INTEGRATING GEOGRAPHIC INFORMATION SYSTEMS
AND ARTIFICIAL NEURAL NETWORKS:
DEVELOPMENT OF A NONLINEAR, SPATIALLY-AWARE
RESIDENTIAL PROPERTY PREDICTION MODEL

by

J. GREGORY CUNNINGHAM

B.A., The University of British Columbia, 1997

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Geography)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
October 2001
© J. Gregory Cunningham, 2001
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Geography

The University of British Columbia
Vancouver, Canada

Date October 12, 2001
Abstract

Mass appraisal of residential real estate is desired and often required for asset valuation, property tax and insurance estimation, sales transactions and urban planning. Multivariate linear regression models, referred to as hedonic pricing functions, have been used to 'unbundle' the characteristics of a dwelling by expressing its price as a function of its mix of attributes. However, the relation between the value of a dwelling and its intrinsic and extrinsic characteristics is complex and generally nonlinear. Consequently, this study attempts to capture this inherently complex relation through the use of Artificial Neural Network (ANN) models and investigates their ability to predict residential real estate values compared to traditional hedonic techniques.

Researchers in the real estate appraisal industry have recently used ANNs to overcome methodological restrictions such as nonlinearity and noise that result from the use of multivariate linear regression techniques. Detailed locational factors, however, have failed to be adequately represented in their models. In my work I extend current research efforts by explicitly incorporating 'space' into ANN models. Through integrating ANN techniques and Geographic Information Systems (GIS), the extraction, transfer and recognition of spatial attributes—such as average family income or secondary school provincial examination performance—can be facilitated. Results indicate that ANN models outperform traditional hedonic models. Further, the inclusion of locational attributes significantly improves the ability of both models to predict the value of a dwelling.
# Table of Contents

Abstract .................................................................................................................. ii

Table of Contents .................................................................................................... iii

List of Figures ......................................................................................................... vi

List of Tables ......................................................................................................... viii

Acknowledgments ................................................................................................... ix

Chapter I: Introduction .......................................................................................... 1

1.1 Assessment of Residential Properties ............................................................... 1

1.2 Intrinsic and Extrinsic Characteristics of a Dwelling ......................................... 2

1.3 Rationale of Approach ....................................................................................... 4

1.3.1 Methodology ................................................................................................ 4

1.3.2 Assumptions ................................................................................................. 6

1.4 Scope .................................................................................................................. 7

1.5 Objectives .......................................................................................................... 7

Chapter II: Hedonic and Artificial Neural Network (ANN) Valuations .................... 9

2.1 Hedonic Pricing Techniques—Literature Review ............................................... 9

2.2 Multivariate Linear Regression Model .................................................................. 11

2.2.1 Spatial Consideration .................................................................................. 13

2.3 Artificial Intelligence (AI) .................................................................................. 15

2.3.1 Artificial Neural Networks (ANN)—Biological Background ....................... 17

2.3.2 Network Functions ...................................................................................... 20

2.3.3 Network Architectures ................................................................................ 24

2.3.4 Training Artificial Neural Networks ......................................................... 30

2.3.5 Strengths and Limitations .......................................................................... 33

2.4 Artificial Neural Network Techniques—Literature Review .............................. 34

2.4.1 Spatial Consideration .................................................................................. 39
### Chapter III: Data Selection and Preparation

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Intrinsic Characteristics—Residential Transaction Attribute Data</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Extrinsic Characteristics—Socioeconomic Attribute Data</td>
<td>44</td>
</tr>
<tr>
<td>3.2.1</td>
<td>The Modifiable Areal Unit Problem (MAUP)</td>
<td>45</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Spatial Scale of Socioeconomic Data</td>
<td>47</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Suppression of Socioeconomic Data</td>
<td>50</td>
</tr>
<tr>
<td>3.3</td>
<td>Data Reduction Techniques—Factor Analysis</td>
<td>50</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Variable Selection</td>
<td>52</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Factor Extraction</td>
<td>54</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Results of Factor Analysis</td>
<td>56</td>
</tr>
<tr>
<td>3.4</td>
<td>Surface Modelling of Socioeconomic Variables</td>
<td>60</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Data Set Development</td>
<td>61</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Non-residential Area Masking</td>
<td>64</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Enumeration Area Centroid Correction</td>
<td>66</td>
</tr>
<tr>
<td>3.4.4</td>
<td>Surface Creation</td>
<td>67</td>
</tr>
<tr>
<td>3.5</td>
<td>Extrinsic Characteristics—Spatial Attribute Data</td>
<td>70</td>
</tr>
<tr>
<td>3.6</td>
<td>Address Matching</td>
<td>73</td>
</tr>
<tr>
<td>3.7</td>
<td>Data Set Division, Spatial Stratification and Data Preprocessing</td>
<td>75</td>
</tr>
</tbody>
</table>

### Chapter IV: Model Development

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Multiple Linear Regression Models</td>
<td>78</td>
</tr>
<tr>
<td>4.2</td>
<td>Artificial Neural Network Models—Network Configuration</td>
<td>79</td>
</tr>
<tr>
<td>4.3</td>
<td>Framework for the Proposed Model</td>
<td>82</td>
</tr>
</tbody>
</table>

### Chapter V: Model Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Model Evaluation Statistics</td>
<td>85</td>
</tr>
<tr>
<td>5.2</td>
<td>Statistical Results</td>
<td>88</td>
</tr>
<tr>
<td>5.3</td>
<td>Graphical Results</td>
<td>90</td>
</tr>
<tr>
<td>5.4</td>
<td>Residual Mapping</td>
<td>94</td>
</tr>
<tr>
<td>5.5</td>
<td>Extrinsic Data Models</td>
<td>99</td>
</tr>
</tbody>
</table>

### Chapter VI: Conclusion

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Basic Framework for the Proposed Model .................................................. 8
Figure 2: Simplified Structure of a Neuron ................................................................. 19
Figure 3: Configuration of a Neuron from an Artificial Neural Network ................. 22
Figure 4: The Standard Logistic Function ................................................................. 24
Figure 5: An Artificial Neural Network ...................................................................... 26
Figure 6: Map of Study Area—Vancouver, British Columbia .................................... 42
Figure 7: Census Tract boundaries for the CMA of Vancouver ............................... 47
Figure 8: Census Tract boundaries for the Municipality of Vancouver ..................... 48
Figure 9: Enumeration Area boundaries for the Municipality of Vancouver ............. 49
Figure 10: Scree Test from Factor Analysis ............................................................... 56
Figure 11: Populated Versus Non-populated Portion of an Enumeration Area .......... 63
Figure 12: Polygons Masking Non-residential Areas ................................................. 65
Figure 13: Correction of Enumeration Area Centroid Placement ............................. 66
Figure 14: Surface Representation for the Percentage of Visible Minority Population .... 68
Figure 15: Secondary School location and District Boundaries ............................... 73
Figure 16: Streets Within the City of Vancouver ...................................................... 74
Figure 17: Three by Four Grid for Spatial Stratified Random Sample ....................... 76
Figure 18: Framework for the Model of Prediction .................................................... 83
Figure 19: Regression Model 1—Intrinsic Attributes Only ....................................... 91
Figure 20: ANN Model 1—Intrinsic Attributes Only ................................................... 91
Figure 21: Regression Model 5—With Extrinsic Attributes (simple random sample) .... 92
Figure 22: ANN Model 5—With Extrinsic Attributes (simple random sample) ............ 92
Figure 23: Regression Model 6—With Extrinsic Attributes (stratified random sample) 93
Figure 24: ANN Model 6—With Extrinsic Attributes (stratified random sample) ......... 93
Figure 25: Surface Representation of MLS Property Values—Orthographic Perspective 95
Figure 26: Regression Model 1 Residuals—Intrinsic Attributes Only ....................... 96
Figure 27: ANN Model 1 Residuals—Intrinsic Attributes Only .................................... 96
Figure 28: Regression Model 5 Residuals—With Extrinsic Attributes (simple random sample) .... 97
Figure 29: ANN Model 5 Residuals—With Extrinsic Attributes (simple random sample) ............ 97
Figure 30: Regression Model 6 Residuals—With Extrinsic Attributes (stratified random sample) ... 98
Figure 31: ANN Model 6 Residuals—With Extrinsic Attributes (stratified random sample) ............ 98
List of Tables

