once a GIS database and additional variables have been derived, regression analysis can be undertaken to model property price. Typically, regression analyses have been performed using some form of geographical boundary to segment the properties into sub-markets. This technique assumes firstly that geographical administrative boundaries, designed for other purposes and the more common boundary used to spatially locate property, are in fact representative of property sub-markets. Secondly, in many cases, regression models are run which depict the various geographical sub-markets through the use of indicator variables rather than performing a regression analysis on each individual sub-market. This assumes that all property influences are the same in each sub-market, when this has often been found to vary (Xu *et al*., 1993).

The incorporation of cluster analysis techniques into property valuation for the classification of property into various sub-markets has recently increased (Bourassa & Hoesli, 1999; Watkins, 1999; Wilhelmsson, 2004). This technique is used to develop sub-markets using a more data-driven approach, rather than through classifying data based on specific structural characteristics or geographical locations. Typically PCA

has been used to derive factors which can then be used in a clustering algorithm, however this can lead to grouping of property into classes where the price influences have little variation and thus reduces the effect or usefulness of regression modelling (Watkins, 1999).

The technique proposed in this research will develop regression models for rural property based on user specified geographical (LGA) sub-markets. Cluster analysis will then be used to derive data driven sub-markets with further regression models being developed which are specific to each cluster region. Finally, a comparison will be made regarding the accuracy of the regression models based on the two methods used to derive sub-markets. Chapter 6 reports on the clustering process undertaken during the research.

Chapter 4 Database Development

4.1 Introduction

This chapter outlines the study areas selected for the property valuation modelling. I also outline the software and hardware used, the data sets obtained from various agencies, the development of the property database and the data integration framework that was developed.

Digital data are crucial to any automated system. Integration of spatial data obtained from various sources can be particularly problematic. The development of the property database is a key stage of this study. In this Chapter I discuss the methods and processes undertaken to convert the various data into one database. In this stage, I use GIS to derive additional variables for the database. Due to the various tasks involved within the database development stage and the issues that arose during this research, I developed a framework to support the future integration of spatial and tabular non-GIS data.

4.2 Study Area Description

Properties in the Wimmera and West Gippsland CMAs were chosen for modelling. These CMAs are located in the west and south-east of Victoria, respectively.

4.2.1 Wimmera

The Wimmera CMA is approximately 33,000 square kilometres and accounts for 13% of the total area of Victoria. The population is approximately 43,000 (Wimmera

Catchment Management Authority, 2000) and although declining in the rural parts of the catchment, there has been a slight population increase in the urban regional centres, particularly Horsham. The Wimmera CMA comprises the LGAs of Horsham, Yarriambiack and Northern Grampians and properties from these LGA's were chosen for study. Land degradation issues within the catchment include water and wind erosion and destruction by pest plants and animals. Salinity is a major threat to agricultural production with 170 square kilometres affected in the region (Wimmera Catchment Management Authority, 2000).

4.2.2 West Gippsland

The West Gippsland CMA is approximately 17,000 square kilometres and accounts for 7% of Victoria's total area. The population is approximately 200,000. A portion of this population increase can be attributed to an increasing movement of people into the region from Melbourne (West Gippsland Catchment Management Authority, 2001). The West Gippsland CMA comprises the LGAs of Latrobe, Wellington, Baw Baw, South Gippsland, East Gippsland, Delatite and Bass Coast. Water erosion, land degradation, salinity and pest plants and animals have the greatest affect in the region with salinity around the Lake Wellington area leading to reduced agricultural production (West Gippsland Catchment Management Authority, 2001). The Wellington LGA was selected as a study region within this CMA as it encompasses the Lake Wellington regional area which is experiencing salinity issues.

4.3 Software

A variety of GIS software packages is currently available. Some specialise in utilities, engineering or mining applications and others are aimed at either novice or advanced

users. ESRI ArcInfo, ESRI ArcView, ESRI ArcGIS, MapInfo Professional, Infomaster and GE Smallworld are some of the more popular GIS software packages available.

ESRIs ArcInfo 8.2 was used throughout this study as the primary software for integration and development of the database due to its superior processing and data handling capabilities over other Microsoft Windows based GIS solutions. ESRI ArcInfo was used in this study from a PC, networked to a Unix system. MapInfo Professional 5.0 and 6.0 were utilised only to export data supplied in MapInfo Professional format into either ESRI ArcView shapefiles or ESRI ArcInfo coverages depending on the supplied data formats. Likewise, ESRI ArcView 3.2 was used only to export those data supplied as ESRI ArcView shapefiles into ESRI ArcInfo coverages or where conversions between MapInfo Professional and ESRI ArcInfo were not successful and required an intermediary process via ESRI ArcView. The maps created for display within the research and for later processing were generated using ESRI ArcGIS 8.1. Microsoft Excel 2000 was used to export tabular data into ArcInfo tables and to assess the valuation models.

Minitab (release 13) was utilised for the statistical testing and regression analysis stage of the research due to its ability to perform the various testing and regression techniques and also due to the availability of the software. Minitab was used through a Microsoft Windows platform. SPSS13 was used in the latter part of the research to develop clustering regions.

4.4 Data Integration Framework

Data integration has long been an issue in the sharing of geographic information amongst different organisations, mainly due to diverse data sets, poor documentation and incompatible data quality. The creation and integration of GIS databases can be a timely and costly exercise (Montgomery & Schuch, 1993). Although data integration can often be quite a lengthy process, it is becoming more essential due to the costs of data acquisition (Devogele *et al.*, 1998). A framework to address data incompatibilities, incomplete data sets, accuracy, scale and format differences that may be an impediment to the effective use of the data is important to help minimise the costs when utilising multiple data sets from various agencies. It has been widely accepted that data sharing adds value to data (Shepherd, 1991; John, 1993; Ralphs & Wyatt, 1998) providing that data quality issues have been addressed and that data quality statements accompany the integrated data so that the reliability of the data can be assessed (Lunetta *et al.*, 1991; Slagle, 1994).

Data integration can be broken down into a number of key tasks. Figure 4.1 indicates the main components of integration as formulated in this research. These include database design where both logical and physical design are considered. A key component of this stage is the need to plan the capacity in which the database will be used. This includes evaluation of the accuracy, completeness and format of all available data to determine the most appropriate data to acquire. In addition, it incorporates the formulation of documentation detailing the processes required to create the database and the attributes to be included within the database. Data set conversion and database creation incorporates conversion of tabular GIS data, tabular non-GIS data, spatial GIS data and conversion between various GIS software formats. This process may also involve geocoding of tabular non-GIS data and reclassification of coding systems used within the data set. Finally, projection transformation is

required to convert existing spatial data sets into the one standard projection and datum so that data sets can be overlayed.

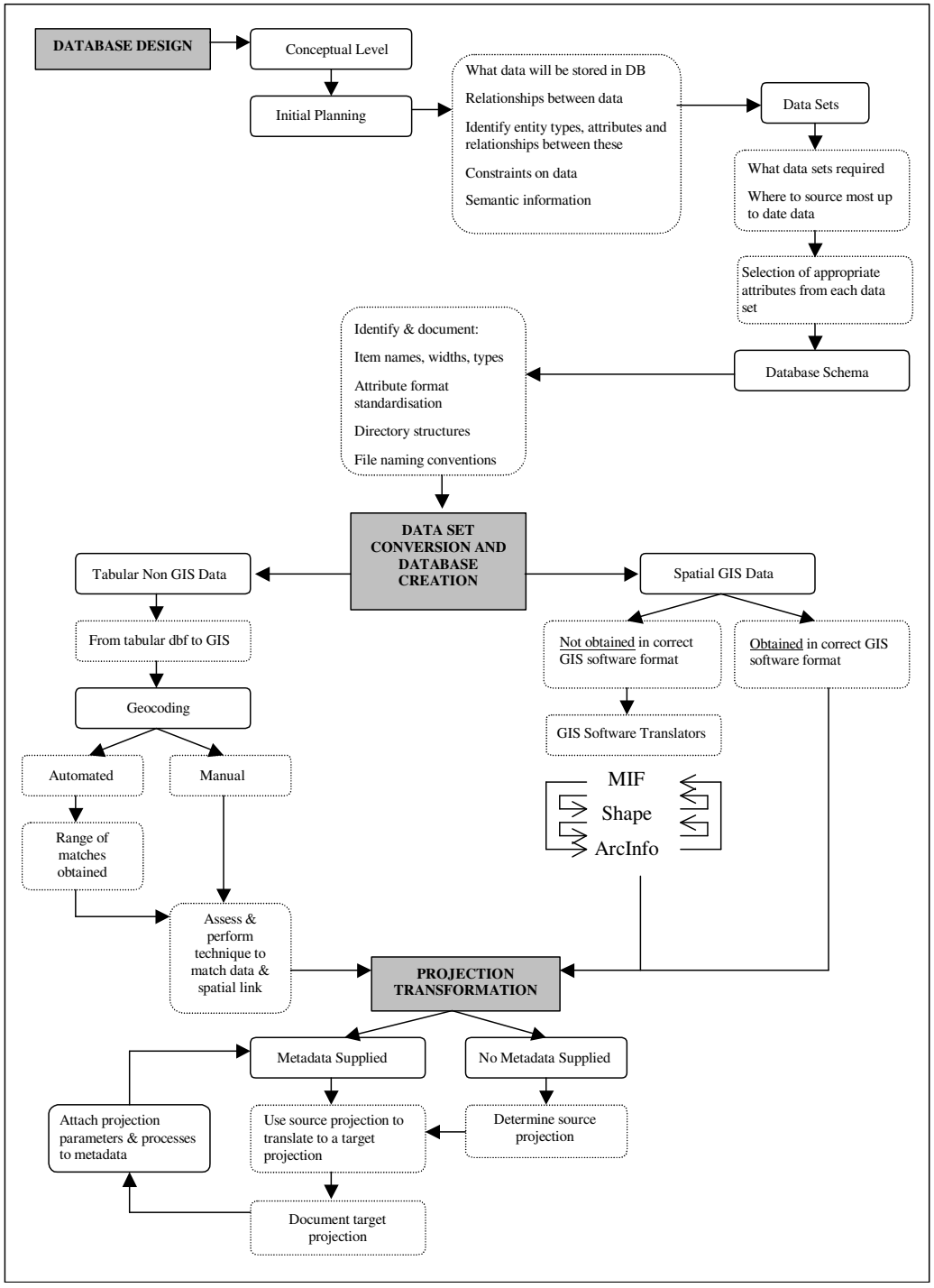

Figure 4.1 Data Integration Framework (after Hayles & Grenfell 2002)

4.4.1 Database Design

The formulation of the database in terms of what data sets are required, where the most up to date data can be sourced, and planning the structure of the database is necessary prior to commencement of any form of conversion between data sets (Cannistra, 1999). Decisions regarding the capacity in which the data are to be used, the contents of the final database and the degree of accuracy required will assist in the selection of appropriate data sets to meet these requirements. The choice of relevant data sets should be determined by the accuracy and currency of the data necessary for the tasks at hand.

The selection of which attributes are needed from each data set is another issue that can help to minimise processing during the integration process. Whilst there may be no immediate need to include all attributes from all data sets, it is necessary to determine the present and future needs of the user(s) (Foster & Hamilton, 1991) and use this knowledge in the overall database design. Documenting the attribute names, widths and types, the location of data themes within a directory structure and the conventions for file naming are also necessary, especially to indicate the processes undertaken during integration. Conforming to current street addressing standards will ensure that the database can be used by different users since different format types and addressing standards are eliminated (Croswell, 2000). The creation of a new table that excludes the elements used in the final database, but retains those not used from the source data set, may help these attributes to be more readily integrated at a later stage providing there is an identifier to link the two data sets and to minimise duplication.

4.4.2 Data Set Conversion and Database Creation

The three types of data referred to in this thesis are tabular non-GIS data, tabular GIS data and spatial data. Tabular non-GIS data refers to a data set, often in a spreadsheet format that has not been used within a GIS. Often the data may contain attributes with (implicit) spatial references. These may require geocoding to link to explicit references in related data sets. Tabular GIS data refers to data used within a GIS and is in a tabular format. Spatial data refers to graphical data with spatial reference attributes which are explicit and have been used within a GIS.

4.4.2.1 Spatial GIS Data

Data that has been derived and stored in various GIS needs to undergo translation such that the data can be used in a single GIS software package. Although there are inbuilt translators in the major GIS packages, there can be difficulties between the parameters and export functions that they use. Where possible the same method for converting data layers from various GIS software should be utilised to minimise any discrepancies between conversion translators. Once the data are in a single GIS software package, integration may still be an issue due to the various standardisation methods and classification schemes used for attributes (Croswell, 2000). Most organisations provide different attribute representations and classifications that may result in the need to reclassify the data into a more common standard for a particular use. Whilst GIS products such as ESRI ArcInfo now have a selection of common data models available providing entity names and types for a particular application, there is still great diversity in the various applications of GIS.

4.4.2.2 Tabular Non-GIS Data

Tabular data that have not previously been used in a GIS can be quite difficult to integrate. Although it may take minimal processing to transfer a spreadsheet into a GIS table, such conversions may be hindered by data inadequacies. Whilst automated geocoding may prove successful in urban regions, this has not been the case in rural regions where many properties have lacked a proper numbered street address. The rural addressing project in Victoria, undertaken by Land Victoria, aims to standardise rural addresses by providing a numbered address to improve property identification using distances along a road to locate the property entrance. Although this is complete in some LGAs, data sets compiled prior to this project will still have limited address details and thus can hinder integration through geocoding due to the lack of this information. Until these address attributes for properties are updated to take into consideration the new rural road numbering scheme, there will need to be some form of manual geocoding to link data sets to a property database. This highlights the need when creating data sets to use current standards and up to date information and in some instances it may also be necessary to have a dual representation of the addresses.

Land use coding may differ between organisations due to their different uses of the data. Sometimes there may be overlap of the coding such that a new coding system needs to be developed to support merges. Consideration should be made as to whether a new system is likely to cause more problems or whether the existing codes from one data set can be translated into an existing classification scheme.

4.4.3 Projection Transformations

Transformations between projections are usually performed within the GIS software package with each having various degrees of customisation. Depending on the software used, and the projections required, the transformation should only require the source and the target projections to be specified. Where the projection is unknown or not a standard projection included in the GIS transformation software package, then the parameters will need to be input manually. For a Universal Transverse Mercator projection, input requirements include the coordinate system, units, spheroid, scale factor, longitude of the central meridian, latitude of the origin and the false eastings and northings. Currently, ANZLIC metadata standards do not provide detailed projection information and thus these projection parameters may be incorrectly obtained and specified by a novice. A proposed solution around these problems would be to append a projection file to the data set or detail all projection parameters within the metadata to minimise incorrect parameter input.

4.5 GIS Data Sets

A combination of individual property sale information, valuation, cadastral, topographic, planning, salinity, pest and fire prone data sets were obtained for use in the study. These data sets were selected based on the conceptual model developed within Section 2.4. It should be noted that not all variables conceptualised to be important to rural value were available digitally. Due to privacy regulations, only some of the State Land Tax Valuation file data was available. In addition, although the PRISM data were available for the purposes of this research it was not in a readily usable format for GIS analysis. Thus, the model was only partially implemented due to the lack of availability in Victoria of property based digital data containing sale price information. Once the data sets were obtained and integrated, these variables were used to assist in obtaining price drivers for the determination of a rural property based on the highest

and best use value. Table 4.1 provides a summary of the data sets obtained and the age, currency, data format and the map projection in which the data were supplied.

Data set name	Data Format	Coordinate System	Age	Currency	Datum
Dryland salinity discharge area (DISCH25_AREA)	ESRI ArcInfo	AMG	Mid 1960's	Current	AGD66
Fire Cover (LASTBURNT100)	ESRI ArcInfo	AMG	Jan 1990	Current	AGD66
LGA Boundary	ESRI ArcInfo	Geographical	Unknown	Unknown	None
Pest Management Infestation Sites (PMIS100)	ESRI ArcInfo	AMG	Unknown	Current	AGD66
Planning Scheme Database	MapInfo Professional	Geographical	1974	Current	AGD66
Property Information and Sale Data (PRISM)	Microsoft Excel	None	1974	Current	None
Psyllid (PYSLLID25)	ESRI ArcInfo	AMG	March 1994	Current	AGD66
Salinity Regions Data Set (SALREG500)	ESRI ArcInfo	AMG	Unknown	Unknown	AGD66
Soils Data	Microsoft Excel	None	Unknown	Unknown	None
State Land Tax Valuation File (SLTVF)	Microsoft Excel	None	Unknown	Unknown	None
Vicmap Digital Property	ESRI ArcView	VICGRID	Unknown	Current	GDA94
Vicmap Digital Topographic	ESRI ArcView	VICGRID	1974	Unknown	GDA94

Table 4.1 Summary of Data Sets obtained

4.5.1 Sale and Valuation Data

The State Land Tax Valuation file (SLTVF) and Property Information and Sale data (PRISM) were obtained to provide valuations and sale prices for the rural properties within the two study areas.

The SLTVF details site valuations for individual properties and was able to be supplied by Wellington and Yarriambiack LGAs in Microsoft Excel format. Although requests were made requesting supply of SLTVF data in the LGAs of Horsham and Northern Grampians, due to data privacy issues by these LGAs, permission was denied. The sale data and the site valuation data are tabular non-GIS data with several of these attributes providing an implicit spatial reference for the data and by geocoding can be used to link to explicit references in the cadastral data set.

The PRISM sale data were derived from the *Notice of Acquisition*, a statutory document completed after settlement by conveyancers within Victoria. Data was obtained for sales occurring between January 1995 and August 2000 based on LGA administrative boundaries used by government departments within Victoria. The PRISM data were obtained for the LGAs of Wellington, Yarriambiack, Horsham and Northern Grampians and were supplied in HTML (Hyper Text Markup Language) files. Attributes within the data set include street number, street name, suburb, postcode, sale amount, sale date, land use code, crown allotment number, area and Melway map reference. The Melway map reference relates to the Melway Street Directory produced in Victoria which encompasses the metropolitan area of Melbourne. This map reference is based on the page number and an alphanumeric reference to a grid within the directory. As the regions obtained for this study are within rural Victoria where

there is no directory coverage of Melways, then this field is redundant in all the regional data obtained. No values were recorded in these instances.

The PRISM data set originated in 1974 and relies on conveyancers to provide sale price information after recent sales. The data set is a tabular non- GIS data set with all information stored as tables that, to date, have no identifier to link it adequately to a geographic area within a GIS database. Although the data set contains address information, which is often suitable to use for geocoding, this is not the case for this data set. In some cases there is no sale price, there are instances of incomplete data, and insufficient address details to adequately identify the properties within the data set. The database, whilst containing sale price information that is updated monthly, does suffer in that some of the sale price values appear questionable. There are some values in the database such as $103,656 or $80,382 which seem odd in that you would not expect that a property price would be dealing with $1 increments in value. A value of $103,600 or $80,300 or even sale prices that are within a $500 range would appear more likely.

The SLVTF details valuations within each LGA. The data were not able to be geocoded at all to the cadastre due to the lack of property identifiers supplied with the data sets. A numeric property number was supplied however as the cadastral data did not have the same unique property identifiers, the two data sets could not be linked.

4.5.2 Topographic and Cadastral Data

Vicmap Digital Topographic data were supplied for Horsham, Wellington, Northern Grampians and Yarriambiack LGAs. A combination of themes were supplied which included hydrographic, relief and vegetation data. Within each theme, a number of GIS

data layers were supplied such as windbreaks, cuttings, embankments, waterbodies, wetlands, launching ramps, wharves and dam walls. From this wide supply of data layers, only waterbody points, watercourses and the dam/well layers were used in this study. All data layers were supplied as ESRI ArcView format (shapefiles) in VICGRID94 coordinates.

Cadastral data were obtained to register site valuation and sale data to individual parcels. Vicmap Digital Property (the cadastral layer) was supplied with a number of GIS layers. An 'Annotation' layer, an 'SDRN' layer (State Digital Road Network) and the 'CadRoad' layer (the polygon boundary data set of properties) were supplied. The CadRoad layer includes land parcels, property identifiers, road centre-lines, easements and various other administrative boundaries. The SDRN layer provided a non-polygon (arc) data set of the road network throughout the four LGAs. This data set was used in the geocoding process to acquire additional properties in the data set. Due to conflicting street names between the CadRoad layer and the PRISM data set many properties were not able to be manually geocoded (Section 4.6.3). Vicmap Digital Property data sets were supplied in ESRI ArcView format for the LGAs of Horsham, Wellington, Northern Grampians and Yarriambiack in VICGRID94 coordinates.

Data were supplied for each LGA boundary within Victoria. The fields supplied with these data were the locality name, LGA name, latitude and longitude, and easting and northing coordinates. Data were supplied as ESRI ArcView shape files in geographical coordinates. These data were used to create a boundary for spatial overlays with data coverages that were not supplied on an LGA bounding basis. The data set was also used for display purposes to show the properties which fell within each LGA.

4.5.3 Planning Data

A variety of data layers were supplied from the Planning Scheme Database. These were supplied in MapInfo format for Horsham, Wellington, Northern Grampians and Yarriambiack LGAs in geographical coordinates. Each LGA had a number of files representing different zones and codes for different planning overlay information. Data layers supplied included airport environments, design and development, development plan, environment audit, erosion management, environmental significance, heritage, rural floodway, restructure, road closure, significant landscape, vegetation protection and zone overlays. Each LGA was comprised of between 7-11 overlays which represented the above mentioned data layers.

Variables from this Planning Scheme data included information regarding the land which is subject to inundation along with the planning code/ land zoning regions within each LGA. The 'land subject to inundation overlay' was represented by a polygon area and depicted those areas that are subject to inundation. Attributes included the area, perimeter, identifier number, zone number, creation date of the object and other identifier related information. This data set can only be used to populate a variable regarding whether a property is included within a 'subject to inundation' region or not. There was no additional attribute information which can be populated regarding inundation from this table of information. The 'land zoning' attribute only allows for introducing one variable into another data set. The field '*ZONE_CODE*' represents the zoning code regions over the LGA. Upon overlay with the property database of cadastral details, the code name can be populated into another data set. As the study was specifically researching rural properties, an indicator variable was employed in the modelling to specify if the property was within a 'rural zone'.

4.5.4 Land Management Data

The Pest Management Infestation Sites data set (PMIS100) details reports of pest infestation over an area (Victorian Department of Sustainability and Environment, 2006). Attributes within this data set are report number and tenure, the code and name of the pest species, report date, extent of the infested area and a location code for property location. The 'pest species name' attribute was represented by either fox, rabbit or dog and the 'pest species code' was a numeric attribute representing these pest types. The 'extent of infested area' was a hectare estimate of the area infested by the pest animal. Upon performing a spatial overlay of these data with the property database, it was found that there were only instances of fox infestations and there were only minimal instances of this occurring on the properties within this database.

The Psyllid affected areas data set (PSYLLID25) defines treed areas affected by Psyllid and fungi (Victorian Department of Sustainability and Environment, 2006). Attributes within this data set include a description of the infestation and an intensity rating, when and by whom the data was surveyed and a description of the region. The intensity of the infestation was rated as either low, medium, high or none. The 'intensity' of the affected areas would be the likely attribute to be populated from this data set to depict if any properties were affected from Psyllid. Upon performing an overlay on this data set with the property database, no instances occurred of Psyllid on any of the study properties.

The dryland salinity discharge area data set (DISCH25_AREA) indicates the location and extent of an area affected by dryland salinity (Victorian Department of Sustainability and Environment, 2006). The attribute '*severity*' was included into the property database and showed that there were only a few instances of properties

affected by some form of dryland salinity. This conflicts with the Wellington and Wimmera Catchment Management information that states that salinity is detrimentally affecting properties in the two catchment areas (West Gippsland Catchment Management Authority, 2001; Wimmera Catchment Management Authority, 2000). This conflict could be due to the location of the properties in the database developed in this research. It may be possible that the properties that were not geocoded may have coincided with the 3000 dryland salt affected sites mapped within Victoria.

The salinity regions data set (SALREG500) contains polygons representing regions affected by salinity (Victorian Department of Sustainability and Environment, 2006). Attributes include the identification number of the polygon area and name for each area. There was no further information within the data tables which provided information regarding the intensity or severity of any geographic area affected by these bounding salinity regions. The only relevant information which could be populated from this data set would be an indicator variable to record either the presence or absence of a saline affected area.

The fire cover data set (LASTBURNT100) details the history of recent bushfire or prescribed burning (Victorian Department of Sustainability and Environment, 2006). Attributes defined include the season, type of fire, dates of the start of the fire, the intensity, types of prescribed burning and location of fire within the fire district. The 'fire-type' attribute details wildfire, prescribed burning and unknown fire types. The 'fire intensity' attribute details the intensity of the fire through 10% gradations. This data set would enable a number of variables to be populated within a property database. The frequency of fires within an area and the effect that they may contribute to crop and livestock destruction are variables which could be derived from the 'intensity' and the

'fire dates' attribute. Data from this data set were redundant in this study as there were no instances of either bush fire or prescribed burning occurring within the four LGA study areas. This was an issue in the population of the property database in that data sets were obtained and converted to the appropriate software format and map projection. However, after performing a spatial overlay, it was found that there were no instances of the variable over any of the properties.

The soils data set did not have attributes recorded for all properties and there were also variations in the way in which the values were specified. As such there was not a complete coverage of consistent information for this data set, and thus the information was not included.

4.6 Development of the Property Database

The development of the property database involved steps to accurately convert the supplied data sets into the desired software format and into a database consisting of the digitally available conceptualised property attributes. This required a combination of re-formatting the way attributes were represented in Microsoft Excel, to abbreviation or elimination of specific fields due to their irrelevance to the project, projection transformations and geocoding. The tabular nature of the valuation and sale data sets supplied ensured that standards usually applied to geospatial data were not present in these instances. A common standard for the representation of variable and attribute names was necessary to ensure consistency in the property database. Once data was integrated into the one software format, further re-formatting of the variable definitions were necessary to enable additional processing using statistical software as detailed in Section 4.6.6.

4.6.1 Data Cleaning and Minimisation

The PRISM data set required cleaning prior to geocoding with the cadastral data set. Initially there were a total of 1710 property sale transactions for the four LGAs, however, after 'filtering', 428 remained. The initial data set contained property that was rural as well as urban, thus any property specified as urban was eliminated during the filtering process. Although *The Valuation of Land Act (1960)* specifies farm land as being not less than 2 hectares, 20 hectares was chosen for this study due to the land use categories in use in the PRISM data set. Hobby farms were classified as being less than 20 hectare in size. It was decided that this size limit be utilised throughout the other land use categories to eliminate any smaller farms that were not classed as hobby farms yet their size limited the agricultural productivity of the property. Land uses designated as vineyards and market gardens were also deemed as non-agricultural land uses for this study and removed from the data set due to their category of being a specialist property (Baxter & Cohen, 1997). Parcels with multiple or incomplete records or with no recorded sale prices were also eliminated from the data set resulting in a total decrease of 75% in the number of records over the four LGAs (Table 4.2).

Whilst the amount of filtering seems quite substantial in terms of the total number of properties supplied within the data set, the data set supplied was not representative of a homogenous rural valuation data set. The 428 properties that remained in the database represented rural properties within the four LGA's that were classed as agricultural and were also over 20 hectares in size.

LGA	Initial number of records	Final number of records	% Decrease
Horsham	426	111	74
Northern Grampians	339	61	82
Yarriambiack	277	89	68
Wellington	668	169	75
Total	***1710***	***430***	***75***

Table 4.2 PRISM data sets detailing record numbers pre and post trimming

4.6.2 Spatial Graphical Data Conversion

Conversion of the spatial graphical data sets involved software format conversion and projection transformations. The topographic (Vicmap Digital Topographic) and cadastral data (Vicmap Digital Property) initially underwent transformation from VICGRID coordinates to Australian Map Grid 1966 (AMG66), zone 54 or 55 depending on which UTM zone the data were located in. Projection parameters for the VICGRID projection were not specified in ESRI ArcInfo so needed to be determined and manually input during the transformation process. The land management data were all supplied with AMG coordinates, so did not require projection transformations.

The Vicmap Digital Property data were supplied as ESRI ArcView (shape) files and required software conversion into ESRI ArcInfo format. When converting between these formats, files with polygon topology required greater handling than linear (arc) files. Linear files only require ESRI ArcInfo's *SHAPEARC* command to be performed, however the polygon files needed to undergo an additional process called *REGIONPOLY*. This process was implemented to generate shape files for use in ESRI ArcView.

The *ARCLINK* function in MapInfo Professional allows an export file to be created (e00) which can then be imported into ESRI ArcInfo. Conversion of the planning scheme data from MapInfo Professional to ESRI ArcInfo resulted in only some of the files converting appropriately when using *ARCLINK*. The translator found line segments with zero length, polygon abnormalities and sliver errors forcing the program to abort due to its inability to fix self intersecting polygons that do not intersect on a point (MapInfo, 1997). The FME (Feature Manipulation Engine) Universal Translator was then used on the remainder of the planning scheme data. The FME Universal translator version used did not allow conversion between MapInfo Professional and ESRI ArcInfo. Thus, an intermediary conversion needed to be made between ESRI ArcView and ESRI ArcInfo.

The time spent integrating the spatial GIS data is estimated to be approximately 15-20 days per LGA and was related to the number and complexity of the variables deemed necessary for inclusion in the property database. Although this was a significant period of time, in the case of the planning scheme data, there were on average 15 files per LGA that required conversion. In addition, there were multiple themes supplied for the topographic, cadastral and hydrographic data sets which needed conversion. The enormity of this task is significant for any deliberation for implementing an automated decision support tool for rural valuation in Victoria.

4.6.3 Tabular Data Conversion

The only tabular non-GIS data supplied were the PRISM sale data and SLTVF data. To utilise the tabular data in a GIS, links were made to the cadastre, a polygon data set. Street number, street name and township are implicit spatial references in the PRISM data set and needed to be linked to an explicit location in the cadastral data set

through geocoding to create the property database. Within each LGA, the PRISM data differed in the completeness of their representation of the street address attributes and in some instances there were multiple parcels represented with the same street name, no street number reported and also instances of incomplete addresses. There is an additional field within the SLTVF data that was not supplied due to privacy issues and enables geocoding to be performed using a unique property identifier thus eliminating the use of the techniques described below.

The PRISM data was converted from HTML into Microsoft Excel and then into ESRI ArcInfo tables. The SLTVF data only required conversion between Microsoft Excel and ESRI ArcInfo. The conversion of the PRISM data set proved quite lengthy and once in ESRI ArcInfo tables, each record was linked (geocoded) to a polygon in the cadastral data set to create the property database. Initially, geocoding using the full street address proved to be unsuccessful in that no matches to the property database were made. A match field was populated as either full or partial based on the number of address identifiers used in the matching process (Table 4.3). After completion of the full and partial match, a further match process was undertaken (manual match) which involved using the State Digital Road Network (SDRN) data to search for street names listed in the PRISM data set. These roads were then located in the property database by displaying the SDRN roads as a background coverage. This resulted in a further increase in the total number of parcels matched as can be seen in Table 4.4. The property sale details together with land use were then populated into the property database if a match occurred.

Match Type	Street Name	Locality	Property Area	Crown Allotment
Full	✔	✔	✔	✔
Partial		✔	✔	✔
Manual	✔	✔	✔	

Table 4.3 Full, partial and manual match attributes

Match Type	Wellington (169 records)		Horsham (111 records)		Yarriambiack (89 records)		Northern Grampians (61 records)	
	Actual	*%*	*Actual*	*%*	*Actual*	*%*	*Actual*	*%*
Full	2	1	34	31	1	1	0	0
Partial	28	17	4	4	40	45	8	13
Manual	28	17	10	9	5	6	1	2
TOTAL	**58**	**35**	**48**	**44**	**46**	**52**	**9**	**15**

Table 4.4 Total number and percentage of records matched

Full, partial or manual matches in the Yarriambiack, Horsham and Wellington LGAs accounted for 52%, 44% and 35%, respectively, of the data sets being matched. The number of full matches was small, especially in the Wellington data set (Table 4.4). Matches in the Northern Grampians LGA were significantly less with only 15% of the data matched.

In this instance, the time spent geocoding the PRISM data to the cadastre to create the property database with full, partial and manual matching techniques was approximately five days per LGA. It should be noted that this was only a subset of the original data

and to geocode the whole data set prior to the minimisation techniques employed would have taken much more time.

The time taken highlights the importance of the Standard Parcel Identifier project (SPI) which is a project developed to address the issues on non-standard parcel identifiers. Standardising parcel identifiers will enable further integration between data sets to link compatible data and thus enable more PRISM and SLVTF data to be geocoded.

This research has highlighted some of the major deficiencies in the PRISM data set in that it cannot readily be incorporated with other geographical data sets. Although the PRISM data are appropriate to enable comparable analysis through a search by suburb to find locationally-similar properties, these data are not readily used in GIS. As highlighted by McCluskey *et al.* (1997), a major deficiency in digital property data for valuation is that although data are available, it is often difficult to 'unlock the significant potential' of the data. In the case of sale price information within Victoria, the information was supplied in HTML format with no identifier; thus it could not easily be linked spatially to a geographic location. The exception is the use of a street address. However this application may also present problems.

4.6.4 Location of Geocoded Properties

Figure 4.2 shows the location of the geocoded properties within the Wimmera CMA. The majority of the study properties were located in the Horsham and Yarriambiack LGAs.

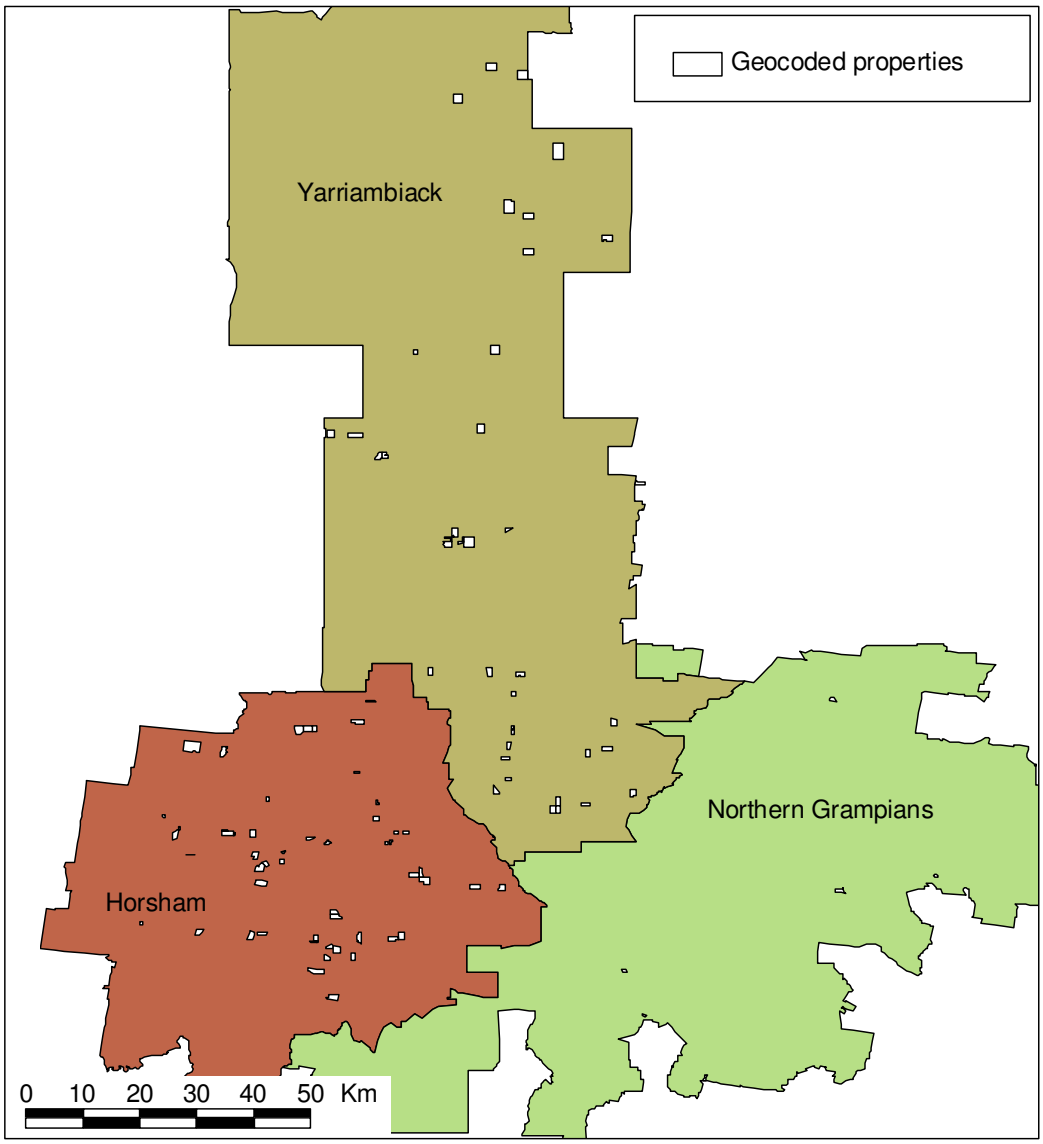

Figure 4.2 Wimmera Geocoded Properties

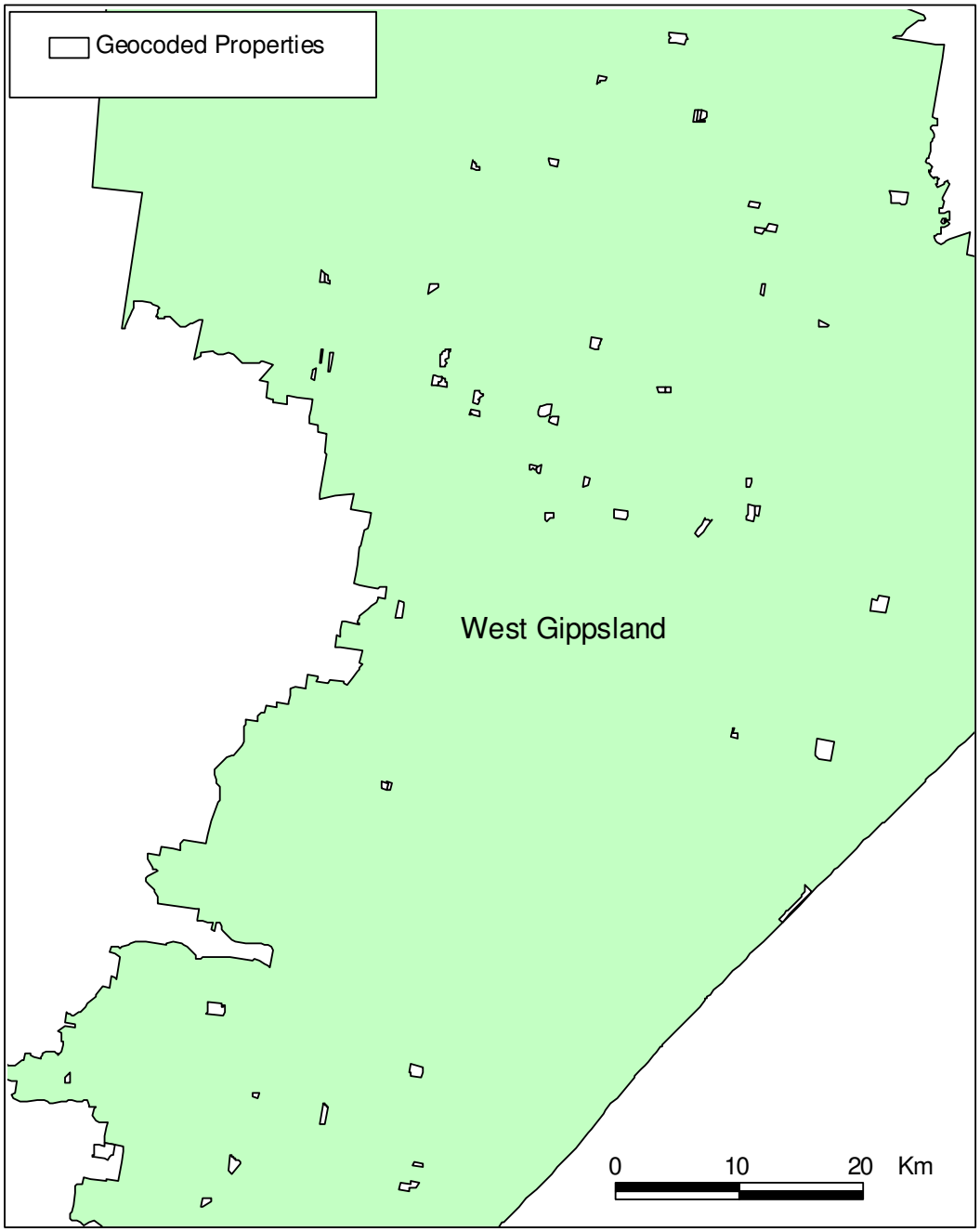

Figure 4.3 West Gippsland Geocoded Properties

Figure 4.3 depicts the location of the geocoded properties within the West Gippsland CMA. The properties are more evenly dispersed across the study region compared to the Wimmera. There is a greater concentration of properties near the town of Sale and a small concentration of properties in the southern area of the CMA.

4.6.5 Database Population with Additional Variables

Once all data sets were converted into ESRI ArcInfo format, additional variables were populated into the property database at an individual property level. These variables were derived from the data sets detailed in Section 4.6.6. Table 4.5 indicates the variables populated in the property database prior to adjustments, with a brief description of them.