Table 1: Descriptive Statistics for Prediction Results from Worzala, Lenk and Silva (1995) ........... 38
Table 2: Descriptives of Intrinsic Characteristics ............................................. 43
Table 3: Variable Listing for Factor Analysis .................................................. 54
Table 4: Structure Matrix Results From an Oblique Factor Analysis ...................... 59
Table 5: Descriptives of Extrinsic Characteristics ............................................ 75
Table 6: Pearson Correlations of the Intrinsic and Extrinsic Characteristics ............ 75
Table 7: ANN model Network Configurations .................................................. 81
Table 8: Performance of the Multivariate Linear Regression Models ....................... 88
Table 9: Performance of the Artificial Neural Network Models ............................. 89
Table 10: Bivariate Statistics for Variable Sequence 6—Model Training Data Set ......... 89
Table 11: Performance of the Extrinsic Regression Model (stratified random sample) .... 99
Table 12: Performance of the Extrinsic ANN Model (stratified random sample) ......... 99
Acknowledgments

I would like to take this opportunity to express my sincere thanks to the many individuals who have contributed to the accomplishment of this project. First of all, I am grateful to my immediate supervisor Dr. Brian Klinkenberg for his approachability, patience, guidance and thoughtful advice during my attendance in the programme. I would like to express appreciation to Dr. Dan Hiebert and Dr. David Ley who were always there for support, advice and assistance. I must also thank Mark MacLean for his constructive criticism, and mentorship.

I am very fortunate to have had the support from an incredible group of friends and colleagues. Finally, a profound note of gratitude to my family for their unconditional support and care, without which this opportunity would have never been possible.
Chapter I: Introduction

1.1 Assessment of Residential Properties

The supply and demand of a housing market is closely linked with the social geography of a city. The analysis of this market is of central importance to the determination of the level of welfare in society, demographic transition and the degree of aggregate economic activity. In many economies, a purchased dwelling not only represents a shelter and a major capital good, but it also accounts for a large proportion of total household wealth, as well as being a status symbol. While the share of income allocated to housing thus represents a substantial proportion of total household expenditure, the aggregation of residences encompasses a considerable share of a municipality's taxation base (Sheppard, 1999). Therefore, it is not unexpected that analysts and governments have devoted considerable time to understanding the structure of the demand for housing.

While it is a relatively easy task to conceptualize the housing characteristics that will be most significantly associated with dwelling value, measuring their magnitude and inter-relations is another matter. Many researchers have relied upon single measures such as the average value of sales transactions to analyse the market. However, such an analysis is complicated by the fact that housing is not a homogeneous good—prices should be standardized for the characteristics of the dwelling. Alternatively, a dwelling can be viewed as consisting of a bundle of intrinsic and extrinsic attributes (Bourne, 1981). More recently, notably in economics, others have implemented regression
techniques to circumvent the challenges associated with the heterogeneous nature of a dwelling by statistically controlling for varying features (Follain and Jimenez, 1985). This methodology is referred to as an hedonic price function and involves unbundling the characteristics of a dwelling by expressing its value as a function of its mix of attributes.

Researchers have also recently explored Artificial Neural Network (ANN) models because of their freedom from traditional statistical assumptions and their ability to overcome methodological restrictions, such as nonlinearity and noise, that result from the use of regression techniques. Detailed locational factors, however, have failed to be adequately represented in the ANN models. While housing markets are undeniably intrinsically spatial, most empirical studies have disregarded the locational attributes of a dwelling altogether. Given the emergence of Geographic Information Systems (GIS) technology, and the real estate dictum of "location, location, location", it is surprising that it has not been more extensively employed in this particular domain.

1.2 Intrinsic and Extrinsic Characteristics of a Dwelling

A dwelling is occupied in a particular place by a specific kind of person, a 'self-sorting' in a sense, matching the supply of an assortment of characteristics and the demand of our unique tastes and preferences. The quality of life for many individuals is intimately connected to the satisfaction with and attachment to the structure and sur-
rounding environment of a dwelling. Describing the condition and inhabitants of districts in the East End of London, Booth (1904: 66) reported that:

Each district has its character—its peculiar flavour. One seems to be conscious of it in the streets. It may be in the faces of the people, or in what they carry—perhaps a reflection is thrown in this way from the prevailing trades—or it may lie in the sounds one hears, or in the character of the buildings.

Beyond the internal attributes of a dwelling exists a complex mesh of positive and negative externalities operating at different scales and magnitudes—most of which are not easily quantified.

The basic factors that encompass the external characteristics can be categorized as:

- **environmental**—physical surroundings (noise and air pollution, view corridor and physical aesthetics, green space and vegetation, landuse);
- **social**—community (access to amenities and transportation, school and neighbourhood quality, cultural influences, demographic composition, historical dimension, market / media trends); and
- **personal**—psychological / gender (proximity to place of work and family, potential for capital gain, size of household, lifestyle change, preference for population group) (Kain and Quigley, 1970; Kanemoto 1980; Zeng and Zhou 2001).

The external characteristics investigated in this study will focus exclusively on selected social aspects.
1.3 Rationale of Approach

The appraisal of residential real estate is desired and often required for asset valuation, property tax and insurance estimation, sales transactions and urban planning. In 1974, on the west coast of Canada, the British Columbia Assessment Authority was created with the mandate of establishing an independent, uniform and efficient framework to evaluate real estate. The primary objective of this crown corporation is to provide tax authorities with an assessment roll (the basis for local and provincial taxation) and individual property owners with a notice of assessment—an estimation of market value as of July first of the year preceding publication of the rolls.

Property assessments are arrived at through random site inspection and comparison of properties with similar characteristics. While the outcomes and techniques implemented in these evaluations have been widely accepted, it is proposed that modern analytical techniques, when combined with the spatial analytical capabilities of Geographic Information Systems (GIS), can be used to greatly improve the quality of mass real estate appraisal and, therefore, the distribution of the property tax burden.

1.3.1 Methodology

Traditionally, price trend analysis of housing markets was merely confined to the use of the average value of a dwelling (of all properties of a particular kind), not standard-
ized for any associated intrinsic or extrinsic attributes. It was eventually recognized that it was crucial to statistically control for the varying characteristics in order to reveal the true structure of demand (Case and Quigley, 1991). However, the intricate and obscure nature of the demand for a dwelling renders the extraction of these separate components a laborious task.

It was thereafter assumed that the challenges associated with determining the differences in dwelling quality could be circumvented by basing an index on the sale of the same property at different times—the repeat sales price index (Bailey, Muth and Nourse 1963). However, by restricting the analysis to properties sold more than once, it is extremely wasteful of transaction information, not to mention the assumed bias toward only those properties that sold two or more times (Palmquist, 1980).

The hedonic price function, which now represents an important component of the urban economists' toolbox, is the most commonly cited alternative (or a combination of the two) to the repeat sales price index. The application of this technique to the analysis of housing is empirically established, and theory backing it has been evolving over the past thirty years. However, the implementation of these modelling techniques in the assessment of residential properties faces a variety of limitations, many of which are still, unfortunately, poorly understood.

Alternatively, the discipline of property appraisal can be viewed as a problem in pattern recognition. Each property has its peculiar intrinsic and extrinsic attributes that togeth-
er furnish a pattern of the property. In recent times, pattern recognizing ANN modelling procedures have produced promising results, and hence have gained increased acceptance in the scientific community. These modern analytical techniques have recently been applied to the problem of real estate appraisal, with varying results, in an attempt to remove the restrictions imposed by traditional methods—particularly methodological constraints associated with linear assumptions.¹

### 1.3.2 Assumptions

A fundamental difficulty with valuing housing markets is that the real world must be represented by a model. Researchers must acknowledge the requisite assumptions while accepting that the model is shaped by various concessions. Results from a property valuation are a function of the discontinuity of the transaction data (during a given period, sales do not occur for all products), while the validity of the output is often compromised by the quality and richness of the input. When modelling, the analyst inevitably encounters insufficient, inadequate and or erroneous data.² It is anticipated, however, that the integration of ANNs and GIS will better explain the heterogeneous

---

¹ For instance, Do and Grudnitski (1998) revealed that a negative relationship between dwelling value and age occurs only within the first sixteen to twenty years. Thereafter, a positive relationship between the variables and an appreciation in values was discovered by the authors.

² For example, micro and macro market fluctuations and market activity of dwellings that are less than average value undoubtedly bias the sales transaction database and thus skew the model’s ability to predict.
nature of the built environment and, therefore, improve upon traditional techniques by providing more accurate valuations of residential properties.

1.4 Scope

It is beyond the scope of this thesis to further develop market indices in particular. While modern methodologies and procedures are considered—so that traditional techniques can be further understood, reinforced and advanced—the temporal dimension is explicitly not considered so that model simplicity can be maintained given practicality and time constraint. An attempt is made to explore, apply and analyse ANN models, but in many cases a more comprehensive study would be required to adequately unpack and explain the dynamics and behaviour of the model.

1.5 Objectives

The objectives associated with this research are twofold: 1) to explore the feasibility of implementing ANNs and to compare the results with multivariate linear regression techniques; 2) to develop an ability to assess spatial, extrinsic attributes and investigate their significance in the prediction models. Accordingly, this thesis is arranged as follows. The relevant theoretical and technical literature for the competing models is summarized in the next chapter along with a review of studies implementing the tech-
niques in the area of property valuation. Intrinsic dwelling characteristics will briefly be discussed in Chapter Three followed by detailed sections outlining how socioeconomic attributes were selected, modelled and appended to a property transaction database. The development and selection of the particular models (regression and ANN) is established in the fourth chapter, while Chapter Five explores the different methods for evaluating the models and presents the results. Finally, the last chapter reviews the findings and offers recommendations for further research. The figure below illustrates a rudimentary framework for the proposed model of prediction.

Figure 1: Basic Framework for the Proposed Model
Chapter II: Hedonic and Artificial Neural Network (ANN) Valuations

2.1 Hedonic Pricing Techniques—Literature Review

A dwelling can be viewed as consisting of a bundle of attributes, some of which are related to its physical features such as the square footage or age, while others are associated with the quality of its location in the city, district or neighbourhood. In a perfect market, the value of a residential dwelling equals the sum of the present value of the costs and benefits associated with it. Consumers maximize their utility by selecting the dwelling that best suits their tastes and preferences, given market availability and income constraints.

A primary difficulty associated with the effort to estimate the demand or preference for unique characteristics is the inherent heterogeneity of a product. A technique developed in the automotive industry by Court (1939) and popularized by Griliches (1971), however, circumvented this difficulty by providing economists with the ability to deal with commodity heterogeneity. The term hedonic was used by Court in his pioneering works to define the weighting of the relative importance of various automotive components, such as horsepower, braking capacity and window area, in constructing an index of usefulness and desirability (Goodman, 1998). Following the work of Griliches in the early 1960's and 70's, hedonic pricing techniques received considerable attention and were quickly added to the micro-econometric tool box and, in particular, became an important tool for analysing housing markets.
The hedonic price function is a methodology for estimating the implicit prices of the characteristics that differentiate closely related products in a product class, such as housing. The notion of an implicit market is defined by Sheppard (1999: 1598) as the process of production, exchange, and consumption of commodities that are primarily (perhaps exclusively) traded in 'bundles'. The explicit market, with observed prices and transactions, is for the bundles themselves. Such a market, however, might be thought of as constituting several implicit markets for the components of the bundles themselves. This is of particular importance when the bundles are not homogeneous, but vary due to the varying amounts of different components that they contain.

Essentially, the hedonic procedure involves unbundling the characteristics of a dwelling by expressing the value as a function of its mix of attributes—recovering the prices of characteristics from information on the bundled price. Freeman (1979), Follain and Jimenez (1985) and Sheppard (1999) provide particularly detailed accounts of the process.

Generally, hedonic prices are evaluated by regressing dwelling values on significant attributes in a linear fashion using ordinary least squares. This approach assumes that the marginal willingness to pay for a particular characteristic is interpreted as the derivative of the hedonic regression with respect to the characteristic itself. It is important to note, however, that the marginal price derived from the equation does not precisely measure what an individual household is willing to pay for additional units, but a valuation that is the result of interactions within a sampled housing market (Follain and Jimenez, 1985).
2.2 Multivariate Linear Regression Model

Multiple regression modelling is a widely used procedure for conducting multivariate analysis when more than two independent variables are concerned. The advantage of multiple regression, over most other techniques, rests in its ability to derive how a single variable is functionally dependent on two or more independent variables. The multivariate linear regression model examines how a single variable is functionally dependent on a set of independent variables by fitting a curve of linear functional form to the data:

\[ Y = b_0 + b_1 x_1 + b_2 x_2 + b_n x_n + e, \]  

(1)

where the left side of the equation (Y) denotes the dependent variable, \( b_0 \) represents the intercept with the y axis (constant) and the estimates of the slope parameters, or coefficients, are expressed as \( b_1, b_2 \) and \( b_n \), for the independent variables of \( x_1, x_2 \) and \( x_n \) respectively. The error term (e) simply speaks for the proportion of the variance in the dependent variable that is unexplained by the equation.

The regression model estimates the coefficients of the equation using least-squares methodology. This technique yields approximations of the coefficients such that the sum of all residuals squared is minimized. The model is structured to determine the response of one variable (Y) under fixed values for other variables (\( x_1, x_2 \) and \( x_n \)). In other words, the coefficients estimate the amount of change that occurs in the dependent variable for a single unit of change in any independent variable. Furthermore, each
coefficient expresses the amount by which the dependent variable changes, with the effect of all the other independent variables partialled out, or controlled (Bryman and Cramer, 1997).

The standard approach to applying regression techniques—based on estimation using a cross-section of data—presents an abundance of difficulties, ranging from determining the proper parametric specification, to coping with idiosyncrasies in the data. For an overview of the caveats and empirical problems with the application of hedonic price functions, see Freeman (1979), Follain and Jimenez (1985). The established methodology has been to adopt some form of parametric approach. The majority of analysts have tended to rely on linear or logarithmic forms, while more recently more flexible forms have been obtained by applying the Box-Cox transformation (see Linne­man, 1980 and Freeman, 1979). Alternatively, in an attempt to introduce generality into the models, Rosen (1974) demonstrated that the function need not be linear. It should be highlighted, nevertheless, that the standard hedonic price function imposes a linear constraint.

While a sound theoretical basis for valuing a dwelling has been established, empirical application of the method has produced results that vary widely with respect to quality (Follain and Jimenez, 1985; Can, 1992). An extensive body of literature, too immense to summarize in this study, has developed on the application of hedonic pricing tech­niques in the domain of housing market valuation. Selected studies include:

- in general: Bailey, Muth and Nourse (1963), Case and Quigley (1991);
• residential services: Kain and Quigley (1970);
• age effects: Clapp and Giaccotto (1998), Grether and Mieszkowski (1974);
• effects of race: Bailey (1966);
• locational amenities: Hoyt and Rosenthal (1997);
• neighbourhood dynamics and spatial effects: Can (1990, 1992);
• landuse and visibility analysis: Lake, Lovett, Bateman and Day (2000);
• effects of residential investment: Ding, Simons and Baku (2000); and

2.2.1 Spatial Consideration

Although hedonic pricing techniques are theoretically well founded and have become an essential tool for economists studying urban housing markets, much of the existing empirical work has neglected important methodological issues that result from the spatial nature of data.\footnote{There is a considerable body of literature concentrating on the impact location has on residential real estate value (see Ratcliff, 1961; Nowlan, 1978; Can, 1990; and Wyatt, 1996).} While in the recent past the power, technology and medium\footnote{Records of real estate transactions simply were not computerized (samples had to be coded and keypunched) while geographic coding was realized through arranging pins on a map (Goodman, 1998).} of resources simply did not exist, however, today it is possible to collect, explore and realize locational behaviour.
Commonly, the estimation of the effect of location on a dwelling’s price is extracted through inclusion of a few small scale characteristics associated with the socio-economic composition of a neighbourhood, linear accessibility measures to amenities and generalised public service provisions. Alternatively, beyond the use of a few scattered spatial variables, market segmentation was the first attempt by economists to realize locational complexity in the hedonic models.