Variable Name	Variable Description
AREA_TYPE	Local Government Area of the property (Horsham, Northern Grampians, Yarriambiack, Wellington)
SALE_DATE	Recorded date of sale, dd/mm/yyyy
SALE_MONTH	Month of sale – numeric value , 1 (Jan 1995) through to 65, with each increment indicating the following month in time
SALE_PRICE	Sale price of property in $
ADJ_PRICE	Sale price of property (adjusted to a year 2000 dollar value)
ADJ_PRICEPHA	Price of property per hectare (adjusted to a year 2000 dollar value)
LAND_USE	Land use type – land with no building, cereal, dairy, beef, sheep, other
PROP_AREA	Size of property in hectares
SEVERITY	Severity of salinity – low, medium, high, unknown
DSTYPE	Dryland salinity type – natural, induced, incipient
SPECIESCD	Pest species – fox, rabbit, dog
DAM_WELL	Presence of dam or well, (1 dam/well, 0 otherwise)
WATERB_PT	Number of waterbody points
WATERB_AREA	Total area of waterbody (square metres)
WATERCRS	Length of watercourses (metres)
ZONE_CODE	Planning scheme code (1 rural, 0 otherwise)
LSIO	Land subject to inundation (1 subject to inundation, 0 otherwise)
HORSH_DIST	Distance to Horsham (metres)
STAW_DIST	Distance to Stawell (metres)
SALE_DIST	Distance to Sale (metres)

Table 4.5 Variable descriptions of property data prior to adjustments

The property database (Appendix A), derived from the cadastre (Vicmap Digital Property), had a number of variables detailing the location of each polygon using street name, number, locality, postcode, town and parish code as well as lot and plan

number, part and section details. These additional variables as shown in *Vicmap Digital Property (Standard) Version 2.01* were removed from the property database as they were no longer necessary in this study.

The fields populated from the PRISM data set were *SALE_PRICE*, *SALE_DATE*, *LAND_USE and PROP_AREA*. The land use category was numerically represented and the property area was provided as a separate field which depicts the property size in metres, acres or hectares. All property areas were converted into hectares with those properties less than 20 hectares excluded. The above variables were created in the property database when the PRISM data was matched to the Vicmap Digital Property data set.

Additional processing of the property database involved merging the attributes of multiple parcels constituting a single property. The merging of parcels further decreased the number of reported properties as some originally had between two to three parcels related to them. Although the parcels were kept as individual polygons, the attributes were taken from all the parcels and combined. The merging of parcels into properties resulted in a further decrease in the number of recorded properties to 143 for the property database. There has been some property valuation research that used small property databases consisting of 94 (Miranowski & Hammes, 1984) and 158 (Gardner & Barrows, 1984) properties. It was decided to use all of the 143 properties in my study, and not take a random sample of the data as this would have further decreased the sample size. This may have affected the results obtained in this study. However, it was decided that the alternative approach of using fewer properties may reduce the potential accuracy of the models.

The fields *SEVERITY* and *DSTYPE* were populated from the DISCH25_AREA data set. The DISCH25_AREA data set was overlayed with the property database and any instances of salinity were recorded by populating the severity of the salinity as either low, medium or high. *DSTYPE* was populated with the type of salinisation if any was present.

SPECIESCD was the only field populated from the PMIS100 data set and again this was performed using a visual overlay of the property data set with the PMIS100 data set.

After conversion of the LASTBURNT100 data set, a spatial overlay indicated that there were no instances of any of the properties in areas where there had been reported fires or prescribed burning so this variable was eliminated. This highlights a major limitation in using data from various sources in that conversion to the appropriate software format and projection is often necessary before the data can be used. It is only once an overlay is performed that it can be ascertained whether there are any instances of the variable overlaying any of the properties within the property database.

Variables created from overlay between the hydrographic theme of the Vicmap Digital Topographic data and the property database included *DAM_WELL, WATERB_PT,* and *WATERCRS*. Descriptions of these variables are located in Table 4.5. *DAM_WELL* was populated as an indicator variable by recording the presence or absence of a dam or well on a property. *WATERB_PT* was populated by totalling the number of waterbodies that fell in each property whilst *WATERCRS* was populated by totalling the length of all watercourses in each property. Where multiple parcels made up one property, then the total length of the *WATERCRS* attribute for all parcels were summed

instead of reporting these separately. This was also the case for the *WATERB_PT* variable.

The variables *ZONE_CODE* and *LSIO* were populated using the Planning Scheme database. *ZONE_CODE* was populated as either in a 'rural planning zone' or 'not part of land classed as a rural zone'. Likewise, *LSIO* was populated as 'land being subject to inundation' or 'land not subject to inundation' using indicator variables once again.

Distance variables were created by measuring the Euclidean distance between the centroid of a property to the nearest major town in the data set. For the Wimmera region, distances to Horsham and Stawell were determined and later the closest distance was taken between the two and populated into a variable called *TOWN_DIST*. A distance was determined in the Wellington data set between the property centroid and the town of Sale. The ESRI ArcInfo command *POINTDISTANCE* was used to measure the distance between the two points.

The adjustment of sale prices involved using the CPI to develop an index to bring all sale prices into a common year representation. The CPI rates were obtained from the Australian Bureau of Statistics website for the years 1995 to 2000. A base index at the year 1995 was chosen and all prices were adjusted so that they were representative of a year 2000 dollar value. As 1995 was selected as the base year, an index value of 100.0 was assigned to this year with increases per year being added to the previous years index figure. A calculation was then performed to adjust each year into a year 2000 value (Table 4.6) and this was stored as the variable *ADJ_PRICE* in the property database. This calculation is similar to that of Net Present Value (NPV) in that the year

2000 index rate is divided by the 1995 index rate to obtain a multiplier to apply to the sale prices based on the year in which they were sold.

Year	CPI Index	Index adjustment calculation
1995	100.0	110.8/100 = 1.108
1996	104.8	110.8/104.8 = 1.057
1997	106.4	110.8/106.4 = 1.041
1998	106.4	110.8/106.4 = 1.041
1999	107.9	110.8/107.9 = 1.027
2000	110.8	110.8/110.8 = 1.000

Table 4.6 CPI Index figure for each year within the data set

The variable *ADJ_PRICEPHA* was determined by dividing the size in hectares of each property by the *ADJ_PRICE* variable for that corresponding property.

The variable *AREA_TYPE* was derived from the LGA to which each parcel belonged by performing a spatial overlay of the property with the LGA boundary data set.

4.6.6 Variable Adjustments

4.6.6.1 Creation of Indicator Variables

Creation of indicator variables is necessary in statistical analysis to perform regression analyses on the data. Where one variable called *AREA_TYPE* is represented by a textual description of either 'Horsham', 'Northern Grampians', 'Yarriambiack' or 'Wellington' based on the LGA to which the property belongs, four new variables are necessary to represent these geographical areas as indicator variables. Within each variable name, a value of either '1' for the presence of the variable or a '0' for the

absence of a variable is provided as shown in Table 4.7. In every case, if there are *k* indicator variables (*k* is the total number of indicator variables), only *k*-1 will enter the model explicitly.

Indicator Variable Name	Indicator Variable Description
AREA1	(1,0) 1 if in Horsham, 0 otherwise
AREA2	(1,0) 1 if in Northern Grampians, 0 otherwise
AREA3	(1,0) 1 if in Yarriambiack, 0 otherwise
AREA4	(1,0) 1 if in Wellington, 0 otherwise

Table 4.7 *AREA_TYPE* Indicator Variables

LAND_USE was another variable in which indicator variables were created. The land use data supplied in the PRISM data set defined land use via a code as seen in Table 4.8.

PRISM Land Use Code	Land Use Type
2	Farm land without buildings
31	Cereal
32	Dairy
33	Beef
34	Sheep
41	Other rural property

Table 4.8 PRISM Land Use Categories

The *LAND_USE* variable was changed to six indicator variables indicating the presence or absence of each of the six different types of land use as can be seen in Table 4.9.

Indicator Variable Name	Indicator Variable Description
LUSE_1	(1,0) 1 if farm land w/o buildings, 0 otherwise
LUSE_2	(1,0) 1 if cereal farm, 0 otherwise
LUSE_3	(1,0) 1 if dairy farm, 0 otherwise
LUSE_4	(1,0) 1 if beef farm, 0 otherwise
LUSE_5	(1,0) 1 if sheep farm, 0 otherwise
LUSE_6	(1,0) 1 if other rural property, 0 otherwise

Table 4.9 Land Use Indicator Variables

A number of other variables were converted into indicator variables with some also undergoing variable name changes. *DSTYPE* was renamed to *NATURAL* as this was the only reported dryland salinity type recorded. Likewise, *SPECIESCD* was changed to *FOX* as there were no reported instances of rabbit or dog pests so the variable was only representing the presence or absence of foxes. Table 4.10 details the *DSTYPE* and *SPECIESCD* variable name changes.

Variable Name	New Indicator Variable Name	Indicator Variable Description
DSTYPE	*NATURAL*	(1,0) 1 if salinity type is natural, 0 otherwise
SPECIESCD	*FOX*	(1,0) 1 if fox, 0 otherwise

Table 4.10 *DSTYPE* and *SPECIESCD* Indicator Variables

4.6.6.2 Other Variable Adjustments

The variable *SEVERITY* was initially represented through a textual description and thus was re-formatted to a numerical range from one through to four indicating levels of salinity. A summary of the additional variable changes are detailed in Table 4.11.

Variable Name	New Variable Name	Variable Description
HORSH_DIST	*TOWN_DIST*	Distance in metres to nearest town
STAW_DIST	*TOWN_DIST*	Distance in metres to nearest town
SALE_DIST	*TOWN_DIST*	Distance in metres to nearest town
SEVERITY	*SEVERITY*	Range of 1-4 of low to high salinity

Table 4.11 Town distance and dryland salinity variable alterations

4.6.7 Land Use Representation throughout Study Areas

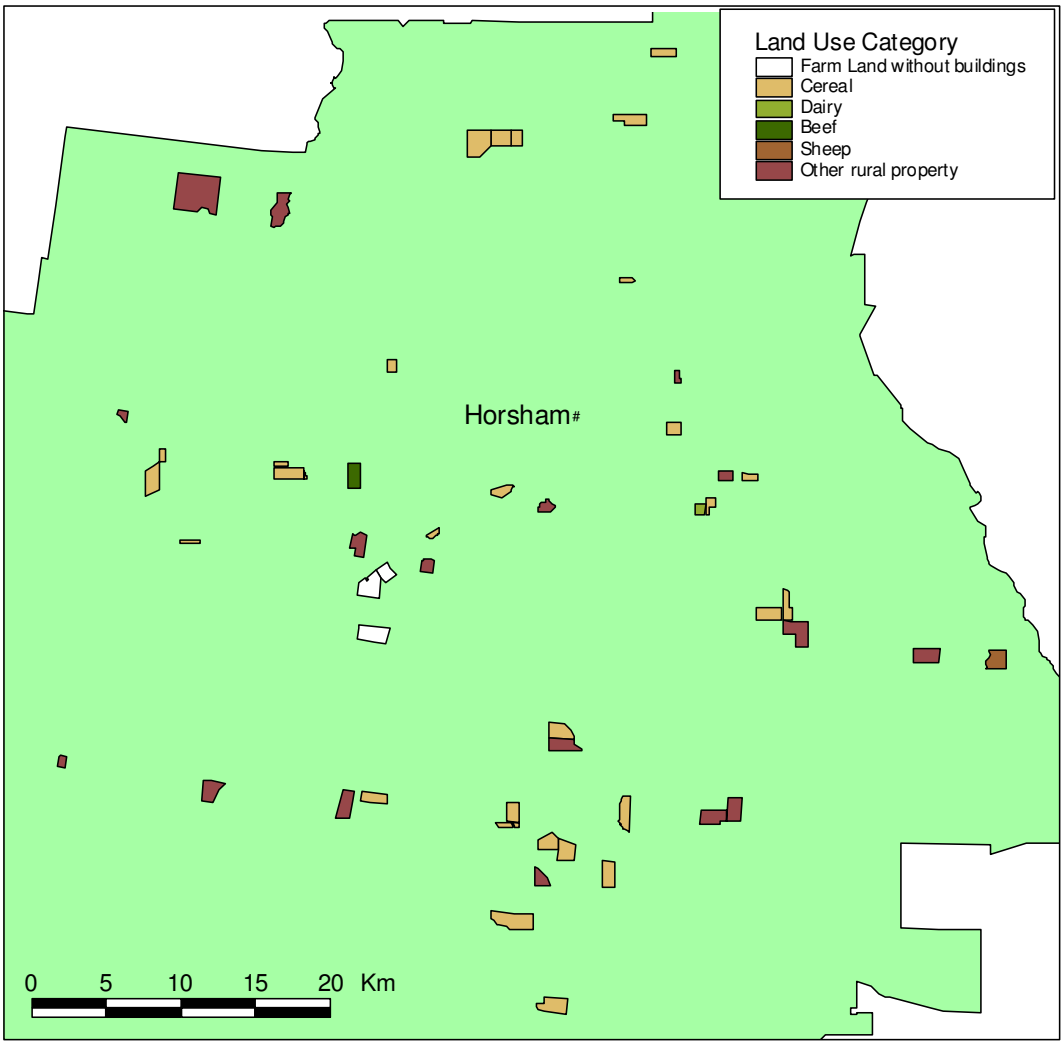

Figure 4.4 Horsham Land Use

All six land uses are represented within the Horsham LGA as can be seen in Figure 4.4. The southern and northern sections of the LGA tend to support cereal and rural land uses classed as 'other'. Nearest to the town of Horsham a wider variety of land use types exists including dairy, beef and grain production.

Within the Northern Grampians LGA, only a small number of properties exist with the majority of these being cereal, other rural, and farm land without buildings. Due to the number of properties within this LGA, not all land uses are represented as can be seen in Figure 4.5.

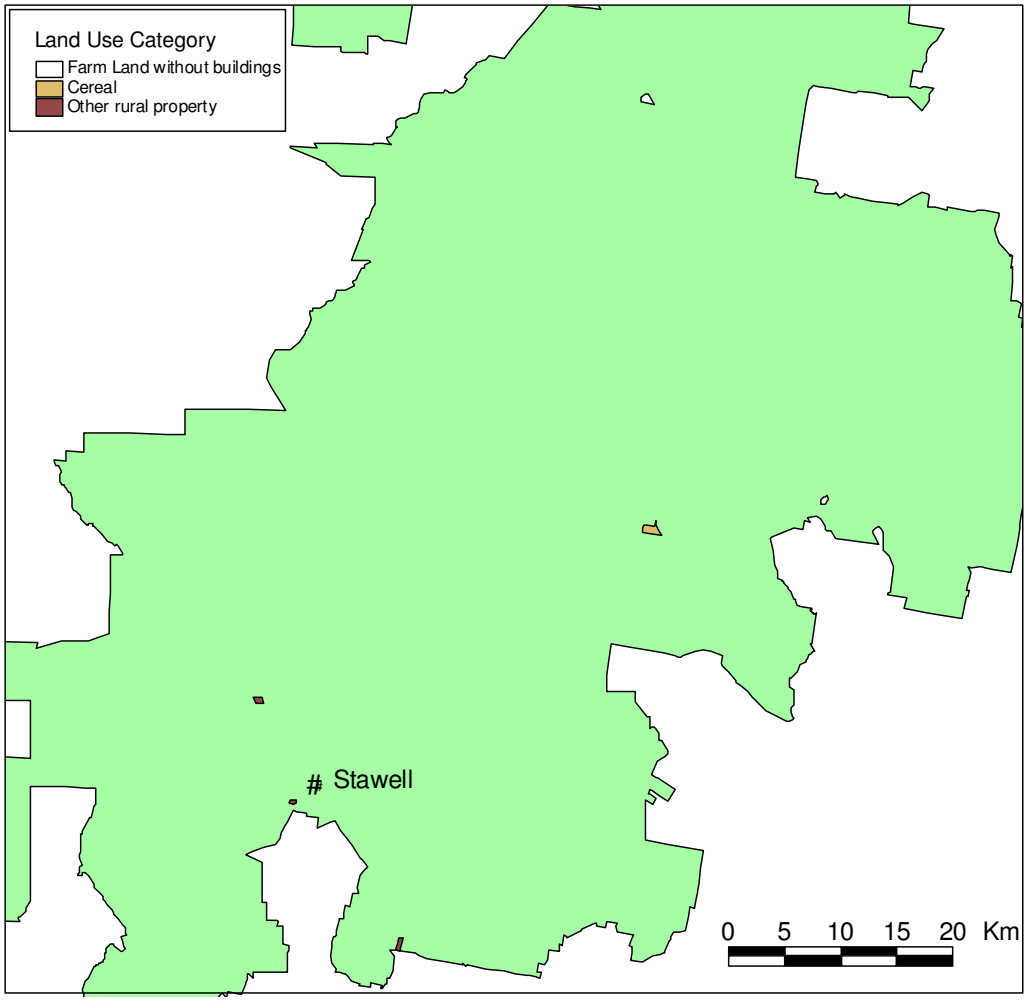

Figure 4.5 Northern Grampians Land Use

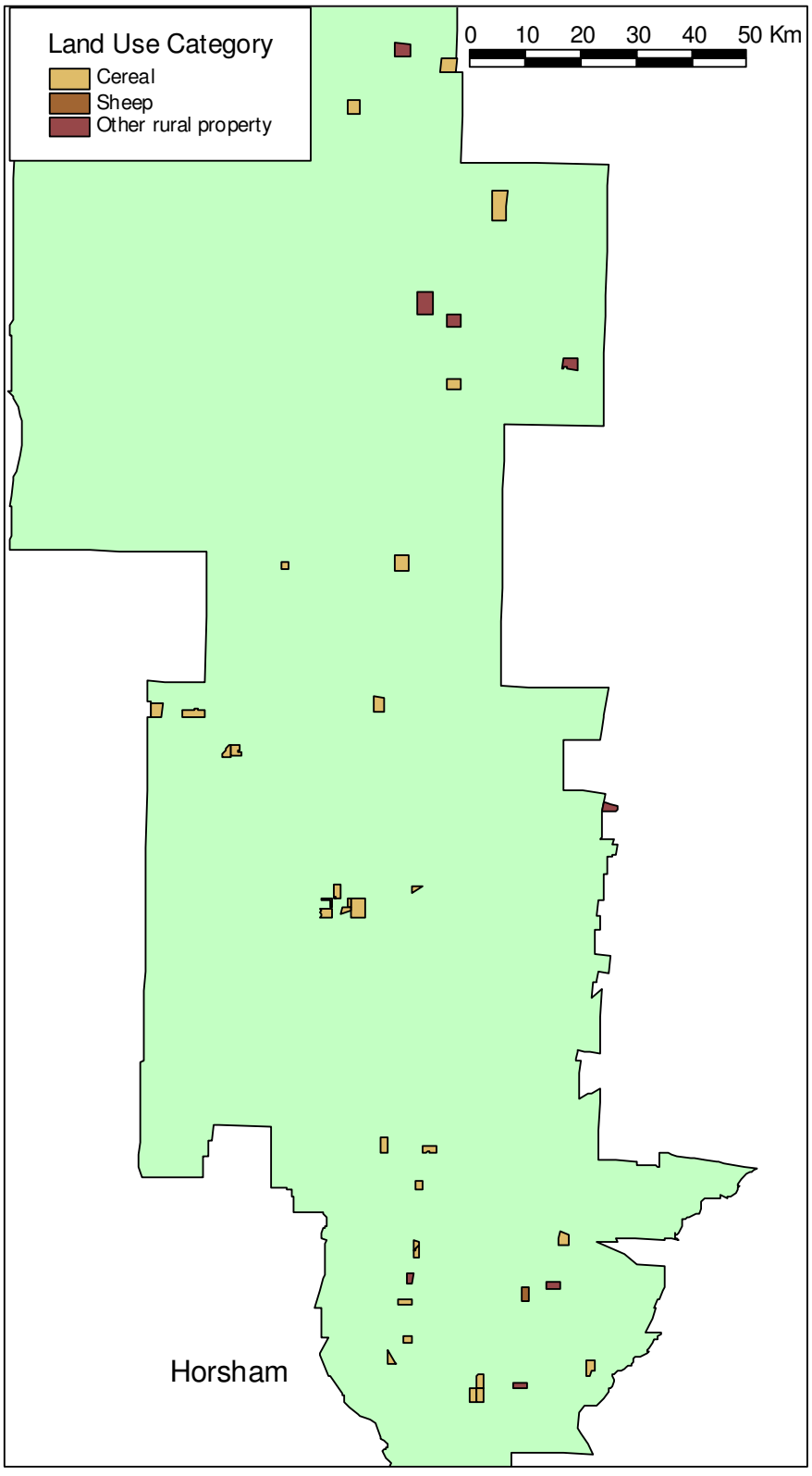

Figure 4.6 Yarriambiack Land Use

Within Figure 4.6 a larger number of properties exist compared to that of the Northern Grampians LGA. Only 3 land use types are represented within this LGA with a concentration of cereal properties within the centre of the LGA. To the north, cereal and other rural property land uses are dominant whilst the south tends to have a combination of land use types.

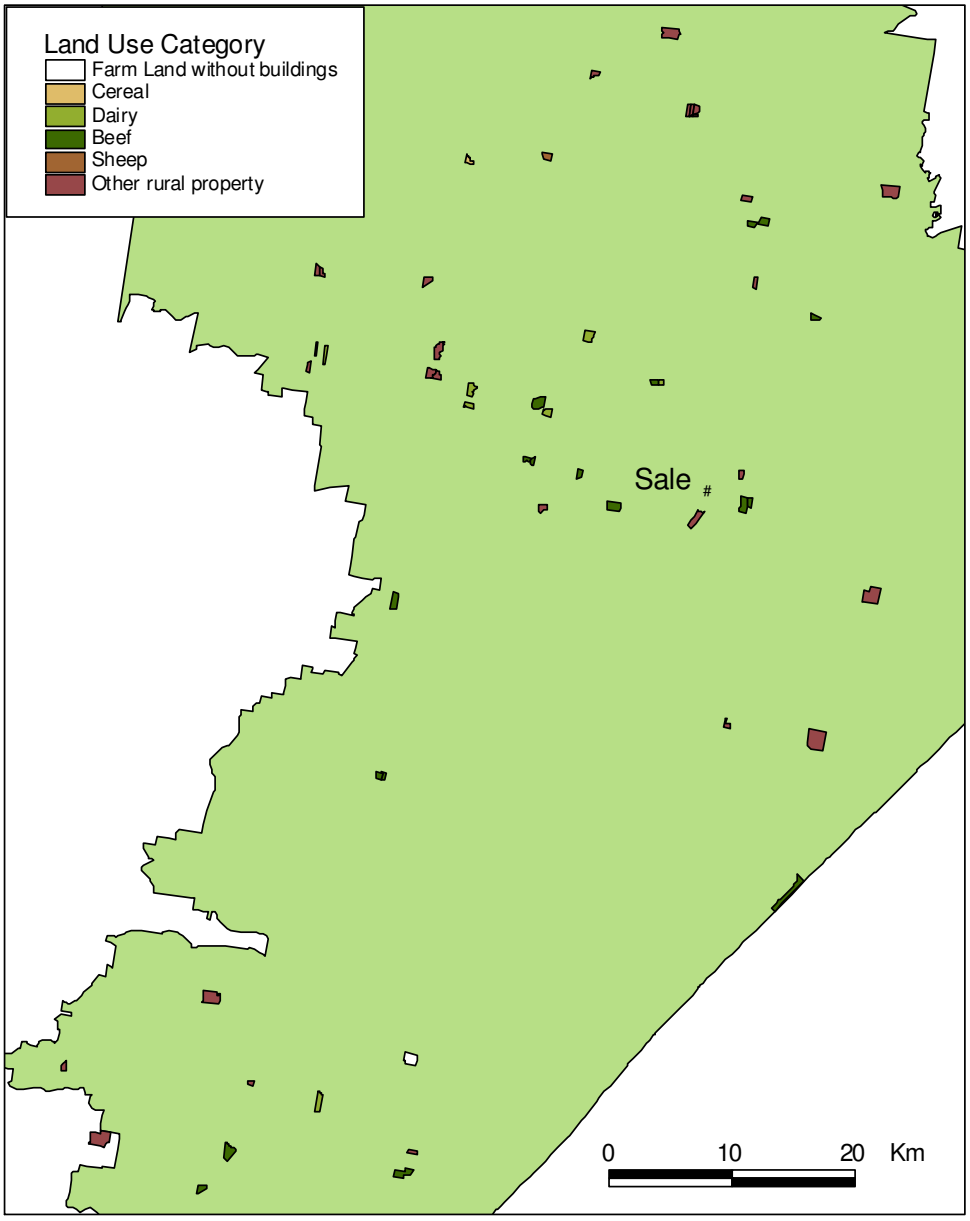

Figure 4.7 Wellington Land Use

Land use types in the Wellington LGA include dairy farming, beef production and sheep grazing (Figure 4.7).

4.7 Comparison between The Rural Property Valuation Model and The Property Database

The conceptual model developed within Section 2.4 details a wide array of property characteristics hypothesised to be significant for rural property valuation in Victoria. The model was developed using texts on rural valuation, peer reviewed research and Valuation Best Practice specifications to provide a table of the significant property characteristics that can be considered in modelling the value of rural properties. In Victoria some of these variables are not available in digital form. Surveys and questionnaires may be required to obtain these data. As the aim of my research was to develop numerical models using only 'publicly available' digital data, the option of undertaking manual fieldwork was discounted.

To ascertain if automated property valuation models were able to be developed within Victoria a number of key data sets were obtained. During the development of the property valuation database, a variety of digital data was available for use within the study, however as mentioned previously, many rural data variables were unable to be derived. Due to the selection of properties within the study regions, there were often no instances of the variable or characteristic being present. Within the 'Structural' characteristics; house presence, house age, house condition, building condition were not available digitally. In addition, fencing type and the proportion of each land use on the property were not available. From the 'Environmental' characteristics, fire prone area was available yet there were no instances of this variable for any of the properties. Flood information was also not available. The 'Accessibility' characteristics were all able to be derived from existing topographic data.

Within the 'Neighbourhood' variables, all characteristics were available digitally however due to the size of the study sample it was felt that incorporating temperature and rainfall into the database would have lead to just two regional variables being populated into most of the data. As the temperature and rainfall data are regional and averaged over a relatively large area, this information would have not been suitable for regression modelling if all instances of the temperature variable were the same or similar.

The 'Economic' characteristics included site valuation which was not used due to the unavailability of the data over the whole study area, and the issues with linking the data to each individual property. The Consumer Price Index (CPI) and sale month were incorporated as one variable as an adjustment factor of the sale prices to account for the time between sales and inflation over the time span of the study properties. Production information was not available on an individual property basis however could have been obtained as an aggregate data set from the annual Farm Surveys conducted by the Australian Bureau of Statistics (ABS).

The justification for including most of the variables was based on whether through GIS analyses a variable could be derived which was specific to each individual property. ABS data were generally not used within this study due to its aggregation to Census collection districts and the authors requirement to have more detailed information which was specific to each individual property. Although the conceptual model depicts a large number of property characteristics that may be significant to rural property, unless the database of properties is particularly large, all variables could not be used in a study which incorporated around 150 properties. Thus, by selecting a portion of this information, based on 'publicly available' data and data specific to the individual

property, it was felt that this would enable adequate model development and testing for these study areas.

The lack of housing and building characteristics may have been an influence on the results of this research, however other research has modelled rural values with minimal (Reynolds & Regalado, 2002) variables representing building or housing information, and some studies used none at all (Gardner & Barrows, 1984). The implications of using a small variety of property characteristics may in fact lead to less accurate models given there may be some other factors that influence price and that could not be modelled due to the lack of variables.

4.8 Summary

The two study areas selected for this research were the Wimmera and West Gippsland CMAs in Victoria. The Wimmera, located in the west of the state and West Gippsland, in the south-east, are major agricultural producing areas in Victoria. These areas have many land degradation issues prevalent, thus making both suitable for use within this research. Within these two CMAs four LGAs were selected for this study.

The data sets obtained included information detailing property sale prices, municipal valuations, the cadastre, topography, hydrography, salinity regions, pest infestation sites, bushfire prone areas, dryland salinity discharge areas and land planning data. After obtaining these data sets, converting them to appropriate GIS software formats, transforming map projections and performing spatial overlays, it became apparent that many of the data sets obtained were redundant for the purposes of my study. This

proved to be a major limitation as much time was spent integrating the data to find specific attributes were no longer relevant. As a result, many of the property characteristics conceptualised to be important for rural valuation could not be tested. These limitations have highlighted the fact that more detailed property data often are not available in Victoria. Moreover, where such data are available their use in the public domain may be limited due to licensing and privacy constraints.

The GIS software used in this study included ESRI ArcInfo, ESRI ArcView, ESRI ArcGIS and MapInfo Professional. Minitab and SPSS13 were used for the statistical regression analyses and cluster analysis stages of the research. Microsoft Excel was used to transfer tabular data into GIS tables and also in the model testing stage of the study.

The database development process involved cleaning and minimisation of the PRISM data set, conversion of the tabular GIS data, tabular non-GIS data, spatial graphical data and projection transformations. Additional processing involved the population of additional variables into the property database and also alteration of the variables into indicator variables for statistical processing.

A framework for use in future data integration tasks was developed based on procedures undertaken during this research and addresses many of the issues encountered during the database development stage of the research. A list of recommendations to facilitate greater ease during integration was developed and will be presented in the final chapter of the thesis. These recommendations include alterations to the elements included in current metadata standards and also a means to

improve the utilisation of the PRISM sale transaction information with other GIS data sets.

Chapter 5 Development of the Numeric Rural Property Valuation Models

5.1 Introduction

The Conceptual Rural Property Valuation Model provides an framework for relating property characteristics to value. The previous Chapter described the practical limitations of fully implementing the model, due to the inability to readily populate a database of these characteristics. In this Chapter, various regression techniques are utilised in an attempt to establish a relationship between those characteristics which have been quantified, and several measures of value. These implemented models are referred to as the Numerical Rural Property Valuation Models. A statistical software package (Minitab) was used to apply regression techniques to the previously described data sets. The objective was to derive mathematical models which represent the highest and best use value of a property based on available digital data describing the property.

This stage of the research involved developing numerical models based on geographically constrained areas as sub-markets. This phase involved the use of an indicator variable to depict the LGA to which each property belonged rather than determining the sub-market statistically or through using other a-priori techniques.

An exploratory research process was followed during this stage, making adjustments to the data set to detect and remove outliers and to re-structure the data into various categories and groupings based on price ranges, land use and area types. Some variables were progressively eliminated due to their lack of price estimation. The

'*ADJ_PRICE*' dependent variable was tested using a number of implementations of the dependent variable. This included applying Logarithms to a base of 10 and using adjusted prices per hectare. This was performed due to the varying dependent variables used in other rural research to ascertain in rural Victoria which was the most appropriate in modelling property price.

Due to the number of models developed, it was decided to only spatially present the results of those models where the R^2 value exceeded 45%. Thus, only Models 1, 7 and 8 were mapped in this stage of the research.

5.2 Development of the Numeric Models

After the variable re-formatting and indicator variable creation discussed in Section 4.6.6, the property database was imported into Minitab for statistical analysis and model development.

The final variable names after the creation of indicator variables are shown in Table 5.1. Indicator variables were created to allow the variables to be used within statistical regression analyses and an explanation of the process can be found in Section 4.6.6.1.

Variable Name	Variable Description
AREA1	(1,0) 1 if in Horsham, 0 otherwise
AREA2	(1,0) 1 if in Northern Grampians, 0 otherwise
AREA3	(1,0) 1 if in Yarriambiack, 0 otherwise
AREA4	(1,0) 1 if in Wellington, 0 otherwise
ADJ_PRICE	Sale price of property (adjusted to a year 2000 dollar value)
ADJ_PRICEPHA	Price of property per hectare adjusted to a year 2000 dollar value
PROP_AREA	Size of property in hectares
SEVERITY	Severity of salinity – range 1 to 4 (low, medium, high, unknown)
NATURAL	(1,0) 1 if natural dryland salinity type, 0 otherwise
FOX	(1,0) 1 if presence of fox, 0 otherwise
WATERB_PT	Number of waterbody points
WATERB_AREA	Total area of waterbody (square metres)
WATERCRS	Length of watercourses (metres)
ZONE_CODE	Planning scheme code (1 rural, 0 otherwise)
LSIO	Land subject to inundation (1 subject to inundation, 0 otherwise)
TOWN_DIST	Distance to nearest town (metres)
LUSE_1	(1,0) 1 if farm land w/o buildings, 0 otherwise
LUSE_2	(1,0) 1 if cereal farm, 0 otherwise
LUSE_3	(1,0) 1 if dairy farm, 0 otherwise
LUSE_4	(1,0) 1 if beef farm, 0 otherwise
LUSE_5	(1,0) 1 if sheep farm, 0 otherwise
LUSE_6	(1,0) 1 if other rural property, 0 otherwise

Table 5.1 The 22 variables used for statistical analysis

Prior to regression analysis, two stem and leaf plots were produced, one using the variable '*ADJ_PRICE* ' (Figure 5.1) the other utilising a Logarithm of this variable (Figure 5.2). The stem and leaf plots are a descriptive statistic and were used to

examine the shape and distribution of the two dependent variables used during this study. Stem and leaf plots are used for visualisation of the data set by showing the shape of the distribution, the range of data values and the magnitude of the data. Figure 5.1 shows that the '*ADJ_PRICE*' value has values ranging from $10,000 to $810,000. It also shows that the majority of the values are concentrated in the lower values between $10,000 to $430,000 and that there are only a small number of values above $430,000. This is indicated by the smaller number of values between 4/6 ($460,000) and 8/1 ($810,000) in Figure 5.2.

```
  11    0 11122233344
  39    0 55555556666667777888899999
 (34)   1 00000000011122222233333333444444
  70    1 5555666677777788889999
  48    2 00001111223333
  34    2 555788888
  25    3 011123
  19    3 5666688
  12    4 02333
   7    4 6
   6    5 2
   5    5
   5    6 004
   2    6 8
   1    7
   1    7
   1    8 1
```

Figure 5.1 Stem and Leaf Plot of '*ADJ_PRICE*', eg: 0/5 = 50000

```
   3    4 001
   4    4 3
   9    4 44445
  20    4 66677777777
  39    4 8888888888999999999
 (37)   5 00000000000000000111111111111111111111
  67    5 22222222222222222233333333333333
  34    5 4444444444455555555555
  12    5 666666777
   3    5 889
```

Figure 5.2 Stem and Leaf Plot of 'Log_{10} *ADJ_PRICE*', eg: 4/0 = 4.0

Figure 5.2 utilises the 'Log_{10} *ADJ_PRICE*' variable. Taking a Logarithm to base 10 of the '*ADJ_PRICE*' variable has altered the distribution of the data so that it is not as concentrated in the lower end of the values as in Figure 5.1. Logarithms have been taken of dependent variables in regression analysis (Reynolds & Regalado, 2002; Gardner & Barrows, 1984)) for property valuation. The values of 'Log_{10} *ADJ_PRICE*' range from 4.0 to 5.9 with the values concentrated nearer to the middle of the data value range. The stem and leaf plot in Figure 5.2 shows a more even distribution of values unlike that of Figure 5.1 where the distribution was not as even.

Four tests were conducted using different regression techniques with each test creating two models for rural valuation as shown in the following sections. Testing of the regression models involved applying the regression coefficients to the property variables from the property database to determine a valuation estimate. A comparison was then made between this value estimate and the actual sale price. It should be noted that due to the limited size of the property database, cross validation of the regression residuals (testing on a separate data set) was not undertaken and thus the residuals were tested on the same data that the regression models utilised. This in no way compromises the results obtained due to the regression models developed having their own verification procedure. This testing highlights the ability of the models to determine effective residuals and also examines the similarities between the estimated price and the actual sale price.

5.3 Outlier Removal

A scatterplot was produced (Figure 5.3) of the '*ADJ_PRICE*' versus the '*ADJ_PRICEPHA*'. This scatterplot was only produced after initial testing indicated that due to the low levels of price prediction, there must be other factors affecting value.

The scatterplot showed that an outlier was present in the data set as can be seen in Figure 5.3. There is only one value plotted at $20,000 per hectare with no other values plotted around this price per hectare value indicating the presence of an outlier. This outlier was found to be a small property of 20 hectares with an uncharacteristically high sale price.

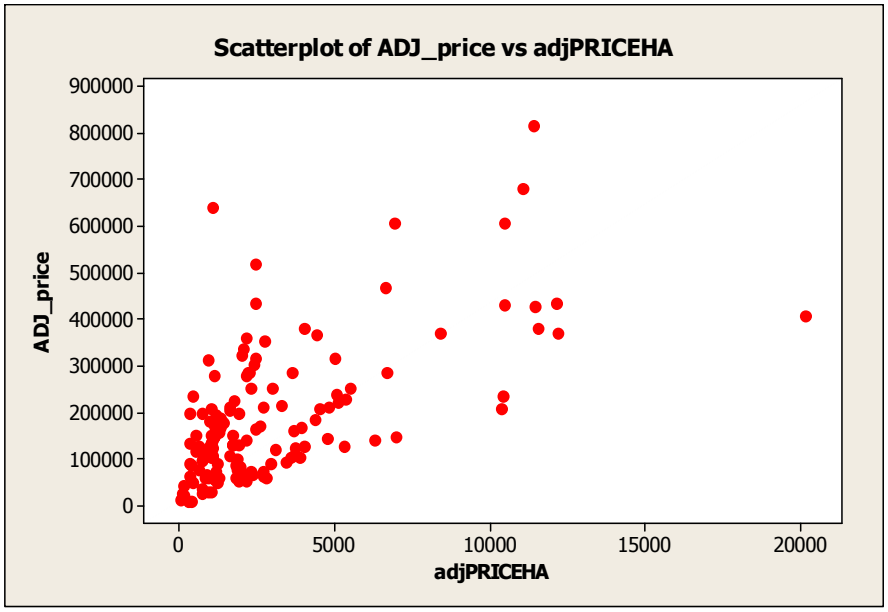

Figure 5.3 Scatterplot of '*ADJ_PRICE*' versus '*ADJ_PRICEPHA*'

Due to the lack of data detailing the characteristics of the buildings on each property, the outlier could be indicative of a lifestyle property which is not used for agricultural purposes or could even be a property with a much larger, newer house than the other houses on nearby agricultural properties. Whichever the case, there is some other underlying characteristic of this property, of which there was no additional information which gave it such a large price per hectare value than its other neighbouring properties and consequently was removed from the database.

5.4 Re-classification of Data Set

5.4.1 Removal of One LGA

The Northern Grampians LGA comprised only six properties and had sale prices ranging between $27,000 to $211,000. The properties also ranged in size between 21 to 107 hectares and included land uses from land use categories one, two and six which are farm land without buildings, cereal and other rural land uses. This LGA was used for Models 1 through 5 and Model 7, but was removed for Models 6 and 8. The six properties within the geographic region of the Northern Grampians LGA (the independent variable *AREA2)*, have prices which are more concentrated in the lower end range of the sale prices (under $300,000) whilst those in other LGAs are as high as $800,000. Initial testing included this LGA, but was removed to determine if estimates improved for the latter testing.

5.4.2 Combining Land Use Categories

Initial testing and model development was undertaken using the land use categories as detailed in Table 5.1. The models developed used separate land use categories (Models 1 through 4). These categories did not prove to have reliable estimates of property values so the land use categories were reduced to three categories for later assessments of the models. Due to similarities in pastoral uses of some of the land use categories and with some land use categories displaying similar price characteristics, the amalgamation was intended to enable more information to be obtained from each regression performed. Land use category one (farm land without buildings) was merged with land use category two (cereal farming) to create *LUSE_12*. Land use categories three (dairy), four (beef) and five (sheep) were merged to create *LUSE_345* and land use categories three and four were merged to create *LUSE_34* which contained only beef and sheep land use types. The amalgamations between

land use categories were examined to determine if utilising these new categories would improve the price estimates. These amalgamations were used for Models 5 through 8.

5.5 Regression Analysis - Testing Phase 1 – Geographically determined sub-markets

5.5.1 Test 1: Dependent variable = '*ADJ_PRICE'*

The entire database using the variables from Table 5.1 underwent the initial regression analyses. 'Best subsets' regression is a technique to identify models using as few predictors as possible (Daniel & Wood, 1980) and six significant variables were identified using this technique. Adjusted sale price (*ADJ_PRICE*) was used as the dependent variable and a regression analysis was performed which yielded an R^2 of 45.9% and an adjusted R^2 of 43.5%. The R^2 value is a measure of the fit of the regression equation. An R^2 of 45.9% indicates that 45.9% of the price is explained by the regression equation and thus there is 54.1% attributed to other factors.

The P value in Table 5.2 indicates the level of significance of the independent variable with any P value under 0.05 being significant. The SE coefficient is the Standard Error in the coefficients whilst the T value is a ratio of the coefficient and the SE coefficient.

Predictor	Coefficient	SE Coefficient	T	P
Constant	132957	18454	7.20	0.000
PROP_AREA	337.3	105.0	3.21	0.002
WATERB_AREA	16.475	2.596	6.35	0.000
WATERCRS	18.070	5.203	3.47	0.001
TOWN_DIST	-1.0704	0.3100	-3.45	0.001
LUSE_3	223080	40206	5.55	0.000
LUSE_4	74534	28725	2.59	0.011

Table 5.2 Regression coefficients for Model 1

The six variables in Table 5.2 are shown to be significant in affecting property value. The sign of the town distance variable (*TOWN_DIST*) is negative which coincides with a property declining in value as you move further from each town.

The testing procedure used the developed regression equation and multiplied each variable in the data set for a particular property by the regression coefficients in the equation. This was performed using Microsoft Excel to determine price estimates for each property using the whole data set. Testing of the regression equation on the data set did not show promising estimation results. Only 16% of price estimates obtained fell within 10% of the actual sale price. Analysis of prices that fell within 20% of the sale price were also low with only 26% falling in this range.