The first to raise the question of market segmentation within a housing market, in the context hedonic pricing functions, was Straszheim (1974). He suggested that many sub-markets exist within an urban area, based on demand and supply inequalities and barriers of participation. Straszheim proposed that the entire market should be broken into segments, and each one furnished with a unique hedonic price function, so as to avoid inaccurate estimates of the implicit prices facing buyers in each independent market. Estimating separate price functions for segments of the San Francisco Bay Area, he illustrated that the sum of the squared errors as a whole could be reduced. Goodman (1978) later emphasized that hedonic price functions revealed differentials of up to twenty percent between suburban and city sub-markets in a metropolitan area.

While the importance of market segmentation has been well recognized in recent literature, it is not the purpose of this study to specifically identify sub-markets through-

3. Taken to the extreme, each block, or unique area, would have its own hedonic function.
4. See, for example, Dale-Johnson (1982) for sorting using Q-factor analysis (conducting a multivariate data grouping, or clustering, technique Dale-Johnson identified 13 natural market segments in Santa Clara County using various intrinsic and extrinsic variables); Thibodeau (1989) for the computation of indexes for metropolitan areas in the United States; and Hamilton and Hobden (1992) for use of sub-markets in the Vancouver metropolitan area.
out the metropolitan area of Vancouver. As proposed in upcoming sections, however, the objective is to devise a single spatially-aware portable price model that accounts for heterogeneity while not requiring the construction of multiple models for each segmented market.

2.3 Artificial Intelligence (AI)

As the subject of Artificial Intelligence (AI) covers such a broad spectrum, with concurrent developments in several disciplines, there is no universally accepted definition. One particularly useful description, nevertheless, is presented by Openshaw and Openshaw (1997: 16):

Al... is an attempt to mimic the cognitive and symbolic skills of humans using digital computers in the context of particular applications. This is not so much a matter of seeking to replicate skills that are already possessed by humans but of attempting to improve and amplify our intelligence in those applications where it is deficient and can benefit from it most.

From this perspective, AI researchers seek to develop systems that simulate human intelligence. A common objective is to enhance our ability to better solve challenges previously encountered, in a problem-specific domain, as well as to extend the breadth of problems that can be undertaken by seeking to understand how the human brain functions.
The evolution of Al invokes several philosophical questions about the nature of intelligence and of humankind itself. While the aim of Al is to develop programs for machines that have characteristics associated with human intelligence, the difficulty is that we still do not fully understand how individuals think or what "human intelligence" actually implies. Thus, there is no absolute benchmark for assessing the development of the so-called intelligence of a machine (apart from a test introduced by Turing\(^5\) in a 1950 paper, which attempted to determine if a computer program had intelligence).

For many, the reference to "human intelligence" introduces uneasiness or overt excitement. Those who claim that machines endowed with Al are truly intelligent, function in the same way as the brain, and will, one day, think like humans, encourage hype—particularly in the popular media. Others who allude to Al's ability to solve all of the world's difficult computational challenges only serve to confound and exasperate the situation. Alternatively, while Al is not a computational panacea, its abilities should not simply be reduced to hype and nonsense; they can frequently provide novel insights into a problem or a better implementation of an existing approach. Problem solving with Al is about computing with machines; its misleading reference to biology has served only as the inspiration to duplicate the observed workings of a brain (Murray, 1995; Openshaw and Openshaw, 1997).

---

5. Alan M. Turing (1912-1954) presumed that one day computers would be programmed to acquire abilities rivalling human intelligence. To support his argument Turing proposed the idea of an 'imitation game', now commonly referred to as the 'Turing test' for intelligence, in which a human and a computer would be blindly interviewed by means of textual messages. Turing asserted that if the interrogator was unable to distinguish them during questioning, then it would not be unreasonable to call the computer intelligent. The test, nonetheless, has been subject to much criticism and has been at the heart of many discussions in Al, cognitive science and philosophy for the past 50 years.
Perhaps the reference to intelligence should be removed altogether. A machine is merely running through a set of operations that it was programmed to accomplish, while not actually thinking about the task itself. In this instance, the machine essentially has no choice in the matter nor exhibits any consciousness of the process. Is this what makes a human more intelligent than a machine? While machines can process vast amounts of data faster and more accurately than humans, what prevents them from being or becoming intelligent; when does intelligence happen? The following sections will attempt to emphasize what an incredibly difficult, if not impossible, task it would be to replicate the true parallel processing that is found in the brain.

2.3.1 Artificial Neural Networks (ANN)—Biological Background

Artificial Neural Networks (ANN) is a subject area that has recently emerged in the field of AI. Although ANNs have been applied successfully to model nonlinear relations, the theory supporting them straddles several disciplines, is often inconsistent and requires an advanced level of mathematical comprehension. The following sections, therefore, attempt to provide a general foundation of the basic structures, but omit a detailed discussion on the specific algorithms and internal operations.

6. Parallel processing basically entails the concurrent execution of two or more processes in a single unit. One of the leading areas of research in cognitive psychology, the parallel-distributed processing model, states that information in biological systems is processed simultaneously by the densely interconnected, parallel structure of the mammalian brain. Computationally, parallel processing involves the execution of program instructions divided up among multiple processors with the objective of running a procedure in less time.
ANNs are biologically inspired, as is much of Al, based on a loose analogy with the presumed workings of a brain. The brain is made up of an intricate network of cells called neurons (a few hundred neurons are present in simple creatures, while an estimated hundreds of billions are present in humans) that are linked to receptors and effectors (Arbib, 1995). Receptors continually convert stimuli from the human body or the external environment into electrical impulses that present information to the network. On the output end, motor neuron cells—controlled by network activity—manipulate muscle movement and gland secretion by converting electrical impulses into discernible responses. In between, the network of neurons constantly combines the signals from the receptors with signals encoded from past experience. This in turn supplies the motor neurons with signals which will yield adaptive interactions with the environment.

The function of the brain is, however, far more complex than the simple stimulus-response sequence outlined above. Instead, the neurons are interconnected in loops and coils, so that signals entering through the receptors interact with billions of signals already spanning the system—these not only generate responses which control the effectors, but also to modify the characteristics of the network so that subsequent behaviour will reflect prior experience (Arbib, 1995; Haykin, 1999).

In order to further understand the processes that occur between the stimulus and response, the basic neuron (simplified in Figure 2) must be examined. Each neuron in a network operates like a simple processor, while the incredible interaction between all
neurons and their parallel processing make the abilities of the brain possible. As indicated by the figure, a single neuron consists of dendrites (for incoming information), synaptic terminals (for outgoing information), a nucleus, cell body and an axon. The size and shape of these neurons vary widely across different parts of the brain.

Interaction or communication of information between neurons is facilitated by the dendrites (so called because of their combined resemblance to a tree), the axon and the synaptic terminals. Incoming messages, in the form of electrical stimulation, are passed through the dendrites and received by the neuron where they are summed. If stimulation exceeds a certain threshold, information is then delivered along the axon to the synaptic terminals where it is then propagated to other neurons. In this instance, the neural cell is classified as “activated”. Conversely, if the incoming stimulation is too low, the neuron remains “inhibited” and information will not be transferred through the axon.
When it is observed that each of the neurons in the human brain may receive input from 5000–15,000 other neurons, we can begin to appreciate that the range of functions for a single neural unit is indeed immense, and the processing enormously efficient (Openshaw and Openshaw, 1997). The rich network of interconnections between neurons are adaptive—meaning that the connective configuration is dynamically changing. It is now commonly acknowledged that this adaptation provides the learning ability or "intelligence" of the human brain (Chester, 1993; Arbib, 1995; Haykin, 1999).