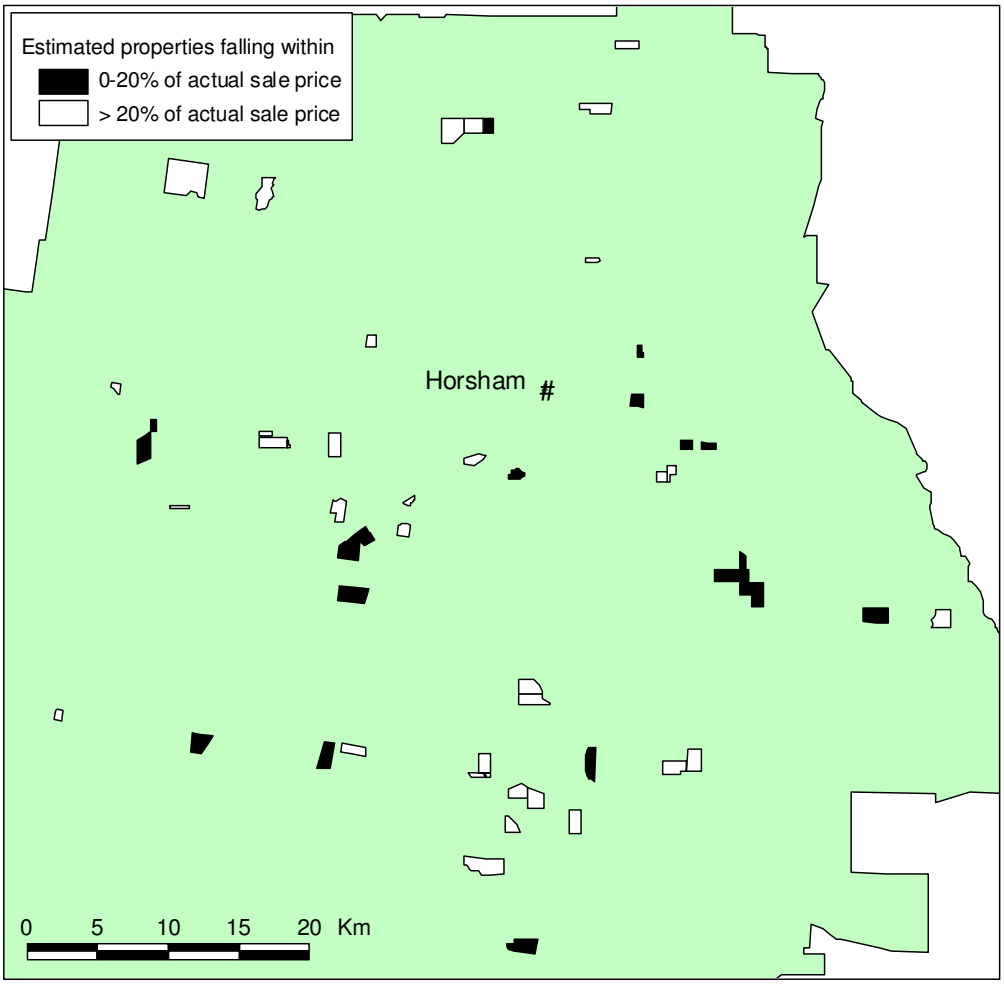

Figure 5.4 Horsham - Model 1 - properties falling within 0-20% of actual sale price

Figure 5.4 shows the properties where the models generated estimates that fell within 0-20% of the actual sale price of the property. There did not appear to be any locational relationship between properties that were estimated with more accuracy than those that were not.

Figure 5.5 Northern Grampians - Model 1 - properties falling within 0-20% of actual sale price

Figure 5.5 shows the properties within the Northern Grampians LGA. None of the models developed for properties within this LGA generated valuations within 0-20% of the actual sale price.

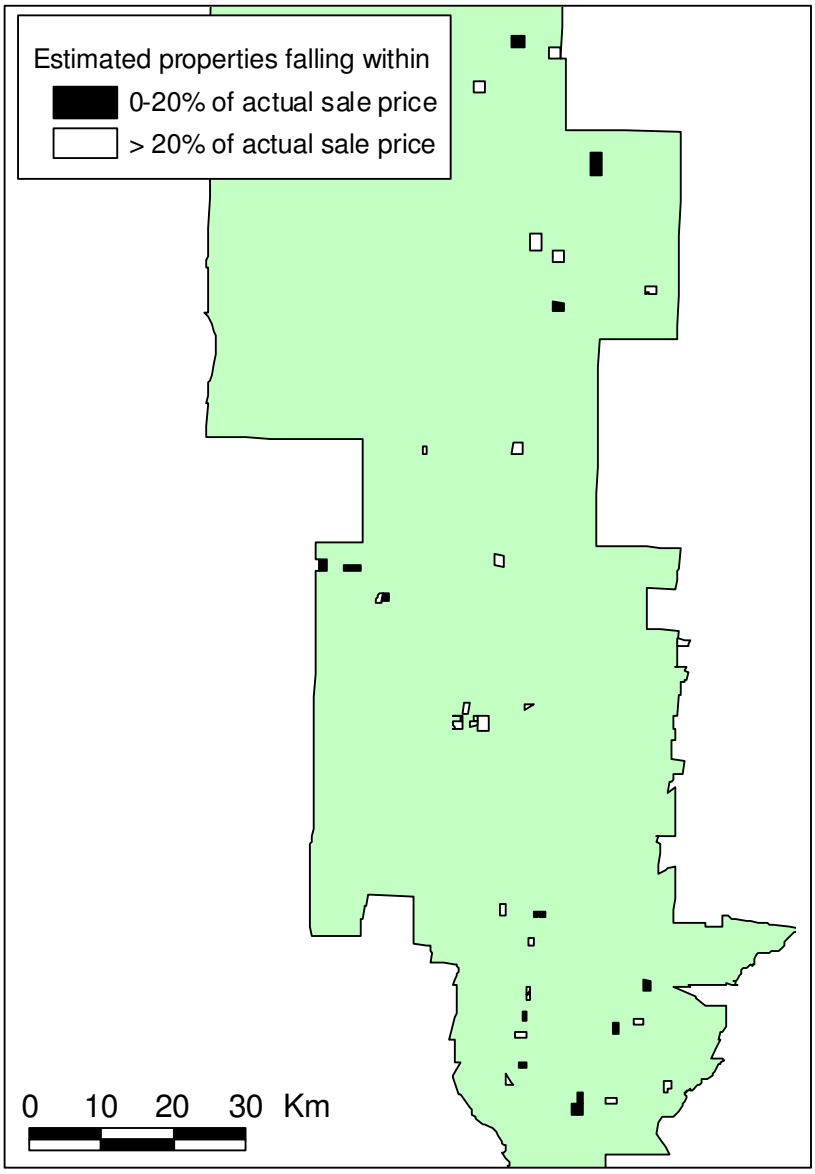

Figure 5.6 Yarriambiack - Model 1 - properties falling within 0-20% of actual sale price

Figure 5.6 shows the location of properties within the Yarriambiack LGA where the model was able to estimate the value of the property to within 20% of the actual sale price. These was no apparent geographic relationship between the location of the property and the accuracy of the models results.

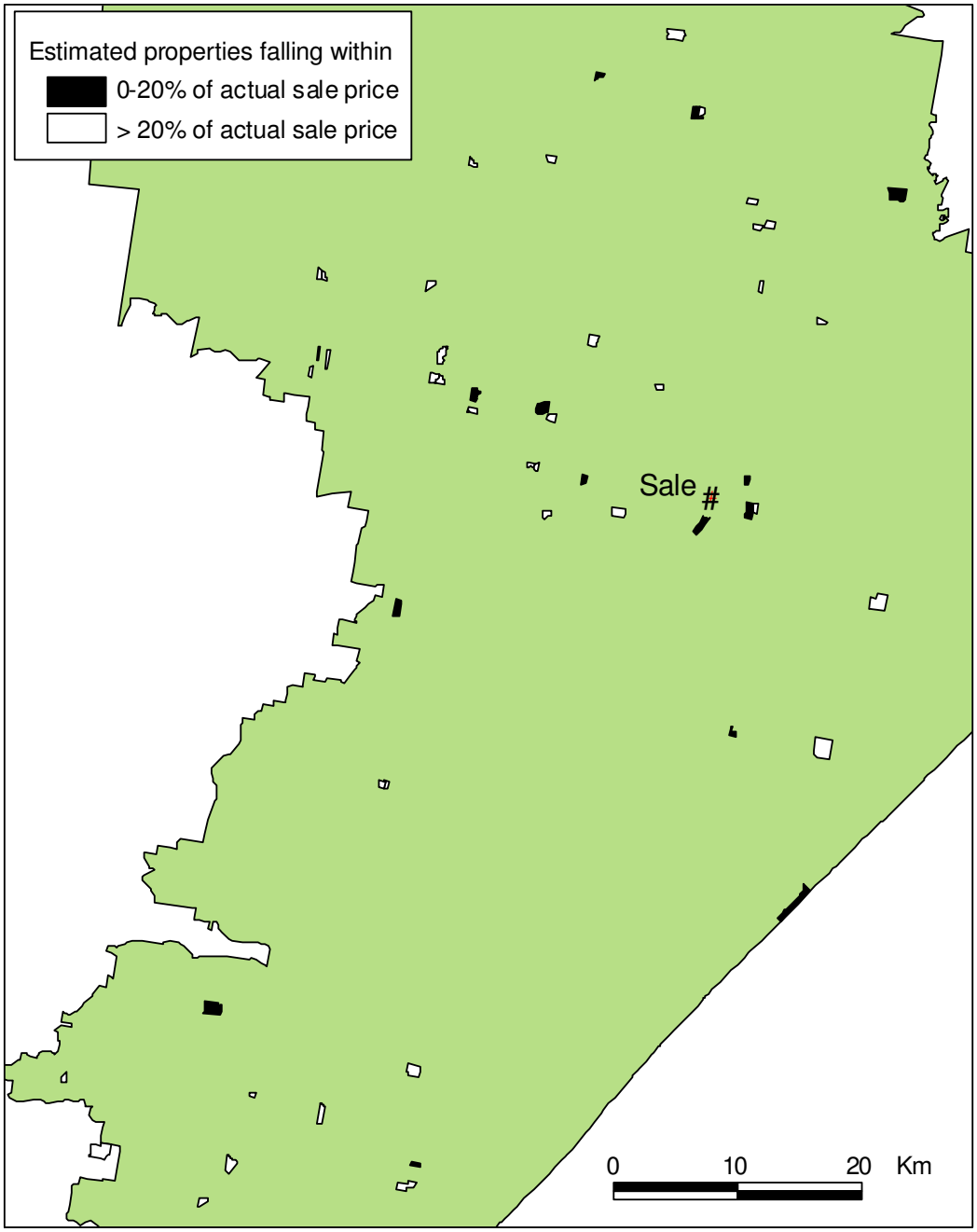

Figure 5.7 Wellington - Model 1 - properties falling within 0-20% of actual sale price

Figure 5.7 shows the results of the Model 1 property valuation estimates for the Wellington LGA. Properties nearby the town of Sale did not have a significantly higher likelihood of the model generating a more reliable estimate of sale price.

Rank regression techniques were then applied using '*ADJ_PRICE*' as the dependent variable and yielded very similar results to the price estimates determined in Model 1. Rank regression minimises a weighted sum of ranks of the residuals and is meant to lessen the impact of outliers in price. Rank regression is affected by outliers in the dependent variable (Cronan *et al*., 1986). Table 5.3 shows the regression coefficients of Model 2 and compares this with coefficients derived from least squares regression.

	Coefficient		**SE Coefficient**	
Predictor	Rank	Least-sq	Rank	Least-sq
Constant	120225	132957	17432	18454
PROP_AREA	361.22	337.3	99.19	105.0
WATERB_AREA	16.486	16.475	2.452	2.596
WATERCRS	15.466	18.070	4.915	5.203
TOWN_DIST	-1.0007	-1.0704	0.2928	0.3100
LUSE_3	203743	223080	37979	40206
LUSE_4	64145	74534	27134	28725

Table 5.3 Regression coefficients for Model 2 compared with least squares

Of the estimates obtained, only 10% fell within 10% of the sale price and 27% of valuation estimates fell within 20% of the sale price, which was slightly lower than using Model 1 determined by the 'best subsets' regression analysis.

5.5.2 Test 2: Dependent variable = '*Log_{10} ADJ_PRICE*'

A Logarithm to the base 10 was taken of the adjusted sale price variable (*ADJ_PRICE*). This dependent variable has been utilised as a dependent variable in research by Reynolds & Regalado (2002) and Theriault & Rosiers (2003).

The same variables from Model 1 and Model 2 were specified and a standard regression was performed using '*Log_{10} ADJ_PRICE*' as the dependent variable using the whole data set. Although the R^2 and Adjusted R^2 were slightly lower, (37.4% and 34.6%) the price estimates obtained were a slight improvement on those determined from Models 1 and 2. Within the 10% sale price range, 12% of the estimates fell within this range whilst 33% of the estimates fell within 20% of the actual sale price. The regression coefficients using '*Log_{10} ADJ_PRICE*' as the dependent variable is shown in Table 5.4. All variables were found to be significant at the 0.05% level for this model.

Predictor	Coefficient	SE Coefficient	T	P
Constant	5.04379	0.04857	103.84	0.000
PROP_AREA	0.0009977	0.0002764	3.61	0.000
WATERB_AREA	0.00002337	0.00000683	3.42	0.001
WATERCRS	0.00005019	0.00001370	3.66	0.000
TOWN_DIST	-0.00000347	0.00000082	-4.26	0.000
LUSE_3	0.4465	0.1058	4.22	0.000
LUSE_4	0.20227	0.07561	2.68	0.008

Table 5.4 Regression coefficients for Model 3

Rank regression was also applied using the dependent variable of '*Log_{10} ADJ_PRICE*' with the following regression coefficients determined in Table 5.5.

	Coefficient		SE Coefficient	
Predictor	Rank	Least-sq	Rank	Least-sq
Constant	5.05612	5.04379	0.04407	0.04857
PROP_AREA	0.0010647	0.0009977	0.0002508	0.0002764
WATERB_AREA	2.1655E-05	2.3373E-05	6.1990E-06	6.8321E-06
WATERCRS	0.00004923	0.00005019	0.00001243	0.00001370
TOWN_DIST	-3.552E-06	-3.475E-06	7.4037E-07	8.1598E-07
LUSE_3	0.43002	0.4465	0.09602	0.1058
LUSE_4	0.19154	0.20227	0.06860	0.07561

Table 5.5 Regression coefficients for Model 4 compared with least squares

Results were poor. For Model 4, of the total number of property estimates generated, only 12% fell within 10% of the actual sale price, and only an additional 22% were within 20% of the actual sale price.

5.5.3 Test 3: Dependent variable = '*ADJ_PRICEPHA*'

Utilising the dependent variable '*ADJ_PRICEPHA*', two further regressions were performed. Significant variables found were *TOWN_DIST, AREA1, AREA3, LUSE_12* and *LUSE_345*. The *LUSE_12* variable was determined by combining the categories *LUSE_1* and *LUSE_2*. Likewise *LUSE_345* was determined by combining the *LUSE_3, LUSE_4* and *LUSE_5* variables as detailed in Section 5.4.2. Testing the regression models resulted in an R^2 value of 34.4% and an adjusted R^2 of 32.5%. As shown in Table 5.6 the variable *LUSE_12* and *TOWN_DIST* were not significant whilst the other variables were.

Predictor	Coefficient	SE Coefficient	T	P
Constant	5868.0	461.9	12.70	0.000
TOWN_DIST	-0.023387	0.008745	-2.67	0.008
AREA1	-3221.0	611.3	-5.27	0.000
AREA3	-2756.6	727.4	-3.79	0.000
LUSE_12	-537.9	561.7	-0.96	0.340

Table 5.6 Regression coefficients for Model 5

Using this equation to estimate prices, 8% of the estimates fell within 10% of the actual sale price while 17% fell within 20% of the actual sale price. Further analysis of the data found that there was one outlier within the data that could be influencing the price values and was possibly a large farmhouse on small acreage which fetched above average prices. The outlier, mentioned in Section 5.3 was removed due to its large influence to the regression model and the lack of additional information detailing why the price per hectare was so high. The outlier was removed and Model 6 was generated. The regression coefficients for Model 6 are shown in Table 5.7.

Predictor	Coefficient	SE Coefficient	T	P
Constant	4973.6	487.0	10.21	0.000
TOWN_DIST	-0.023198	0.007786	-2.98	0.003
AREA1	-2694.8	557.4	-4.84	0.000
AREA3	-2252.6	679.4	-3.32	0.001
LUSE_12	-85.2	496.9	-0.17	0.864
LUSE_345	1513.9	548.9	2.76	0.007

Table 5.7 Regression coefficients for Model 6

The removal of the outlier and all properties belonging to the Northern Grampians LGA gave slightly better results. The R^2 was 44.5% and adjusted R^2 was 42.4% however upon testing with the data set, only 11% of estimates fell within 10% of the actual sale price, and 22% fell within 20%. Although the R^2 was slightly better than the other models, this was not overly apparent in the testing. Within this model, *LUSE_12* was not significant.

5.5.4 Test 4: Dependent variable = '*Log_{10} ADJ_PRICEPHA'*

Utilising the whole data set again with the variables *TOWN_DIST, AREA1, AREA3* and *LUSE_12* and with a dependent variable of the '*Log_{10} ADJ_PRICEPHA'* yielded Model 7 with coefficients shown below in Table 5.8 for this model.

Results obtained from Model 7 included an R^2 of 49.9%. Upon testing of the model, 13% of the estimates obtained fell within 10% of the actual sale price which was the second highest value determined in this category. In the 20% of the actual sale price

range category, 33% of estimates fell within this range performing as well as Model 3. Again the variable *LUSE_12* was not significant.

Predictor	Coefficient	SE Coefficient	T	P
Constant	3.78630	0.05510	68.71	0.000
TOWN_DIST	-0.00000613	0.00000104	-5.88	0.000
AREA1	-0.52297	0.07292	-7.17	0.000
AREA3	-0.41380	0.08678	-4.77	0.000
LUSE_12	0.00654	0.06701	0.10	0.922

Table 5.8 Regression coefficients for Model 7

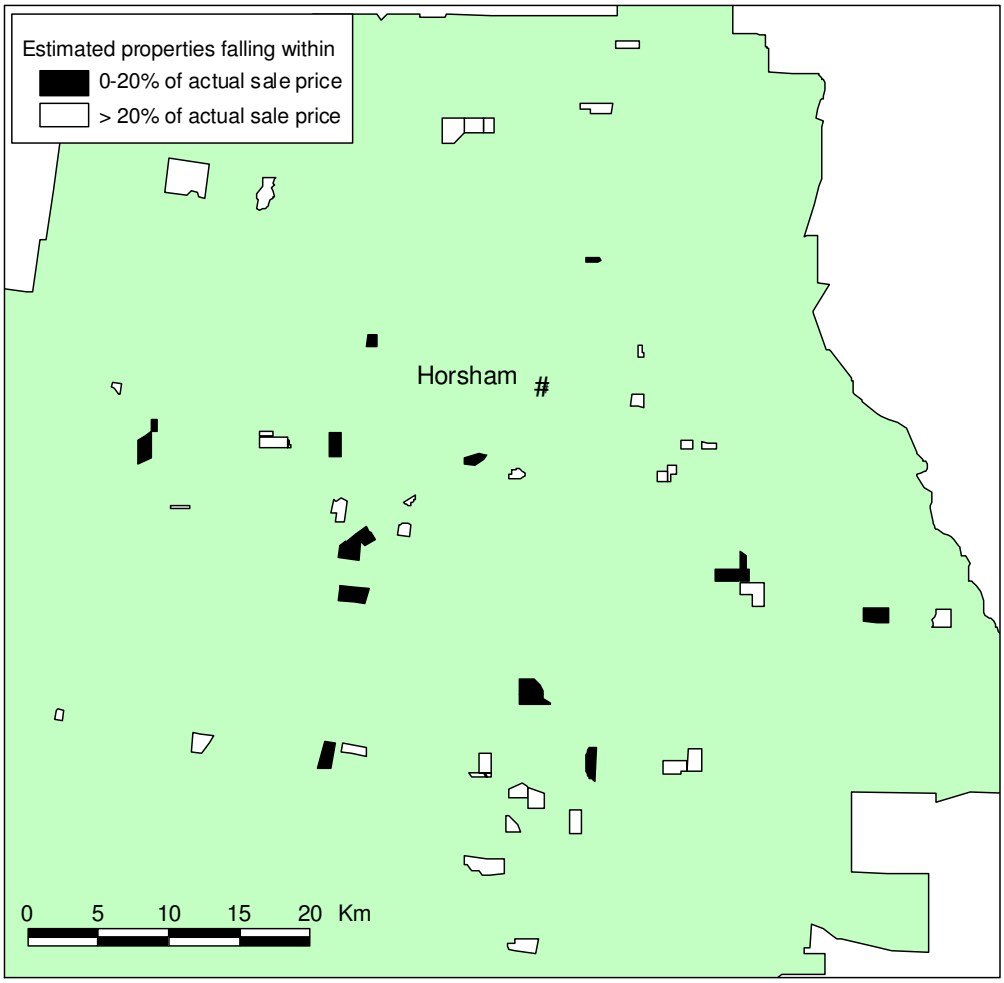

Figure 5.8 Horsham - Model 7 - properties falling within 0-20% of actual sale price

Figure 5.8 shows the properties within Horsham LGA with estimates that fell within 0-20% of the actual sale price. Towards the north of the LGA, numerous properties did not have accurate estimates (ie. greater than 20% of the sale price). Similarly, the value of two clusters of properties towards the south and also to the east of the town of Horsham could not be modelled reliably.

Figure 5.9 Northern Grampians - Model 7 - properties falling within 0-20% of actual sale price

Within the Northern Grampians LGA (Figure 5.9), all properties were estimated within the same accuracy range (>20%) with the exception of the most northern property within the LGA. Overall the properties remained the same between models with no clustering of properties that had estimates achieving higher accuracy within the LGA.

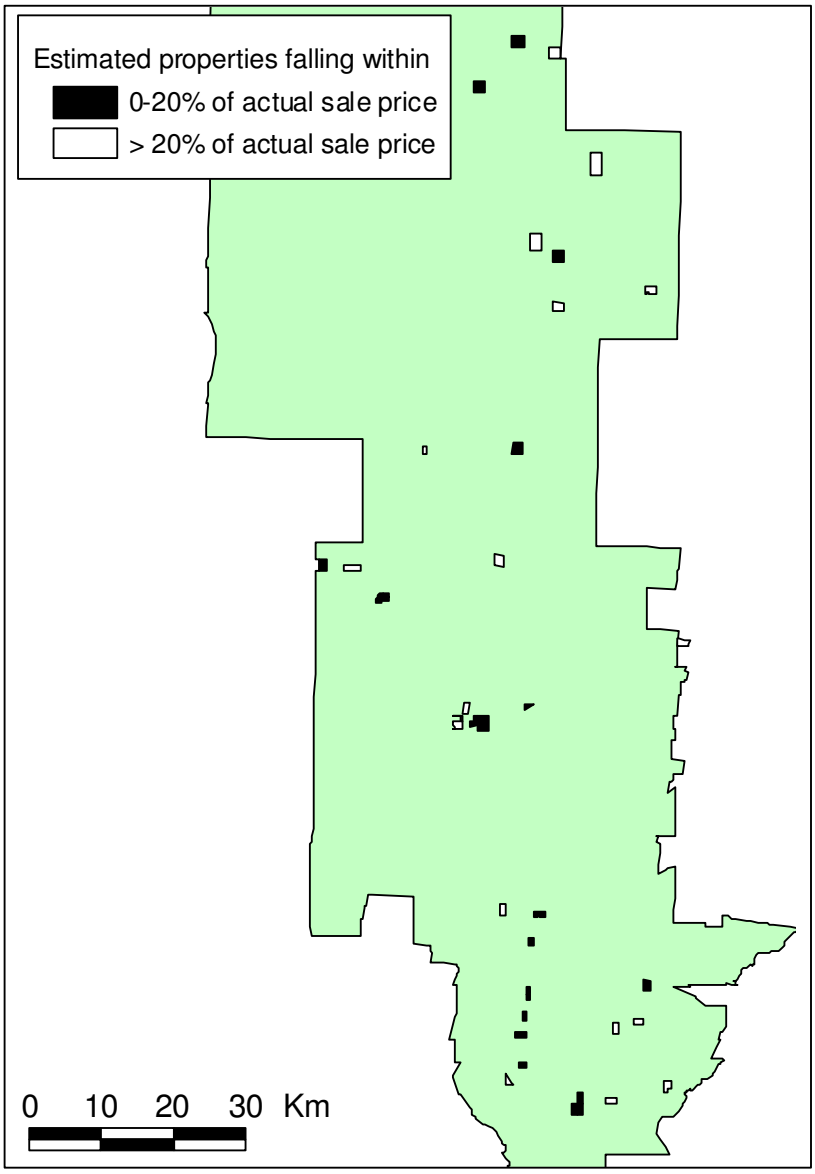

Figure 5.10 Yarriambiack - Model 7 - properties falling within 0-20% of actual sale price

Figure 5.10 depicts the results of properties situated within the Yarriambiack LGA. There are a few more properties with more accurate price estimates than in Model 1 - these are found towards the south of the LGA (to the east of Horsham).

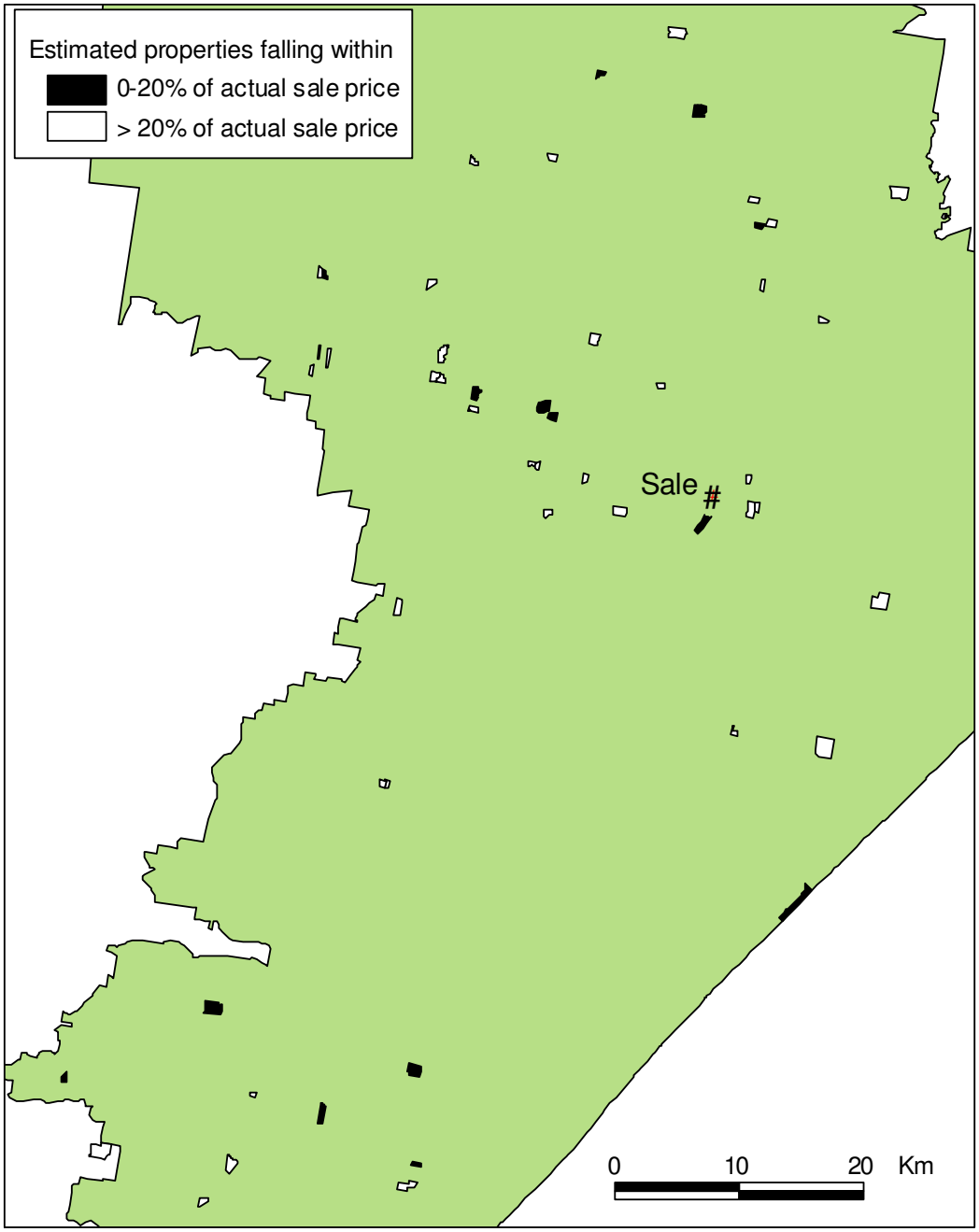

Figure 5.11 Wellington - Model 7 - properties falling within 0-20% of actual sale price

Figure 5.11 shows the properties for Model 7 within the Wellington LGA. More properties within the south of the LGA, in particular, had their valuation estimates improved in Model 7 compared to Model 1.

Model 8 differed to Model 7 in that the LGA of Northern Grampians was excluded (Section 5.4.1). In addition one outlier was removed from the data set (Section 5.3). The regression coefficients for Model 8 are shown in Table 5.9.

Predictor	Coefficient	SE Coefficient	T	P
Constant	3.71128	0.06559	56.58	0.000
TOWN_DIST	-0.00000597	0.00000105	-5.70	0.000
AREA1	-0.48976	0.07507	-6.52	0.000
AREA3	-0.39115	0.09151	-4.27	0.000
LUSE_12	0.05718	0.06693	0.85	0.394
LUSE_345	0.13525	0.07393	1.83	0.070

Table 5.9 Regression coefficients for Model 8

Results obtained from Model 8 included an R^2 of 54.9% and Adjusted R^2 of 53.2% which was the highest obtained in any of the tested models to date. However, only 11% of the estimates obtained fell within 10% of the actual sale price which was lower than what was determined in most of the other models. In the 20% of the actual sale price range category, 33% of estimates fell within this range, therefore performing as well as Model 3 and Model 7. The P value in Table 5.10 shows that the variables *LUSE_12* and *LUSE_345* were not significant.

Figure 5.12 shows the results from Model 8 for the Horsham LGA. Apart from the inclusion of a few properties that now have valuation estimates within 0-20% of the actual sale price, there are no significant differences between the properties and their results from Model 7. Although this model had higher R^2 values, only 33% of estimates

fell within the 0-20% range which was the same as Model 7 and Model 3. Model 1 had a higher proportion of properties falling in the 0-10% range.

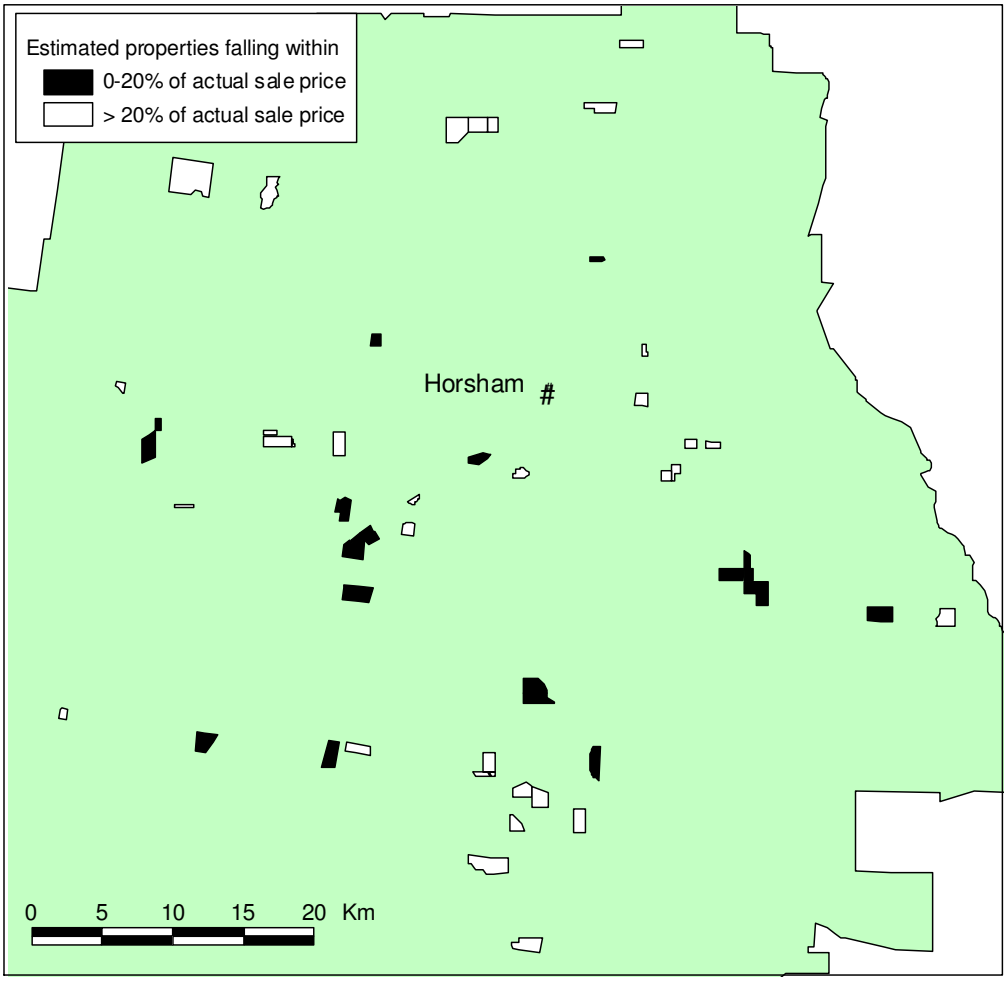

Figure 5.12 Horsham - Model 8 - properties falling within 0-20% of actual sale price

Models 7 and 8 (Figure 5.13) for Yarriambiack LGA are very similar in terms of the properties for which reasonable valuation estimates could be generated. As properties within the Northern Grampians LGA were excluded in this model, there is no mapping of results for this geographic region.

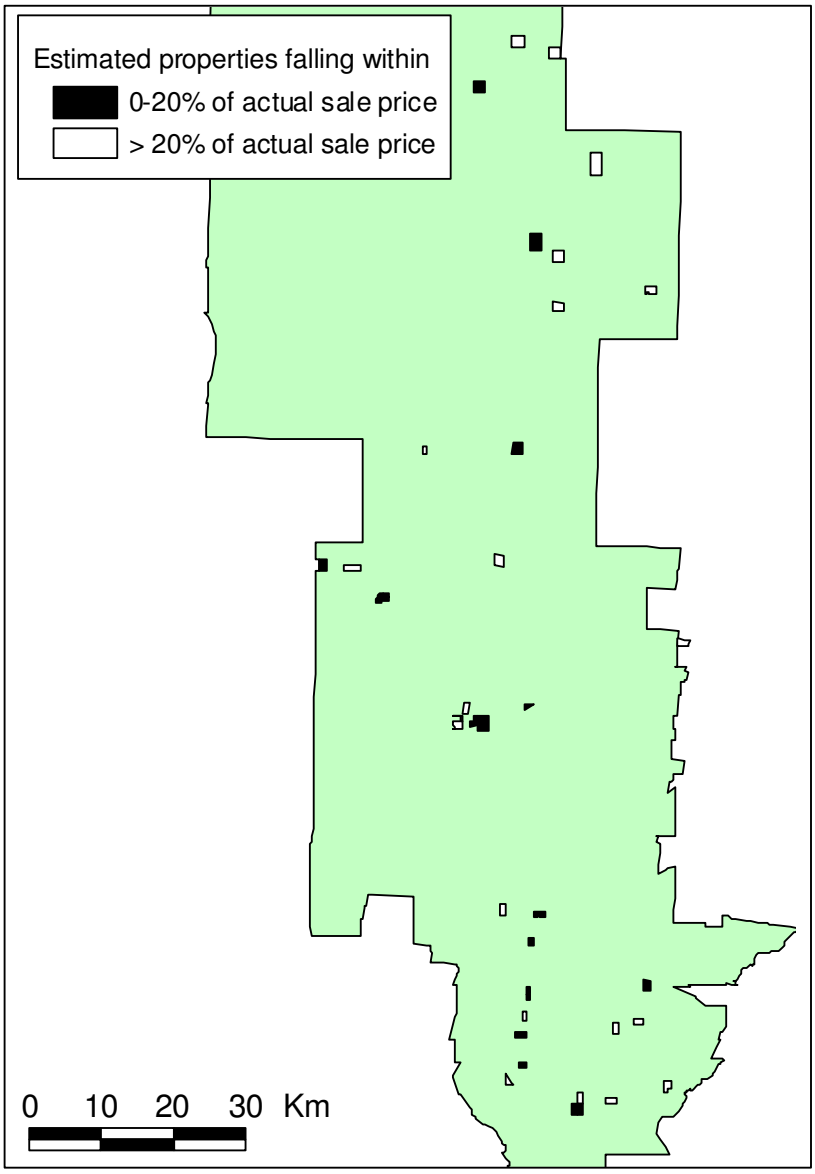

Figure 5.13 Yarriambiack - Model 8 - properties falling within 0-20% of actual sale price

Model 8 (Figure 5.14) for Wellington LGA reported the same as the Yarriambiack LGA in that there were few changes of property with respect to the accuracy of the value estimates.

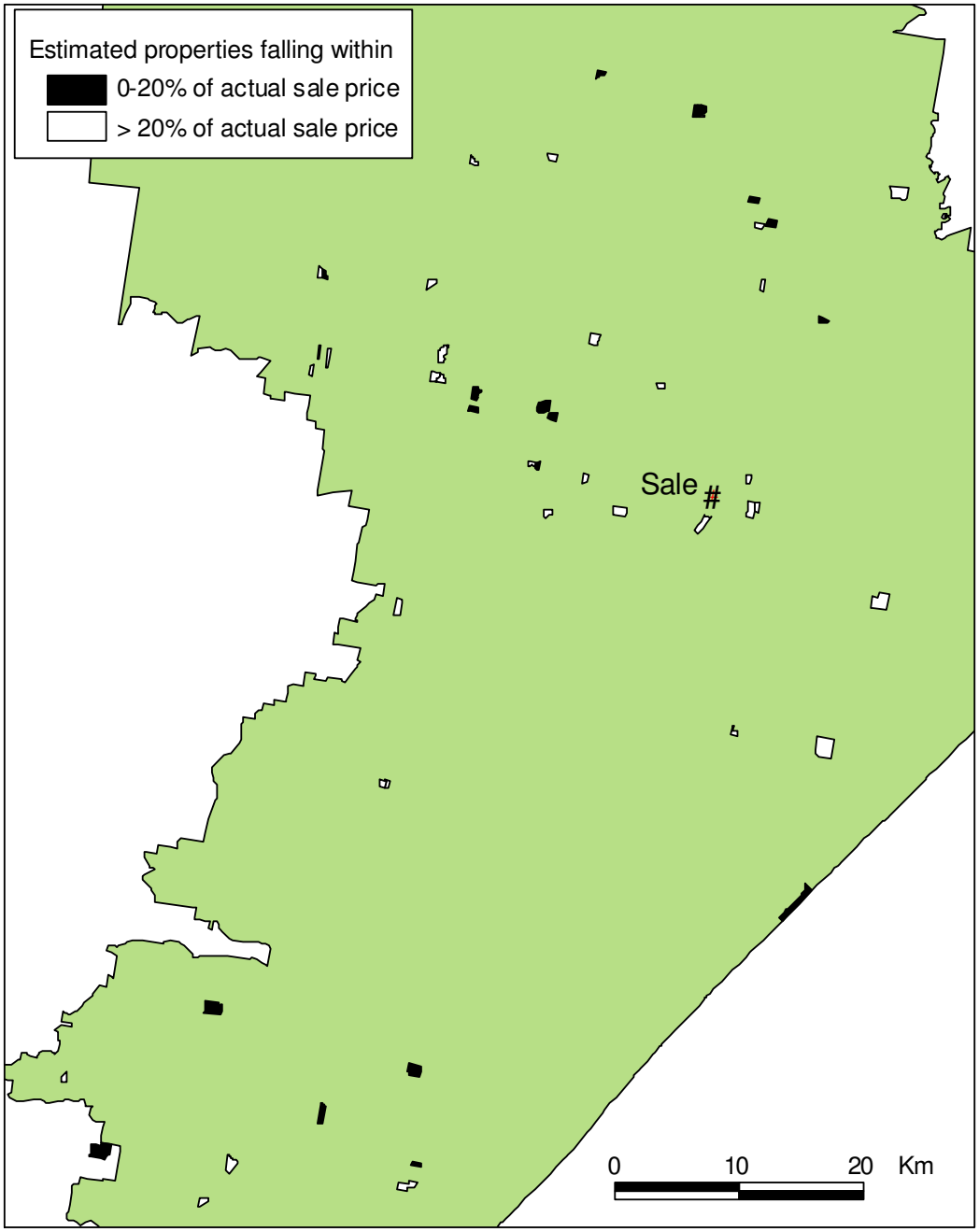

Figure 5.14 Wellington - Model 8 - properties falling within 0-20% of actual sale price

5.6 Discussion and Summary

The results of the eight models developed with the available digital data and using geographically defined sub-markets are shown in Table 5.10.

					Actual Price	
Model	**Regression Type**	**Dependent Variable**	**Data Set Used**	**R^2**	**10%**	**20%**
1	Best subsets	*ADJ_PRICE*	Whole	45.9	16%	26%
2	Rank	*ADJ_PRICE*	Whole	-	10%	27%
3	Regression	*Log_{10} ADJ_PRICE*	Whole	37.4	12%	33%
4	Rank	*Log_{10} ADJ_PRICE*	Whole	-	12%	22%
5	Regression	*ADJ_PRICEPHA*	Whole	34.4	8%	17%
6	Regression	*ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed	44.5	11%	22%
7	Regression	*Log_{10} ADJ_PRICEPHA*	Whole	49.9	13%	33%
8	Regression	*Log_{10} ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed	54.9	11%	33%

Table 5.10 Regression Models Summary – Geographically defined sub-markets with available digital data

Two dependent variables, sale price and sale price per hectare were tested with a total of eight models developed for rural valuation in Victoria during the initial phase of the project. Various implementations of these dependent variables were made and resulted in applying Logarithms to the base of 10 and using adjusted prices to account for sales occurring over a wide time span. The eight different models that were developed applied various regression techniques and eliminated outliers from the data set. The Northern Grampians LGA was also removed from the data set for Model 6 and 8. Of the independent variables used from Table 5.1, between four and six of these were utilised in each model. Although R^2 values were not as high as in other research, the results suggest that a larger number of variables may be required to

achieve a similar accuracy to that achieved by Gardner & Barrows, (1984) and Xu *et al.* (1993).