Similar to biological networks, ANNs also consist of 'neurons' and 'connectors'. While these are examples of how artificial networks attempt to emulate the structure of a neural network, it is critical to recognize that the structural composition found in the human brain is totally unique. Its highly organized architecture is not present in computers, and humankind is far from recreating them and, therefore, modelling human intelligence with ANN applications.

2.3.2 Network Functions

If the presumed workings of the brain are to be simulated using ANNs, drastic simplifications are obviously required; it is essentially impossible to reproduce the true parallel processing of all neural cells. Although computers exist that have parallel processing capabilities, the extraordinary number of processors that would be neces-
sary to realize this can not be afforded given existing hardware constraints. An addi-
tional limitation is the inability to modify the internal structure of a computer while it is
operating (Fröhlich, URL; Haykin, 1999). Thus, how is electrical stimulation, or learn-
ing, emulated in a computer program? Decoding the puzzle behind the learning mech-
anisms of the brain has perplexed many researchers. These issues require the
development of an abstracted model.

An artificial neuron, depicted in Figure 3, is purely an information processing unit that
is critical to the operation of the network. Its overall function is to transport incoming
information \((x_1, x_2, \ldots, x_n)\) through outgoing connections \((y_k)\) to coupled neurons. An ex-
cited neuron is activated if its combined input exceeds a particular level. However, be-
fore output is propagated, each neuron manipulates the information by executing a
number of procedures.
Firstly, each input is multiplied by a weight matrix, signified by $w_{k1}$, $w_{k2}$ and $w_{kn}$, that is associated with neuron $k$. The net of these input signals, or activation potential ($v_k$), is computed at the summing junction ($\Sigma$), and in turn processed by an activation function $\varphi(\cdot)$. The activation function simply transforms the received input by imposing a limit on the amplitude of the output. Typically, the range of the output is scaled from 0.0 to 1.0 and can be transformed using a variety of different functions (e.g., sigmoidal).

If a node does not transform its net input it is simply known to have a linear function—an arrangement of linear activation functions is still a linear function. The most common type of transformation used in an ANN is a nonlinear sigmoid function, whose
curve roughly resembles the shape of an S (Openshaw and Openshaw, 1997; Sarle, 1997). The standard logistic function (an example of a sigmoid function), rendered in Figure 4, can be expressed in mathematical terms as:

\[ \varphi(v) = \frac{1}{1 + e^{-av}}, \]  

where \( \varphi(v) \) represents the output of the activation function, parameter \( a \) determines the slope and \( v \) is the activation potential (Haykin, 1999).

The transformation of the input signal in an artificial neuron is fundamental if we desire to introduce nonlinearity into the model. A network of neurons with nonlinear activations, such as the logistic function, is itself nonlinear. With logistic neurons distributed throughout the model, the nonlinearity is extraordinary; this is what makes an ANN so powerful.
2.3.3 Network Architectures

There are over 50 ANN architectures that have been developed, but generally they all have the same components. While much writing has been devoted to selecting appropriate structures and sizes, the procedure remains abstruse and requires much effort (Openshaw and Openshaw, 1997). The fundamental goal is to design a network robust enough to tackle the problem at hand, yet basic enough to train efficiently and generalize favourably. Although a rather time consuming procedure, one can determine the best architecture for a specific problem through training several different net-
work configurations and retaining the one that yields the best results (Reed and Marks, 1995).

The neurons, or nodes in an ANN, are aligned in rows called layers. This type of network is often referred to as feed-forward Multi-Layer Perceptron (MLP). A three-layer network is pictured in Figure 5; note how each neuron in the hidden layer\(^7\) (also known as the feature-detection layer) is connected to all neurons of the input and output layers. When data is presented to the ANN, signals are fed forward from the input layer to the output layer through either none, one, or more hidden layers (Dayhoff, 1990; Davalo and Naim, 1991; Leung, 1997). The function of the hidden layer is to encode concepts or hypotheses which are not directly observable in the input layer by intervening between the independent and dependent variables. The addition of one, or more, of these layers introduces stability and enables the network to extract higher-order statistics. Thus, despite the local connectivity of a neuron, the network acquires a global perspective in a rather loose sense, due to the additional dimension of interaction and the added set of connections (Carling, 1992; Garson, 1998; Haykin, 1999).

\[7\] This layer is hidden in the sense that it does not directly interact with its outside environment.
The inclusion of hidden layers in an ANN requires the execution of a training, or learning algorithm. The most popular algorithm implemented by ANNs is known as back-propagation (more accurately termed error backpropagation). It can be defined as, "a procedure for efficiently calculating the derivatives of some output of a nonlinear system, with respect to all inputs and parameters of that system, through calculations proceeding backwards from outputs to inputs" (Werbos, 1995: 135).

After initially being proposed by Werbos in his 1974 doctoral dissertation, the backpropagation learning algorithm was revived by Rumelhart, Hinton and Williams (1986) with applications of supervised learning and pattern recognition. The unique feature of backpropagation is that it organizes the internal structure of an ANN by transferring error terms (derivatives) backwards from the output layer to the connections of the input layer. That is, the error terms are used as the basis for a strategy that modifies the
strengths of the interconnections between the network nodes (Werbos, 1995). The benefits of the backpropagation algorithm, in the domain of ANNs, have been realized in applications of prediction, classification and function approximation (Rzempoluck, 1998).

The general structure of an ANN is usually dictated by the sum of the input attributes and the number of required outputs. For instance, each of the independent variables requires a unique node in the input layer, and if the intention is to predict a solitary dependent variable, a single node will be necessary in the output layer. The requisite number of nodes in the hidden layer depends in a complex way on the number of input and output nodes, the type of activation function employed, the number of cases in the data set and the intricacy of the problem at hand (Sarle, 1997). Garson (1998) suggests that parsimony is a good rule of thumb when deciding upon the properties of the hidden layer(s), highlighting that simple models are as or more efficient than those that are complex.

Certain characteristics of the hidden layer can significantly impact the precision of the model. Overfitting, for example, may occur if the configuration of the network is overly complex. In this circumstance, the ANN may fit the noise of the data set and not just the signal, resulting in model predictions that are greater than the range of the training data. Conversely, a network that is not sufficiently complex can altogether fail at detecting the signal in a complicated data set, leading to what is known as underfitting (Sarle, 1997; Garson, 1998).
Unfortunately, there are no definitive guidelines or well-defined theory to help with this stage of designing the ANN. While several researchers have proposed hypothetical foundations, they are often specific to the parameters of the problem. Openshaw and Openshaw (1997) recommend the inclusion of two hidden layers, suggesting that a second layer increases the strength of the network, therefore permitting more complicated decision surfaces to be modelled. It has been demonstrated, however, that a network with a single hidden layer with a sufficient number of nodes is capable of approximating any continuous function to any degree of accuracy. This outcome is referred to as the universal approximation theorem (Funahashi, 1989 and Hornik, Stinchcombe and White, 1989 in Rzempoluck, 1998). In addition, Garson (1998) has noted that it is unusual, save for special circumstances, to find that networks with more than three layers greatly improves the effectiveness of a model. He reasons that as the number of layers increase, the meaningfulness of backpropagated error terms decrease.