The level of price estimation in each of the eight models varied with an R^2 value between 34.4% and 54.9%. This indicates that between 34.4% and 54.9% of prices were explained by the developed models. These models achieved similar R^2 values to Miranowski & Hammes (1984) and Boisvert *et al.* (1997) however there have been other studies which have obtained a greater level of accuracy for rural automated valuation (Gardner & Barrows, 1984; Xu *et al.*, 1993; Reynolds & Regalado, 2002). Testing of each model using the developed regression equation helped to determine the closeness of the estimates to the actual sale price. Analysis of the percentage of estimated prices falling within 10% of the actual prices showed that between 8-16% of estimates fell within this range. Likewise, broadening the range to include estimates falling within 20% of the actual sale price increased the percentage to between 17-33%. Although there was an increase in the percentage of properties falling within the 20% range, these levels are still low in terms of the ability of each model to accurately estimate property values using the determined regression equation.

The following Chapter discusses the processes involved in using cluster analysis to determine sub-markets. Cluster analysis was undertaken using a 'two step' clustering procedure. The number of clusters were constrained to 3 and then to 4 clusters. This phase was undertaken to determine if the use of clustering techniques can assist in re-defining properties into different sub-markets to then enable more accurate automated models to be determined rather than using a-priori geographical techniques for sub-market definition.

Chapter 6 Cluster Analysis and Numerical Model Development

6.1 Introduction

Cluster analysis has been used in various disciplines to segment data into more meaningful classes or groups. The use of cluster analysis for rural property market segmentation has had minimal use. Cluster analysis has generally been confined to residential (Day, n.d.; Bourassa *et al.*, 1997; Goetzmann *et al.*, 1998; Watkins, 1998; Wilhelmsson, 2004), commercial (Dunse *et al.*, 2001), storage space (O'Roarty, 1997) and rental market applications (Smith & Kroll, 1989).

The approach taken during this phase of the research was to use the variables from Stage 1 of the research (Chapter 5) and apply a clustering algorithm to segment the properties into homogenous sub-markets. The method undertaken was to use a 'two step' approach which is available within SPSS, a statistical software package commonly used in environmental applications. The two step approach allows both continuous and categorical variables to be used within the development of the clustering model. The number of clusters was pre-set to be 3 and then set to be 4 clusters as part of an iterative process to determine the best cluster solution to use.

The regression models developed in Stage 1 (Chapter 5) were the result of different dependent variables being used, removal of outliers and the removal of one of the LGAs. Testing of the cluster process involved using the parameters which defined the initial regression models however this time segmenting the data based on which cluster group each property fell into and developing regression models for each cluster group.

The results of the models which use statistically derived sub-markets are presented in this Chapter.

6.2 Cluster Analysis Techniques

6.2.1 Variable Selection and Standardisation for Cluster Analysis

The variables used for cluster analysis as shown in Table 6.1 have been reduced from those specified within Table 5.1. The reduction of these variables was based on a few parameters. The geographical areas (*AREA1, AREA2, AREA3* and *AREA4*) were removed as there were to be no geographical constraints during the clustering process. The variables, *SEVERITY*, *NATURAL*, *FOX* and *LSIO* were eliminated as there was little variation within each of these variables. Likewise, *ZONE_CODE* was eliminated as all properties were classed as a rural zone thus there was no variation within this variable. The *ADJPRICEPHA* was removed as it was decided to only use the adjusted sale price as the adjusted price would have been correlated with the property area (*PROP_AREA*) variable. The land use variables (*LUSE_1, LUSE_2, LUSE_3, LUSE_4*, *LUSE_5* and *LUSE_6*) were amalgamated into one categorical variable for the analysis. Using 6 variables to depict land use via an indicator variable would have almost doubled the number of variables used in this analysis and led to a clustering algorithm which contained a high number of land use variables rather than a more even spread of different variable types. The clustering algorithm used (two step cluster analysis) is capable of working with mixed mode data.

Variable Name	Variable Description
ADJ_PRICE	Sale price of property (adjusted to year 2000 dollar value)
PROP_AREA	Size of property in hectares
WATERB_PT	Number of waterbody points
WATERB_AREA	Total area of waterbody (square metres)
WATERCRS	Length of watercourses (metres)
TOWN_DIST	Distance to nearest town (metres)
LANDUSE	Land use category, 1 to 6

Table 6.1 Descriptions of variables used for cluster analysis

Standardisation was undertaken for all continuous variables due to issues regarding the determination of similarity measures during the clustering process with data that has different scales (Everitt *et al*., 2001). A number of standardisation procedures can be used such as auto-scaling or dividing each variable by its standard range (Everitt *et al*., 2001). The standardisation procedure used was based on the inter-decile range standardisation which can overcome the outlier issues that have been associated with the range technique (National Statistics, 2005). All continuous variables were standardised using the equation below and can be seen in Appendix B.

$$\frac{X_i - X_{med}}{X_{90} - X_{10}}$$

Equation 2 Inter-decile Range Standardisation equation for continuous variables (National Statistics, 2005)

The equation compares each variable X_i to the median X_{med} and divides the value by the distance between the 90th percentile X_{90} and the 10th percentile X_{10} (National Statistics, 2005).

6.2.2 Two Step Cluster Analysis

A two step cluster analysis was used primarily as it can handle both continuous and categorical data and also provides the opportunity to pre-define or constrain the number of clusters to be developed during the modelling. The total number of clusters was constrained during this research. This was primarily due to the size of the property database and the number of variables that would be used in regression modelling and that too few properties would mean that regression analyses could not be performed. A 3 cluster solution was undertaken and then a 4 cluster solution. Results of the cluster analysis are shown below.

6.2.2.1 Three Cluster Solution

Constraining of the two step cluster process to 3 clusters led to the following cluster groups being specified as can be seen in Table 6.2.

Cluster	Number of Cases
1	14
2	84
3	54

Table 6.2 Two Step Cluster Analysis - Three cluster solution

Cluster 1 is not capable of being regressed due to the small sample and would ideally be suited to manual valuation. The frequencies of land use type for each of the 3 clusters is shown in Appendix C. Cluster 1 is comprised of land use type 1,3 and 6 (farm land without buildings, dairy and other rural property). Cluster 2 comprised land use type 1,3,4,5 and 6 (the same as Cluster 1 with the inclusion of beef and sheep land use types). Cluster 3 comprised land use 2 only (cereal). As can be seen in Figure 6.1 and Figure 6.2 the clusters have been segregated over a wide spatial area however there does appear to be smaller groups of Cluster 3 that have segregated spatially in the Wimmera region. The Wellington study area is only comprised of Cluster 1 and Cluster 2 (Figure 6.2) with all of the Cluster 3 property being segregated into the Wimmera study area.

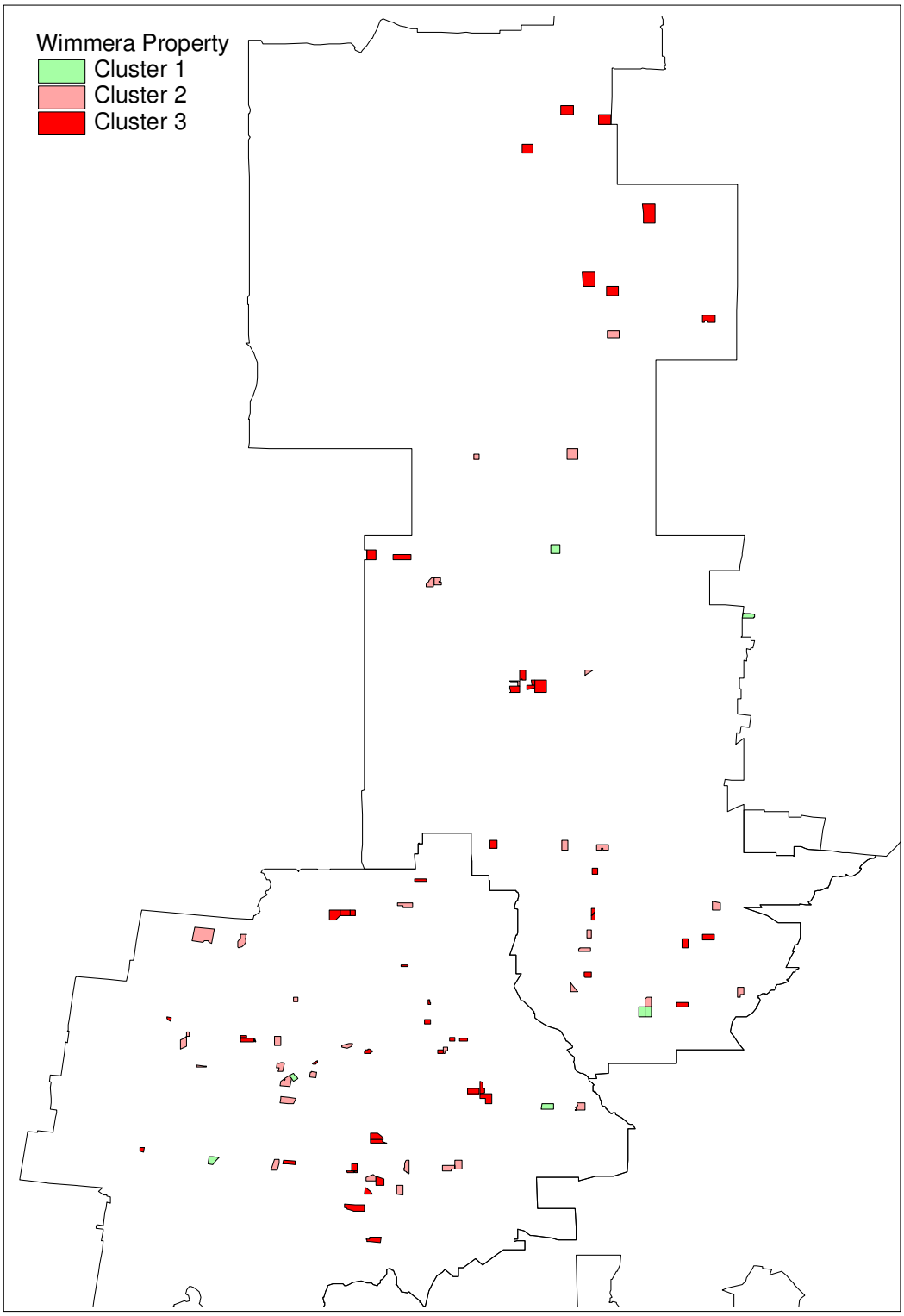

Figure 6.1 Cluster groups within the Wimmera study area

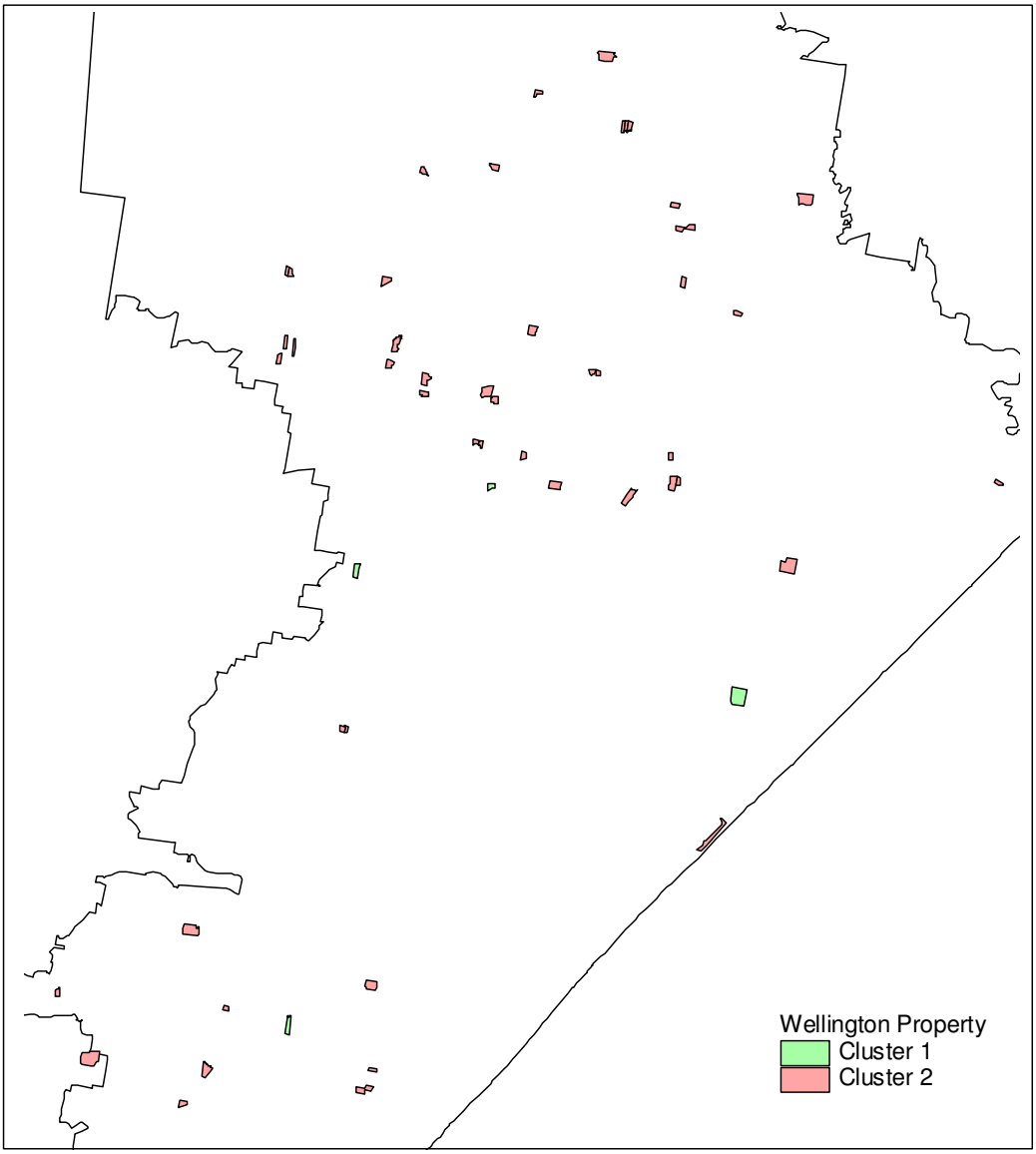

Figure 6.2 Cluster groups within the Wellington study area

Within the Cluster Profiles table of Appendix C, the mean and standard deviations are shown for each cluster for each variable. The difference between the means indicates the degree of variation between clusters. Ideally the means should have variation as this indicates different more distinct clusters have been developed. Cluster 1 typically is the most distinct cluster in comparison to the other two clusters. The mean is generally more distinct between Cluster 1 and Clusters 2 and 3, whilst Clusters 2 and 3

in many cases are very similar. Cluster 1 is most distinct for all variables with the exception of the price and the waterbody point variable. Analysis of Clusters 2 and 3 show that there is less variation amongst these. This is especially the case for the waterbody area and watercourse variables, however there still is a distinction, albeit small, between the two clusters for the other variables.

6.2.2.2 Four Cluster Solution

A four cluster solution was then applied using the same two step clustering algorithm and the same standardised variables as shown in Table 6.1. The developed clusters and the number of cases in each cluster is shown in Table 6.3

Cluster	Number of Cases
1	12
2	52
3	34
4	54

Table 6.3 Two Step Cluster Analysis - Four cluster solution

Again, Cluster 1 is not capable of being regressed due to its small sample size and would be a region that would require valuation using manual techniques. Cluster 1 is the same as Cluster 1 from the 3 cluster solution however two properties have been assigned to Cluster 3 in this cluster solution. Clusters 2 and 3 are a split of Cluster 2 from the 3 cluster solution with the extra two properties taken from Cluster 1 (3 cluster solution). Cluster 4 is the same as Cluster 3 from the 3 cluster solution.

Analysis of the frequencies in Appendix D, shows that Cluster 1 and Cluster 4 have the same land uses as Cluster 1 and Cluster 3 from the previous 3 cluster solution. Cluster 2 is now comprised solely of land use type 6 (other rural property) although some properties with this land use are also included in Cluster 1. Cluster 3 is comprised of land use types 1,3,4 and 5. The difference between the two cluster analyses undertaken is that the land use 6 category has been removed from one cluster to form a separate cluster in the 4 cluster solution.

The Cluster Profile in Appendix D shows the mean and standard deviations for each cluster for each variable. Cluster 1 and Cluster 4 are identical to Cluster 1 and Cluster 3 respectively from the 3 cluster solution. As such the variation is similar and has only altered marginally between the two stages. Property area, waterbody point and town distance tends to have the most variation between Cluster 1 and the other clusters. The sale price variable has two sets of similar clusters. Cluster 1 and 3 are quite similar and Cluster 2 and 4 have less distinction between the clusters for the sale price variable. Clusters 2,3 and 4 tend to vary about their similarity and as such there is often little distinction between the clusters. Waterbody point, watercourse, waterbody area and town distance are examples of where there is little variation between the clusters. This implies that a 3 cluster solution is possibly the best solution for these data. Increasing the number of clusters has decreased the variation between them.

6.3 Regression Analysis – Testing Phase 2 – Statistically determined sub-markets

6.3.1 Test 1: Dependent variable = '*ADJ_PRICE*'

Test 1 involved using all data with the dependent variable '*ADJ_PRICE* (adjusted price). Cluster 2 and then Cluster 3 were tested using Best Subsets regression. The regression coefficients obtained for both clusters is shown in Table 6.4.

Model 9	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	155567 *(0.000)*	35881 (4.34)	109697 *(0.000)*	26946 *(4.07)*
TOWN_DIST	-1.1331 *(0.122)*	0.7248 (-1.56)	-1.2384 *(0.011)*	0.4716 *(-2.63)*
PROP_AREA	494.2 *(0.120)*	314.5 (1.57)	751.8 *(0.000)*	186.4 *(4.03)*
WATERCRS	36.33 *(0.003)*	11.76 (3.09)	7.97 *(0.626)*	16.24 *(0.49)*
LUSE_3	185386 *(0.000)*	48789 (3.80)	-	-

Table 6.4 Regression coefficients for Model 9

For Cluster 2, the R^2 value was 29.8% and adjusted R^2 was 26.3%. Analysis of the price estimates which fell within 0-10% of the actual price was only 11%, whilst 20% of property fell within the 0-20% range. For Cluster 3 the R^2 value was 28.4% and adjusted R^2 was 24.1%, both being slightly lower than those determined for Cluster 2. Within the 0-10% range, there were 18.5% of estimates falling in this category, whilst 30% falling within the 0-20% range. Within Cluster 2, the watercourse length and land use type 3 variables were significant at the 0.05 level. Within Cluster 3, only the

property area and town distance variables were significant. The differences between the clusters were that land use type 3 was not specified in the Cluster 3 model due to there not being any property which fell into this land use type.

Rank regression was then applied using the same parameters as specified for Model 9. The regression results for Cluster 2 and Cluster 3 are shown in Table 6.5. No R^2 values are detailed for this method however analysis of the within percentage range of the two clusters shows that within Cluster 2, 9% of the estimates fell within 0-10% of the actual sale price, with 26% falling within 0-20% of the actual. For Cluster 3 this was considerably higher with 27% falling within the 0-10% range and 38% falling within the 0-20% range. The variables specified were also varied between the two clusters. Waterbody area (*WATERB_AREA*) was not included in the Cluster 2 model whilst water course length (*WATERCRS*) and land use type 3 were excluded from the Cluster 3 model.

Model 10	**Cluster 2**		**Cluster 3**	
	Coefficient *(SE Coefficient)*		**Coefficient *(SE Coefficient)***	
Predictor	Rank	Least-sq	Rank	Least-sq
Constant	130043 *(40804)*	155567 *(35881)*	86672 *(20896)*	109697 *(26946)*
TOWN_DIST	-0.7352 *(0.8243)*	-1.1331 *(0.7248)*	-0.9033 *(0.3657)*	-1.2384 *(0.4716)*
PROP_AREA	516.8 *(357.6)*	494.2 *(314.5)*	796.6 *(144.6)*	751.8 *(186.4)*
WATERCRS	35.38 *(13.38)*	36.33 *(11.76)*	-	-
WATERB_AREA	-	-	-7.97 *(12.60)*	7.97 *(16.24)*
LUSE_3	165689 *(55484)*	185386 *(48789)*	-	-

Table 6.5 Regression coefficients for Model 10 compared with least squares

6.3.2 Test 2: Dependent variable = '*Log_{10} ADJ_PRICE*'

Within Test 2, a different dependent variable was used compared to the previous test, this being a Logarithm to a base of 10. The same data set was in the regression analyses. Best subsets regression was used with the regression coefficients and are shown in Table 6.6.

Model 11	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	4.9770 *(0.000)*	0.1004 *(49.58)*	4.98998 *(62.63)*	0.07967 *(0.000)*
TOWN_DIST	-0.00000300 *(0.106)*	0.00000183 *(-1.63)*	-0.00000340 *(0.014)*	0.00000134 *(-2.54)*
PROP_AREA	0.0018455 *(0.023)*	0.0007951 *(2.32)*	0.0020806 *(0.000)*	0.0005528 *(3.76)*
WATERCRS	0.00008010 *(0.008)*	0.00002930 *(2.73)*	-	-
WATERB_AREA	0.00004092 *(0.085)*	0.00002346 *(1.74)*	-	-
LUSE_3	0.4239 *(0.001)*	0.1244 *(3.41)*	-	-
LUSE_4	0.17245 *(0.048)*	0.08563 *(2.01)*	-	-

Table 6.6 Regression coefficients for Model 11

Cluster 2 had an R^2 value of 33% and an adjusted R^2 of 27.8%. Only 14% of estimates fell within the 0-10% range whilst 23% fell within the 0-20% range. A larger number of variables were specified for Cluster 2 than Cluster 3. Both variables specified within the Cluster 3 model were significant. Within the Cluster 2 model, property area, watercourse length and both land use type variables were significant. The R^2 and adjusted R^2 values were slightly lower in the Cluster 3 model being 24.8% and 21.8% respectively. The percentage of properties falling within the 0-10% range was also

lower in Cluster 3 (18.5%) however the number of properties falling within the 0-20% range was higher at 38%.

Rank regression was then applied using the same parameters as the previous model to yield the regression coefficients shown in Table 6.7.

Model 12	**Cluster 2**		**Cluster 3**	
	Coefficient *(SE Coefficient)*		**Coefficient *(SE Coefficient)***	
Predictor	Rank	Least-sq	Rank	Least-sq
Constant	4.98418 *(0.09543)*	4.9770 *(0.1004)*	4.94525 *(0.07149)*	4.98998 *(0.07967)*
TOWN_DIST	-2.89823E-06 *(0.000001744)*	-2.99808E-06 *(0.000001835)*	-3.78272E-06 *(0.000001201)*	-3.39757E-06 *(0.000001339)*
PROP_AREA	0.0019447 *(0.0007559)*	0.0018455 *(0.0007951)*	0.0026610 *(0.0004960)*	0.0020806 *(0.0005528)*
WATERCRS	0.00007437 *(0.00002785)*	0.00008010 *(0.00002930)*	-	-
WATERB_AREA	0.00003821 *(0.00002230)*	0.00004092 *(0.00002346)*	-	-
LUSE_3	0.4224 *(0.1183)*	0.4239 *(0.1244)*	-	-
LUSE_4	0.17067 *(0.08141)*	0.17245 *(0.08563)*	-	-

Table 6.7 Regression coefficients for Model 12 compared with least squares

The models specified for Cluster 2 and 3 used the same property characteristics as those specified in Model 11 for the respective clusters. Results from Cluster 2 were slightly lower in terms of percentage estimates than those from Cluster 3. Within the 0-10% range, Cluster 2 had 11% of properties estimated within this range whilst 29% were estimated within 0-20%. Within Cluster 3, 13% of property were estimated within

0-10% and 36% within the 0-20% range. In comparison with the previous model, Model 11 had a higher proportion of properties falling in the 0-10% range than Model 12. Within the 0-20% range, this was the opposite for Cluster 2 where percentages were higher in Model 12 (29%) than for Model 11 (23%). The opposite occurred for Cluster 3 with Model 11 reporting a higher percentage of estimates within the 0-20% range at 38% than Model 12 at 36%.

6.3.3 Test 3: Dependent variable = '*ADJ_PRICEPHA*'

This test used adjusted price per hectare as the dependent variable. Within this model, land use categories were combined as mentioned in Section 5.4.2 to create the variables *LUSE_12* and *LUSE_345*. The regression coefficients for Model 13 are shown in Table 6.8.

Model 13	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	3854.1 *(0.000)*	840.7 *(4.58)*	1961.4 *(0.000)*	244.6 *(8.02)*
TOWN_DIST	-0.01686 *(0.414)*	0.02051 *(-0.82)*	-0.01373 *(0.008)*	0.004940 *(-2.78)*
WATERCRS	0.0975 *(0.750)*	0.3047 *(0.32)*	0.04145 *(0.617)*	0.08237 *(0.50)*
WATERB_AREA	0.0531 *(0.835)*	0.2547 *(0.21)*	-0.0097 *(0.956)*	0.1734 *(-0.06)*
LUSE_12	-314 *(0.931)*	3630 *(-0.09)*	-	-
LUSE_345	2521.3 *(0.005)*	877.6 *(2.87)*	-	-

Table 6.8 Regression coefficients for Model 13

R^2 values were considerably lower in Model 13 for Cluster 2 and 3 compared to previous models. For Cluster 2, the R^2 value was 12.4% and adjusted R^2 value 6.8%. For Cluster 3, the R^2 value was 15.9% and adjusted R^2 value was 10.9%, thus slightly

higher in Cluster 3 than Cluster 2. Analysis of the estimates falling within 0-10% of the actual price show that Cluster 2 had 14% and Cluster 3 had 18.5% of property falling within this range. Within the 0-20% range, a larger percentage of properties fell within this range, 21.4% for Cluster 2 and 35% for Cluster 3.

To create Model 14, the outlier identified in Section 5.3 and located within Cluster 2 was removed. To replicate the Test 3 performed in Section 5.5.3, property belonging to the Northern Grampians LGA were also removed. The combined land use categories 1 and 2 (*LUSE_12*) were removed from this equation as all values were identical. The resulting regression coefficients are shown in Table 6.9.

Model 14	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	4270.4 *(0.000)*	687.1 *(6.22)*	1934.2 *(0.000)*	250.8 *(7.71)*
TOWN_DIST	-0.04797 *(0.011)*	0.01842 *(-2.60)*	-0.013022 *(0.014)*	0.005127 *(-2.54)*
WATERCRS	0.2515 *(0.320)*	0.2512 *(1.00)*	0.04988 *(0.556)*	0.08422 *(0.59)*
WATERB_AREA	0.1278 *(0.538)*	0.2067 *(0.62)*	-0.0126 *(0.943)*	0.1747 *(-0.07)*
LUSE_345	2651.7 *(0.000)*	710.2 *(3.73)*	-	-

Table 6.9 Regression coefficients for Model 14

For Cluster 2, the R^2 value was 26% and adjusted R^2 value 22%. This was somewhat lower for Cluster 3 with a 14.5% R^2 value and an adjusted R^2 of 9.3%. Cluster 2 had 9.5% of properties estimated within 0-10% and 24% within 0-20% of the actual price. For Cluster 3, 22% of estimates were within 0-10% and 36% within 0-20% of the actual price. The significant variables in the Cluster 2 model were town distance and the land

use category 3,4,5 which was an amalgamation of 3 land use types. Town distance was the only significant variable within the Cluster 3 model. The land use variable was not included in the Cluster 3 model due to all values being the same.

6.3.4 Test 4: Dependent variable = '*Log_{10} ADJ_PRICEPHA*'

Logarithms to a base of 10 were taken of the adjusted sale price per hectare variable with all properties being used for this test. The regression coefficients for Model 15 are shown in Table 6.10.

Model 15	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	3.41783 *(0.000)*	0.09178 *(37.24)*	3.27581 *(0.000)*	0.07127 *(45.96)*
TOWN_DIST	-0.00000218 *(0.333)*	0.00000224 *(-0.97)*	-0.00000375 *(0.012)*	0.00000144 *(-2.61)*
WATERCRS	0.00002582 *(0.440)*	0.00003326 *(0.78)*	-0.00002565 *(0.290)*	0.00002400 *(-1.07)*
WATERB_AREA	0.00002362 *(0.398)*	0.00002780 *(0.85)*	-0.00001875 *(0.712)*	0.00005053 *(-0.37)*
LUSE_12	0.0936 *(0.814)*	0.3963 *(0.24)*	-	-
LUSE_345	0.27238 *(0.006)*	0.09581 *(2.84)*	-	-

Table 6.10 Regression coefficients for Model 15

The R^2 and adjusted values were quite low for this model. Cluster 2 had an R^2 value of 13.7% and an adjusted R^2 value of 8.2%. Cluster 3 had an R^2 value of 15.5% and an adjusted R^2 value of 10.4%. The percentage of properties falling within the 0-10% range was 10% for Cluster 2 and 16% for Cluster 3. The percentage falling within the 0-20% range was 20% for Cluster 2 and 39% for Cluster 3. For Cluster 3, land use 1

and 2 (*LUSE_12*) is constant and was removed from the equation. Again the amalgamation of land use types 3, 4 and 5 (*LUSE_345*) is zero and was also removed.

Model 16 was generated by removal of the outlier and the Northern Grampians LGA. The regression coefficients for Model 16 are shown in Table 6.11.

Model 16	**Cluster 2**		**Cluster 3**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**	**Coefficient** ***(P value)***	**SE Coefficient** **(T value)**
Constant	3.45534 *(0.000)*	0.08721 *(39.62)*	3.26340 *(0.000)*	0.07273 *(44.87)*
TOWN_DIST	-0.00000470 *(0.048)*	0.0000023 *(-2.01)*	-0.00000343 *(0.025)*	0.00000149 *(-2.30)*
WATERCRS	0.00003977 *(0.216)*	0.00003189 *(1.25)*	-0.00002180 *(0.376)*	0.00002442 *(-0.89)*
WATERB_AREA	0.00002887 *(0.275)*	0.00002624 *(1.10)*	-0.00002010 *(0.693)*	0.00005065 *(-0.40)*
LUSE_345	0.27766 *(0.003)*	0.09014 *(3.08)*	-	-

Table 6.11 Regression coefficients for Model 16

The R^2 values were low for this model, 20% for Cluster 2 and 12.7% for Cluster 3. Adjusted R^2 values were 15.6% (Cluster 2) and 7.4% (Cluster 3). The percentage of properties estimated within 0-10% of the actual price was 12% for Cluster 2 and 22% for Cluster 3. For Cluster 2, 24% of properties fell within the 0-20% range whilst this was 36% for Cluster 3. For Cluster 2, significant variables were town distance and the land use variable. Town distance was the only variable significant within the Cluster 3 model.

6.4 Discussion and Summary

Cluster analysis was undertaken to determine non a-priori sub-markets for rural property. The technique involved using both categorical variables (land use type) and continuous variables (adjusted price, property area, water body point, water body area, water course length and town distance). The clustering process was undertaken twice, firstly constraining the number of clusters to three and then constraining to four clusters. A three cluster solution was more appropriate due to the similarity between clusters in the four cluster process undertaken whilst in the three cluster solution there was more differentiation between cluster groups. From the three clusters determined, one of these, Cluster 1 was not used in the regression analyses due to its small sample size. As a result, only Clusters 2 and 3 were then tested.

Testing involved replication of the four tests undertaken during the numerical model development stage in Section 5.5. The models developed for Cluster 3 tended to yield a higher percentage of estimates falling in both the 0-10% and 0-20% range than the models developed for Cluster 2 as shown in Table 6.12. Between 9-27% of price estimates fell within 0-10% of the actual price, whilst 20-39% of estimates fell within 0-20% of the actual price. The R^2 values ranged between 12.4% and 33% thus were not considerably high. However, this phase of modelling showed that there was a consistently higher percentage of estimates falling within 0-20% of the actual price compared to those models defined geographically (Table 5.10).

				Cluster 2 (models depicted by 'a')			Cluster 3 (models depicted by 'b')		
Model	**Regression Type**	**Dependent Variable**	**Data Set Used**	**R^2**	**10%**	**20%**	**R^2**	**10%**	**20%**
9	Best subsets	*ADJ_PRICE*	Whole	29.8	11	20	28.4	18.5	30
10	Rank	*ADJ_PRICE*	Whole	-	9	26	-	27	38
11	Regression	*Log_{10} ADJ_PRICE*	Whole	33	14	23	24.8	18.5	38
12	Rank	*Log_{10} ADJ_PRICE*	Whole	-	11	29	-	13	36
13	Regression	*ADJ_PRICE PHA*	Whole	12.4	14	21.4	15.9	18.5	35
14	Regression	*ADJ_PRICE PHA*	Outlier, Northern Grampians LGA Removed	26	9.5	24	14.5	22	36
15	Regression	*Log_{10} ADJ_PRICE PHA*	Whole	13.7	10	20	15.5	16	39
16	Regression	*Log_{10} ADJ_PRICE PHA*	Outlier, Northern Grampians LGA Removed	20	12	24	12.7	22	36

Table 6.12 Regression Models Summary – Statistically defined sub-markets with available digital data

Additional restricted data were obtained and their use is reported in the following chapter. It should be noted that these data are not generally available to the public. The restricted data were tested using the procedures developed in Chapters 5 and 6. The first stage involved using sub-markets developed from the LGA to which each property belongs whilst the second stage involved defining sub-markets using cluster analysis. The results and processes undertaken using this additional data are presented in Chapter 7.

Chapter 7 Numerical Model Development with Restricted Digital Data

7.1 Introduction

The addition of a restricted data set into the property database led to the conceptual model being able to be fully tested. Restricted data are data held by Land Victoria and used for their biannual valuations in Victoria. Unfortunately, these data were not available at the commencement of this research, but became available at a late stage. The introduction of these variables into the property database allowed for additional testing to evaluate the predictions from the earlier modelling phases. This Chapter replicates the modelling methods outlined in the previous two chapters to enable a true comparison to be made between the different sub-market property valuation techniques.

7.2 Restricted Data Variables

Data were supplied in two files, one for the Wellington CMA and one for the Wimmera CMA in Microsoft Excel format. The new data variables linked to the existing property database with an identifier 'PROP_ENTID' that is a primary identifier in the Vicmap Property data sets. The restricted data set is a selection of rural property variables held by the Land Victoria and which is used for their rating valuations. The variables for which additional data were obtained are shown in Table 7.1.

Variable Name	Variable Description
CONS YEAR	Year of building construction
BCC	Building condition code; numerical values are 1 through 5 (1,2 poor condition, 3 average condition and 4,5 above average condition)
ALL IMPROVEMENTS	Presence of specific improvements; garage, carport, dairy, hayshed, shed, stables, yards, outbuilding, milk shed, verandah
ACCESS CODE	Property access; numerical values are 1 through 5 (1,2 poor access, 3 average access and 4,5 above average access)
WATER SUPPLY CODE	Water supply code; numerical values are 1 through 5 (1,2 poor supply, 3 average supply and 4,5 above average supply)
FENCING CONDITION CODE	Condition of fencing; numerical values are 1 through 5 (1,2 poor condition, 3 average condition and 4,5 above average condition)
PCC	Condition of pasture; numerical values are 1 through 5 (1,2 poor condition, 3 average condition and 4,5 above average condition)

Table 7.1 Additional variables supplied as the Restricted Data Set

Given the nature of the variables supplied, some of the variables required alteration to derive new property characteristics or required conversion to indicator variables. The variables which were derived from the 'restricted' data are shown in Table 7.2 and were used for the 'restricted data' statistical analysis phase of the project. The '*CONS YEAR*' variable was used to derive two variables depicting house presence, and presence of a new house. The 'BCC' variable was used to create an indicator variable to depict above average house condition. All property classed as 4 and above were classed as 'above average'. A variable was derived for farm building presence using the '*ALL IMPROVEMENTS*' variable. Farm buildings were deemed to be any type of shed such as milking sheds, shearing sheds, hay sheds, dairy and stables. Properties with these improvements were coded as having a farm building on the property. The WATER SUPPLY CODE variable was converted to an indicator variable called 'WATER' which indicated using 1, the presence of above average water supply.

Likewise the FENCING CONDITION CODE variable was used to create an indicator variable 'FENCE' indicating the presence or absence of average and above average fencing condition. The 'PASTURE' indicator variable was created using PCC and indicates a value of 1 for average and above average pasture condition. As can be seen in Appendix F and G, there were only three properties with an 'ACCESS CODE' rated as poor and thus this variable had little or no variation in the attributes representing this feature. The 'ACCESS CODE' variable was therefore not used in modelling.

Variable Name	Variable Description
HOUSE	Presence of house (1 for yes, 0 for no)
HOUSE_NEW	Presence of new house, less than 20 years old (1 for yes, 0 for no)
HOUSE_COND	Presence of above average building condition code (1 for yes, 0 for no)
FARMB	Presence of farm building (1 for yes, 0 for no)
WATER	Presence of above average water supply (1 for yes, 0 for no)
FENCE	Presence of average and above fencing (1 for yes, 0 for no)
PASTURE	Presence of average and above pasture (1 for yes, 0 for no)

Table 7.2 Restricted Data Variables used for statistical analysis

During integration of the restricted data, there were 14 properties in the original data set which could not be linked to the restricted data and thus there were no additional variables available for them. This was due to discrepancies in the original data obtained and that which was acquired at the later stage. The earlier data were more prone to data and address matching issues than the restricted data.

7.3 Regression Analysis – Restricted Data -Testing Phase 3 – Geographically determined sub-markets

The following statistical processing involved replicating the processes undertaken within Chapter 5. This involved applying the data used in Chapter 5 with the above additional data variables shown in Table 7.2.

7.3.1 Test 1: Dependent variable = '*ADJ_PRICE'*

This test used the adjusted price dependent variable and best subsets regression. The resultant coefficients and P values are shown in Table 7.3. All variables were significant at the 0.05% level except for the town distance variable. This model included the new variables relating to the farm building condition, water access and fencing condition. The R^2 value was 48.7% with the adjusted R^2 being 44.2%. The percentage of estimates falling within 10% of the actual sale price was 12% and 26% for those within 20% of the actual sale price. The model developed used a large number of independent variables to model price, however upon testing, accuracy was not high despite the higher R^2 value. Previously during this test, two models were developed, the second using rank regression techniques. This was not performed during this phase as the previous chapters had highlighted no significant difference between the two techniques to warrant its inclusion with the restricted data.

Predictor	Coefficient	SE Coefficient	T	P
Constant	309050	47886	6.45	0.000
AREA1	-179295	50942	-3.52	0.001
AREA2	-168629	66577	-2.53	0.013
AREA3	-168023	53415	-3.15	0.002
LUSE_3	170310	42411	4.02	0.000
WATERB_PT	14297	5850	2.44	0.016
WATERB_AREA	14.956	2.749	5.44	0.000
WATERCRS	17.955	5.753	3.12	0.002
TOWN_DIST	-0.8243	0.4293	-1.92	0.057
FARMB	-55465	23769	-2.33	0.021
WATER	70831	33154	2.14	0.035
FENCE	-134835	53729	-2.51	0.013

Table 7.3 Regression coefficients for Model 17

7.3.2 Test 2: Dependent variable = '*Log_{10} ADJ_PRICE*'

This test and all those following used best subsets regression however applied a different dependent variable during the regression analysis. In all of the models generated during this phase, a large number of independent variables were required to predict price. This test differed from the previous in terms of the dependent variable used; in this model it was a Logarithm of the adjusted sale price. The R^2 value was 36.9% and the adjusted R^2 was 32.5%. The P values shown in Table 7.4 show that all variables were significant except for the distance to town variable. Of the restricted variables used, only the farm building condition was specified in the regression model. The percentage of estimates falling within 10% of the actual price was 10% and 21% for those properties within 20%, which was comparable to Model 17.

Predictor	Coefficient	SE Coefficient	T	P
Constant	5.27481	0.06797	77.60	0.000
AREA1	-0.24815	0.06797	-3.65	0.000
AREA2	-0.2100	0.1418	-1.48	0.141
AREA3	-0.18339	0.09216	-1.99	0.049
LUSE_3	0.3323	0.1135	2.93	0.004
WATERB_PT	0.04595	0.01566	2.94	0.004
WATERB_AREA	0.00001873	0.00000732	2.56	0.012
WATERCRS	0.00004454	0.00001521	2.93	0.004
TOWN_DIST	-0.00000300	0.00000115	-2.62	0.010
FARMB	-0.12955	0.06147	-2.11	0.037

Table 7.4 Regression coefficients for Model 18

7.3.3 Test 3: Dependent variable = '*ADJ_PRICEPHA*'

The first part of test 3 used the dependent variable adjusted price per hectare, the whole data set and combined land use types. The coefficients are shown in Table 7.5 and show that all variables were significant except the town distance, pasture and farm building condition variables. The R^2 value was 49.9% with the adjusted R^2 value 45.6%. There were 15.9% of estimates falling within 10% and 31.8% within 20% of the actual sale price.

Predictor	Coefficient	SE Coefficient	T	P
Constant	7626	1013	7.53	0.000
PROP_AREA	-6.720	2.974	-2.26	0.026
AREA1	-5126	1052	-4.87	0.000
AREA2	-4299	1418	-3.03	0.003
AREA3	-5383	1136	-4.74	0.000
LUSE_345	2347.6	610.2	3.85	0.000
WATERB_AREA	0.13578	0.05892	2.30	0.023
TOWN_DIST	-0.004761	0.009541	-0.50	0.619
FARMB	-506.9	501.5	-1.01	0.314
WATER	2480.1	795.7	3.12	0.002
FENCE	-3401	1178	-2.89	0.005
PASTURE	-1519.8	833.2	-1.82	0.071

Table 7.5 Regression coefficients for Model 19

Model 20 used only a selection of the data set and not the whole data set as in the previous models. The outlier and the Northern Grampians LGA were removed prior to processing. This resulted in the following coefficients being specified as shown in Table 7.6. Water supply, fence condition and pasture were specified from the new restricted variables. They were not significant at the 0.05% significance level. The R^2 value decreased compared to the previous model and was 44.7% with the adjusted R^2 value 40%. There were 18.3% of estimates falling within 10% of the actual price and 29% that were within 20% of the actual sale price. This model produced poor estimates compared to the others developed.

Predictor	Coefficient	SE Coefficient	T	P
Constant	5154.6	924.9	5.57	0.000
AREA1	-2395.0	983.9	-2.43	0.016
AREA3	-2388.0	1046.0	-2.28	0.024
LUSE_345	2517.8	522.2	4.82	0.000
WATERB_PT	-233.0	107.2	-2.17	0.032
WATERB_AREA	0.13239	0.0499	2.65	0.009
WATERCRS	0.1029	0.1045	0.98	0.327
TOWN_DIST	-0.021254	0.0078	-2.69	0.008
FARMB	-425.7	424.8	-1.00	0.318
WATER	1725.4	599.3	2.88	0.005
FENCE	-1308.0o	1035.0	-1.26	0.209

Table 7.6 Regression coefficients for Model 20

7.3.4 Test 4: Dependent variable = '*Log_{10} ADJ_PRICEPHA*'

This test used the dependent variable adjusted price per hectare however a Logarithm to a base 10 was taken. In this first model developed, Model 21; all of the data was used. Results of this model are shown in Table 7.7. The R^2 value was 59.9% with the adjusted R^2 value 57.8%. AREA2 and the waterbody area variable were not significant in this model. There were 13.8% of estimates falling within 10% of the actual price and 31.9% falling within 20% of the actual sale price. Within this model there were no variables specified from the restricted data set.

Predictor	Coefficient	SE Coefficient	T	P
Constant	3.72474	0.06008	62.00	0.000
PROP_AREA	-0.0016219	0.0003620	-4.48	0.000
AREA1	-0.39182	0.06525	-6.01	0.000
AREA2	-0.1632	0.1333	-1.22	0.223
AREA3	-0.31875	0.08247	-3.87	0.000
WATERB_AREA	0.00001162	0.00000705	1.65	0.102
TOWN_DIST	-0.00000354	0.00000116	-3.06	0.003
LUSE_345	0.20462	0.07338	2.79	0.006

Table 7.7 Regression coefficients for Model 21

Model 22 used the same dependent variable as in the previous model however the outlier and the Northern Grampians LGA were removed. The R^2 value was 64.1% with an adjusted R^2 value of 62.1%. Water course length, farm building presence, *AREA3* and the combined land use category *LUSE_12* were not significant in this model. There were 3% of estimates falling within 10% and 12.9% of estimates within 20% of the actual sale price.

Predictor	Coefficient	SE Coefficient	T	P
Constant	3.70777	0.06568	56.45	0.000
AREA1	-0.42242	0.07090	-5.96	0.000
AREA3	-0.3410	0.1084	-3.15	0.002
LUSE_345	0.23612	0.07039	3.35	0.001
LUSE_12	0.07497	0.07089	1.06	0.292
WATERB_AREA	0.00001210	0.00000668	1.81	0.073
WATERCRS	0.00001134	0.00001451	0.78	0.436
TOWN_DIST	0.00000425	0.00000114	-3.73	0.000
FARMB	-0.02595	0.05616	-0.46	0.645
PROP_AREA	-0.0014731	0.0003544	-4.16	0.000

Table 7.8 Regression coefficients for Model 22

7.4 Regression Analysis – Restricted Data - Testing Phase 4 – Statistically determined sub-markets

Modelling using Cluster 2 and Cluster 3 did not generate models with an increased accuracy. This was due to the restricted variables having no variation especially in Cluster 3. It was decided to re-cluster the data to determine if the cluster areas changed and led to more accurate modelling.

7.4.1 Cluster Analysis constraining to 2 Clusters

A two step cluster analysis was undertaken using the land use categories. The number of clusters were chosen to be two for this stage of modelling to ascertain if this improved modelling results. It was decided to use only the land use types from the original data variables (Table 5.1) and amalgamate these with the restricted data variables (Table 7.2) for clustering during this stage as these were standard variables

in use by VBP. The number of properties assigned to each cluster is shown in Table 7.9. Appendix E shows the frequencies and cluster profiles for this two cluster process. The means for each variable for both Cluster 1 and Cluster 2 are more distinct compared to the three and four cluster solution previous undertaken. In terms of the land use type, each cluster supported all land use types. This is in contrast to the three cluster solution (Appendix C) where Cluster 1 comprised land uses 1,2 and 6 whilst Cluster 3 comprised only land use 2.

Cluster	Number of Cases
1	55
2	83

Table 7.9 Two Step Cluster Analysis - Two cluster solution

7.4.2 Test 1: Dependent variable = '*ADJ_PRICE*'

This phase of processing involved testing using the newly segregated data which was split into two clusters. Using the adjusted price as the dependent variable, the following regression coefficients were developed as shown in Table 7.10. Cluster 1 had an R^2 value of 47.3% and an adjusted R^2 value of 39.4%, Cluster 2 in comparison had an R^2 value of 35% and an adjusted R^2 of 30.8%. Model testing showed that only 14.5% of properties in Cluster 1 and 18% of property estimates in Cluster 2 fell within 10% of the actual sale price. Analysis of values estimating within 20% of the actual price improved somewhat with 29.1% for Cluster 1 and 27.7% for Cluster 2. The variables which were not significant in the Cluster 1 model were the land use type 1, the distance to town and the water supply variable. In Cluster 2, the land use classification 1 variable and town distance were not significant. Both cluster areas varied in the variables specified in their models.

Model 23	Cluster 1		Cluster 2	
Predictor	**Coefficient** *(P value)*	**SE Coefficient** *(T value)*	**Coefficient** *(P value)*	**SE Coefficient** *(T value)*
Constant	256199 *(0.000)*	55418 *(4.62)*	141609 *(0.000)*	20948 *(6.76)*
LUSE_1	-225381 *(0.329)*	228599 *(.99)*	89345 *(0.256)*	78119 *(1.14)*
LUSE_3	200950 *(0.001)*	56206 *(3.58)*	-	-
WATERB_AREA	11.240 *(0.004)*	3.739 *(3.01)*	22.466 *(0.001)*	6.342 *(3.54)*
WATERCRS	30.42 *(0.041)*	14.49 *(2.10)*	15.247 *(0.045)*	7.474 *(2.04)*
WATERB_PT	-	-	6262 *(0.280)*	5758 *(1.09)*
TOWN_DIST	-1.861 *(0.090)*	1.077 *(.73)*	-0.6366*(0.035)*	0.2968*(-2.14)*
FARMB	-102563 *(0.017)*	41383 *(2.48)*	-	-
WATER	55873 *(0.145)*	37699 *(1.48)*	-	-

Table 7.10 Regression coefficients for Model 23

7.4.3 Test 2: Dependent variable = 'Log_{10} *ADJ_PRICE'*

These two models differed from Model 23 in that a Logarithm to a base 10 was taken of the dependent variable. Model results are shown in Table 7.11. For Cluster 1 the R^2 value was 40.5% and adjusted R^2 value was 33.1%. For Cluster 2, the R^2 value was 25.5% and adjusted R^2 value was 20.7%. The percentage of estimates falling within 10% of the actual price was 14.5% for Cluster 1 and 10.8% for Cluster 2. This increased slightly when analysing the percentage of estimates falling within 20% of the actual price. Cluster 1 had 23.6% and Cluster 2 estimated 30.1% of property within 20% of the actual. The variables specified in the models were also varied. The variables not significant in Cluster 1 were the waterbody area and the fence condition variable, whilst in Cluster 2 variables not significant were the property area and the farm building presence.

Model 24	Cluster 1		Cluster 2	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***
Constant	5.3360 *(0.000)*	0.1105 *(48.27)*	5.09674 *(0.000)*	0.06660*(76.52)*
PROP_AREA	-	-	0.0007639*(0.094)*	0.0004501*(1.70)*
LUSE_3	0.3246 *(0.005)*	0.1101 *(2.95)*	-	-
WATERB_AREA	0.00001272 *(0.087)*	00000728 *(1.75)*	0.00005455*(0.004)*	0.00001854 *(2.94)*
WATERCRS	0.00004417 *(0.015)*	00001743*(2.53)*	0.00006462*(0.010)*	0.00002438 *(2.65)*
TOWN_DIST	0.00000601 *(0.007)*	00000215 *(-2.79)*	-0.00000381 *(0.002)*	0.00000118 *(-3.23)*
FARMB	-0.16829*(0.038)*	0.07882 *(-2.14)*	-0.2170 *(0.062)*	0.1148 *(-1.89)*
FENCE	0.14067 *(0.136)*	0.09271 *(1.52)*	-	-

Table 7.11 Regression coefficients for Model 24

7.4.4 Test 3: Dependent variable = '*ADJ_PRICEPHA*'

Model 25 was created by using the adjusted price dependent variable with the whole data set and incorporating combined land uses into the models as shown in Table 7.12. The R^2 value for Cluster 1 was 40.5% and adjusted R^2 value was 31%. The percentage of estimates falling within 10% of the actual was 10.9% and 27.3% for those within 20% of the actual sale price. The Cluster 2 model had an R^2 value of 12.1% and an adjusted R^2 value of 6.5%. The percentage of estimates falling within 10% of the actual price was 9.6% and for those within 20% of the actual, this was 21.6% thus lower than Cluster 1 results. Apart from the land use variables, all other variables specified in the cluster models were the same. Water course length and the

farm building presence were not significant in both cluster models. In addition, the waterbody point and the town distance variables were not significant for Cluster 2.

Model 25	**Cluster 1**		**Cluster 2**	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***
Constant	7449 (0.000)	1147 (6.50)	3715.5 *(0.000)*	613.5 *(6.06)*
LUSE_12	-	-	-1156.1 *(0.039)*	551.5 *(-2.10)*
LUSE_345	3122.4 (0.000)	822.0 (3.80)	-	-
WATERB_PT	-864.8 (0.007)	305.2 (-2.83)	-203.2 *(0.152)*	140.5 *(-1.45)*
WATERCRS	0.1282 (0.506)	0.1915 (0.67)	0.1375 *(0.484)*	0.1954 *(0.70)*
TOWN_DIST	-0.06659 (0.006)	0.02297 (-2.90)	-0.014791 *(0.064)*	0.007857 *(-1.88)*
FARMB	-1051.7 (0.207)	822.8 (-1.28)	-1017.6 *(0.234)*	848.8 *(-1.20)*

Table 7.12 Regression coefficients for Model 25

Model 26 involved removing the outlier and the Northern Grampians LGA. The outlier was present in Cluster 2 and therefore removed. There were also no properties from the Northern Grampians LGA within Cluster 1. Therefore, the Cluster 1 data set contained the same properties as that produced with the whole data set (Model 25) thus a model was not created for Cluster 1. The regression coefficients are shown in Table 7.13. Three of the five variables specified in this model were not significant. These were the combined land use variable and the farm building presence and water body area variables. The R^2 value for this model was 27.6% and the adjusted R^2 value was 22.4% The percentage of estimates falling within 10% of the actual was 15.7% whilst there were 21% of properties falling within 20% of the actual sale price.

Model 26	Cluster 2	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***
Constant	2786.1 *(0.000)*	291.0 *(9.5)*
LUSE_345	460.8 *(0.547)*	762.1 *(0.547)*
WATERB_PT	-160.21 *(0.034)*	74.12 *(0.034)*
WATERB_AREA	0.13962 *(0.073)*	0.07672 *(0.073)*
TOWN_DIST	-0.017492 *(0.000)*	0.004020 *(0.000)*
FARMB	-505.7 *(0.249)*	435.2 *(0.249)*

Table 7.13 Regression coefficients for Model 26

7.4.5 Test 4: Dependent variable = 'Log_{10} *ADJ_PRICEPHA'*

Test 4 used the adjusted price per hectare variable and applied a Logarithm to the base of 10. Model 27 used all the data set, whilst Model 28 removed the outlier and the Northern Grampians LGA thus replicating Test 3 with a different dependent variable. Regression coefficients for Model 27 are shown in Table 7.14. The R^2 value for Cluster 1 was 44.7% and the adjusted R^2 value was 37.4%. The percentage of estimates falling within 10% of the actual price was 14.5% and this increased to 29% when examining those properties which fell within 20% of the actual price. The house condition and the waterbody area variables were not significant in the Cluster 1 model.

For Cluster 2, the R^2 value was 30.1% and adjusted R^2 value was 26.5% thus lower than the Cluster 1 model. The percentage of estimates falling within 10% of the actual was 15.6% and was 33.7% for those falling within 20% of the actual. The waterbody area and the farm building presence variables were not significant in Cluster 2.

Model 27	Cluster 1		Cluster 2	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***
Constant	3.4961 *(0.000)*	0.1113 *(31.41)*	3.48695 *(0.000)*	0.08126 *(42.91)*
LUSE_345	0.27022 *(0.001)*	0.07484 *(3.61)*	-	-
WATERB_PT	-0.05730 *(0.034)*	0.02620 *(-2.19)*	-0.05407 *(0.010)*	0.02057 *(-2.63)*
WATERB_AREA	0.00000997 *(0.149)*	0.00000680 *(1.47)*	0.00003470 *(0.117)*	0.00002191 *(1.58)*
TOWN_DIST	-0.00000601 *(0.004)*	0.00000201 *(-2.98)*	-0.00000572 *(0.000)*	0.00000113 *(-5.09)*
HOUSE_COND	0.06036 *(0.440)*	0.07753 *(0.78)*	-	-
FENCE	0.25612 *(0.008)*	0.09246 *(2.77)*	-	-
FARMB	-	-	-0.2173 *(0.083)*	0.1236 *(-1.76)*

Table 7.14 Regression coefficients for Model 27

As was the case in test 3, no model could be developed for Cluster 1. The Cluster 2 model had poor estimation results. The R^2 value was 27.5% and the adjusted R^2 value was 23.4%. Table 7.15 shows the variables specified. Of the four variables specified in modelling, the farm building presence variable was not significant at the 0.05% level. The estimates produced had only 1.2% falling within 10% of the actual sale price and 2.5% for those falling within 20%.

Model 28	Cluster 2	
Predictor	**Coefficient** ***(P value)***	**SE Coefficient** ***(T value)***
Constant	5.11202 *(0.000)*	0.06490 *(78.77)*
WATERB_AREA	0.00005412 *(0.004)*	0.00001813 *(2.98)*
WATERCRS	0.00005412 *(0.001)*	0.00002382 *(3.35)*
TOWN_DIST	-0.00007989 *(0.004)*	0.00000102 *(-2.98)*
FARMB	-0.1693 *(0.120)*	0.1075 *(-1.57)*

Table 7.15 Regression coefficients for Model 28

7.5 Discussion and Summary

This Chapter presents the results of the numerical model development using restricted digital data. The data supplied typically encompassed building information and quality, farm building presence, fence and pasture condition, improvements and water supply information. Modelling involved replicating the processes undertaken during the geographical phase and the statistical phase (Chapter 5 and Chapter 6).

Cluster analysis was performed again, this time using a two cluster solution with the new variables and land use category variables. The results obtained showed an increase in R^2 values for the geographical phased models which used restricted data compared to the models produced which used the publicly available digital data. A summary of the model results is shown in Table 7.16 and Table 7.17.

					Actual Price	
Model	**Regression Type**	**Dependent Variable**	**Data Set Used**	**R^2**	**10%**	**20%**
17	Best subsets Regression	*ADJ_PRICE*	Whole	48.7	12	26
18	Regression	*Log_{10} ADJ_PRICE*	Whole	36.9	10	21
19	Regression	*ADJ_PRICEPHA*	Whole	49.9	15.9	31.8
20	Regression	*ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed	44.7	18.3	29
21	Regression	*Log_{10} ADJ_PRICEPHA*	Whole	59.9	13.8	31.9
22	Regression	*Log_{10} ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed	64.1	3	12.9

Table 7.16 Regression Models Summary – Geographically defined sub-markets with restricted digital data

				Cluster 1			Cluster 2		
Model	**Regression Type**	**Dependent Variable**	**Data Set Used**	**R^2**	**10%**	**20%**	**R^2**	**10%**	**20%**
23	Best subset Regression	*ADJ_PRICE*	Whole	47.3	14.5	29.1	35	18	27.7
24	Regression	*Log_{10} ADJ_PRICE*	Whole	40.5	14.5	23.6	25.5	10.8	30.1
25	Regression	*ADJ_PRICE PHA*	Whole	40.5	10.9	27.3	12.1	9.6	21.6
26	Regression	*ADJ_PRICE PHA*	Outlier, Grampians LGA Removed	-	-	-	27.6	15.7	21
27	Regression	*Log_{10} ADJ_PRICE PHA*	Whole	44.7	14.5	29	30.1	15.6	33.7
28	Regression	*Log_{10} ADJ_PRICE PHA*	Outlier, Grampians LGA Removed	-	-	-	27.5	1.2	2.5

Table 7.17 Regression Models Summary – Statistically defined sub-markets with restricted digital data

Analysis of the estimates falling within 20% of the actual sale price showed that in the geographically defined phase, values ranged from 12.9-31.9%. For Cluster 1, 23.6-29.1% of estimates fell within 20% of the actual price, whilst this was 2.5-33.7% for Cluster 3. There was a marginal difference between the two techniques. A discussion of these model results and the significance and use of the property characteristics in each model is presented in Chapter 8.

Chapter 8 Assessment of the Numeric Rural Property Valuation Models

8.1 Introduction

The 8 models developed and presented in Chapter 5 were created through regression analyses using a variety of dependent and independent variables. Alterations to the classification of variables and elimination of outliers and particular LGAs varied the size of the data set used for Models 1 through 8.

Models 1 through 8 encompassed the geographically determined sub-market phase of testing which assigned property into sub-markets based on the LGA to which each property belonged. Cluster analysis, a second phase in the research, was undertaken to develop statistically derived sub-markets (Chapter 6). Two clusters that were suitable for regression modelling were used to develop a further 16 models (ie. eight models for each of the two clusters). The results of the geographically determined sub-markets were then compared to those determined using cluster analysis to examine the affect of the two techniques for modelling.

Many variables from the conceptual model were not implemented in the initial phase of the regression analyses as suitable data sets were not available in digital form (Chapter 5 and Chapter 6). A restricted digital data set was acquired later in the study. The additional data was used to help ascertain the influence of selected variables on rural property prices in Victoria. Chapter 7 presented a re-analysis of the processing

from Chapter 5 and Chapter 6 with the inclusion of the restricted data to the property database.

This Chapter considers the data integration issues that arose whilst compiling the property database. It discusses the implications of using statistical sub-market derivation techniques and the affects that sub-market grouping has on the models developed over those developed using a-priori techniques. The Chapter discusses the results obtained after the inclusion of the restricted data and uses this information to assess the conceptual model developed earlier in the thesis.

8.2 Data Integration

Most of the difficulties encountered in the data integration stage of the research were due to incomplete data sets, addressing standards problems and difficulties in matching the cadastre to the PRISM sale price data. Population of data that already existed as a GIS data set required less integration than for those data that are supplied as spreadsheet files with no spatial reference which cannot be linked spatially.

Sale prices have been widely utilised in many hedonic regression studies as a dependent variable (Elad *et al.*, 1994, Lake *et al.*, 1998, Mahan *et al.*, 2000). The PRISM data set was the only data set available for use in Victoria which contains such information. Many problems were encountered in linking the sale price information (PRISM) to the cadastre as there were no identifiers to geocode the PRISM data set to the cadastre. This is highlighted by the fact that valuers in Victoria tend not to use the PRISM sale price data set because of the difficulties in linking these data to existing

data sets. Instead, they often rely on prior valuations obtained in each LGA over past years (Connie Spinoso 2003, pers. comm., 24 September). The Victorian government has recognised these limitations and has, more recently, put in place a measures to address them. For example, the Property Information Program (PIP) has been established in Victoria to improve the match rates between Local and State government property data sets, and governments are investing more in building the skills and expertise of staff responsible for the maintenance of such information.

The PRISM data set was also found to have many incomplete records, and since there was no identifier to geocode the data to a spatial data set, this proved a lengthy process to merge the PRISM data to the Cadastre from which the property database was derived. An automated approach linking the PRISM data to the property database would have been faster than the methods undertaken within this research, however the automated geocoding procedure undertaken proved unsuccessful as there were no records that were matched to the data set. The inability to utilise automated geocoding was mostly due to the storage of incomplete address attributes in the PRISM data.

The wealth of information in the PRISM data set in terms of the time span which the data set encompassed and the record of numerous sales transactions seems limited by the way in which the data are stored. Although registered valuers can access the data from a web site connection after registration, it only allows for query of properties based on specific search terms. The ability to use these data in a GIS and be able to relate it to geographic property boundaries such as a cadastre would enable greater use of this data set. Incorporation into the data set of a property identifier would allow

for more spatial representations of the sales prices to be better utilised and allow for the property information to be used for different purposes.

Other data integration issues that arose were mainly concerned with the format in which the data was supplied. When the map projection type is not specified during data supply, it can be difficult to determine the projection of the data, particularly if it is not in a format the user is familiar with. Where coordinate systems are relatively new and cannot yet be found under the customised projections of the GIS software, the parameters are required to be input manually. This can cause errors in the coordinates generated if a parameter is input incorrectly and often will not be found until an overlay of multiple themes is performed. The Vicmap data supplied for this research was in VICGRID and thus required some knowledge about the projection parameters and conversion from one projection system to another. In the latest version of ESRI (ArcGIS 9.0), the VICGRID projection parameters are now inbuilt but they were not at the time of this research. A novice user may not understand the caution required to specify and input false easting and northing coordinates or standard parallels and may not even be aware of the significance of these projection parameters if they are not customised within a GIS software package. A recommendation of this research would be either to include projection parameter files or projection files to enable transformation between projections so that the data supplier and user are fully aware what each parameter represents.

The intermediary step used in this research between ESRI ArcInfo and MapInfo Professional via ESRI ArcView, along with differences in data concepts between the GIS software packages led to topological errors being generated. The user should be

aware that a converted file may have topological errors or different topology than in the original file (MapInfo, 1997). This could also affect the accuracy of the data set if topological errors are prevalent. Reporting of data quality and processes undertaken during integration to derive a new data set is necessary to ensure subsequent users are aware of any limitations and errors which may have been generated in the data set.

Much time was spent converting the data sets into the one software format and projection, however until this was complete, there was no other means to verify if a particular data set was necessary and needed conversion. In some instances, once the data were converted and overlayed with the property database, there were no instances of overlap of the data sets and thus some data sets or portions of them were not required (LASTBURNT100 and Planning Scheme Database). If all data were supplied in the one format and projection it would have been easier to check each data set to see if there were instances of the features over any of required properties used in the property database and thus the integration time would have been greatly reduced.

This raises an issue regarding a more wider specification within the metadata of the data sets. If data concentration over an area can be more widely detailed in the metadata, then this may enable a more effective way to select appropriate data and ensure that data that are purchased are current and not redundant. Having more detailed descriptions within the metadata regarding the fields of the tables and the measures used within each field may assist in providing a better description of the attributes and the way in which they are measured within the data sets. A method to depict the extent of areal coverage of a data set may provide another technique to

examine if the data set may actually overlap the specified area that a user is interested in. An example of this would be to detail the percentage area of the geographic feature of the data set over the total geographic area of the whole data set so that a user could predict if the data they are interested in is likely to overlap the specific regions that they are concerned with. Thus, for polygon features (lakes) this may provide a better determination of the likelihood of spatial overlay between data sets however this technique would not work effectively for point data sets (centroid of a dam). In these cases it may be more effective to provide a count of the number of the point features to ascertain the extent of the coverage. These descriptors of the data would provide a greater insight into the appropriateness of the data set and is another possible research area that could be undertaken to improve reporting of metadata elements.

8.3 Performance of the Numeric Rural Property Valuation Models

Table 8.1 presents a summary of the models and their respective model numbers along with the dependent variables used, regression type and the extent of data used for each model.

Models				Regression Type	Dependent Variable	Data Set Used
1	9	17	23	Best subsets	*ADJ_PRICE*	Whole
2	10			Rank	*ADJ_PRICE*	Whole
3	11	18	24	Regression	Log_{10} *ADJ_PRICE*	Whole
4	12			Rank	Log_{10} *ADJ_PRICE*	Whole
5	13	19	25	Regression	*ADJ_PRICEPHA*	Whole

6	14	20	26	Regression	*ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed
7	15	21	27	Regression	Log_{10} *ADJ_PRICEPHA*	Whole
8	16	22	28	Regression	Log_{10} *ADJ_PRICEPHA*	Outlier removed, Northern Grampians LGA Removed

Table 8.1 Summary of Dependent Variables used for each Regression Model

The two tables following, summarize the variables specified within each model. The bold '**X**' within Table 8.2 depicts the variables which were not significant at the 0.05% level in the geographical phase with publicly available data. Variables denoted by an italic, underlined '*X*' had no significance level computed due to the technique used during regression not reporting this information. The remaining capital 'X ' represents the significance of the variables at 0.05% level. Table 8.4 represents the variables and their significance during the further processing with the restricted data. The 'a' and 'b' following the model numbers represent Cluster 2 and Cluster 3 for the one model number.

Property Characteristic	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
PROP_AREA	X	*X*	X	*X*				
AREA1					X	X	X	X
AREA3					X	X	X	**X**
LUSE_3	X	*X*	X	*X*				
LUSE_4	X	*X*	X	*X*				
LUSE_12					**X**	**X**	**X**	**X**
LUSE_345						X		**X**
WATERB_AREA	X	*X*	X	*X*				
WATERCRS	X	*X*	X	*X*				
TOWN_DIST	X	*X*	X	*X*	**X**	X	X	X

Table 8.2 Variables specified during publicly available data phase (Geographical Models)

Property Characteristic	**9a**	**9b**	**10a**	**10b**	**11a**	**11b**	**12a**	**12b**	**13a**	**13b**	**14a**	**14b**	**15a**	**15b**	**16a**	**16b**
PROP_AREA	**X**	X	*X*	*X*	X	X	*X*	*X*								
AREA1																
AREA3																
LUSE_3	X		*X*		X		*X*									
LUSE_4					X		*X*									
LUSE_12									**X**				**X**			
LUSE_345									X		X		X		X	
WATERB_AREA				*X*	**X**		*X*		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
WATERCRS	X	**X**	*X*		X		*X*		**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**
TOWN_DIST	**X**	X	*X*	*X*	**X**	X	*X*	*X*	**X**	X	X	X	**X**	X	X	X

Table 8.3 Variables specified during the publicly available data phase (Statistical Models)

	Geographical Models						Statistical Cluster Models									
Property Characteristics	**17**	**18**	**19**	**20**	**21**	**22**	**23a**	**23b**	**24a**	**24b**	**25a**	**25b**	**26b**	**27a**	**27b**	**28b**
PROP_AREA			x		x	x				**X**						
AREA1	x	x	x	x	x	x										
AREA2	x	**X**	x		**X**											
AREA3	x	x	x	x	x	x										
LUSE_1							**X**	**X**								
LUSE_3	x	x					x		x							
LUSE_12						**X**						x				
LUSE_345			x	x	x	x					x		**X**	x		
WATERB_PT	x	x		**X**				**X**			x	**X**	x	x	x	
WATERB_AREA	x	x	x	x	**X**	**X**	x	x	**X**	x			**X**	**X**	**X**	x
WATERCRS	x	x	x	**X**		**X**	x	x	x	x	**X**	**X**				x
TOWN_DIST	**X**	x	**X**	**X**	x	x	**X**	x	x	x	x	**X**	x	x	x	x
FARMB	x	**X**	**X**	**X**		**X**	x		x	**X**	**X**	**X**	**X**		**X**	**X**
WATER	x		x	x			**X**									
FENCE	x		x	x					**X**					x		
PASTURE			**X**													
HOUSE_COND														**X**		

Table 8.4 Variables specified during restricted data phase (X denotes significant at 0.05%)

8.3.1 Results using Publicly Available Digital Data

Table 8.5 presents the R^2 and percentage of properties estimated within 10% and within 20% of the actual sale price. The table summarises the results of the models developed using the available digital data phase and provides a comparison of the two techniques used to determine sub-markets (geographical and statistical).

GEOGRAPHICALLY DEFINED SUB-MARKETS					STATISTICALLY DEFINED SUB-MARKETS					
					CLUSTER 2			CLUSTER 3		
Model	**R^2**	**0-10%**	**0-20%**	**Model**	**R^2**	**0-10%**	**0-20%**	**R^2**	**0-10%**	**0-20%**
1	45.9%	16%	26%	9	29.8%	11%	20%	28.4%	18.5%	30%
2	-	10%	27%	10	-	9%	26%	-	27%	38%
3	37.4%	12%	33%	11	33%	14%	23%	24.8%	18.5%	38%
4	-	12%	22%	12	-	11%	29%	-	13%	36%
5	34.4%	8%	17%	13	12.4%	14%	21.4%	15.9%	18.5%	35%
6	44.5%	11%	22%	14	26%	9.5%	24%	14.5%	22%	36%
7	49.9%	13%	33%	15	13.7%	10%	20%	15.5%	16%	39%
8	54.9%	11%	33%	16	20%	12%	24%	12.7%	22%	36%

Table 8.5 Model Results using Available Digital Data

In addition to the above results, a number of summary statistics were calculated based on Valuation Best Practice Standards (Valuation Best Practice, 2006) and the Standard on Ratio Studies (IAAO, 1999). More detailed examples of calculations can be obtained from either of the above publications. A sales ratio was calculated for each property for each model and also a ratio based on the 'estimated sale price' divided by

the ‘actual sale price’. The median sale ratio is the mid-point of these ratios when they are ranked in order of magnitude.

For the models developed using geographical constraints (the LGA to which each property belongs), the median sale ratios range from 0.91 to 1.27 (Figure 8.1). Using the models defined using cluster analysis, models 9a to 16a (Cluster 2) had median sale ratios ranging from 0.24 to 1.22 whilst models 9b to 16b (Cluster 3) had median sale ratios between 0.90 and 1.08 (Figure 8.2). The tolerances set by Valuation Best Practice Standards (Valuation Best Practice, 2006) are between 0.9 to 1.0. A median sale ratio value closer to 1.0 indicates the estimates are similar to the actual sale prices. Values over 1.0 indicate the property estimates are higher (over-valued) than the actual sale prices. Analysis of the median sales ratios for the two cluster groupings and the geographically defined sub-markets show that the models developed using Cluster 3 (statistically defined sub-market) had a closer range of ratios to those set by VBP compared to those developed using Cluster 2 or those developed using the geographical constraints.

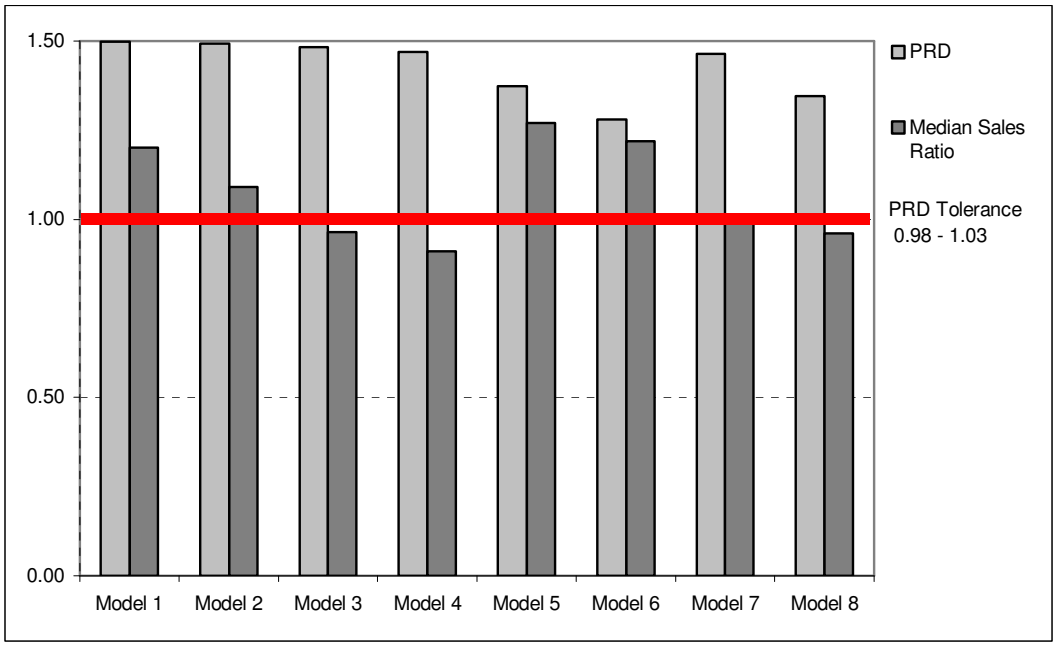

Figure 8.1 Summary Statistics – Publicly Available Data - Geographically defined sub-markets - Median Sales Ratio with PRD tolerances – Models 1-8

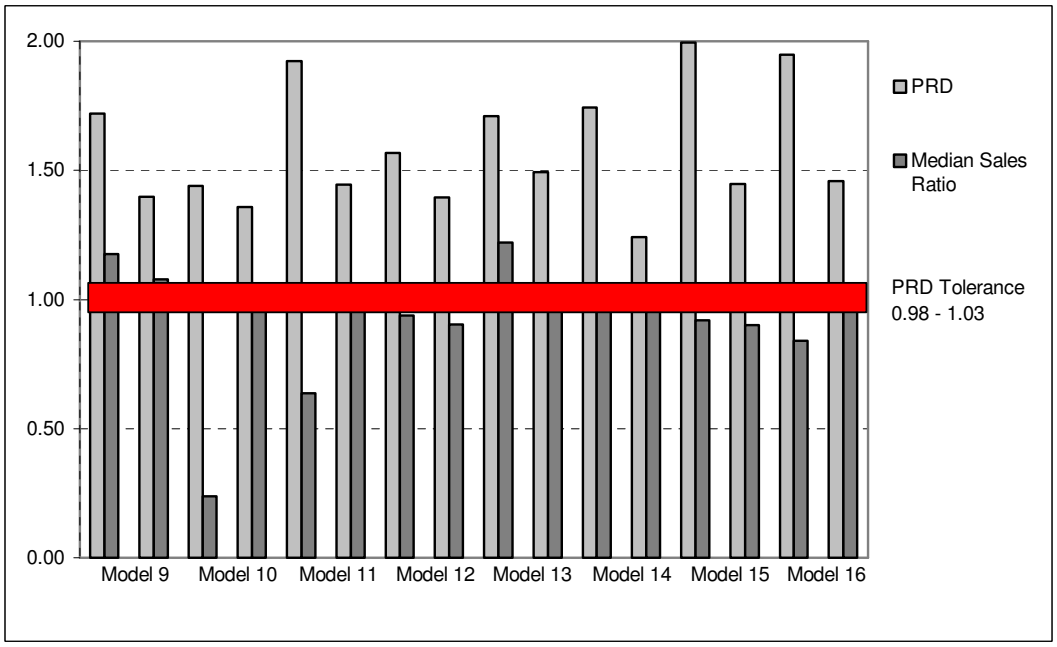

Figure 8.2 Summary Statistics – Publicly Available Data - Statistically defined sub-markets - Median Sales Ratio with PRD tolerances - Models 9-16

The COD and COV are measures of variability of the sale ratios. The COD measures the average percentage deviation of the sale ratio from the median sale ratio and is classed as the most useful of the two measures (IAAO, 1999). The COD tolerances were all above the '<15' set by VBP and range between 19-39 for the geographically defined models (Models 1-8) as shown in Figure 8.3. COD values were between 41-66 for Cluster 2 (Models 9a-16a) and between 29-43 for Cluster 3 (Models 9b-16b). The high COD values obtained in this research (Figure 8.3 and Figure 8.4) indicate that there are outliers or properties which are not exhibiting the same behaviour as others in terms of their estimates. The COV value which indicates normality, showed that these values were extremely high given the presence of the outlier ratios computed. COV values ranged from 96-137 for the geographically defined sub-markets (Models 1-8), between 100-138 for Cluster 2 (Models 9a-16a) and between 95-164 for Cluster 3 (Models 9b-16b).

The PRD values are also higher than the 0.98-1.03 tolerance set by VBP and this is clearly shown in Figure 8.1 and Figure 8.2. For the geographically defined sub-market models, the PRD ranged from 1.28 to 1.50. For Cluster 2, the PRD ranged from 1.13 and 1.99 whilst Cluster 3 achieved similar results to the geographically defined sub-markets with the PRD ranging from 1.24 to 1.49.

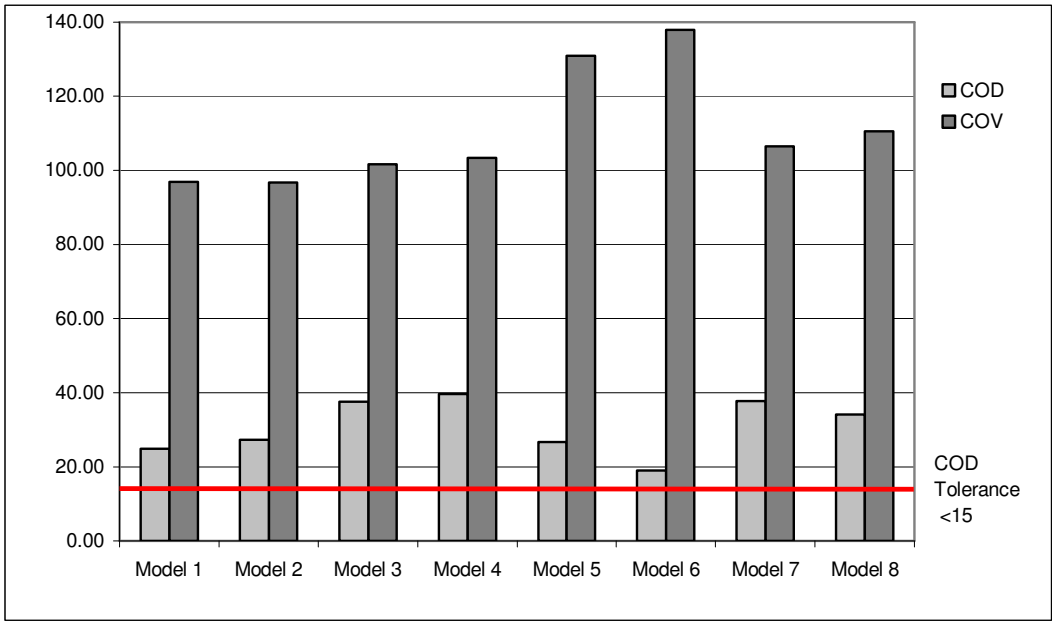

Figure 8.3 Summary Statistics – Publicly Available Data - Geographically defined sub-markets – COD and COV – Models 1-8

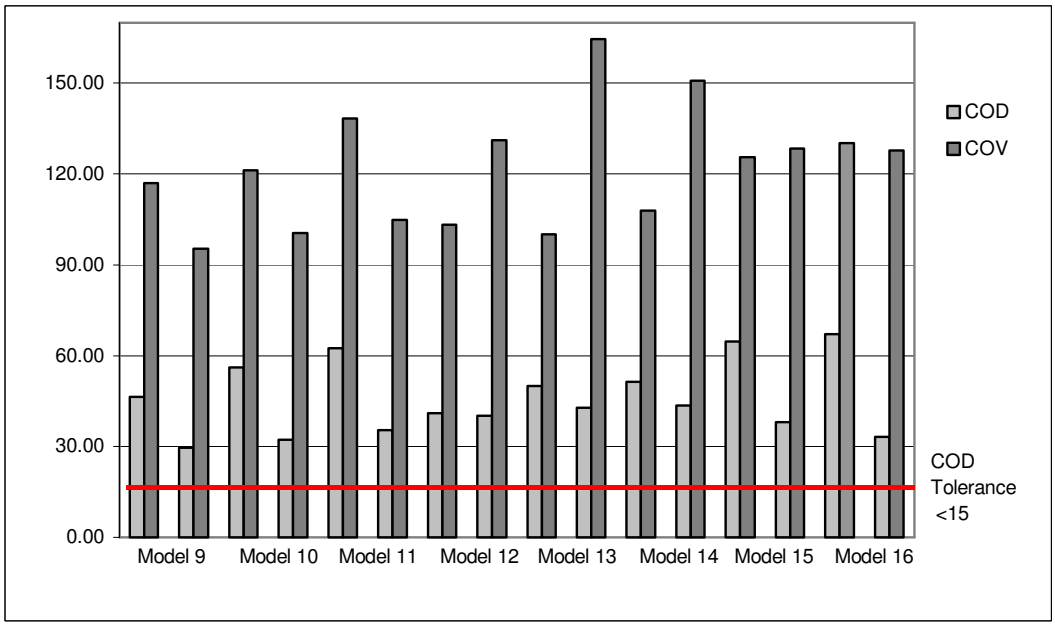

Figure 8.4 Summary Statistics – Publicly Available Data - Statistically defined sub-markets – COD and COV – Models 9-16

After producing price estimates for each of the models using the regression equations, Cluster 3 tended to have a greater percentage of properties which estimated at a closer range to the actual sale prices (Table 8.6). This occurred in both the 0-10% and 0-20% price ranges. An example is that Cluster 3 had between 30-39% of properties estimating within 20% of the actual sale price whilst this was 20-29% for Cluster 2 and 17-33% for the geographically defined sub-markets. Using a greater accuracy level (estimates within 10% of the actual price), Cluster 3 still had the highest percentage of estimates falling within 10% of the actual at 13-27%. For Cluster 2 between 9-14% of properties fell within 10% of the actual price and 8-16% of properties fell within 10% of the actual price for the geographically defined sub-markets (Table 8.6).