Reed and Marks (1995) suggest that small, simple models are preferred to those that are larger and more elaborate, if their performance is equal. They explain how poor generalization often occurs when the network modifies itself to account for extreme data points, and add that simple systems are more capable of generalization because they possess fewer degrees of freedom and are better constrained by the available data. Regardless, there is no straightforward way of determining the optimal number of layers without training several networks and comparing their generalization errors.
Deciding upon the correct number of neurons to begin with in the hidden layer can be critical, but deciphering the optimum number is abstract. Including too many nodes enables the ANN to memorize the training data rather than excerpt the general pattern, which empowers the net to manage unobserved situations (Fahlman, 1988; Openshaw and Openshaw, 1997). Although no theoretical standards exist for determining the number of neurons required in the hidden layer(s), many individuals have offered rules of thumb:

- between the size of the input and output layers (Blum, 1992 in Sarle, 1997);
- never more than twice the number of nodes in the input layer (Berry and Linoff, 1997 in Sarle, 1997); and
- average of the input and output nodes (Stanley, 1988 and Ripley, 1993 in Garson 1998).

However, Sarle (1997) argues that there are existing examples that could easily disprove the above rules of thumb simply because they fail to acknowledge any properties beyond the basic structure of the net. In essence, trial and error is required in order to scrutinize the effect that changing the number of hidden neurons has on the accuracy of the model.
2.3.4 Training Artificial Neural Networks

Perhaps the hardest and most computationally intensive component of ANN modelling is the training procedure. The central goal of training is to extract the underlying pattern of a data set and reproduce this signal in the internal configuration of a network. Through an input / output mapping function the network basically attempts to construct associations based upon the outcomes of historical situations. In other words, an ANN learns from experience much like a child learns to recognize different types of toys from examples of toys.

In order to facilitate network learning, data are typically divided into training and testing subsets. The training data set is simply an aggregation of cases that is used by the ANN program to learn the underlying pattern in the data. A testing data set is produced so that a quasi-independent measure on the predictive capabilities can be acquired. Periodically the training process is interrupted so that the current network can generate error responses associated with characteristics in the unseen training data (Rzempoluck, 1998). Openshaw and Openshaw (1997) recommend that the testing set of data should amount to a 20-50% random sample of the entire data set. Alternatively, as a rule of thumb, Garson (1998) and Gopal (URL) indicate that the ratio of the training to test set should be 80:20.

In more elaborate designs, a third set of data, known as the verification or cross validation set, may also be randomly derived from the sample. This data set, contrary to
the training set, is not directly involved with assisting the network establish an internal representation of the data. Instead, the verification set continually checks the training progress while the learning process evolves, ensuring the network is learning an accurate signal while providing the user with an independent benchmark for model architectures. This approach is often referred to as the hold out method (Sarle, 1997). Since this procedure can itself be subject to overfitting, the performance of the final network should be verified by the test set. The critical point is that the test set should never be used, solely, to select among different network designs (Bishop, 1995).

There is currently no outstanding theory available with regard to the appropriate division of the data into the aforementioned sets. Gopal and Fischer (1996), however, mention a random division of their data into a 70, 20 and 10% split for the training, cross validation and test sets respectively. Although ANN methodology suggests the division of the sample, the sparseness of some data sets often precludes the possibility of extracting the third cross validation set.

Primarily, there are only two different types of learning strategies available for ANNs—supervised and non-supervised. The prevailing paradigm, and of central concern to this thesis, is supervised learning. During training, the ANN continually adapts its weight matrix\(^8\) \((w_{k1}, w_{k2}, \ldots, w_{kn})\) so that the estimates produced by the model approximate the target outputs \((y_k)\) for a given set of data \((x_1, x_2, \ldots, x_n)\) which contain a particular pattern. That is, each case in the training set possesses a unique input signal and a

---

8. Initially, random weights are assigned to the matrix. As the ANN learns the relations and patterns from incoming data, the weights adjust accordingly.
corresponding output, or target. As each of the cases in the training set are presented to the network, the weights (parameters) of the network are adjusted in a direction that is expected to minimize the differences between the actual and desired responses. The adjustment of these weights facilitates an alteration in the connection structure. It is here that the learning capability of an artificial neuron is achieved (Clothiaux and Bachmann, 1994; Werbos, 1995).

Over the course of the training process, individual cases in the training set are presented to the input layer of the network. Once every unique case from the data set has been presented, often expressed as an *epoch*, a subsequent submission of the data set is initiated. By using error terms supplied by the back propagation algorithm the network attempts to generalize for the entire domain through adjusting its internal settings after each cycle of the data. This process is repeated up to the point where a potentially desirable network is obtained. A typical training procedure will often entail hundreds or thousands of epochs (Dayhoff, 1990; Werbos, 1995; Haykin, 1999). Once the ANN is trained, the individual weights in the matrix indirectly represent the influence input has upon the output.

An assessment of generalization performance can be conducted by calculating the error on the testing set. Typically, the resulting error terms descend early on in the training procedure as the model learns the main characteristics of the data. However, as the network begins to concentrate on negotiating abnormalities in the training data (idiosyncrasies in the variables that are of progressively lesser significance in relation to
the sought-after function) the error term from the testing data starts to decrease at a significantly lower rate. This is referred to as overtraining and implies that it is theoretically possible for a network to learn relations so well that it merely captures a pattern specific to the training data while unsuccessfully recognizing the pattern in a new set of data (Sarle, 1997; Rzempoluck, 1998). By monitoring the fluctuation of the test set error term, at some recurrent interval, the user can suspend training and, therefore, help prevent overtraining.

2.3.5 Strengths and Limitations

Obviously ANNs do not present the ultimate solution for modelling every nonlinear complex behaviour, but the main advantages of this technique, over traditional models, is its potential ability to approximate nonlinear relations and discover patterns and relations instead of relying solely on researchers to define them. Moreover, in the context of prediction, ANNs:

- can perform input / output mapping—an ability to learn from examples;
- contain no critical assumptions about the nature of the underlying data;
- can adapt to changing environments by adjusting the weight matrix;
- can be particularly effective when faced with noisy data; and
- can recognize complex patterns and relations that are not otherwise known (Openshaw, 1993; Openshaw and Openshaw, 1997; Haykin, 1999).
Although ANNs have been successfully trained to readily recognize patterns and generalize for unseen data, it is important to identify the limitations of this relatively young technique. In particular ANNs:

- may be redundant if there are strong statistical alternatives;
- represent a "black-box", data-driven technology;
- will only operate on unseen data that is within the bounds of the training set;
- offer limited understanding of the methodology or phenomena being modelled;
- have over 50 architectures to choose from and encompass subjective design;
- use training strategies that are ad hoc and often not technically sound; and
- retain a misleading analogy with the brain (Openshaw, 1993; Openshaw and Openshaw, 1997; Hewitson and Crane, 1994).

2.4 Artificial Neural Network Techniques—Literature Review

Nonlinear ANN models are fairly well understood, and the procedures required for successful implementation in several application domains have been documented. While the use of ANNs in the field of mass appraisal has increased over the last decade, literature assessing their capacity is relatively thin. Cited results indicate that feasibility in the application of these models varies from group to group. As a whole, the existing literature weighs heavily on the positive qualities of ANNs and their ability to outperform traditional assessment techniques. The following section will provide a brief overview of the research.
Exploring the practicability of ANNs through examination of four separate cases,\(^9\) Borst (1991) proposed that this new technique deserves strong consideration by the assessment community. By investigating "clean" and "dirty" samples, and including a price range output layer with 16 neurons from transformed and non-transformed original sales data sets of 310 transactions, the author inferred that ANNs can model a training set to an error no larger than 10%.

Studying 47 residences in England and Wales, Evans, James and Collins (1992) concluded that ANNs are able to estimate property values with a high level of accuracy and statistical significance. After removing outliers from the sample (distributed across 14 homogenous streets and covering a six month period) the results from their study are as follows: average error 5.03%, maximum error 11.1% and a Spearman Coefficient of 0.746.