	Geographically defined sub-markets	**Statistically defined sub-markets**	
		Cluster 2	**Cluster 3**
Percentage of estimates within 10% of actual sale price	8-16%	9-14%	13-27%
Percentage of estimates within 20% of actual sale price	17-33%	20-29%	30-39%
COD	19.02-39.75	41.03-66.99	29.57-43.57
COV	96.80-137.95	100.12-138.41	95.31-164.65
PRD	1.28-1.50	1.13-1.99	1.24-1.49

Table 8.6 Summary Ranges of Models – Publicly Available Digital Data - COD, COV, PRD and Percentage of estimates within 10 and within 20% of actual price

8.3.2 Results using Restricted Digital Data

A summary of the regression R^2 values and the percentage of estimates falling within 10% and within 20% of the actual price are shown in Table 8.7. Analysis of the R^2 values showed a higher value for those models developed using geographical sub-

markets than those developed with clustering. The Cluster 1 and Cluster 2 models also had significant differences with Cluster 2 not modelling as well if you examine the R^2 values. The downfall of the models was in their estimation ability in that all models performed poorly compared to the models developed using the publicly available data, alone.

GEOGRAPHICALLY DEFINED SUB-MARKETS					STATISTICALLY DEFINED SUB-MARKETS					
					CLUSTER 1			CLUSTER 2		
Model	**R^2**	**0-10%**	**0-20%**	**Model**	**R^2**	**0-10%**	**0-20%**	**R^2**	**0-10%**	**0-20%**
17	48.7	12	26	23	47.3	14.5	29.1	35	18	27.7
18	36.9	10	21	24	40.5	14.5	23.6	25.5	10.8	30.1
19	49.9	15.9	31.8	25	40.5	10.9	27.3	12.1	9.6	21.6
20	44.7	18.3	29	26	-	-	-	27.6	15.7	21
21	59.9	13.8	31.9	27	44.7	14.5	29	30.1	15.6	33.7
22	64.1	3	12.9	28	-	-	-	27.5	1.2	2.5

Table 8.7 Model Results using Restricted Digital Data

Due to the extremes in values when predicting using the regression equations for each model, it was decided that there was no gain in producing COD, COV and PRD values for all of these models. This was due to the fact firstly that the percentage of estimates falling within 10 or 20% of the actual sale price was worse than in the publicly available data phase. The second point is that the poor COD, COV and PRD results determined in the publicly available data phase did not add any value to the analysis of model accuracy.

The following table, Table 8.8, shows the ranges of values for the percentage of estimates falling within 10% and within 20% of the actual sale price. The table shows that at the top end, each technique had some models which had around 30% of properties within 20% of the actual sale price, therefore, no real difference between the sub-market techniques used. Looking at the bottom end, Cluster 2 performed the worst in one model which only estimated 2.5% within 20% of the actual sale price. The other models still varied in their estimation power and there were some models that performed poorly.

	Geographically defined sub-markets	Statistically Defined Sub-markets	
		Cluster 1	Cluster 2
Percentage of estimates within 10% of actual sale price	3 – 18.3	10.9-14.5	1.2-18
Percentage of estimates within 20% of actual sale price	12.9 – 31.8	23.6-29.1	2.5-33.7

Table 8.8 Summary Ranges of Models – Restricted Data - Percentage of estimates within 10 and within 20% of actual price

8.4 Discussion

8.4.1 Numerical Model Development using Publicly Available Digital Data

The development of the numerical models was influenced by research that has applied both 'sale price' and 'sale price per hectare' or acre as dependent variables in models (Miranowski & Hammes, 1984; Xu *et al.*, 1993; Bastian *et al.*, 2001). In addition, modelling has been undertaken that has also applied Logarithms (generally natural Logarithms) (Reynolds & Regalado, 2002). The process of model development in my research followed an exploratory process in which different dependent variables; sale

price and sale price per hectare were tested and then Logarithms were applied to test the performance of the models with these dependent variables. The process also involved the removal of outliers and the removal of specific LGAs during the model development.

Analysis of the results of the models will now focus on the percentage of estimates falling within 20% of the actual sale price rather than within 10% of the actual price. The models developed using geographical LGA constraints (Models 1 – 8) had the highest results in 3 models. These models (Models 3, 7 and 8) had 33% of estimates falling within 20% of the actual sale price. Models 7 and 8 use the same dependent variable as Model 3 however an outlier is removed along with the Grampians LGA for Models 8. Overall the models developed using geographical sub-markets did not yield a high level of price prediction (33% of estimates fell within 20% of the actual sale price). Although some of the models increased in their accuracy upon removing outliers and removing specific LGAs, this did not dramatically increase results. The technique applied to segregate property based on the LGA to which each property belongs involves creating indicator variables for each LGA. The numerical models were developed using the whole study area and these indicator variables were used to determine if the geographical location that each property was assigned to was in fact significant to price.

During regression modelling, the latter models of the geographical phase, Models 5-8, had two of these LGA regions in the model equations which were significant at the 0.05% level. These were the variables '*AREA1*' and '*AREA3*'. The initial four models did not include the area variables (location of property into a specific LGA) indicating

that they were not an influence to price. Results of the developed models did not show a higher level of accuracy for models that included the '*AREA1*' and '*AREA3*' variables.

Due to the number of properties available to test, regression modelling would not have been possible if properties had been segmented into their four LGAs and the areas had been processed separately. Hence this was not attempted. Furthermore, modelling property values using a significantly smaller sample size is likely to have led to the introduction of more error (McCluskey & Deddis, n.d.).

Although modelling each LGA separately could yield a higher level of accuracy, the selection of the geographical region to model may not necessarily lead to more accurate results. Wilhelmsson (2004) argued, for example, that rural areas will not be homogenous across a large region and therefore many factors are likely to influence the creation of sub-markets. Selecting regions to model based on an administrative boundaries may not actually replicate the market forces of specific sub-markets in operation within an area.

Sub-markets may not necessarily be geographically constrained or if they are, they may not be to such a large geographical area. The location of a property and its apparent situation in a particular sub-market may be influenced by a combination of factors or could be constrained to a larger scaled area. Property constrained to a specific ranged distance to town may prove to be a more appropriate means to assign property to sub-markets rather than an administrative boundary to which a property may fall. This simplistic natured technique of deriving sub-markets assume that sub-markets and buyer preferences are related to one characteristic and as such can be

modelled with ease. In essence, the buyer preferences and the delineation of property into sub-markets is a more complex process which involves multiple criteria.

Various techniques have been used to segregate property into sub-markets using a-priori techniques (Bourassa & Hoesli, 1999; Xu *et al.,* 1993). This can involve splitting the properties based on physical characteristics such as land use type, property type by a range of values, or irrigation presence. Other techniques involved amalgamation of counties (Xu *et al.,* 1993) or LGAs into sub-markets, or using a nested approach utilising a geographical boundary in addition to structural constraints (Adair *et al*., 1996). It is hypothesised that the sub-market influences are a combination of both geographical and structural constraints (Bourassa & Hoesli, 1999) or may include socio-economic factors (Dunse *et al*., 2001). In some cases these factors may not fit easily into an arbitrary division of property or locational characteristics. Therefore, it may be difficult to isolate the composite factors of specific sub-markets and to develop reliable means to minimise, if not eliminate possible (a-priori) biases.

Attempts to model or segregate property into divisions may not actually mirror the true segregation process that occurs in rural property. Where a greater number of property is being purchased for lifestyle purposes, then different influences and thus sub-markets have a greater force. Properties being sold to lifestyle buyers are more likely to reflect a desire for a property rather than a purchase made which places more importance on the agricultural aspects of the property such as buildings, land use, salinity, irrigation, property size and location to transport hubs.

Cluster analysis is a statistical technique used to delineate sub-markets for the valuation of property (Bourassa *et al.* , 1997; Dunse *et al.,* 2001; O'Roarty, 1997; Wilhelmsson, 2004). A simplistic approach, it measures distances between the values of specific variables or property characteristics in an attempt to create similar groups of items or properties in this instance. The technique used in this research was to cluster the property based on a selection of property characteristics and using a pre-determined number of clusters. The division of clusters into three groupings was the most valid for the property in the study sample. Any further number of divisions between the cluster groups devised would of lead to more groups requiring manual processing rather than using a more automated regression analysis approach.

During the three Cluster solution, the properties were split into three distinct clusters with the clusters not splitting into geographically defined regions. In the first cluster, Cluster 1; the land uses classed into this cluster were land use 1, 2 and 6. Cluster 2 also included land uses from land use type 1 and 6 along with land use type 3, 4 and 5. This was the only cluster which had property with these types of land uses (*LUSE3, LUSE4, LUSE5*). Cluster 3 was solely comprised of land use type 2.

The clusters defined led to groups of property being specified that were partly based on land use type due to the 'two step' cluster process utilising categorical variables. However, analysis of the variables used for the cluster analysis did not show clear divisions between the variables used to cluster. General trends can be seen in Table 8.9 in that Cluster 1 tends to have the extremes of data for each variable compared to the other two clusters. It has the properties with the largest property area, the highest adjusted price, highest number of water body points on a property, longest length of

watercourses on a property etc. However, the distinction is not clear enough for one to manually divide the properties into clusters (eg: if all property had a property area of 450 or larger than they could be assigned to a cluster). In essence, there may be overlap in the range of each variable or property characteristic such that this distinction is not always obvious.

	CLUSTER 1	**CLUSTER 2**	**CLUSTER 3**
PROP_AREA	71-593 ha	20-251 ha	31-335 ha
ADJ_PRICE	$13,200-$785,000	$10,047-$683,968	$26,500-$521,538
WATERB_PT	1-11	1-5	1-6
WATERB_AREA	1,281-3,2491 sq m	1,045-7,906 sq m	384-3,245 sq m
WATERCRS	1,258-13,939 m	7-6,310 m	41-5,810m
TOWN_DIST	2.5-151 km	3-82 km	6-116 km
ADJ_PRICEPHA	$73-$10,959	$177-$12,169	$107-$5,488

Table 8.9 Range of variables within each cluster – Three Cluster solution

Clusters 2 and 3 were processed using regression however Cluster 1 would require manual valuation as there were only 14 properties assigned to this cluster. Analysis of the results of the clusters showed an improvement in the properties which estimated with 20% of the actual sale price as can be seen in Table 8.10.

				Geographical	Statistical	
					Cluster 2	Cluster 3
Model		Dependent Variable	Data Set Used	Within 20% of actual price	Within 20% of actual price	Within 20% of actual price
1	9	*ADJ_PRICE*	Whole	26%	20%	30%
2	10	*ADJ_PRICE*	Whole	27%	26%	38%
3	11	Log_{10} *ADJ_PRICE*	Whole	33%	23%	38%
4	12	Log_{10} *ADJ_PRICE*	Whole	22%	29%	36%
5	13	*ADJ_PRICE PHA*	Whole	17%	21.4%	35%
6	14	*ADJ_PRICE PHA*	Outlier, Northern Grampians LGA Removed	22%	24%	34%
7	15	Log_{10} *ADJ_PRICE PHA*	Whole	33%	20%	36%
8	16	Log_{10} *ADJ_PRICE PHA*	Outlier, Northern Grampians LGA Removed	33%	24%	36%

Table 8.10 Model results of the Percentage of Estimates within 20% of the actual sale price (Restricted Data Phase)

The distinction between the geographically defined sub-markets and Cluster 2 from the statistically derived ones was not as apparent. In five models, Model 1, 2, 3, 7, 8, the geographically defined models had a higher percentage of estimates within 20% of the sale price with only Models 4, 5 and 6 of Cluster 2 with a higher percentage of estimates within 20% of the actual price. Thus, comparison of the geographical models and Cluster 2 did not show that one technique, in particular the statistically derived models estimated more accurately. Comparison of Cluster 3 with the geographical models showed that all the statistically derived models using Cluster 3 had a higher percentage of estimates within 20% of the actual price. This indicates that the models

developed from Cluster 3 using Cluster analysis had a higher proportion of accurate results when examining the percentage of estimates falling within 20% of the actual sale price.

The rural property market place can be segmented using cluster analysis and models can be developed which are more accurate than geographically derived sub-markets with regard to estimates falling within 20% of the actual price. The cluster analysis technique is, however, subject to having a large data set to enable an adequate split of property into groups or sub-markets such that regression modelling can then be undertaken. The technique also requires that there is some variation amongst the property characteristics to ensure that regression modelling can be undertaken after the sample has been split into more homogenous regions. A resultant cluster may have little variation amongst its property characteristics and thus when using a small number of variables to categorise a regression model, most of these may be similar. In the case of the land use type variable, when developing Model 14, the combined land use category '*LUSE12*' was removed as all values were identical. This is one of the issues when using either geographically or statistically derived sub-markets in that the resultant properties assigned to each sub-market group may have little variation in their values if too similar groups of properties are created to form a sub-market. Cluster analysis may be a better technique as in this case it has not clustered into groups with little variation. If the resultant clusters had been containing only the one type of land use or the same number of waterbody points then modelling error or inadequacy of models may have been introduced.

The increase in accuracy of those models developed on the Cluster 3 properties over the geographically defined models can be explained firstly by the fact that now different variations in characteristics exist over the sample area. Initially the geographically defined models are constraining and assuming that all the property characteristics are the same over the whole study area. In effect, it is highly likely that this is not the case (Wilhelmsson, 2004; Xu *et al*., 1993) and that sub-markets exist that are not defined by an administrative boundary alone. Even though a variable denotes the different administrative boundary and the regression model takes into account the variations in spatial location, the constraint on these models is that all characteristics specified in the regression will be the same across the whole study area.

The clustering techniques and segregation used in this research could be altered to develop a hybrid approach which would endeavour to amalgamate the segregation so that it considers geographical location at a finer spatial scale in addition to property characteristics. By segregating property into spatial sub-markets first and then clustering, a more definitive sub-market may be derived than what was used in this research. Although it is unlikely that sub-markets split on specific administrative boundaries such as post code areas (a finer resolution than LGA boundary), a smaller geographical region or other spatial constraints being applied would create a clustered region where the location of the properties were considered. The clusters developed in this research were segregated over a large spatial area even though there were location based variables used during model development. In effect, geographical location did not play a part when the properties were segregated.

For property tax rating purposes in Victoria, different rates of taxation are in operation over different spatial areas. Thus, to cluster property that does not first segregate or take into account geographical location would not lead to accurate valuation modelling to occur or enable a saving in time by using automated techniques if there are in fact differences in model specification over these taxation regions.

During the geographically derived sub-market phase of the research, 8 models were created. Models 1 to 4 varied to Models 5 to 8 in terms of the property characteristics found to be significant during the regression modelling (Table 8.2). Models 1 to 4 contained the water variables (waterbody area and water course length) whilst in Models 5 to 8, these variables were not significant. The other variation was in the property area variable (*PROP_AREA*) in that it was significant in the first 4 models and not in Models 5 through 8. Models 5 to 8 had a dependency on the LGA to which each property belonged in addition to the amalgamated land use types (LUSE_12 and LUSE_345).

The statistically derived models had more variation between each model in terms of the property characteristics found to be significant (Table 8.2) than those of the geographically derived models. However, overall there was no clear set of variables which were significant in all models developed. In comparison, in the geographically derived sub-market models, there were some similarities between the variables specified in those models defined geographically. Analysis of the cluster models showed that all the property characteristics specified were the same amongst these models, however models using Cluster 2 (Model 9a, 10a, 11a, 12a) all tended to have more characteristics specified in their model equations than those of Cluster 3 (Models

9b, 10b, 11b, 12b). Models 13a-16b had few significant variables specified. These models specified waterbody area, watercourse length and the town distance variables in their equations even though they were not statistically significant at the 0.05% level. Comparison of the two sub-market derivation techniques showed that variable use differed between Models 5 to 8 and Models 13 to 16. Models 13-16 used the above mentioned water variables whilst Models 5 to 8 did not.

The cluster analysis stage of the process obviously enabled slightly varied groupings of property to be developed which had different price drivers between the clusters which was evident in the model testing undertaken. This was also the case when comparing the cluster models to those developed using geographical constraints. Although there was no clear distinction between the variables specified in model equations and those found to be significant in both sets of models developed, it highlights that the same types of variables are being used in modelling however with variations in their amounts of influence. It seems to be evident that similar types of property characteristics are influential no matter which sub-market derivation technique is used. However, there are still small differences in their levels of influence and their statistical significance depending on whether the models were developed using cluster analysis or as a whole data set which segmented property based on geographical constraints. The property characteristics specified during both modelling phases are presented in Table 8.11. This similarity between the techniques indicate that they are valid in that they are performing in similar ways and using mostly similar variables which is what you would expect in that they are modelling price using the same pricing influences. It does, however, re-iterate that depending on the technique and dependent variable used, there will be variations in the property characteristics modelled leading to the question as to why the models and their differing techniques vary in such a way. The answer to

that is partly the use of different geographical regions in that the cluster technique is attempting to model only a segment of the property that was modelled within Models 1 to 8. Thus, a higher level of accuracy is achieved in Cluster 3 models than those where geographical constraints were used to create sub-markets.

Property Characteristics
PROP_AREA
AREA1
AREA3
LUSE_3
LUSE_4
LUSE_12
LUSE_345
WATERB_AREA
WATERCRS
TOWN_DIST

Table 8.11 Variables specified during numerical modelling (Publicly Available Data Phase)

In comparison to the variables that were available within this phase of the research for use, the ones which were not used in models were typically those where there was little variation (*SEVERITY, NATURAL, FOX, LSIO, ZONE_CODE*). The other variables not specified in models were *AREA2, AREA4, WATERBPT, LUSE_1, LUSE_2, LUSE_5* and *LUSE_6*. In contrast to the conceptual model (Figure 2.1), many variables were not available digitally and hence were not able to be tested to verify their existence in the model. The model testing undertaken allowed the available property characteristics to be tested and assessed as to their suitability in the conceptual model. The

ZONE_CODE, LSIO and WATERB_PT were not valid in this testing region and can be excluded from the model. The variables SEVERITY, NATURAL and FOX could not adequately be excluded due to the little variation existing within each variable. However, they were always excluded during regression modelling and thus their significance could not be verified. If a larger sample of properties could be obtained and thus there was greater variation within each characteristic or variable, then their inclusion in the conceptual model could be verified.

This phase of the research tested the importance of using sub-markets for rural property and used two techniques to ascertain the effect on using geographically derived versus statistically derived sub-markets. The research has shown that in only one of the cluster regions developed, the results were higher than those developed geographically. I concluded that sub-markets do exist and that different sub-markets exhibit different price drivers and characteristics and therefore a study region requires sub-division into appropriate areas to adequately reflect the variety of pricing influences that exist in different areas.

8.4.2 Numerical Model Development using Restricted Digital Data

The processing undertaken using the restricted digital data produced mixed results as can be seen in Table 8.12. The data supplied required categorisation into indicator variables.

				Geographical	Statistical	
					Cluster 1	Cluster 2
Model		Dependent Variable	Data Set Used	Within 20% of actual price	Within 20% of actual price	Within 20% of actual price
17	23	*ADJ_PRICE*	Whole	26	29.1	27.7
18	24	*Log_{10} ADJ_PRICE*	Whole	21	23.6	30.1
19	25	*ADJ_PRICE PHA*	Whole	31.8	27.3	21.6
20	26	*ADJ_PRICE PHA*	Outlier, Grampians LGA Removed	29	-	21
21	27	*Log_{10} ADJ_PRICE PHA*	Whole	31.9	29	33.7
22	28	*Log_{10} ADJ_PRICE PHA*	Outlier, Grampians LGA Removed	12.9	-	2.5

Table 8.12 Model results of the Percentage of Estimates within 20% of the actual sale price (Restricted Data Phase)

The data set was re-clustered to include a two step clustering process which segregated the areas based on the restricted variables as well as land use type. A two cluster solution was used rather than the three cluster solution in the available data phase, mainly due to one cluster requiring manual valuation, it was thought it may yield better results to cluster into two areas rather than three. The properties within Cluster 1 did not include any properties from the Northern Grampians LGA and with the outlier present in Cluster 2, there were no properties to remove when this phase of processing was undertaken. The clustering process segregated the data to such a homogenous state, in terms of variables, that for Cluster 2 the values were the same for all properties for the *PASTURE, WATER* and *HOUSE_COND* variables. As such they could not be included in any modelling and may of resulted in better models if in fact these variables were a significant influence to property price.

A hybrid clustering technique may be necessary to better account for both the structural characteristic influences and the location based factors which distinguish sub-markets. The use of just structural characteristics can lead to extremely homogenous areas being created in terms of their property characteristics which will effect regression modelling as many characteristics will be the same.

In the previous clustering, the three cluster solution; the data was not visibly different in terms of property characteristics due to the inclusion of a wider range of both qualitative and quantitative variables. In the two cluster solution undertaken in Section 7.4 the sub-markets had more visible segregation in terms of property characteristics, largely due to the number of indicator variables. However, the models produced in the restricted data phase with geographical sub-markets did not show an increase in model accuracy, thus it was not solely the clustering technique which effected model results.

The models generated during the clustering process generally resulted in the inclusion of only a few restricted variables, with the majority of the variables used coming from the publicly available data set. In contrast, the models determined geographically resulted in the inclusion of a greater number of restricted variables, in addition to a larger number of publicly available variables. Comparing the two sub-market methods, the geographical models specified a larger number of variables in each model than the clustering models where fewer variables were specified. The clustering models had more variables which were the same, ie: the *WATER* and *PASTURE* variables which were identical in Cluster 2, therefore, models could not utilise these property characteristics.

The variables used in the restricted data phase are shown in Table 8.13. All of the restricted variables were specified in one or multiple models. The variables *AREA2*, *LUSE_1* and *WATERB_PT* were not specified during the publicly available data phase however were specified in the restricted phase. The specification of different variables in models was not standard across the models. There were only five variables which one could conclude are significant to rural property valuation. The basis for this is that the *TOWN_DIST* variable was used in all models, albeit with different levels of statistical significance. Likewise the *WATERB_AREA* variable was used in nearly all models from the restricted data phase. The *WATERCRS* variable was used in both research phases to a limited extent. The *FARMB* variable was used regularly in the restricted models.

Property Characteristics	
PROP_AREA	*WATERB_AREA*
AREA1	*WATERCRS*
AREA2	*TOWN_DIST*
AREA3	*FARMB*
LUSE_1	*WATER*
LUSE_3	*FENCE*
LUSE_12	*PASTURE*
LUSE_345	*HOUSE_COND*
WATERB_PT	

Table 8.13 Variables specified during numerical modelling (Restricted Data Phase)

The variables reflect a consideration in modelling for water based property characteristics along with a location based variable. The research suggests that the

rural market place is so complex that it is difficult to model and that sub-markets do exist that have different pricing characteristics in different areas. The research also shows that there are still data issues involved with the modelling of rural property values. The restricted data supplied differed in the standards of attribute representation and coding systems as shown in Appendices F and G. The *SOIL* and *VEG* variables had different classification schemes for the representation of data. There were some properties where information on variables such as *UNUSEDROAD/WF* variable were absent (Appendices F and G).

It is possible that utilising a geographically smaller study area with these restricted variables may yield increased model accuracy. A difficulty still remains in that it is necessary to have sufficient property characteristics for the sample properties for an adequate analysis. The research suggests that due to the results obtained using geographical sub-markets in both the restricted and available data phases, these restricted variables are not actually influential as model accuracy did not increase. It is possible that other characteristics may be more influential than those suggested and tested. Changes to the set of variables may be necessary to improve results. For example, the presence of water features may be more significant than merely the length of a water course on a property.

8.5 Summary

This Chapter discussed the data integration and numerical modelling stages of the research. The PRISM data set which holds sale price information, and was the primary data set used to acquire sale prices was found to be inconsistent in its recording of attributes. Many instances were found of incomplete records or missing values within

the data set. The data set is accessible to a restricted set of users online through a query function, yet for this research was supplied in HTML files which were then used as tabular files in a GIS. The present use of the data set is limited in that valuers and real estate agents do not have an ability to readily integrate the sale price information within a GIS and thus is limiting the use of the data set which does hold a wide array of sale price information. Issues arose within the data integration stage regarding the areal extent of a data set in that many instances arose where a data set was obtained and no information was populated from it as the features did not overlay the property database used in this research. The major outcomes of my research are considered in further detail in Chapter 9.

Chapter 9 Conclusions

9.1 Introduction

The aim of this research was to develop a rural property valuation model for Victoria that used automated techniques to arrive at valuation estimates. The research also examined the extent to which sub-market grouping of property into more homogenous regions through the use of cluster analysis can enhance property valuation estimates.

Chapter 2 highlighted both manual and automated techniques currently used to value residential and rural property both internationally and in a local context. Manual techniques are discussed with reference to Valuation Best Practice (Valuation Best Practice, 2006) and the influences that these specifications have had on rural valuation within Victoria. The Chapter discussed the use of automated techniques, in particular the use of decision support, regression analysis, criteria ranking, case based reasoning, expert systems and artificial neural networks. Each technique has its own limitations within a rural market and this has been presented with reference to other research that has applied these automated techniques to rural valuation. The use of specific property characteristics and the accuracy of other rural models are discussed. A common problem with rural valuation is finding a method to select appropriate property characteristics prior to applying an automated technique. Thus, a model for rural property valuation estimation was presented in Section 2.4 to highlight the variables considered most likely to influence rural property values based on the reviewed literature.

Chapter 3 provided an overview of the data integration issues associated with integrating data from multiple sources and the use of metadata to assist in locating suitable data sets. The use of GIS within valuation is examined, in particular the minimal use it has had in the rural valuation industry. The Chapter discussed the use of GIS for variable creation and discusses some of the GIS techniques which could enhance rural valuation variable determination. I also considered the process of cluster analysis and its use in determination of sub-markets for property valuation.

For the validation of the conceptual model, a GIS property valuation database was created utilising data sets obtained from a combination of sources as detailed in Chapter 4. Information regarding the rationale for selecting the two study areas and their geographical location within Victoria, along with the GIS and statistical software used within the study was presented. This Chapter outlined the software formats and coordinate systems of the acquired data sets along with the methodology undertaken to convert these data sets into the one GIS database. The use of GIS for the derivation of additional variables was shown, and data integration issues that arose during the course of the database development stage were discussed and a framework for data integration was presented.

Chapter 5 highlighted the concepts leading to the development of the initial eight numeric models known as the geographically derived sub-market phase. A variety of regression techniques were applied using the available variables from the property database. Based on the initial statistical results of each model, further testing and refinement was conducted using an exploratory process. In all, 8 numeric models were developed during this stage.

The numeric models were also tested using an additional procedure. This involved using the regression equation determined from each model and applying each property's variables into the equation to determine a property valuation estimate. The estimates obtained were compared to the actual sale price. Within each model, the percentage of property estimates falling within 10% and within 20% of the actual sale price was determined. There were between 8-16% of property estimates falling within the smaller 10% range. This improved slightly with between 17-33% of estimates falling within the 20% range. These values are much lower than previously thought by the author to be necessary to develop a robust and workable model.

Chapter 6 presented the clustering processes used during the statistically derived sub-market phase of the project. The results of the two step clustering procedure suggest that the 3 cluster solution is more ideal for this data. One cluster had few properties assigned to it, thus is a cluster which would require manual valuation. Using the other two clusters, numerical models were created which replicate the processes used to derive the geographically derived models in Chapter 5. Chapter 7 utilised additional restricted data and created new models to help evaluate those developed in Chapters 5 and 6. The inclusion of these data came about at a late stage of the research and the models were developed to ascertain the affect of model accuracy with these additional property characteristics.

Chapter 8 discussed the models developed and their success at determining rural property valuation estimates in Victoria. This Chapter drew comparisons between the two techniques used to derive sub-markets and discussed the use of cluster analysis for market segmentation as opposed to segmentation using geographical influences. The Chapter discussed the additional models developed with the inclusion of restricted

digital data and the affect that this has had on the initial models where property characteristics were limited.

9.2 Findings and Synthesis

Rural valuation is becoming more automated as researchers aim to develop more effective and accurate means to value such properties. Automated techniques aim to reduce biases in the present manual techniques and provide more consistent valuations. Within Victoria, the advent of Valuation Best Practice (Valuation Best Practice, 2005) has seen reforms in the valuation techniques used throughout the State. Although the Victorian valuation process is still largely manual more automated techniques are being used to assist this process. GIS is used for display of previous valuations and validation of current values for the detection of anomalies. In addition, statistical measures are used to evaluate results to ensure consistency and accuracy in manually derived values.

A summary of the most important property variables used in the studies reported in the literature and during my study is shown in Table 9.1 (significant, U – unclear, X – not significant). In most models, property size, LGA location and the land use type variables were significant.

In Chapter 2, Table 2.1 I summarised the property characteristics most frequently reported in the literature as significantly affecting rural property values. In comparison to those characteristics, in my research property size and county/LGA location were found to be significant. Characteristics such as pasture condition and house condition were found to be not significant in the modelling presented in the thesis. The affect of

characteristics such as water supply, fencing condition and farm building on rural property values is unclear, a result consistent with the literature review (Table 2.1).

Significant Property Characteristics reported in the reviewed Literature	Significance in my Research
Property size	
Distance to town	U
LGA location	
Land use type	
Number of waterbody points	U
Area of waterbodies	U
Length of watercourses	U
Presence of a farm building	U
Average and above average water supply	U
Average and above average fencing	U
Pasture condition	X
House condition	X

Table 9.1 Summary of important explanatory variables used to estimate property values based on a review of the property valuation literature and the new modelling undertaken in this study

The numerical models developed from the conceptual model did not lead to highly accurate models. The research found that whilst data can be obtained and integrated, there is still some degree of limitation in the property data that are available to the general public. As such, during the first stages of the research, the conceptual model was not able to be fully implemented and tested.

The creation of models based on (1) geographically derived sub-markets and (2) statistically derived sub-markets allowed for the examination of the effectiveness of the

two sub-market grouping techniques. Results of the cluster analysis stage led to a number of models being developed and these were able to be compared to those models which were derived using geographical sub-markets. Of the 3 clusters developed, Clusters 2 and 3 were suitable for regression modelling. Cluster 3 proved to be a region where automated modelling was more accurate than for the models developed using geographical constraints. As such, every model developed using Cluster 3 properties was more accurate than the geographically derived ones in terms of the percentage of property estimates falling within 20% of the actual sale price. The inference of this is that the clustering did enable a more homogenous region to be developed automatically and this did lead to more accurate valuation estimates than those determined geographically. Unfortunately, this was not the case for Cluster 2 as it did not perform as well as Cluster 3. The research has indicated that sub-market identification is necessary during property valuation and that cluster analysis can lead to an improvement in model accuracy.

The improvements reported when using cluster analysis, in particular with cluster 3, would warrant further research using clustering techniques to further examine the effect of model accuracy using this method. However, additional data in terms of sample size and an increase in the number of property characteristics, would need to be acquired to effectively conclude on the effectiveness of the clustering techniques.

Access to restricted data was provided in the later stages of this research. The additional information obtained included data on improvements to property, water access quality, farm building presence, house presence and condition, fence condition and pasture condition. A further 12 models were developed using these data and the R^2 results of these models were higher compared to those developed earlier in the

thesis. However, modelling of the regression equations showed that the models did not have as high a percentage of estimates falling within 20% of the actual sale price compared to those models developed using the publicly available data, alone.

One observation from the research regarding the use of digital data for rural valuation is that in many instances data on property characteristics are not available. Another finding is that when data are available they may not actually encompass the areal extents of the properties of interest. The development of a technique to report on the areal extents of a data set for point, line and area features would greatly improve the use of some data and allow users to more adequately determine if the characteristics held within the data set are actually suitable prior to acquisition.

This research has proposed a new framework for data integration (Figure 4.1, Figure 9.1). The framework addresses data incompatibilities and incompleteness, and accuracy, scale and format differences that arose during the study. The framework comprises a database design suited to GIS data integration and details the processes necessary to convert tabular non-GIS and spatial GIS data from various GIS software formats into the one homogeneous data set. It details the projection transformations required in Victoria to convert the existing data sets into the one standard projection and datum so that they can be overlayed reliably and effectively.

Figure **9.1** shows a schematic diagram depicting the data integration process and the importance of metadata to accompany the supply of digital data. The diagram is a modified version of Figure 4.1 and focuses on the metadata specifications which

should be attached to digital data sets. In particular, it highlights the importance of defining co-ordinates systems parameters, regardless of how the data are supplied (in a geographical or coordinate based system). The metadata should also define any classification scheme used in any of the attribute tables within the data set and provide details of any Look Up Tables (LUTs) and attach this information to the metadata for the user to refer to. Any processing undertaken to the data set should also be mentioned in the metadata to allow the user to decide how suitable for use the data are after any additional processing. The development of these additional variables in the metadata is important, especially if automated of rural property valuation continues throughout Victoria or Australia to allow valuers to ascertain the validity of the valuation results obtained using a particular data set. Section 9.4 presents in more detail the recommendations for the above additional elements to be included in the metadata for the supply of digital data.

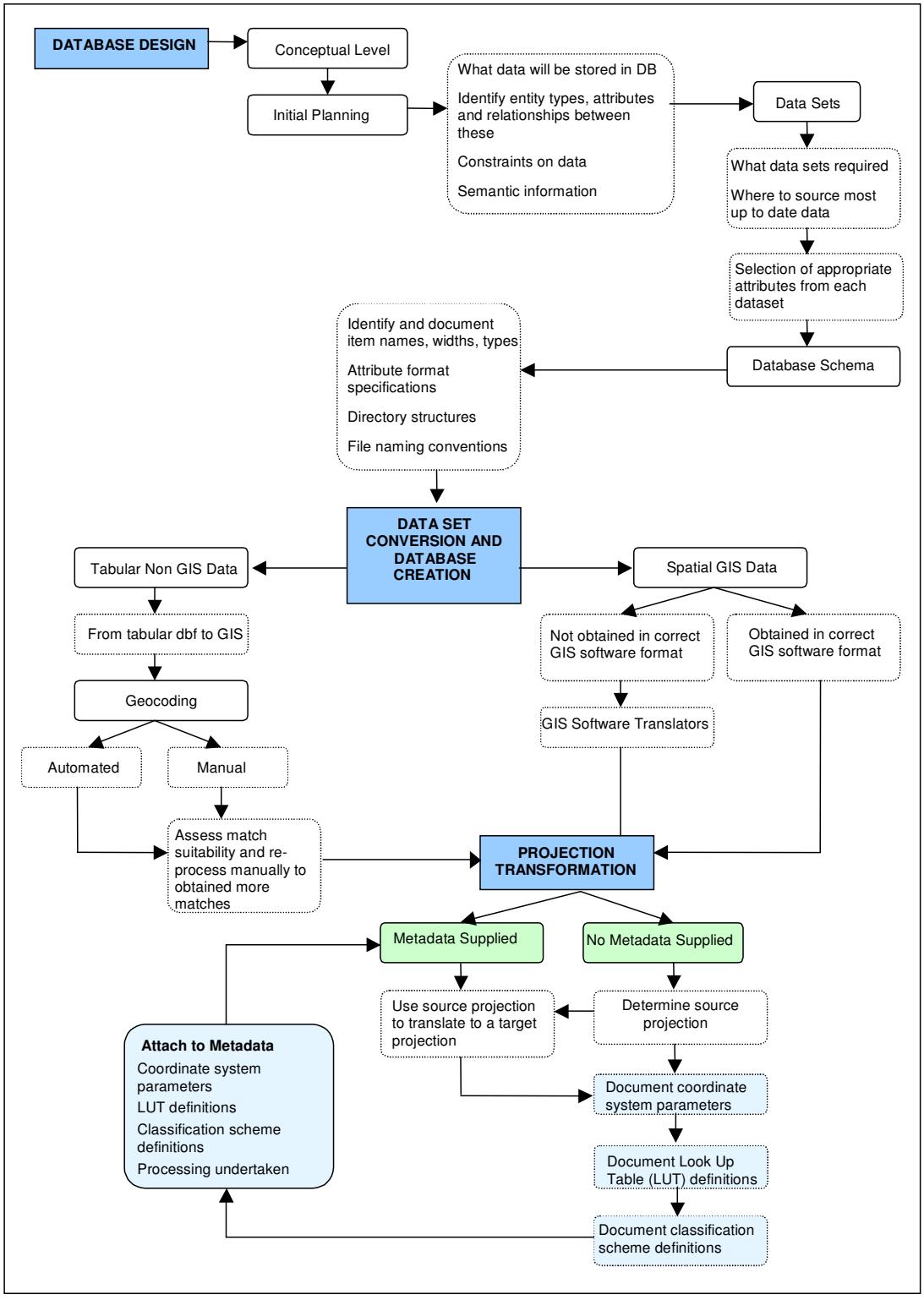

Figure 9.1 Schematic of the new data integration framework developed during the thesis emphasising the significance of suitable metadata standards to support an automation of property valuation using quantitative models.

Overall, my research has highlighted the limitations and issues associated with the use of current digital data in Victoria for property valuation. The research has shown that automated modelling of rural property valuation is feasible. However, at this stage a high level of accuracy is unlikely given the issues outlined earlier in the thesis. In addition, the research has shown an increase in accuracy in some of the models developed using statistical techniques (cluster analysis). The research concludes that to enhance regression models for rural valuation, further research needs to be undertaken using a larger database of properties in a smaller geographical area. In addition more manual methods of data attainment may be necessary to test if automated modelling to a higher degree of sophistication and accuracy can be achieved in Victoria in the medium term.

9.3 Limitations of the Research

After integrating numerous data sets and then developing numerical models, it became apparent that obtaining sufficient information pertaining to each property was a limitation of this research. A variety of publicly available digital data was obtained and integrated into the property database. However, in many cases the utility of the data sets obtained was limited as they did not cover the geographical extents of the sample properties.

Another limitation of the research was the quality of the sale price data and the format in which it was supplied. In many instances there were incomplete records - in particular no recorded sale price information or property address. Thus, these properties had to be removed from the data set. The sale price data set was supplied

as a spatial non-GIS data set and required an address identifier to be present to enable geocoding to a spatial GIS database. A manual technique was employed as automated geocoding was not successful and lead to lengthy integration times and a database of fewer properties than initially supplied. Some options to overcome these types of issues and constraints are outlined in the next section.

9.4 Recommendations

Due to the integration issues that arose during this research it is recommended that the following Metadata elements be added as additional parameters to be included within the current ANZLIC Metadata Guidelines as core metadata elements.

1. Co-ordinate System Parameters – where appropriate this would include the projection units, spheroid, scale factor, longitude of the central meridian, latitude of the origin, false easting and false northing coordinates to be defined within the metadata and thus be supplied with the data. It should be noted that depending on the projection not all of these parameters are necessary and some additional parameters would be required for different projections.

2. Look Up Table Definitions – this would include documenting the look up table files that accompany the data so that the user is aware of the additional files associated with data sets.

3. Classification Scheme Definitions – this would document the classification schemes used within the data and provide definitions of these (ie: land use coding – define rural, residential, commercial).

These additional parameters would enable more information to be obtained regarding the projection parameters of the data set, and would also enhance the knowledge on the attributes and coding systems in use.

I also recommend that for future property sale transactions, a unique ID is assigned to enable linking of the sale information with other data sets used by Land Victoria. Currently the PRISM data is not identified by a unique ID. This inclusion would enable cadastral information and other data sets to be linked together with sale information (PRISM data) for each individual property. This would then enable a data set to be created that utilises both municipal valuation information in addition to sale results.

Drawing on the data availability and data quality issues that arose from this research, the following are recommendations for future research to be undertaken prior to the development of further modelling techniques for rural property valuation estimates within rural Victoria.

a) The analysis of time series trends to address the issues associated with the infrequent sales of rural properties. This would aim to identify patterns within the data sets to account for the lack of sales occurring within a similar time frame and provide some means to identify any trends in sale prices over time,

b) The identification of price trends by utilising both sale price and valuation data for each property. With valuations performed every two years in Victoria, trends may be identified for each property,

c) The development of a technique to provide more detailed information regarding the appropriateness of a data set. This would entail documenting more thoroughly within metadata the extents of the features within each data set.

Although bounding coordinates are given for the overall coverage, this does not provide information regarding the number of instances of a feature within a data set,

d) The development of a technique to enhance the sale price data set (PRISM) to enable it to be more widely used in valuation and property modelling through an identifier that allows for integration with existing digital property data,

e) That further exploratory numerical modelling is undertaken using a smaller geographical area than employed in my study and a larger sample size.

References

ABARE 2006, see http://www.abareconomics.com

Adair, A, Berry, J and McGreal, W 1996, 'Hedonic modelling, housing submarkets and residential valuation', *Journal of Property Research*, vol. 13, pp. 67-83.

Adair, A and McGreal, W 1988, 'Computer assisted valuation of residential property', *Real Estate Appraiser*, vol. 54, no. 4, pp. 18-21.

Aldenderfer, M and Blashfield, R 1987, *Cluster Analysis*, in series 'Quantitative Applications in the Social Sciences', Sullivan, J L and Niemi, R G (ed). Sara Miller McCune, Sage Publications, Inc, London.

American Institute of Real Estate Appraisers, 1983, *The appraisal of rural property*, American Institute of Real Estate Appraisers, Chicago, USA.

Ameson, P, Garvin, T and Neubauer, B 1996, 'Reconciliation of Valuation Indications In The Real Estate Appraisal Process: A Multicriteria Approach', *Proceedings of the 27th Annual Meeting of the Decision Sciences Institute*, Orlando, Florida, vol. 2 of 2, Texas A&M University, Texas, pp. 981-983.

Ausway 2004, *Melway Edition 32 2005*, Ausway Publishing, Mt Waverly, Victoria, Australia.

Bastian, C, McLeod, D, Germino, M, Reiners, W and Benedict, B 2001, 'Environmental amenities and agricultural land values: a hedonic model using geographical information systems data', *Ecological Economics*, vol. 40, pp. 337-349.

Baum, S, Trapp, C and Weingarten, P 'Typology of rural areas in the CEE new Member States', *87th EAAE - Seminar: Assessing rural development of the CAP*.

Baxter, J and Cohen, R 1997, 'Rural Property', in *Valuation principles and practice*, (ed), Australian Institute of Valuers and Land Economists, Deakin, ACT, pp. 218-256.