Do and Grudnitski (1992) demonstrated that ANNs are particularly well suited for finding accurate solutions in a field such as property value estimation. They concluded that ANNs present a superior method over multiple regression in their ability to estimate the value of residential real estate. The Multiple Listing Service (MLS) data used in their study came from a fairly homogeneous neighbourhood with respect to income, population density, access to major public facilities and its residential-commercial

---

9. Case 1—outliers removed, single output neuron, transformed data; Case 2—outliers removed, single output neuron, non-transformed data; Case 3—outliers included, single output neuron, non-transformed data; Case 4—outliers removed, 16 output neurons, transformed data.
composition. Consequently, the range of the average selling prices of their data was fairly small—$105,000 to $288,000—from a sample size of 163.

The two competing models were used to appraise 105 independent units within the sample set. The ANN model produced a mean error of -1.31% and a mean absolute error of 6.9%. The corresponding error values were 2.73% and 11.26% for the multiple regression model. Other results indicated that approximately 20% of the regression estimates have absolute errors of less than 5%, while over 40% of the ANN estimates contain errors of less than 5%. Finally, the highest error estimate yielded by the ANN was 19%, whereas just over 10% of the regression estimates had errors of 20% or more.

Lu and Lu (1992) reported that the models developed in their study perform quite well compared to the current system of assessing dwelling values with the help of computer-based cost component systems, which assign costs to various characteristics of a residential property. The data used in their study spanned two years and included 336 cases for a midwestern city in the United States.

The results of their proposed forecast model performed favourably to previously assessed values. The results, measured by mean absolute error, were 6.99% and 9.56% respectively. With regard to the validity of the data, for the ANN model, the mean absolute deviation was 5,327 and the mean square error 52,953,694 compared to 7,570 and 100,989,898 obtained from human assessments aided by a computer based cost
component system. They emphasized, however, that the City Assessor decided that ANN models are best left in the back room (due to the unfamiliar, black-box nature of the technology) to help reinforce existing techniques by providing an objective checking device.

Examining the performance of ANN techniques, Tay and Ho (1992) explored 1,055 residential apartment sales in two compact, geographically detached sub-markets of Singapore. The results indicated that their ANN model outperformed the predictive capabilities of traditional multiple regression techniques. Specifically, the derived mean absolute error of 3.9%, with an associated standard deviation of 31.9%, corresponded to an error of 7.5% and a standard deviation of 44.4% for the multiple regression model. Even after subsequent removal of outliers, Tay and Ho (1992) highlighted that their ANN model produced better estimates and, therefore, presented a good alternative to the traditional measures.

While the results from the studies mentioned above provide evidence that ANN models can outperform multiple regression models, Worzala, Lenk and Silva (1995) were not as confident about the ability of ANNs to estimate property values. In their paper, the authors argued that ANNs are not a superior tool for appraisal and advised caution in the implementation of this new technology due to the problems they encountered.

Precisely, three cases were constructed and two separate ANN software packages (@Brain and NeuroShell) were tested against multiple regression with 288 cases from
a small town in Colorado, United States. A summary of the results are illustrated in the table below.

<table>
<thead>
<tr>
<th></th>
<th>Case 1 (entire original data set)</th>
<th>Case 2 ($105,000-288,000)</th>
<th>Case 3 (homogeneous sub-set)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple Regression</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean abs. error</td>
<td>15.20%</td>
<td>11.10%</td>
<td>12.80%</td>
</tr>
<tr>
<td>&lt;5% error</td>
<td>32.40%</td>
<td>37.20%</td>
<td>24.10%</td>
</tr>
<tr>
<td><strong>@Brain Software</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean abs. error</td>
<td>13.20%</td>
<td>10.00%</td>
<td>11.70%</td>
</tr>
<tr>
<td>&lt;5% error</td>
<td>29.60%</td>
<td>41.90%</td>
<td>31.00%</td>
</tr>
<tr>
<td><strong>Neuroshell Software</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean abs. error</td>
<td>14.40%</td>
<td>13.10%</td>
<td>11.60%</td>
</tr>
<tr>
<td>&lt;5% error</td>
<td>32.40%</td>
<td>32.60%</td>
<td>34.50%</td>
</tr>
</tbody>
</table>

Table 1: Descriptive Statistics for Prediction Results from Worzala, Lenk and Silva (1995)

As emphasized by Worzala, Lenk and Silva (1995), observed differences in the statistics above are not significant and fail to compare to the results published by Do and Grudnitski. The authors stated, however, that the results may be a function of the software, the model architecture or the sample data used in the examination. They indicated the opportunity and potential for improving the accuracy of ANNs as, “when statistical packages were first introduced there were often inconsistencies due to programming errors and other complications” (Worzala, Lenk and Silva, 1995: 200).

Regardless of the results presented above, several methodological issues are neglected, such as the homogeneity of neighbourhoods, an inadequate number of cases for training, improper training procedures, failure to benchmark the results or even unsatisfactory evaluation techniques. More importantly, there is no inclusion of signif-
icant extrinsic characteristics, suggesting that the models fall short of their true potential.

2.4.1 Spatial Consideration

In 1998 Lewis and Ware presented a paper that addressed the need for a locationally-portable ANN model. They highlighted the lack of attention within the current ANN residential real estate prediction literature on the effect of location upon property value. The authors demonstrated that use of census data can significantly improve prediction accuracy in real estate appraisal.

The primary objective of their research was to alleviate the limitation of single, homogeneous-area ANN estimations. They proposed that a heterogeneous market area can be modelled independently with the use of several homogeneous sub-market ANN models, implementing a Kohonen Self Organizing Map (SOM) to classify sub-market areas using census data for the UK.¹¹

¹⁰ The number of training cases required primarily depends on the amount of noise in the data and the complexity of the function being modelled. A liberal rule of thumb is to have at least ten times as many cases as nodes in the input layer. A more conservative rule of thumb is to have a minimum of ten times the number of nodes in the input and hidden layer(s). However, this still may not be enough to avoid overfitting. Thus, some researchers have proposed that there should be thirty times as many cases as connections in the network (Garson, 1998; Sarle, 1997; StatSoft, URL).
Compared to multiple regression analysis and conventional ANN models, Lewis and Ware discovered that ANN techniques reinforced with a SOM clustering algorithm improve upon the predictive capacity of traditional approaches. When tested on roughly three years of data from a large, heterogeneous residential area in Cardiff, mean absolute errors of 8%, 18% and 24% were derived from the enhanced nonlinear, nonlinear and linear models respectively. Furthermore, 22%, 74% and 79% of the predictions generated by the same three models had errors greater than 10%. A substantive question thus arises—how much do locational attributes impact the value of a dwelling? Taking this into consideration, how can a spatial component improve the results? These components will be explored in the following chapters.

11. The SOM algorithm, one of the best known ANN algorithms, was developed by a Finnish researcher, Teuvo Kohonen. A SOM is unique in the sense that it constructs a topology-preserving map where the location of a neuron carries semantic information. Essentially, the neural cells organize themselves into groups, according to incoming information—the output is a two-dimensional display of the input space enabling easy visualization. These maps are mostly used for the classification or clustering of data; with respect to census data, the most important aspects of SOMs are in exploratory data analysis and pattern recognition (Davalo and Naim, 1991; Kohonen, 1997). Classifications from census data using SOMs have also been explored, in a similar fashion, in recent years by a few researchers (Winter and Hewitson, 1994; Openshaw, 1994; Openshaw and Wymer, 1995). The categories of census data used in the classification by Lewis and Ware (1998) included occupation, employment status, education, housing type, tenure, ethnicity and others. However, they stipulate that their methods for determining the independent census variables relied heavily on \textit{a priori} knowledge.
Chapter III: Data Selection and Preparation

3.1 Intrinsic Characteristics—Residential Transaction Attribute Data

The residential property transaction data for this study was kindly provided by Stanley Hamilton, Faculty of Commerce, University of British Columbia. The core of the database was established from Multiple Listing Services (MLS) data and augmented with supplementary property sources.¹

The records used in this analysis were restricted to the City of Vancouver,² British Columbia (see figure below) primarily because of the established and contiguous housing market there. The municipality of Vancouver represents a geographic region small enough to keep data development manageable while encompassing an area large enough to explore the flexibility and power of the competing models. In order to maintain temporal consistency with statistical socioeconomic variables, a six month sample was cropped from the database straddling census day, May 14, 1996. The database was further limited to detached residential parcels flagged as houses.