Boisvert, R, Schmit, T and Regmi, A 1997, 'Spatial, productivity and environmental determinants of farmland values', *American Journal of Agricultural Economics*, vol. 79, no. 5, pp. 1657-1664.

Bonissone, P and Cheetham, W 1997, 'Financial Applications Of Fuzzy Case-Based Reasoning To Residential Property Valuation', *Proceedings of 6th IEEE International Conference On Fuzzy Systems*, 1-5 July, Barcelona, Spain, IEEE, pp. 37-44.

Bourassa, S and Hoesli, M 1999, 'The structure of housing submarkets in a metropolitan region', *Ecole des Hautes Etudes Commerciales, Universite de Geneve-Papers*, vol. 99.15, 28pp.

Bourassa, S, Hoesli, M, Hamelink, F and MacGregor, B 1997, 'Defining housing sub-markets: evidence from Sydney and Melbourne', *RICS Cutting Edge*, Dublin.

Bryant, K 1991, 'The integration of qualitative factors into expert systems for evaluation agricultural loans', *10th Australian Conference on Information Systems*, 1-3 December, 1999, Wellington, New Zealand, pp.110-122.

Bui, T (ed) 1987, *Co-oP: a group decision support systems for cooperative multiple criteria group decision making*, Springer-Verlag, New York.

Cannistra, J 1999, 'Converting utility data for a GIS', *American Water Works Association*, vol. 91, no. 2, pp. 55-64.

Carver, S 1991, 'Integrating multi-criteria evaluation with geographical information systems', *International Journal of Geographical Information Systems*, vol. 5, no. 3, pp. 321-339.

Castle, G 1994, 'GIS In Real Estate Valuation', *GIS In Business '94 Conference proceedings*, San Francisco, California, GIS World Inc, pp. 133-136.

Chicoine, D 1981, 'Farmland values at the urban fringe: an analysis of sale prices', *Land Economics*, vol. 57, no. 3, pp. 353-362.

Clayton, D and Waters, N 1999, 'Distributed knowledge, distributed processing, distributed users: integrating case-based reasoning and GIS for multicriteria decision making', in *Spatial multicriteria decision making and analysis-a geographic information sciences approach*, Thill, J (ed), Ashgate Publishing Ltd, Aldershot, England, pp. 275-307.

Connellan, O and James, H 1998, 'Estimated realisation price (ERP) by neural networks: forecasting commercial property values', *Journal of Property Valuation and Investment*, vol. 16, no. 1, pp. 71-86.

Connolly, T, Begg, C and Strachan, A (eds) 1999, *Database Systems: A practical approach to design, implementation and management*, Addison-Wesley, England.

Cronan, T P, Epley, D R and Perry, L G 1986, 'The Use of Rank Transformation and Multiple Regression Analysis in Estimating Residential Property Values With A Small Sample', *Journal of Real Estate Research*, vol. 1, no. 1, pp. 19-31.

Croswell, P 2000, 'The role of standards in support of GDI', in *Geospatial data infrastructure: concepts, cases and good practice*, Groot, R and McLaughlin, J (eds), Oxford University Press, Oxford, pp. 57-83.

Czernkowski, R 1989, 'Expert systems in real estate valuation', *Journal of Valuation*, vol. 8, no. 4, pp. 376-393.

Dale-Johnson, D 1982, 'An alternative approach to housing marekt segmentation using hedonic price data', *Journal of Urban Economics*, vol. 11, no. 3, pp. 311-332.

Daniel, C and Wood, F 1980, *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd Edition, Wiley, New York.

Day, B 'Submarket identification in property markets: A hedonic housing price model for Glasgow,' CSERGE Working Paper EDM 03-09, Norwich, 44pp.

Department of Primary Industries. (2005). Dry Seasonal Conditions in Rural Victoria, Report 48, 23 June, 2005. [online report]. 2005 (15 October), Available: http://www.lawlex.com.au/files/LAWLEX%20Water%20Newsfeed.htm#1

Devogele, T, Parent, C and Spaccapietra, S 1998, 'On spatial database integration', *International Journal of Geographical Information Science*, vol. 12, no. 4, pp. 335-352.

Division of Property Assessments 1992, *Computer-Assisted Appraisal System Rural Land Procedures Manual*, Division of Property Assessments, Nashville, USA.

Do, Q and Grudnitski, G 1992, 'A neural network approach to residential property appraisal', *The Real Estate Appraiser*, vol. 58, December, pp. 38-45.

Donnelly, W 1988, 'Calculating the component value for floodplain property,' Australian National University, Centre for Resource and Environmental Studies, Canberra, CRES working paper, 12 pp.

Dunse, N, Leishman, C and Watkins, C 2001, 'Classifying office submarkets', *Journal of Property Investment and Finance*, vol. 19, no. 3, pp. 236-250.

Elad, R, Clifton, I and Epperson, J 1994, 'Hedonic estimation applied to the farmland market in Georgia', *Journal of Applied Economics*, vol. 26, no. 2, pp. 351-366.

Evans, A, James, H and Collins, A 1992, 'Artificial Neural Networks: an application to residential valuation in the UK', *Journal of Property Valuation and Investment*, vol. 11, June, pp. 195-204.

Everitt, B S, Landau, S and Leese, M 2001, *Cluster Analysis*, Arnold, London.

Faux, J and Perry, G 1999, 'Estimating irrigation water value using hedonic price analysis: a case study in Malheur county, Oregon', *Land Economics*, vol. 75, no. 3, pp. 440-452.

Fletcher, M, Gallimore, P and Mangan, J 2000, 'The modelling of housing submarkets', *Journal of Property Investment & Finance*, vol. 18, no. 4, pp. 473-487.

Foster, S and Hamilton, S 1991, 'Appraisal database and information', *Journal of Property Valuation and Investment*, vol. 10, no. 1, pp. 413-423.

Gallimore, P, Fletcher, M and Carter, M 1996, 'Modelling the influence of location on value', *Journal of Property Valuation and Investment*, vol. 14, no. 1, 6pp.

Gardner, K and Barrows, R 1984, 'The impact of soil conservation investments on land prices', *American Journal of Agricultural Economics*, vol. 67, December, pp. 943-947.

Goetzmann, W N, Spiegal, M and Wachter, S M 1998, 'Do cities and suburbs cluster?', *Cityscape: A Journal of Policy Development and Research*, vol. 3, no. 3, pp. 193-202.

Gonzalez, A and Laureano-Ortiz, R 1992, 'A case-based reasoning approach to real estate property appraisal', *Expert Systems With Applications*, vol. 4, pp. 229-246.

Goodchild, M 1995, 'Sharing imperfect data', in *Sharing geographic information*, Onsrud, H and Rushton, G (eds), Center for urban policy, New Brunswick, New Jersey, pp. 413-425.

Guariso, G and Werthner, H 1989, *Environmental Decision Support Systems*, in series 'Computers and their applications', Meek, B (ed). Ellis Horwood, Chichester, England.

Hardester, K P 2002, 'An enterprise GIS solution for integrating GIS and CAMA', *Assessment Journal*, vol. 9, no. 6, pp. 27-35.

Harrison, D and Rubinfield, D 1978, 'Hedonic housing prices and the demand for clean air', *Journal of Environmental Economics and Management*, vol. 5, pp. 81-102.

Hayles, K and Grenfell, R 2002, 'Data integration: an application in rural property valuation', *Cartography*, vol. 31, no. 2, pp. 39-50.

Holt, A and Benwell, G 1996, 'Case-based reasoning and spatial analysis', *URISA Journal*, vol. 8, no. 1, pp. 27-36.

IAAO 1999, 'Standard on Ratio Studies', *Assessment Journal*, September/October, pp. 23-65.

IAAO 2003, 'Standard on Automated Valuation Models (AVM's)', 35 pp.

Isakson, H 2001, 'Using multiple regression analysis in real estate appraisal', *Appraisal Journal*, vol. 69, no. 4, pp. 424-430.

Jankowski, P 1995, 'Integrating geographical information systems and multiple criteria decision-making methods', *International Journal of Geographical Information Systems*, vol. 9, no. 3, pp. 251-273.

John, S 1993, 'Data integration in a GIS - the question of data quality', *Aslib proceedings*, vol. 45, no. 4, pp. 109-119.

Just, R and Miranowski, J 1993, 'Understanding farmland price changes', *American Journal of Agricultural Economics*, vol. 75, no. 1, pp. 156-168.

Kershaw, P and Rossini, P 1999, 'Using neural networks to estimate constant quality house price indices', *Fifth Annual Pacific-Rim Real Estate Society Conference*, 26-30 January, 1999, Kuala Lumpur, Malaysia, pp. 1-8.

Kettani, O, Oral, M and Siskos, Y 1998, 'A multiple criteria analysis model for real estate evaluation', *Journal of Global Optimization*, vol. 12, no. 2, pp. 197-214.

Kim, T 1999, 'Metadata for geo-spatial data sharing: a comparative analysis', *The Annals of Regional Science*, vol. 33, no. 2, pp. 171-181.

Klein, M and Methlie, L 1990, *Expert Systems: A decision support approach: with applications in management and finance*, in series 'Insight series in Artificial Intelligence', Morgan, T (ed). Addison-Wesley Publishing Company, Workingham, England.

Kolodner, J (ed) 1993, *Case-based reasoning*, Morgan Kaufmann Publishers, San Mateo, California.

Kuhn, W 1994, 'Defining Semantics For Spatial Data Transfers', *Advances in GIS Research : Proceedings of The Sixth International Symposium On Spatial Data Handling*, vol. 2 of 2, Waugh, T and Healey, R (eds), Taylor & Francis, London, pp. 973-987.

Kwon, O and Kirby, E J 1997, *Farm Appraiser: a neural network for agricultural appraisal*, viewed 23 July, 2003 , <http://hsb.baylor.edu/ramsower/ais.ac.97/papers/kwon.htm>

Laaribi, A, Chevallier, J and Martel, J 1996, 'A spatial decision aid: a multicriterion evaluation approach', *Computers, Environment and Urban Systems*, vol. 20, no. 6, pp. 351-366.

Lake, I, Lovett, A, Bateman, I and Langford, I 1998, 'Modelling environmental influences on property prices in an urban environment', *Computers, Environment and Urban Systems*, vol. 22, no. 2, pp. 121-136.

Lake, I R, Lovett, A A, Bateman, I J and Day, B 2000, 'Using GIS and large-scale digital data to implement hedonic pricing studies', *International Journal of Geographical Information Science*, vol. 14, no. 6, pp. 521-541.

Land Victoria, Victorian Department of Sustainability and Environment, 2006; see http://www.land.vic.gov.au

Land Victoria 2004, 'Valuation Best Practice 2004 Specification Guidelines', Melbourne.

Land Victoria 2004, Revaluation, VBP 2004 Fact Sheet. [online]. 2005 (15 October), Available: http://www.melbourne.vic.gov.au/rsrc/PDFs/Revaluation/revaluation2004.pdf

Lenk, M, Worzala, E and Silva, A 1997, 'High-tech valuation: should artificial neural networks bypass the human valuer?', *Journal of Property Valuation and Investment*, vol. 15, no. 1, pp. 8-26.

Local Government Act 1989 (Act No. 11/1989)

Longhorn, R 1998, 'Data integration for commercial information products: experiences from the EC's IMPACT-2 programme', in *European Geographic Information Infrastructures: opportunities and pitfalls*, Burrough, P and Masser, I (eds), Taylor & Francis, London, pp. 101-111.

Longley, P, Higgs, G and Martin, D 1994, 'The predictive use of GIS to model property valuations', *International Journal of Geographical Information Systems*, vol. 8, no. 2, pp. 217-235.

Lunetta, R, Congalton, R, Fenstermaker, L, Jensen, J, McGwire, K and Tinney, L 1991, 'Remote sensing and geographic information system data integration: error sources and research issues', *Photogrammetric Engineering & Remote Sensing*, vol. 57, no. 6, pp. 677-687.

Mahan, B, Polasky, S and Adams, R 2000, 'Valuing urban wetlands: a property price approach', *Land Economics*, vol. 76, no. 1, pp. 100-113.

Malczewski, J 1999a, *GIS and multicriteria decision analysis*, John Wiley & Sons, New York.

Malczewski, J 1999b, 'Spatial multicriteria decision analysis', in *Spatial multicriteria decision making and analysis-a geographic information sciences approach*, Thill, J (ed), Ashgate Publishing Ltd, Aldershot, England, pp. 11-48.

Malczewski, J, Pazner, M and Zaliwska, M 1997, 'Visualization of multicriteria location analysis using raster GIS: a case study', *Cartography and Geographic Information Systems*, vol. 24, no. 2, pp. 80-90.

MapInfo 1997, *Arclink - bidirectional conversion utility for MapInfo and ARC/INFO files for Windows*, WWW document, Columbia University, viewed 23 July 2003, <http://www.columbia.edu/acis/eds/gis/mapinfo65_pdf/ARCLINK3.PDF>

Marano, W 2000, 'The market value of remnant native vegetation on rural holdings in a clearance regulated environment', *Pacific Rim Real Estate Society (PRRES) Conference*, Sydney, 23-27 January 2000, pp. 1-15.

Mathieson, K and Dreyer, B 1993, 'Improving the effectiveness and efficiency of appraisal reviews: an information systems approach', *Appraisal Journal*, vol. 63, no. 3, pp. 414-418.

MathSoft 1997, *SPlus 4 Guide to Statistics*, Data Analysis Products Division, MathSoft Inc, Seattle, Washington.

McCluskey, W 1996, 'Predictive accuracy of machine learning models for the mass appraisal of residential property', *New Zealand Valuers Journal*, July, pp. 41-47.

McCluskey, W and Anand, S 1999, 'The application of intelligent hybrid techniques for the mass appraisal of residential properties', *Journal of Property Valuation and Investment*, vol. 17, no. 3, pp. 218-238.

McCluskey, W and Deddis, W 'The application of spatially derived location factors within a GIS environment', 11pp.

McCluskey, W, Deddis, W, Mannis, A, McVurney, D and Borst, R 1997, 'Interactive application of computer assisted mass appraisal and geographic information systems', *Journal of Property Valuation and Investment*, vol. 15, no. 5, pp. 448-465.

McGreal, S, Adair, A, McBurney, D and Patterson, D 1998, 'Neural networks: the prediction of residential values', *Journal of Property Valuation and Investment*, vol. 16, no. 1, pp. 57-70.

McSherry, D 1993, 'Case-Based Reasoning Techniques For Estimation', *IEEE Colloquim On Case-Based Reasoning*, (eds), Colloquium Digest No. 1993/036, London, February 1993, 6/1-6/4, 4 pages.

Miranowski, J and Hammes, B 1984, 'Implicit prices of soil characteristics for farmland in Iowa', *American Journal of Agricultural Economics*, vol. 66, no. 5, pp. 745-749.

Montgomery, G and Schuch, H 1993, *GIS data conversion handbook*, GIS World, Fort Collins.

National Statistics 2005, National Statistics, United Kingdom, viewed 26 September, 2005,

<http://www.statistics.gov.uk/about/methodology_by_theme/area_classification/wards/methodology.asp>

Nattagh, N and Ross, D 2000, 'An updated appraisal of automated valuation', *Mortgage Banking*, vol. 61, no. 2, pp. 79-83.

Nawawi, A H, Jenkins, D and Gronow, S 1996, 'Computer assisted rating valuation of commercial and industrial properties in Malaysia,' RICS Research Paper Series, 22 pp.

Nind, C 2002, 'EMS and Land Valuation, the potential for land valuation to drive the adoption of environmental management systems in agriculture,' Department of Agriculture, Western Australia, report for RIRDC, publication number 00/040, 86 pp.

O'Roarty, B 1997, 'Identifying the impact of store space requirements on property selection using cluster analysis,' RICS Cutting Edge, 1997, Aberdeen, 9 pp.

O'Rourke, A 1998, *Automated valuation models - threat and opportunity*, WWW Document, viewed 23 July, 2003, <http://www.appraisaltoday.com/avms.htm>

Palmquist, R 1984, 'Estimating the demand for the characteristics of housing', *Review of Economics and Statistics*, vol. 66, August, pp. 394-404.

Palmquist, R B and Danielson, L E 1989, 'A hedonic study of the effects of erosion control and drainage on farmland values', *American Journal of Agricultural Economics*, vol. 71, February, pp. 55-62.

Paris, S, Ware, J, Jenkins, D and Wilson, I 2001, 'Adding value to desktop valuation: the use of artificial intelligence techniques to assist in residential property price forecasting', *Cutting Edge 2001*, 5-7 September 2001, Oxford.

Poor's, S 2004, 'Guidelines for the Use of Automated Valuation Models for U.K. RMBS Transactions', 6 pp.

Powe, N, Garrod, G, Brunsdon, C and Willis, K 1997, 'Using a geographic information system to estimate an hedonic price model of the benefits of woodland access', *Forestry*, vol. 70, no. 2, pp. 139-149.

Ralphs, M and Wyatt, P 1998, 'The application of geographic and land information systems to the management of local authority property', *Property Management*, vol. 16, no. 2, pp. 83-91.

Rayburn, W and Tosh, D 1995, 'Artificial intelligence: the future of appraising', *Appraisal Journal*, vol. 63, no. 4, pp. 429-437.

Reid Schott, L and White, F 1977, 'Multiple regression analysis of farmland values by land classes', *Appraisal Journal*, vol. 45, July, pp. 427-434.

Reynolds, J and Regalado, A 2002, 'The effects of wetlands and other factors on rural land values', *Appraisal Journal*, vol. 70, no. 2, pp. 182-190.

Roka, F and Palmquist, R 1997, 'Examining the use of national databases in a hedonic analysis of regional farmland values', *American Journal of Agricultural Economics*, vol. 79, no. 5, pp. 1651-1656.

Rosen, S 1974, 'Hedonic prices and implicit markets: product differentiation in pure competition', *Journal of Political Economy*, vol. 82, no. 1, pp. 34-55.

Rosiers, F, Theriault, M and Villeneuve, P 2000, 'Sorting out access and neighbourhood factors in hedonic price modelling', *Journal of Property Investment & Finance*, vol. 18, no. 3, pp. 291-315.

Ross, D and Nattagh, N 1996, 'The future of automated appraisals', *Mortgage Banking*, vol. 56, no. 11, pp. 59-62.

Rossini, P 1997, 'Artificial Neural Networks versus Multiple Regression in the valuation of residential property', *Australian Land Economics Review*, vol. 3, no. 1, pp. 1-10.

Rossini, P 1998, 'Improving the results of Artificial Neural Network models for residential valuation', *Fourth Annual Pacific Rim Real Estate Society Conference*, 19-21 January, 1998, Perth, Western Australia, 18 pp.

Rossini, P 1999, 'Accuracy issues for automated and artificial intelligent residential valuation systems', *International Real Estate Society Conference 1999*, 26-30 January, 1999, Kuala Lumpur, Malaysia, pp. 1-10.

Sauter, B 1985, 'Computers and comparable sales', in *Introduction to computer assisted valuation*, Woolery, A and Shea, S (eds), Oelgeschiager, Gunn & Hain Publishers Inc, Boston, pp. 141-147.

Shepherd, I 1991, 'Information integration and GIS', in *Geographical information systems: principles*, Maguire, D (ed), Longman Scientific and Technical, Essex, England, pp. 337-360.

Simon, H 1977, *The new science of management decision*, rev. edn, Prentice-Hall, Englewood Cliffs, New Jersey.

Slagle, R 1994, 'Standards for integration of multi-source and cross-media environmental data', in *Environmental information management and analysis: Ecosystem to global scales*, Michener, W, Brunt, J and Staford, S (eds), Taylor & Francis, London, pp. 221-233.

Smith, C A and Kroll, M J 1989, 'Utility theory and rent optimization: utilizing cluster analysis to segment rental markets', *The Journal of Real Estate Research*, vol. 4, no. 1, pp. 61-71.

Soto, P 2004, 'Spatial economietric analysis of Louisiana rural real estate values,' unpublished PhD thesis, *The department of Agricultural Economics and Agribusiness*, Louisiana State University and Agricultural and Mechanical College, 174 pp.

Suter, R 1974, *The appraisal of farm real estate*, The Interstate Printers & Publishers Inc., Danville, USA.

Tay, D and Ho, D 1991, 'Artificial intelligence and the mass appraisal of residential apartments', *Journal of Property Valuation and Investment*, vol. 10, pp. 525-540.

Theriault, M and Rosiers, F 2003, 'Modelling interactions of location with specific value of housing attributes', *Property Managment*, vol. 21, no. 1, pp. 25-62.

Thomas, G 2000, 'Achieving enterprise GIS data integration in Fairfax county using GIS everyday', *URISA*, 19-23 August, pp. 346-354.

Valentine, L 1999, 'Automated valuation models speed the appraisal process', *ABA Banking Journal*, vol. 91, no. 1, pp. 46-48.

Valuation of Land Act 1960

Vandeveer, L, Bruner, W, Henning, S, Niu, H and Kennedy, G 2000, 'Rural land values at the urban fringe', *Louisiana Agriculture*, vol. 44, no. 2, pp. 13-15.

Walker, J 1994, 'Accurate appraisals in a rural market: some problems and solutions', *Appraisal Journal*, vol. 62, no. 2, pp. 289-295.

Waller, B 1999, 'The impact of AVMs on the appraisal industry', *Appraisal Journal*, vol. 67, no. 3, pp. 287-292.

Ward, R D, Weaver, R and German, J G 1999, 'Improving CAMA models using Geographic Information Systems/Response surface analysis location factors', *Assessment Journal*, vol. 6, no. 1, pp. 30-38.

Watkins, C 1998, 'The definition and identification of housing submarkets', *Aberdeen Papers in Land Economy*, vol. 98-10, December 1998, 36 pp.

Watkins, C 1999, 'Property valuation and the structure of urban housing markets', *Journal of Property Investment and Finance*, vol. 17, no. 2, pp. 157-175.

West Gippsland Catchment Managment Authority 2001, West Gippsland Catchment Managment Authority, Traralgon, viewed 30 July, 2003, <http://www.wgcma.vic.gov.au>

Wilhelmsson, M 2004, 'A method to derive housing sub-markets and reduce spatial dependency', *Property Management*, vol. 22, no. 4, pp. 276-288.

Wimmera Catchment Management Authority 2000, Wimmera Catchment Management Authority, Horsham, viewed 30 July, 2003, <http://www.wca.vic.gov.au>

Wong, D and Wu, C 1996, 'Spatial metadata and GIS for decision support', *Proceedings of 29th Annual Hawaii International Conference on System Sciences*, vol. 3 of 5, Nunmaker, J F and Sprague, R H (eds), IEEE-CS Press, Washington, D.C., pp. 557-566.

Wooton, J (ed) 1989, *The glossary of property terms*, Estates Gazette Ltd, London.

Worzala, E, Lenk, M and Silva, A 1995, 'An exploration of neural networks and its application to real estate valuation', *The Journal of Real Estate Research*, vol. 10, no. 2, pp. 185-201.

Wyatt, P 1997, 'The development of a GIS-based property information system for real estate valuation', *International Journal of Geographical Information Science*, vol. 11, no. 5, pp. 435-450.

Xu, F, Mittelhammer, R and Barkley, P 1993, 'Measuring the contributions of site characteristics to the value of agricultural land', *Land Economics*, vol. 69, no. 4, pp. 356-369.

Zhu, X, Aspinall, R and Healey, R 1996, 'ILUDSS: a knowledge-based spatial decision support system for strategic land-use planning', *Computers and Electronics in Agriculture*, vol. 15, no. 4, pp. 279-301.

Personal Communication Notes

Dr Connie Spinoso is a Spatial Information Analyst for Land Victoria and is a part of the Valuation Best Practice Team.

Brett Reed is a valuer for K A Reed (Group) Pty Ltd and is a contract valuer for the Municipality of Kingston, Victoria, Australia.

Appendices

APPENDIX A: Property Valuation Database – Available Data

Area1	Area2	Area3	Area4	SALE_MONTH	ADJ_price	adjPRICEHA	PROP_AREA	SEVERITY	NATURAL	FOX	WATERB_PT	WATERB_AR	WATERCRS	ZONE_CODE	LSIO	TOWN_DIST	LUSE_1	LUSE_2	LUSE_3	LUSE_4	LUSE_5	LUSE_6
1	0	0	0	50	640581	1091	587	0	0	0	11	12544.781	2732.165	1	0	30709.000	1	0	0	0	0	0
1	0	0	0	66	26500	107	247	0	0	0	6	0.000	3004.000	1	0	20523.000	0	1	0	0	0	0
1	0	0	0	13	58135	1817	32	0	0	0	0	0.000	0.000	1	0	18472.264	0	1	0	0	0	0
1	0	0	0	38	60378	1285	47	0	0	0	0	760.078	589.234	1	0	32068.018	0	1	0	0	0	0
1	0	0	0	51	63674	368	173	0	0	0	3	0.000	5810.076	1	0	32839.766	0	1	0	0	0	0
1	0	0	0	63	65445	2111	31	0	0	0	0	0.000	0.000	1	0	36692.000	0	1	0	0	0	0
1	0	0	0	6	88640	1231	72	0	0	0	2	0.000	0.000	1	0	18888.000	0	1	0	0	0	0
1	0	0	0	36	102995	1084	95	0	0	0	3	0.000	1620.051	1	0	6251.557	0	1	0	0	0	0
1	0	0	0	11	110084	847	130	0	0	0	3	0.000	329.959	1	0	9934.829	0	1	0	0	0	0
1	0	0	0	47	107329	1073	100	0	0	0	3	0.000	60.772	1	0	11370.061	0	1	0	0	0	0
1	0	0	0	26	135516	1034	131	0	0	0	6	0.000	3195.130	1	0	23595.000	0	1	0	0	0	0
1	0	0	0	47	139585	1082	129	0	0	0	3	0.000	45.899	1	0	6651.965	0	1	0	0	0	0
1	0	0	0	42	150348	1030	146	0	0	0	4	0.000	0.000	1	0	17856.000	0	1	0	0	0	0
1	0	0	0	59	151900	1726	88	0	0	0	2	0.000	807.241	1	0	13739.000	0	1	0	0	0	0
1	0	0	0	32	159512	1286	124	0	0	0	4	0.000	0.000	1	0	26685.334	0	1	0	0	0	0
1	0	0	0	27	166347	2446	68	0	0	0	1	0.000	1299.625	1	0	19562.387	0	1	0	0	0	0
1	0	0	0	27	177306	1396	127	0	0	0	2	0.000	4751.721	1	0	17300.123	0	1	0	0	0	0
1	0	0	0	27	180551	1389	130	0	0	0	3	0.000	4751.721	1	0	18251.143	0	1	0	0	0	0
1	0	0	0	2	194870	1188	164	0	0	0	1	0.000	0.000	1	0	22713.000	0	1	0	0	0	0
1	0	0	0	59	176000	1354	130	0	0	0	3	0.000	364.463	1	0	17097.598	0	1	0	0	0	0
1	0	0	0	38	208200	4526	46	0	0	0	5	0.000	4687.000	1	1	25559.000	0	1	0	0	0	0
1	0	0	0	26	212572	2725	78	0	0	0	1	0.000	472.000	1	0	22692.000	0	1	0	0	0	0
1	0	0	0	39	252443	5488	46	0	0	0	4	0.000	1823.496	1	0	6048.919	0	1	0	0	0	0
1	0	0	0	52	301835	2415	125	0	0	1	2	0.000	2201.072	1	0	30134.176	0	1	0	0	0	0
1	0	0	0	27	317796	2445	130	0	0	0	1	0.000	1445.688	1	0	20397.107	0	1	0	0	0	0
1	0	0	0	48	336856	2079	162	0	0	0	6	0.000	3731.417	1	0	14471.000	0	1	0	0	0	0
1	0	0	0	27	521583	2449	213	0	0	0	3	2065.000	542.442	1	0	25647.000	0	1	0	0	0	0
1	0	0	0	43	220805	5135	43	0	0	0	2	0.000	521.283	1	0	9554.401	0	0	1	0	0	0
1	0	0	0	17	147980	1138	130	3	0	0	4	2830.344	835.177	1	0	27425.902	0	0	0	1	0	0
1	0	0	0	63	123163	955	129	0	0	0	5	0.000	1018.767	1	0	5906.596	0	0	0	0	1	0
1	0	0	0	42	10270	311	33	0	0	0	1	0.000	0.000	1	0	18539.471	0	0	0	0	0	1
1	0	0	0	51	22029	173	127	0	0	0	1	0.000	0.000	1	0	18220.596	0	0	0	0	0	1
1	0	0	0	7	59932	856	70	0	0	0	0	0.000	0.000	1	0	23154.963	0	0	0	0	0	1
1	0	0	0	40	64542	2689	24	0	0	0	0	0.000	0.000	1	0	32219.295	0	0	0	0	0	1
1	0	0	0	23	93170	3451	27	0	0	0	0	3497.000	293.992	1	0	38523.406	0	0	0	0	0	1
1	0	0	0	6	127420	5309	24	0	0	0	0	0.000	239.626	1	1	26856.832	0	0	0	0	0	1
1	0	0	0	24	124920	1096	114	0	0	0	3	2694.375	0.000	1	0	17931.930	0	0	0	0	0	1
1	0	0	0	21	130125	1942	67	0	0	0	3	5512.156	0.000	1	0	26850.947	0	0	0	0	0	1
1	0	0	0	47	147391	1143	129	0	0	0	3	0.000	690.206	1	0	17444.000	0	0	0	0	0	1
1	0	0	0	25	161355	3667	44	0	0	0	5	0.000	2311.158	0	0	27985.615	0	0	0	0	0	1
1	0	0	0	40	187380	1320	142	0	0	0	3	0.000	351.029	1	0	16081.000	0	0	0	0	0	1
1	0	0	0	24	197790	334	593	3	1	0	10	0.000	0.000	1	0	19376.230	0	0	0	0	0	1
1	0	0	0	52	197033	758	260	0	0	0	4	0.000	5155.332	1	0	17545.000	0	0	0	0	0	1
1	0	0	0	36	226508	1756	129	0	0	0	5	0.000	739.689	1	0	5849.670	0	0	0	0	0	1
1	0	0	0	50	230012	5349	43	0	0	0	3	0.000	3313.521	1	0	9972.504	0	0	0	0	0	1
1	0	0	0	59	278995	1112	251	0	0	0	2	0.000	4050.100	1	0	21086.000	0	0	0	0	0	1
0	1	0	0	65	141375	2142	66	0	0	0	0	0.000	0.000	1	0	82896.000	1	0	0	0	0	0
0	1	0	0	63	50000	467	107	3	0	0	0	469.000	2797.000	1	0	82209.000	0	1	0	0	0	0
0	1	0	0	33	27587	766	36	0	0	0	0	0.000	1037.000	1	0	51691.000	0	0	0	0	0	1
0	1	0	0	2	48198	1236	39	3	0	0	2	0.000	315.000	1	0	55519.000	0	0	0	0	0	1
0	1	0	0	7	146256	6965	21	0	0	0	0	0.000	414.000	1	0	63438.000	0	0	0	0	0	1
0	1	0	0	24	211400	4805	44	3	0	0	0	0.000	371.000	1	1	79038.000	0	0	0	0	0	1
0	0	1	0	63	30210	944	32	0	0	0	0	0.000	0.000	1	0	58736.000	0	1	0	0	0	0
0	0	1	0	63	34800	740	47	0	0	0	2	0.000	0.000	1	0	58134.000	0	1	0	0	0	0
0	0	1	0	25	42681	166	257	0	0	0	3	0.000	1459.000	1	0	150984.000	0	1	0	0	0	0
0	0	1	0	47	56188	1147	49	0	0	0	1	0.000	0.000	1	0	63146.000	0	1	0	0	0	0
0	0	1	0	63	58805	933	63	0	0	0	0	0.000	0.000	1	0	57897.000	0	1	0	0	0	0
0	0	1	0	30	66541	876	76	0	0	0	1	414.000	0.000	1	0	25746.000	0	1	0	0	0	0

Area1	Area2	Area3	Area4	SALE_MONTH	ADJ_price	adjPRICEHA	PROP_AREA	SEVERITY	NATURAL	FOX	WATERB_PT	WATERB_AR	WATERCRS	ZONE_CODE	LSIO	TOWN_DIST	LUSE_1	LUSE_2	LUSE_3	LUSE_4	LUSE_5	LUSE_6
0	0	1	0	52	72432	1207	60	0	0	0	0	0.000	0.000	1	0	93286.000	0	1	0	0	0	0
0	0	1	0	52	77436	662	117	0	0	0	0	2585.000	519.000	1	0	47639.000	0	1	0	0	0	0
0	0	1	0	51	78278	1864	42	0	0	0	0	0.000	751.000	1	0	32976.000	0	1	0	0	0	0
0	0	1	0	5	85965	384	224	0	0	0	2	0.000	3460.000	1	0	144555.000	0	1	0	0	0	0
0	0	1	0	63	96300	747	129	0	0	0	4	0.000	815.000	1	0	56860.000	0	1	0	0	0	0
0	0	1	0	51	101837	1886	54	0	0	0	0	384.000	64.000	1	0	32547.000	0	1	0	0	0	0
0	0	1	0	48	103788	1028	101	0	0	0	0	0.000	507.000	1	0	72255.000	0	1	0	0	0	0
0	0	1	0	52	107075	1647	65	0	0	0	0	1337.000	113.000	1	0	37090.000	0	1	0	0	0	0
0	0	1	0	48	109692	1025	107	0	0	0	0	3245.000	81.000	1	0	71921.000	0	1	0	0	0	0
0	0	1	0	4	127930	630	203	0	0	0	3	0.000	41.000	1	0	75943.000	0	1	0	0	0	0
0	0	1	0	53	119740	611	196	0	0	0	4	1191.000	1906.000	1	0	116824.000	0	1	0	0	0	0
0	0	1	0	43	130125	1735	75	0	0	0	1	0.000	43.000	1	0	28178.000	0	1	0	0	0	0
0	0	1	0	49	151701	570	266	0	0	0	2	2755.000	1149.000	1	0	96278.000	0	1	0	0	0	0
0	0	1	0	46	166560	1343	124	0	0	0	1	0.000	1715.000	1	0	35308.000	0	1	0	0	0	0
0	0	1	0	40	171109	1133	151	0	0	0	1	0.000	1267.000	1	0	47980.000	0	1	0	0	0	0
0	0	1	0	46	171765	1342	128	0	0	0	3	0.000	945.000	1	0	34497.000	0	1	0	0	0	0
0	0	1	0	15	176652	1436	123	0	0	0	1	0.000	710.000	1	0	40543.000	0	1	0	0	0	0
0	0	1	0	41	183253	1001	183	0	0	0	2	0.000	1074.000	1	0	80583.000	0	1	0	0	0	0
0	0	1	0	47	196975	1931	102	0	0	0	2	0.000	922.000	1	0	29296.000	0	1	0	0	0	0
0	0	1	0	46	207218	1041	199	0	0	0	2	0.000	2886.000	1	0	76252.000	0	1	0	0	0	0
0	0	1	0	51	236510	466	508	0	0	0	1	2744.000	4436.000	1	0	137267.000	0	1	0	0	0	0
0	0	1	0	3	288080	2216	130	0	0	0	3	0.000	881.000	1	0	59479.000	0	1	0	0	0	0
0	0	1	0	53	280217	2156	130	0	0	0	1	0.000	0.000	1	0	37389.000	0	1	0	0	0	0
0	0	1	0	51	314539	939	335	0	0	0	4	0.000	2602.000	1	0	58488.000	0	1	0	0	0	0
0	0	1	0	59	133510	1043	128	1	1	0	3	0.000	669.000	1	0	42021.000	0	0	0	0	1	0
0	0	1	0	9	206197	1611	128	0	0	0	2	0.000	1666.000	1	0	35982.000	0	0	0	0	0	1
0	0	1	0	26	10047	386	26	0	0	0	0	0.000	7.000	1	0	82171.000	0	0	0	0	0	1
0	0	1	0	14	13952	78	179	0	0	0	0	0.000	0.000	1	0	123186.000	0	0	0	0	0	1
0	0	1	0	55	91925	355	259	0	0	0	3	0.000	1258.000	1	0	151612.000	0	0	0	0	0	1
0	0	1	0	33	117384	524	224	0	0	0	2	1281.000	2546.000	1	0	123612.000	0	0	0	0	0	1
0	0	1	0	15	131295	1728	76	0	0	0	0	0.000	0.000	1	0	30650.000	0	0	0	0	0	1
0	0	1	0	37	134914	347	389	0	0	0	4	0.000	3166.000	1	0	124738.000	0	0	0	0	0	1
0	0	1	0	8	212661	1649	129	0	0	0	0	0.000	209.000	1	0	45264.000	0	0	0	0	0	1
0	0	1	0	31	251897	2311	109	0	0	0	1	0.000	2291.000	1	0	39801.000	0	0	0	0	0	1
0	0	0	1	50	253669	2987.15	85	0	0	0	3	0	13939	1	0	52977.000	1	0	0	0	0	0
0	0	0	1	4	208304	10373.71	20	0	0	0	1	0	646	1	0	9820.000	0	0	1	0	0	0
0	0	0	1	48	140535	6313.34	22	0	0	0	1	0	871	1	0	33661.000	0	0	1	0	0	0
0	0	0	1	5	380266	11561.74	33	0	0	0	2	0	1212	1	0	21027.000	0	0	1	0	0	0
0	0	0	1	15	433370	10447.69	41	0	0	0	0	0	1282	1	0	14873.000	0	0	1	0	0	0
0	0	0	1	20	608832	10448.46	58	0	0	0	2	3308	1108	1	0	16234.000	0	0	1	0	0	0
0	0	0	1	2	683968	11046.00	62	0	0	0	0	0	584	1	0	21348.000	0	0	1	0	0	0
0	0	0	1	17	317100	5007.11	63	0	0	0	1	1321	2230	1	0	59861.000	0	0	1	0	0	0
0	0	0	1	1	236004	10415.00	23	0	0	0	2	0	405	1	0	10097.000	0	0	0	1	0	0
0	0	0	1	2	105260	3587.59	29	0	0	0	1	0	438	1	0	10783.000	0	0	0	1	0	0
0	0	0	1	20	369950	12169.41	30	0	0	0	2	0	2120	1	0	3542.000	0	0	0	1	0	0
0	0	0	1	32	145740	4770.54	31	0	0	0	1	0	0	1	0	16899.000	0	0	0	1	0	0
0	0	0	1	4	124096	3735.58	33	0	0	0	1	2308	1359	1	0	22477.000	0	0	0	1	0	0
0	0	0	1	2	121880	3077.78	40	0	0	0	5	0	767	1	0	71298.000	0	0	0	1	0	0
0	0	0	1	35	83280	1978.62	42	0	0	0	1	0	1475	1	0	22791.000	0	0	0	1	0	0
0	0	0	1	1	186144	4380.89	42	0	0	0	5	0	639	1	0	35850.000	0	0	0	1	0	0
0	0	0	1	24	285390	6683.61	43	0	0	0	0	0	0	1	0	15028.500	0	0	0	1	0	0
0	0	0	1	2	216060	3296.61	66	0	0	0	2	0	679	1	0	7906.000	0	0	0	1	0	0
0	0	0	1	8	170521	2596.24	66	0	0	0	2	0	980	1	0	27487.000	0	0	0	1	0	0
0	0	0	1	39	468450	6652.23	70	0	0	0	2	0	911	1	0	3089.000	0	0	0	1	0	0
0	0	0	1	19	132125	1736.66	76	0	0	0	0	0	992	1	0	61829.000	0	0	0	1	0	0
0	0	0	1	2	368964	4449.10	83	0	0	0	2	1372	5097	1	0	67396.000	0	0	0	1	0	0
0	0	0	1	8	609400	6936.03	88	0	0	0	2	0	4067	1	0	15881.000	0	0	0	1	0	0
0	0	0	1	8	96485	766.60	126	0	0	0	1	0	0	1	0	33733.000	0	0	0	1	0	0
0	0	0	1	64	169000	3940.31	43	0	0	0	1	0	320	1	0	30836.000	0	0	0	0	1	0
0	0	0	1	16	406945	20115.92	20	0	0	0	1	0	932	1	0	61914.000	0	0	0	0	0	1
0	0	0	1	43	58296	2824.42	21	0	0	0	0	0	27	1	0	59907.000	0	0	0	0	0	1
0	0	0	1	14	52850	2184.79	24	0	0	0	1	0	199	1	0	19444.000	0	0	0	0	0	1
0	0	0	1	20	103058	3877.26	27	0	0	0	0	0	984	1	0	3069.000	0	0	0	0	0	1

Area1	Area2	Area3	Area4	SALE_MONTH	ADJ_price	adjPRICEHA	PROP_AREA	SEVERITY	NATURAL	FOX	WATERB_PT	WATERB_AR	WATERCRS	ZONE_CODE	LSIO	TOWN_DIST	LUSE_1	LUSE_2	LUSE_3	LUSE_4	LUSE_5	LUSE_6
0	0	0	1	32	51321	1909.99	27	0	0	0	3	0	584	1	0	35785.000	0	0	0	0	0	1
0	0	0	1	26	65063	2374.54	27	0	0	0	1	0	699	1	0	71411.000	0	0	0	0	0	1
0	0	0	1	14	73990	2690.55	28	0	0	0	1	0	1658	1	0	34620.000	0	0	0	0	0	1
0	0	0	1	15	105700	3684.21	29	0	0	0	1	0	0	1	1	32628.000	0	0	0	0	0	1
0	0	0	1	52	30810	1028.71	30	0	0	0	0	0	2157	1	0	71242.000	0	0	0	0	0	1
0	0	0	1	3	91410	2931.69	31	0	0	0	1	0	1364	1	0	24502.000	0	0	0	0	0	1
0	0	0	1	61	126500	4009.51	32	0	0	0	2	0	1699	1	0	13740.000	0	0	0	0	0	1
0	0	0	1	27	74941	2315.12	32	0	0	0	1	0	499	1	0	17766.000	0	0	0	0	0	1
0	0	0	1	55	436475	12117.57	36	0	0	0	0	0	2188	1	0	24371.000	0	0	0	0	0	1
0	0	0	1	41	429659	11448.42	38	0	0	0	1	0	853	1	0	34529.000	0	0	0	0	0	1
0	0	0	1	51	369720	8381.77	44	0	0	0	1	3598	1867	1	0	24948.000	0	0	0	0	0	1
0	0	0	1	9	238220	5052.39	47	0	0	0	1	7906	92	1	0	29001.000	0	0	0	0	0	1
0	0	0	1	33	88485	1822.18	49	0	0	0	3	1667	1693	1	0	36940.000	0	0	0	0	0	1
0	0	0	1	33	817185	11408.42	72	0	0	0	1	32491	1852	1	0	25212.000	0	0	0	0	0	1
0	0	0	1	4	288080	3642.89	79	0	0	0	1	16943	1618	1	0	2501.000	0	0	0	0	0	1
0	0	0	1	16	380520	4018.59	95	0	0	0	5	7954	2763	1	0	31400.000	0	0	0	0	0	1
0	0	0	1	5	109138	953.17	115	0	0	0	0	0	4834	1	0	38073.000	0	0	0	0	0	1
0	0	0	1	7	288080	2259.45	128	0	0	0	3	12028	4419	1	0	29125.000	0	0	0	0	0	1
0	0	0	1	33	353940	2741.81	129	0	0	0	2	0	6310	1	0	58621.000	0	0	0	0	0	1
0	0	0	1	41	325417	2038.57	160	0	0	0	3	0	1516	1	0	22325.000	0	0	0	0	0	1
0	0	0	1	41	361019	2184.02	165	0	0	0	4	0	104	1	0	15947.000	0	0	0	0	0	1
0	0	0	1	62	436000	2471.09	176	0	0	0	0	0	4759	1	0	73482.000	0	0	0	0	0	1

APPENDIX B: Standardised available data for Cluster Analysis

Note: ID, NEWID, PROP_NO and AREA_TYPE are variables used for identification only and were not used within the clustering algorithms

ID	NEWID	PROP_NO	AREA_TYPE	L_USE	ADJ_PRICE	PROP_AREA	WATERB_PT	WATERB_AR	WATERCRS	TOWNDIST	3 Clusters	4 Clusters
1	26	26	1	6	-0.27135	-0.30462	-0.25000	0.0000	-0.22223	0.021308	2	2
2	4	4	1	6	-0.28868	-0.04043	-0.25000	0.0000	-0.22223	-0.11219	2	2
3	1	1	1	2	0.101137	0.499426	0.00000	0.0000	-0.22223	-0.1187	3	4
4	2	2	1	2	-0.21247	-0.02895	0.25000	0.0000	-0.22223	-0.17503	3	4
5	28	28	1	6	-0.44144	-0.25293	0.00000	0.0000	-0.22223	-0.18016	2	2
6	29	29	1	2	0.004709	0.396049	0.75000	0.0000	-0.22223	-0.19023	3	4
7	10	10	1	6	-0.08163	0.212267	0.50000	1.0123	-0.22223	-0.18911	2	2
8	7	7	1	4	-0.01621	0.304158	0.75000	1.0634	0.006591	-0.04929	2	3
9	34	34	1	2	0.598737	0.487939	1.25000	0.0000	0.800083	-0.24008	3	4
10	31	31	1	6	-0.00471	0.298415	0.50000	0.0000	-0.03313	-0.1963	2	2
11	21	21	1	6	0.237577	0.298415	1.00000	0.0000	-0.01957	-0.36706	2	2
12	22	22	1	2	-0.15052	0.103147	0.50000	0.0000	0.221625	-0.36114	3	4
13	13	13	1	2	-0.04833	0.309901	1.25000	0.0000	0.653155	-0.10571	3	4
14	47	47	1	2	-0.38747	0.976108	1.25000	0.0000	0.60079	-0.15095	3	4
15	15	15	1	2	0.093173	0.304158	0.50000	0.0000	1.079619	-0.18441	3	4
16	16	16	1	2	0.082977	0.286929	0.25000	0.0000	1.079619	-0.19842	3	4
17	5	5	1	2	-0.13348	0.304158	0.50000	0.0000	-0.13183	-0.30689	3	4
18	32	32	1	2	-0.13231	0.131863	0.50000	0.0000	-0.20558	-0.28575	3	4
19	33	33	1	2	-0.02957	0.298415	0.50000	0.0000	-0.20965	-0.35524	3	4
20	44	44	1	5	-0.07128	0.298415	1.00000	0.0000	0.05689	-0.36622	2	3
21	35	35	1	6	0.25844	-0.1955	0.50000	0.0000	0.685591	-0.30634	2	2
22	25	25	1	2	0.319067	-0.17827	0.75000	0.0000	0.277364	-0.36412	3	4
23	30	30	1	3	0.229115	-0.1955	0.25000	0.0000	-0.07941	-0.31249	2	3
24	41	41	1	2	0.022715	0.062945	0.25000	0.0000	-0.00106	-0.25086	3	4
25	42	42	1	6	0.438442	0.999081	0.25000	0.0000	0.887394	-0.14266	2	2
26	27	27	1	6	0.11463	0.373076	0.50000	0.0000	-0.12605	-0.21637	2	2
27	39	39	1	6	0.153401	1.05077	0.75000	0.0000	1.190198	-0.19481	1	2
28	43	43	1	2	0.101546	0.304158	0.50000	0.0000	-0.12237	-0.2014	3	4

ID	NEWID	PROP_NO	AREA_TYPE	L_USE	ADJ_PRICE	PROP_AREA	WATERB_PT	WATERB_AR	WATERCRS	TOWNDIST	3 Clusters	4 Clusters
29	6	6	1	2	-0.29424	-0.25867	-0.25000	0.0000	-0.22223	-0.18115	3	4
30	14	14	1	2	0.193788	0.005513	0.00000	0.0000	-0.09291	-0.11901	3	4
31	11	11	1	6	0.14734	2.963244	2.25000	0.0000	-0.22223	-0.16784	1	1
32	37	37	1	6	-0.40399	0.286929	0.00000	0.0000	-0.22223	-0.18486	2	2
33	40	40	1	2	0.487197	0.275442	0.25000	0.0000	0.380809	-0.0094	3	4
34	17	17	1	2	0.04854	-0.05192	0.00000	0.0000	0.133837	-0.1651	3	4
35	18	18	1	2	0.52442	0.304158	0.00000	0.0000	0.173854	-0.15281	3	4
36	9	9	1	6	-0.18139	-0.28739	-0.25000	1.3139	-0.14168	0.114153	2	2
37	38	38	1	2	-0.27135	0.551114	0.50000	0.0000	1.36958	0.030446	3	4
38	23	23	1	2	-0.28443	-0.17252	-0.25000	0.2856	-0.06079	0.01908	3	4
39	12	12	1	6	0.032855	-0.18975	1.00000	0.0000	0.41097	-0.04104	2	2
40	3	3	1	6	-0.09798	-0.30462	-0.25000	0.0000	-0.15657	-0.05767	2	2
41	24	24	1	2	0.18005	-0.17827	1.00000	0.0000	1.061887	-0.07678	3	4
42	19	19	1	2	1.16475	0.780841	0.50000	0.7759	-0.07361	-0.07549	3	4
43	20	20	1	2	0.027066	0.269699	0.75000	0.0000	-0.22223	-0.06019	3	4
44	8	8	1	6	-0.06527	-0.05766	0.50000	2.0710	-0.22223	-0.05776	2	2
45	46	46	1	6	-0.17158	0.694693	-0.25000	0.0000	0.091259	-0.03441	2	2
46	36	36	1	1	1.566101	2.928785	2.50000	4.7133	0.526315	-0.00094	1	1
47	45	45	1	2	-0.26008	-0.26442	-0.25000	0.0000	-0.22223	0.08718	3	4
48	48	2	2	6	-0.38747	-0.2357	-0.25000	0.0000	0.061885	0.308081	2	2
49	50	4	2	6	-0.33186	-0.21847	0.25000	0.0000	-0.13592	0.364458	2	2
50	51	5	2	6	0.18005	-0.18975	-0.25000	0.0000	-0.12058	0.710838	2	2
51	52	8	2	6	-0.04238	-0.32185	-0.25000	0.0000	-0.1088	0.481087	2	2
52	53	9	2	1	-0.01171	-0.0634	-0.25000	0.0000	-0.22223	0.767658	2	3
53	54	10	2	2	-0.3106	0.172065	-0.25000	0.1762	0.544078	0.75754	3	4
54	55	2	3	2	-0.24345	-0.09786	-0.25000	0.0000	-0.22223	0.920678	3	4
55	56	3	3	2	-0.37533	-0.25867	-0.25000	0.0000	-0.22223	0.411837	3	4
56	57	4	3	2	-0.2818	-0.08063	-0.25000	0.0000	-0.22223	0.399481	3	4
57	58	5	3	6	-0.06784	-0.00597	-0.25000	0.0000	-0.22223	-0.0018	2	2
58	59	6	3	6	-0.43097	0.585573	-0.25000	0.0000	-0.22223	1.361036	1	1
59	60	7	3	2	0.418342	0.304158	0.00000	0.0000	-0.22223	0.097446	3	4

ID	NEWID	PROP_NO	AREA_TYPE	L_USE	ADJ_PRICE	PROP_AREA	WATERB_PT	WATERB_AR	WATERCRS	TOWNDIST	3 Clusters	4 Clusters
60	61	8	3	2	-0.2976	-0.16104	0.00000	0.0000	-0.22223	0.476786	3	4
61	62	10	3	2	-0.36032	-0.17252	0.25000	0.0000	-0.22223	0.402971	3	4
62	63	11	3	2	-0.26507	-0.00597	0.00000	0.1555	-0.22223	-0.07403	3	4
63	64	13	3	6	-0.18137	1.045026	0.50000	0.0000	0.122433	1.779684	1	1
64	65	14	3	2	-0.34004	1.03354	0.50000	0.0000	0.177502	1.770435	1	1
65	66	15	3	2	-0.22037	0.844016	0.25000	0.0000	0.725722	1.675751	1	1
66	67	16	3	2	0.279135	2.475075	0.00000	1.0310	0.99312	1.568416	1	1
67	69	18	3	6	-0.05023	1.791638	0.75000	0.0000	0.645174	1.383893	1	1
68	70	19	3	6	-0.10531	0.844016	0.25000	0.4813	0.47531	1.36731	1	1
69	71	20	3	2	-0.09278	0.683207	0.75000	0.4475	0.299968	1.267338	3	4
70	72	21	3	2	0.00902	1.085229	0.25000	1.0351	0.09257	0.964743	3	4
71	73	22	3	2	0.176966	0.700436	0.25000	0.0000	0.568461	0.669807	3	4
72	74	24	3	2	-0.09648	0.723409	0.50000	0.0000	-0.21099	0.665256	3	4
73	75	25	3	2	-0.12948	0.172065	-0.25000	1.2192	-0.20003	0.606022	3	4
74	76	26	3	2	-0.14803	0.137606	-0.25000	0.0000	-0.08332	0.610941	3	4
75	77	27	3	2	0.101664	0.608546	0.25000	0.0000	0.072022	0.733593	3	4
76	78	28	3	2	0.37631	0.304158	0.50000	0.0000	0.019145	0.42278	3	4
77	79	29	3	2	-0.15915	0.298415	0.75000	0.0000	0.001063	0.384208	3	4
78	80	30	3	2	0.527659	1.481507	0.75000	0.0000	0.490653	0.408185	3	4
79	81	31	3	6	-0.44258	-0.29313	-0.25000	0.0000	-0.22031	0.75698	2	2
80	82	32	3	2	0.07252	0.263956	0.00000	0.0000	-0.0277	0.143897	3	4
81	83	33	3	2	-0.13311	-0.06915	-0.25000	0.5023	-0.19127	0.093042	3	4
82	84	34	3	2	-0.22483	-0.20124	-0.25000	0.0000	-0.01647	0.032452	3	4
83	84	35	3	2	-0.1498	-0.13232	-0.25000	0.1443	-0.20469	0.026134	3	4
84	85	36	3	2	0.144779	0.143349	0.25000	0.0000	0.030378	-0.02175	3	4
85	86	38	3	2	-0.06527	-0.01172	0.00000	0.0000	-0.21044	-0.03821	3	4
86	87	39	3	2	0.063505	0.424765	0.00000	0.0000	0.124899	0.253426	3	4
87	88	41	3	5	-0.04892	0.292672	0.50000	0.0000	-0.03894	0.165664	2	3
88	89	42	3	6	0.15366	0.298415	-0.25000	0.0000	-0.16496	0.213426	2	2
89	90	43	3	6	0.317353	0.183552	0.00000	0.0000	0.405447	0.132969	2	2
90	91	44	3	2	-0.22752	0.229497	-0.25000	0.9712	-0.08003	0.248404	3	4

ID	NEWID	PROP_NO	AREA_TYPE	L_USE	ADJ_PRICE	PROP_AREA	WATERB_PT	WATERB_AR	WATERCRS	TOWNDIST	3 Clusters	4 Clusters
91	92	45	3	1	0.134577	0.292672	0.25000	0.0000	0.234214	0.076724	2	3
92	93	46	3	2	0.065565	0.292672	0.50000	0.0000	0.036679	0.054853	3	4
93	93	47	3	2	0.04921	0.269699	0.00000	0.0000	0.247639	0.066797	3	4
94	1	2	4	4	0.40902	-0.19722	-0.25000	0.0000	-0.22223	-0.23537	2	3
95	1	13	4	4	0.40902	-0.19722	-0.25000	0.0000	0.00709	-0.22837	2	3
96	2	3	4	4	-0.18931	0.280381	0.00000	0.0000	-0.22223	0.043601	2	3
97	3	4	4	6	0.703409	0.101367	0.00000	0.0000	-0.22223	0.004691	2	2
98	3	30	4	6	0.703409	0.101367	0.00000	0.0000	-0.08332	0.012467	2	2
99	3	31	4	6	0.703409	0.101367	0.00000	0.0000	-0.18223	0.004573	2	2
100	4	5	4	4	0.075378	-0.19843	0.50000	0.0000	-0.22223	0.076974	2	3
101	4	58	4	4	0.075378	-0.19843	0.25000	0.0000	-0.04716	0.0726	2	3
102	5	6	4	6	-0.14705	-0.27768	0.00000	0.0000	-0.22223	0.027327	2	2
103	6	7	4	4	-0.01621	-0.267	0.00000	0.0000	-0.22223	-0.20432	2	3
104	7	8	4	6	-0.29097	-0.32391	-0.25000	0.0000	-0.21483	0.429083	2	2
105	8	9	4	6	-0.37602	-0.27045	-0.25000	0.0000	0.368735	0.596021	2	2
106	9	10	4	6	-0.15196	0.215139	-0.25000	0.0000	1.102161	0.107519	2	2
107	10	11	4	3	1.545036	-0.08684	-0.25000	0.0000	-0.06222	-0.1388	2	3
108	11	12	4	3	0.866959	-0.20423	-0.25000	0.0000	0.129008	-0.23416	2	3
109	12	14	4	6	-0.15523	-0.2898	-0.25000	0.0000	0.047364	-0.40801	2	2
110	13	15	4	6	0.952005	0.570871	-0.25000	0.0000	1.081613	0.629011	2	2
111	14	16	4	6	0.660232	0.506892	0.75000	0.0000	-0.19373	-0.21835	2	2
112	15	17	4	5	0.078649	-0.19613	0.00000	0.0000	-0.13455	0.000935	2	3
113	16	18	4	4	-0.21247	-0.20072	0.00000	0.0000	0.181885	-0.11755	2	3
114	17	19	4	6	-0.23867	-0.25655	0.00000	0.0000	-0.08551	-0.19156	2	2
115	18	20	4	6	-0.24518	-0.28452	0.00000	0.0000	0.232022	0.056665	2	2
116	19	21	4	6	0.875912	-0.22691	0.00000	0.0000	0.011474	0.055324	2	2
117	20	22	4	3	-0.03256	-0.31461	0.00000	0.0000	0.016405	0.042541	2	3
118	21	23	4	4	-0.1634	-0.27395	0.00000	0.0000	-0.10223	-0.2944	2	3
119	22	24	4	4	-0.06527	-0.00551	-0.25000	0.0000	0.049556	0.45739	2	3
120	23	25	4	6	0.916024	-0.23558	-0.25000	0.0000	0.377228	-0.09428	2	2
121	24	26	4	6	-0.19611	-0.16357	0.00000	0.0000	-0.02277	0.094014	2	2

ID	NEWID	PROP_NO	AREA_TYPE	L_USE	ADJ_PRICE	PROP_AREA	WATERB_PT	WATERB_AR	WATERCRS	TOWNDIST	3 Clusters	4 Clusters
122	25	27	4	6	-0.20429	-0.26338	0.00000	0.0000	0.151474	-0.09235	2	2
123	26	28	4	6	-0.31289	-0.28813	0.50000	0.0000	-0.06222	0.073822	2	2
124	27	29	4	4	-0.15686	-0.28113	0.50000	0.0000	0.059145	0.043793	2	3
125	28	32	4	4	1.324898	0.062141	0.25000	0.0000	0.892024	-0.21932	2	3
126	29	33	4	4	0.997799	-0.03802	0.25000	0.0000	0.027364	-0.40771	2	3
127	30	34	4	4	0.670699	-0.26786	0.25000	0.0000	0.358598	-0.40104	2	3
128	31	35	4	6	-0.06037	-0.26126	0.25000	0.0000	0.243255	-0.25085	2	2
129	32	36	4	4	0.163695	-0.06605	0.25000	0.0000	-0.0362	-0.33677	2	3
130	33	37	4	3	0.648456	-0.25356	0.25000	0.0000	0.10983	-0.14353	2	3
131	34	38	4	4	0.222573	-0.31231	0.25000	0.0000	-0.11127	-0.3045	2	3
132	35	39	4	3	0.140798	-0.32713	0.00000	0.0000	-0.04524	-0.30858	2	3
133	36	40	4	3	0.50715	-0.07874	0.00000	0.4963	0.388735	0.428406	2	3
134	37	41	4	4	0.615092	0.033827	0.25000	0.5155	1.174216	0.539379	2	3
135	38	42	4	6	0.37631	0.011716	0.00000	6.3658	0.221063	-0.41637	1	1
136	39	43	4	6	0.229115	-0.17166	0.00000	2.9704	-0.19702	-0.02609	2	2
137	40	44	4	6	0.703409	-0.18912	0.00000	1.3518	0.289283	-0.08578	2	2
138	41	45	4	6	2.093581	-0.03107	0.00000	12.2075	0.285173	-0.08189	1	1
139	42	46	4	6	-0.19611	-0.16357	0.25000	0.6263	0.042159	0.087652	2	2
140	43	47	4	6	0.37631	0.2898	0.50000	4.5191	0.988462	-0.02426	1	1
141	44	48	4	3	1.409944	-0.1078	0.25000	1.2429	0.081337	-0.21412	2	3
142	45	49	4	4	-0.1078	-0.25167	0.00000	0.8672	0.150104	-0.12217	2	3
143	46	50	4	6	0.703409	0.101367	0.00000	2.5958	0.108186	0.012349	2	2
144	47	51	4	6	0.703409	0.101367	0.00000	0.3926	0.025446	0.012143	2	2
145	48	52	4	6	0.637989	0.298932	0.25000	0.0000	1.506546	0.410144	2	2
146	49	53	4	6	-0.26971	-0.28509	0.00000	0.0000	-0.03072	0.59851	2	2
147	50	54	4	6	0.785184	-0.32627	0.00000	0.0000	0.033118	0.458642	2	2
148	51	55	4	1	0.333787	0.045256	0.50000	0.0000	3.596687	0.327021	1	3
149	52	56	4	4	-0.11434	-0.21502	1.00000	0.0000	-0.01209	0.596846	2	3
150	53	57	4	4	0.029257	-0.06524	0.25000	0.0000	0.046268	-0.04839	2	3
151	54	59	4	6	-0.3106	-0.30353	0.00000	0.0000	-0.1677	-0.16684	2	2
152	55	60	4	6	0.548364	0.474328	0.50000	0.0000	0.193118	-0.12441	2	2

APPENDIX C: Three Cluster Solution – Available Data

Frequencies

LANDUSE

		1		2		3		4		5		6	
		Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	2	50.0%	3	5.3%	0	.0%	0	.0%	0	.0%	9	15.0%
	2	2	50.0%	0	.0%	8	100.0%	20	100.0%	3	100.0%	51	85.0%
	3	0	.0%	54	94.7%	0	.0%	0	.0%	0	.0%	0	.0%
	Combined	4	100.0%	57	100.0%	8	100.0%	20	100.0%	3	100.0%	60	100.0%

Cluster Profiles

Centroids

		Price_std		Prop_area		Waterb_pt		Watercrs		Towndist		Waterb_area	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	.28709892	.281125893	.44409217	.313210028	.31168831	.293420726	.23592847	.247217541	.52615895	.397633739	.17154592	.298037268
	2	.24243069	.179614846	.08292087	.083820424	.12987013	.120807231	.07713435	.091428398	.19855809	.133496488	.01610991	.045553455
	3	.17589924	.115790690	.17046139	.109427807	.18350168	.155456643	.08412449	.105449284	.24687875	.173603398	.00866626	.022808828
	Combined	.22290869	.175007020	.14728657	.164848277	.16566986	.163463007	.09424346	.125777670	.24589841	.206193865	.02778193	.105324395

APPENDIX D: Four Cluster Solution – Available Data

Frequencies

LANDUSE

		1		2		3		4		5		6	
		Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	1	25.0%	3	5.3%	0	.0%	0	.0%	0	.0%	8	13.3%
	2	0	.0%	0	.0%	0	.0%	0	.0%	0	.0%	52	86.7%
	3	3	75.0%	0	.0%	8	100.0%	20	100.0%	3	100.0%	0	.0%
	4	0	.0%	54	94.7%	0	.0%	0	.0%	0	.0%	0	.0%
	Combined	4	100.0%	57	100.0%	8	100.0%	20	100.0%	3	100.0%	60	100.0%

Cluster Profiles

Centroids

		Price_std		Prop_area		Waterb_pt		Watercrs		Towndist		Waterb_area	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	.28985603	.305144530	.47377906	.323978570	.31060606	.318378530	.16109576	.106845386	.57723510	.405428480	.20013691	.314219875
	2	.21451665	.166833242	.09819232	.105095737	.11888112	.124125150	.08376314	.107285481	.21440191	.128354099	.01943055	.053863729
	3	.28677720	.187077414	.07033217	.059298506	.15775401	.119076553	.10274859	.175772056	.17557013	.138942001	.01008365	.026388299
	4	.17589924	.115790690	.17046139	.109427807	.18350168	.155456643	.08412449	.105449284	.24687875	.173603398	.00866626	.022808828
	Combined	.22290869	.175007020	.14728657	.164848277	.16566986	.163463007	.09424346	.125777670	.24589841	.206193865	.02778193	.105324395

APPENDIX E: Two Cluster Solution – Restricted Data

Frequencies

LUSE

		1.00		2.00		3.00		4.00		5.00		6.00	
		Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Cluster	1	1	33.3%	4	8.2%	7	87.5%	17	89.5%	1	33.3%	25	44.6%
	2	2	66.7%	45	91.8%	1	12.5%	2	10.5%	2	66.7%	31	55.4%
	Combined	3	100.0%	49	100.0%	8	100.0%	19	100.0%	3	100.0%	56	100.0%

Cluster Profiles

Centroids

		HOUSE		HOUSE_N		HOUSE_C		FARMB		WATER		FENCE		PASTURE	
		Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Cluster	1	.7636	.42876	.2000	.40369	.3455	.47990	.6727	.47354	.5636	.50050	.8000	.40369	.6364	.48548
	2	.1446	.35381	.0000	.00000	.0000	.00000	.1325	.34113	.0000	.00000	.0723	.26054	.0000	.00000
	Combined	.3913	.48982	.0797	.27183	.1377	.34582	.3478	.47802	.2246	.41886	.3623	.48242	.2536	.43667

APPENDIX F: Restricted Data - Wellington

PROP_NO	ConsYear	BCC	Gas	Water	Elec	Phone	Sewerage	All improvements	Arable Area	Non arable area	Access code	Water supply code	Fencing condition code	Water rights	Unused roads/WF	PCC	Veg	Soil
2	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
3	0		N	N	Y	Y	N	0	0.00	115	3			0	WSC		BUSH	0
4	0		N	N	N	Y	N	0	0.00	27	3			0	WSC		BUSH	0
5	1882	4	N	N	Y	Y	N	, , , MS, , , ,	94.70	0	3	4	3	WSC	WSC	4	FLATS	N/A
6	1882	4	N	N	Y	Y	N	, , , MS, , , ,	94.70	0	3	4	3	WSC	WSC	4	FLATS	N/A
7	1882	4	N	N	Y	Y	N	, , , MS, , , ,	94.70	0	3	4	3	WSC	WSC	4	FLATS	N/A
8	1882	4	N	N	Y	Y	N	, , , MS, , , ,	94.70	0	3	4	3	WSC	WSC	4	FLATS	N/A
9	1882	4	N	N	Y	Y	N	, , , MS, , , ,	94.70	0	3	4	3	WSC	WSC	4	FLATS	N/A
10	1991	3	N	N	Y	Y	N	0	43.22	0	3	4	3	WSC	WSC	4	CLEARED	N/A
11	1982	3	N	N	N	Y	N	, GARAGE, CARPORT, , , , ,	28.09	0	3			0	WSC			N/A
12	1940	3	N	N	Y	Y	N	, GARAGE, , SHED, SHEARING SHED, YDS, ,	127.00	N/A	3	3	3	WSC	WSC	3	0	0
13	0		N	N	Y	Y	N	0	93.13	N/A	3	1	2	0	WSC	1	0	0
14	0		N	N	Y	Y	N	0	42.09	N/A	4	3	3	WSC	WSC	3	0	0
15	2000	4	N	N	Y	Y	N	, , , YDS, SHED, , ,	32.54	N/A	4	4	3	WSC	WSC	3	0	0
16	0		N	N	N	Y	N	0	48.56	0	4	4	3	WSC	WSC	4	UND	N/A
17	0		N	N	N	Y	N	0	48.56	0	4	4	3	WSC	WSC	4	UND	N/A
18	0		N	N	Y	Y	N	0	32.42	N/A	4	3	3	WSC	WSC	2	0	0
19	1981	3	N	N	Y	Y	N	, GARAGE, , MS, , , HAYSHED,	0.00	47	3			0	WSC		BUSH	0
20	1960	3	N	N	Y	Y	N	, GARAGE, , SHED, , , ,	35.00	N/A	4	3	3	WSC	WSC	3	0	0
21	1960	2	N	N	Y	Y	N	, GARAGE, , , , , ,	58.30	0	3	4	3	WSC	WSC	4	IRRIGATED	N/A
22	0		N	N	Y	Y	N	, , , MS, MS, , ,	23.27	0	4	4	3	WSC	WSC	4	0	0
23	1978	3	N	N	Y	Y	N	, , , BLOW, DAIRY, , ,	71.60	0	4	4	3	WSC	WSC	4	0	0
24	1880	3	N	N	Y	Y	N	, GARAGE, CARPORT, MS, , , ,	22.20	0	4	4	3	WSC	WSC	4	0	0
25	0		N	Y	Y	Y	N	0	26.65	0	4	4	3	WSC	WSC	4	0	0
26	1955	3	N	N	Y	Y	N	, GARAGE, , , SHED, , MS,	44.35	0	4	3	3	WSC	WSC	3	0	0
27	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
28	1940	4	N	N	Y	Y	N	, GARAGE, , , HAYSHED, , ,	22.74	0	4	4	3	WSC	WSC	4	IRRIGATED	N/A
29	1920	3	N	N	Y	Y	N	, GARAGE, , MS, , , , VER, , , SHED, , , ,	57.33	0	4	4	3	WSC	WSC	4	IRRIGATED	N/A
30	1970	3	N	N	Y	Y	N	, GARAGE, , , DAIRY, HAYSHED, MS, MS	61.92	0	4	4	3	WSC	WSC	5	0	0
31	1988	3	N	N	Y	Y	N	, GARAGE, CARPORT, GARAGE, DAIRY, HAYSHED, MS, POOL	87.86	0	4	4	3	WSC	WSC	4	0	0
32	0		N	N	Y	Y	N	0	31.81	0	4	4	3	WSC	WSC	4	0	0
33	0		N	N	Y	Y	N	0	41.68	0	4	4	3	WSC	WSC	4	0	0

PROP_NO	ConsYear	BCC	Gas	Water	Elec	Phone	Sewerage	All improvements	Arable Area	Non arable area	Access code	Water supply code	Fencing condition code	Water rights	Unused roads/WF	PCC	Veg	Soil
35	1957	3	N	N	Y	Y	N	, GARAGE, , HAYSHED, SHED, HAYSHED, MS,	42.49	0	4	4	3	WSC	WSC	4	0	0
36	1995	3	N	N	Y	Y	N	, GARAGE, GARAGE, SHED, , , ,	27.56	0	4	4	3	WSC	WSC	4	0	0
37	1996	4	N	N	Y	Y	N	GARAGE, CARPORT, SHED, STABLES, POOL, FLAT, HAYSHED	36.98	N/A	4	5	3	WSC	WSC	4		
38	1856	5	N	N	Y	Y	N	, GARAGE, , MS, HAYSHED, SHED, ,	70.40	N/A	4	4	3	WSC	WSC	3	0	0
39	1880	4	N	N	Y	Y	N	, GARAGE, , POOL, SHEARING SHED, SHED, ,	30.35	N/A	4	4	3	WSC	WSC	3	0	0
40	1964	3	N	N	Y	Y	N	, GARAGE, , OB, , , ,	0.00	0	4	3	3	WSC	WSC	2	0	0
41	1940	2	N	Y	Y	Y	N	, GARAGE, , , , , ,	65.54	0	4	3	3	WSC	WSC	4	0	0
42	1922	2	N	N	Y	Y	N	, , , BLOW, , HAYSHED, ,	33.00	0	4	3	3	WSC	WSC	4	0	0
43	0		N	N	Y	Y	N	0	82.45	0	3	3	3	WSC	WSC	3	0	0
44	0		N	N	Y	Y	N	0	165.29	0	3	4	3	WSC	WSC	3	0	0
45	0		N	N	Y	Y	N	0	65.67	0	3	4	3	WSC	WSC	4	0	0
46	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
47	0		N	N	N	Y	N	0	210.84	0	3	4	3	WSC	WSC	4	0	0
48	1985	3	N	N	Y	Y	N	, GARAGE, , MS, , , ,	18.55	0	4	4	3	WSC	WSC	4	0	0
49	1979	3	N	N	Y	Y	N	, , CARPORT, BLOW, MS, HAYSHED, ,	24.28	0	3	4	3	WSC	WSC	4	0	0
50	1960	3	N	N	Y	Y	N	0	464.00	N/A	4	3	3	WSC	WSC	3	CLEARED	N/A
51	1992	4	N	N	Y	Y	N	, , , DAIRY, , ,	129.10	N/A	4	3	3	WSC	WSC	4	CLEARED	N/A
52	0		N	N	Y	Y	N	, , HAYSHED, , , ,	84.92	N/A	3	3	3	WSC	WSC	3	CLEARED	N/A
53	1950	2	N	N	Y	Y	N	0	27.40	N/A	3	3	3	WSC	WSC	3	CLEARED	N/A
54	0		N	N	Y	Y	N	0	20.23	N/A	4			0	WSC		CLEARED	N/A
55	1914	4	N	N	Y	Y	N	, CARPORT, MS, DAIRY, HAYSHED, ,	63.34	N/A	3	3	3	WSC	WSC	4	CLEARED	N/A
56	0		N	N	Y	Y	N	0	176.54	N/A	3	3	3	WSC	WSC	4	CLEARED	N/A
57	1950	3	N	N	Y	Y	N	, , , DAIRY, , ,	82.93	N/A	3	3	3	WSC	WSC	5	IRRIGATION	N/A
58	0		N	N	Y	Y	N	0	20.91	N/A	3	3	3	WSC	WSC	4	CLEARED	N/A
59	0		N	N	Y	Y	N	0	0.00	76	3	3	3	WSC	WSC	3	SWAMP	N/A
60	0		N	N	Y	Y	N	0	39.60	N/A	4	3	3	WSC	WSC	4	CLEARED	N/A

APPENDIX G: Restricted Data - Wimmera

NEWID	ConsYear	BCC	ELEC	SEW	WATS	GAS	All improvements	Arable Area	Non arable area	Access code	Water supply code	Fencing condition code	Water rights	Unused roads/WF	PCC	Veg	Soil
64.00	0							262.6	0	2	2	2			2		
65.00	0							270.3	0	2	2	2			2		
66.00	0							226.2	0	3	3	2			2		
69.00	0							0.00	0.00								
70.00	0							0.00	0.00								
59.00	0							307.6	0	4	3	2			2		
71.00	0							582.8	0	4	3	2			2		
72.00	1920							265.5	0	3	3	2			2		
55.00	0							59.9	0	3	3	2			2		
77.00	0							185.8	0	3	3	2		UN/RD ADJ 13	2		
74.00	0							203.6	0	3	3	2			2		
75.00	0							60.7	0	3	3	2			2		
76.00	0							416.8	20.24	3	3	2			2		
81.00	#N/A	#N/A					#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
78.00	0							128.7	0	4	3	2			2		
61.00	0							0.00	0.00								
80.00	0							335.1	0	3	3	2			2		
56.00	1950							312	0	4	3	2			2		
62.00	0							0.00	0.00								
57.00	1950							312	0	4	3	2			2		
79.00	0							263.1	0	4	3	2			2		
60.00	0							0	0	4	3	2			2		
82.00	0							0	0	4	3	2			2		
83.00	0							0	0	3	3	2			2		
84.00	0							0	0	4	3	2			2		
84.00	0							0	0	4	3	2			2		
58.00	0							0	0	4	3	2			2		
89.00	0							0	0	4	3	2			2		
88.00	0							0	0	3	3	2		Pt. WATER RESERVE	2		
86.00	0							0	0	3	3	2			2		
91.00	0							0	0	3	3	2			2		
92.00	0							0	0	4	3	2			2		
90.00	0							0	0	4	3	2			2		
93.00	#N/A	#N/A					#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
94.00	0							0	0	4	3	2			2		
87.00	0							0	0	3	3	2			2		
67.00	0							0.00	0.00								
63.00	0							0	0	4	3	2			2		
85.00	0							0	0	4	3	2			2		
73.00	0							0.00	0.00								

NEWID	ConsYear	BCC	ELEC	SEW	WATS	GAS	All improvements	Arable Area	Non arable area	Access code	Water supply code	Fencing condition code	Water rights	Unused roads/WF	PCC	Veg	Soil
38.00	1999	4	Y	N	N	N	Db Garage										
4.00		0	Y	N	N	N		70	0	4	3	2			2	CG	MX9
0.00	1980	3	Y	N	N	N	Db Garage, MSx3, Barn	294	0	4	3	2			2	CG	BL9
0.00	0	0	Y	N	N	N	0	72	0	3	3	2	0	0	2	CG	0
2.00	0	0	Y	N	N	N	0	72	0	3	3	2	0	0	2	CG	0
28.00	1950	5	Y	N	N	N	DT Garage, MS, SS, Barn, Hayshed	31.97	0	0	0	0	0	0	0	0	MX3
0.00	0	0	Y	N	N	N	SS, MS	146	0	4	3	2	0	0	2	CG	MX9
29.00	0	0	Y	N	N	N	SS, MS	146	0	4	3	2	0	0	2	CG	MX9
10.00	0	0	Y	N	N	N	SS, SI	114	0	2	3	2	0	0	2	CG	MX6
7.00	0	0	Y	N	N	N	0	130	0	3	0	2	0	0	2	CG	BX4
0.00			Y	N	N	N		274	0	3		2			2	CG	
34.00			Y	N	N	N		274	0	3		2			2	CG	
0.00			Y	N	N	N		274	0	3		2			2	CG	
0.00	1970	3	Y	N	N	N	Bk House, Garage, MSx4, Barn	1291	25	3	0	2	0	0	2	CG	MX2
31.00	1970	3	Y	N	N	N	Bk House, Garage, MSx4, Barn	1291	25	3	0	2	0	0	2	CG	MX2
21.00	1950	4	Y	N	N	N	Garage, MSx3, SS, Hay shed	129	0	4	3	2	0	0	2	CG	MX8
22.00	0	0	N	N	N	N	0	95	0	3	3	2	0	0	2	CG	MX2
0.00	0	0	Y	N	N	N	0	131	0	4	3	2	0	0	2	CG	MX8
0.00	0	0	Y	N	N	N	0	131	0	4	3	2	0	0	2	CG	MX8
13.00	0	0	Y	N	N	N	0	131	0	4	3	2	0	0	2	CG	MX8
0.00			N	N	N	N		117.8	0	3	3	2		UR 57399	2	CG	MX5
47.00			N	N	N	N		117.8	0	3	3	2		UR 57399	2	CG	MX5
15.00	0	0	Y	N	N	N	0	257	0	3	3	2	0	0	2	CG	MX2
16.00	0	0	Y	N	N	N	0	257	0	3	3	2	0	0	2	CG	MX2
5.00	0	0	N	N	N	N	Shed	121	8	3	3	2	0	UR 07669	2	CG	MX8
32.00			Y	N	N	N		99	0	3		2			2	CG	
33.00	0	0	Y	N	N	N	Hay shed	229	0	3	3	2	0	WATER FTG - CA 25APT BTWN CA 11 & 22 ROAD EAST CA 22	2	CG	MX8
44.00	1930	1	Y	N	N	N	MSx2, SS, LA,	255	0	2	3	2		ROAD STH CA 202 203 & 206PT WARRANOOK ABUTTING CA 1 WARRA WARRA	2	CG	MX3
35.00	1945	3	Y	N	N	N	Garage, DA, SS	43	0	3	3	2	0	0	2	CG	GB1
25.00	1950	4	Y	N	Y	N	MS, Hay shed, SP, BU	0	0	3	3	2	0	0	2	0	0
30.00	1950	3	Y	N	N	N	Garage	0	0	3	3	2	0	0	2	0	0
41.00	0	3	Y	N	N	N	0	0	0	0	0	0	0	0	0	0	0
0.00			Y	N	N	N		1990	0	3		2			2	CG	
0.00			Y	N	N	N		1990	0	3		2			2	CG	

NEWID	ConsYear	BCC	ELEC	SEW	WATS	GAS	All improvements	Arable Area	Non arable area	Access code	Water supply code	Fencing condition code	Water rights	Unused roads/WF	PCC	Veg	Soil
0.00			Y	N	N	N		1990	0	3		2			2	CG	
42.00			Y	N	N	N		1990	0	3		2			2	CG	
27.00	0	0	Y	N	N	N	0	104	0	3	3	2	0	0	2	CG	MX3
0.00	0	0	Y	N	N	N	0	104	0	3	3	2	0	0	2	CG	MX3
39.00	0	0	N	N	N	N	0	129.4	0	2	3	2	0	0	2	CG	MX8
0.00	0	0	N	N	N	N	0	129.4	0	2	3	2	0	0	2	CG	MX8
43.00			Y	N	N	N		1990	0	3		2			2	CG	
6.00	0	0	Y	Y	Y	Y	0	31	0	4	3	2	0	0	2	CG	BL7
0.00	0	0	Y	N	N	N	Barn, shed	115	0	3	3	2	0	0	2	CG	BL9
11.00	1960	2	Y	N	N	N	MSx4, SS	296	0	3	3	2			2	CG	MX4
37.00	1950	3	Y	N	N	N	Garage, MS, Barn, SS	75	0	4	3	2	0	0	2	CG	BL5
0.00	1980	3	Y	N	N	N	Db Garage, MSx3, Barn	294	0	4	3	2	0	0	2	CG	BL9
1.00	1980	3	Y	N	N	N	Db Garage, MSx3, Barn	294	0	4	3	2	0	0	2	CG	BL9
14.00	0	0	Y	N	N	N	Barn, shed	115	0	3	3	2	0	0	2	CG	BL9
40.00	0	0	Y	Y	Y	Y	0	161	0	4	3	2	0	0	2	CG	BL6
17.00	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
18.00	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
9.00	1950	4	Y	N	N	N	Garage, MSx2	105	0	4	3	2	0	0	2	CG	MX2
26.00	0	0	Y	N	N	N	0	173	0	1	3	2	0	ROAD NORTH & WEST CA 52 SEC A	2	CG	MX4
23.00	0	0	N	N	N	N	0	40	7	3	3	2	0	0	2	CG	BX5
12.00	1988	3	Y	N	N	N	Single garage	90	0	4	3	2	0	0	2	CG	MX8
3.00	0	0	Y	N	N	N	0	24	0	3	3	2	0	WF 07779 OVER CA 86APT ABUTTING CA 34	2	CG	0
0.00	0	3	Y	N	N	N	0	65	0	3	3	2	0	0	2	CG	BX9
24.00	0	3	Y	N	N	N	0	65	0	3	3	2	0	0	2	CG	BX9
0.00	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
19.00	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
0.00	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
0.00	1950	3	Y	N	N	N	Garage	46	0	3	3	2	0	0	2	CG	GB1
20.00	0	0	0	0	N	0	0	0	0	0	3	0	0	WATER FTG ABUTTING CA 81PT & 82PT	2	CG	0
8.00	1900	5	Y	N	N	N	Garage, shed	67	0	3	3	2	0	0	2	CG	MX2
46.00			Y	N	N	N				4		2			2	CG	
0.00			Y	N	N	N		184	0	3		2			2	CG	
0.00			Y	N	N	N		184	0	3		2			2	CG	
36.00	1980	4	Y	N	N	N	Db Garage, MS, SS, Barn	589.2	0	3	3	2	0	ROADS SOUTH CA 120 82 & 81 & WEST CA 120 & 121	2	0	MX6
45.00	0	0	Y	N	N	N	0	159.9	0	4	3	2	0	0	2	CG	BX5
0.00	0	0	Y	N	N	N	0	159.9	0	4	3	2	0	0	2	CG	BX5