---

¹ The MLS database contains over 80% of the actual market transactions and represents the single most comprehensive and accurate data set available (excluding transactions that are not conducted through a real estate board, such as a sale within a family and sale by owner) (Hamilton and Hobden 1992).

² Located in the southwest corner of Canada, Vancouver represents the third largest urban centre in Canada with a 1996 population of 1,831,665 for the Census Metropolitan Area (CMA - a statistical classification defined as a large urban conglomeration, with a population of at least 100,000, of rural and urban areas that have a high degree of socioeconomic interaction with the urban core). The last official population count for the municipality of Vancouver—a sub-area of the CMA covering slightly greater than 113 Km²—coinciding with the time period of this study, was just over 514,000 (Statistics Canada).
For several reasons only six particular intrinsic dwelling attributes were extracted from the MLS database (the short-listed variables are outlined in Table 2 below). Firstly, there was an ultimate desire to keep the models as simple as possible. Secondly, these variables are consistent with the intrinsic variable sets used in several other studies (see, for example, Hamilton and Hobden, 1992; and Hoyt and Rosenthal 1997), thus providing past and future comparability. Finally, the logic behind my variable selection is based on the view that additional attributes simply contribute noise with minimal statistical explanation.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sale Price</td>
<td>$148,624</td>
<td>$3,000,000</td>
<td>$514,565</td>
<td>$411,500</td>
</tr>
<tr>
<td>Lot Dimension</td>
<td>1,716</td>
<td>20,000</td>
<td>5,043</td>
<td>4,148</td>
</tr>
<tr>
<td>Square Footage of Dwelling</td>
<td>636</td>
<td>8,800</td>
<td>2,277</td>
<td>2,168</td>
</tr>
<tr>
<td>Dwelling Age</td>
<td>0</td>
<td>30</td>
<td>22.07</td>
<td>30</td>
</tr>
<tr>
<td># of Bedrooms</td>
<td>1</td>
<td>9</td>
<td>4.19</td>
<td>4</td>
</tr>
<tr>
<td># of Fireplaces</td>
<td>0</td>
<td>5</td>
<td>1.11</td>
<td>1</td>
</tr>
<tr>
<td># of Bathrooms</td>
<td>1</td>
<td>5</td>
<td>2.70</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: Descriptives of Intrinsic Characteristics

Note: Cell shading indicates the dependent variable.

Due to the negative impact of outliers on any model performance, slight alterations were made to the number of records of the final variables. Specifically, cases with dwelling values over $3,000,000 were removed (still leaving 153 cases worth $1,000,000 or more) and the lot dimensions (after recalculation and cleaning), were constrained to those under 20,000 square feet. Further, original age of dwelling values ranged from 0 to 98 years with a spurious recurring cell label of OT (Old Timer). Given the subjectivity and uncertainty of the OT label, subsequent recoding was required—along with cell values greater than or equal to 30 years—assigning them a value of 30 years (or more). After removing outliers, questionable duplicate property listings and missing or severely suspect data, the transaction data set comprised 1908 cases.
3.2 Extrinsic Characteristics—Socioeconomic Attribute Data

Preference for residential location, a partial function of the socioeconomic fabric and quality of a neighbourhood, is also a critical component to estimating the value of a dwelling. Clearly, housing in different locations will deliver very different services to the inhabitants. However, little empirical research in the domain of housing price analysis has been done on detailed locational externalities. Attention has generally been concentrated on intrinsic attributes while treating the market as spatially uniform.

Existing models tend to contain a similar range of structural attributes, while the range of locational attributes is diverse and marginal in comparison. As a result, there has seldom been consensus on the types of locational attributes that influence price, while empirical evidence that does exist has been inconclusive. This could be a function of a fundamental lack of understanding of how spatial and socioeconomic processes operate at different scales, and how this relates to the structure of the collective environment. Furthermore, data relating to the location of a property are much harder to quantify and conceptually less tangible than structural data—their derivation requires significantly more time and effort.

In an attempt to extend current research, efforts have been made to explore and incorporate supplemental extrinsic data in a multi-level framework. The following sections will illustrate how GIS can link socioeconomic data across spatial scales,
facilitating the enrichment of the housing transaction database and, therefore, resolving many technical and conceptual problems.

3.2.1 The Modifiable Areal Unit Problem (MAUP)

Geographic Information Systems (GIS) are extensively implemented for, among a range of other things, mapping socioeconomic phenomenon. However, dissimilar to features in the physical environment, such as a road intersection or a river mouth, individuals are almost always referenced via some larger spatial object such as a postal code or collection unit due to issues of confidentiality\(^3\) (elaborated upon below). Therefore, researchers are rarely able to define the precise location of an individual by merely providing a map reference. This has far reaching implications when attributes such as average family income varies considerably within small localities.

A challenge researchers face when analysing census variables is that patterns in the data may have as much to do with the collection units as to the underlying conditions. This problem, known as the Modifiable Areal Unit Problem (MAUP), was initially identified by Gehlke and Biehl (1934),\(^4\) and stems from the fact that changing the areal boundary configuration for which individual counts are aggregated may yield transformations in the data patterns that are observed. Using data for the smallest available

---

3. By definition, several social phenomena, such as average family income, can only be identified in aggregated entities.
4. Openshaw and Taylor (1979, 1981) and Openshaw (1984a) later added insights to the MAUP.
areal unit will minimize the impact of aggregation damage; as the units become smaller they become closer to the representation of individuals. Openshaw and Wymer (1995: 244) elaborate:

If there is concern about ecological inference errors then the smallest possible geographical areas will be 'best'... Strictly speaking, even the smallest census areas are too large and are possessed of unpredictable levels of social heterogeneity. The best that can be done at present is to hope that these data artefacts are not too damaging and to bear this deficiency in mind when interpreting or using the results.

In this context, a geographically related issue that should be mentioned is the predicament referred to as the ecological fallacy—spurious associations (or conversely, the individualistic fallacy). This implies that it is incorrect to infer that an association exists at a smaller scale from evidence of association at a larger scale (Openshaw, 1984b); all relations may hold only for that particular aggregation of data. More often than not, there is more variation within an areal unit than there is between areal units. The aforementioned concerns can have dramatic impacts on the results of socioeconomic analyses, but it is beyond the scope of this thesis to empirically or theoretically evaluate the consequences (see, for example, Fotheringham and Wong, 1991; Green and Flowerdew, 1996; and Wrigley, Holt, Steel and Tranmer, 1996).

5. The inappropriate assigning of an average or characteristic of areal data to individuals and point locations within those areas.
3.2.2 Spatial Scale of Socioeconomic Data

The spatial scale of data provided by Statistics Canada imposes restrictions on the analysis of socioeconomic data. The boundaries depicted in Figure 7 represent Census Tracts (CT) for the CMA of Vancouver and outline the general area of interest for the proposal.

Figure 7: Census Tract boundaries for the CMA of Vancouver

6. The boundaries of these units, or polygons, are designed to be compact and follow permanent and recognizable physical features. Revision of CT polygons is discouraged to enable comparison between censuses. The population that comprise them range from 2,500 to 8,000, with a preferred average of 4,000 (Statistics Canada, 1999).
Outside the Municipality of Vancouver (illustrated in Figure 6), pockets of developed low to medium-density clusters are dispersed among the metropolitan area. This is highlighted by the clusters of small polygons embedded in less populated, undeveloped rural fringes that are enclosed in the larger polygons. To reduce the study area to one that is more manageable and relatively homogeneous, with regard to density, data for the Municipality of Vancouver (CT map shown below) will be isolated and extracted.

Readily apparent at this scale are the arbitrary boundaries delineated by the CTs. Ideally, more relevant boundary configurations should have been developed by the statistics agency. Nevertheless, social scientists must live with this and should therefore use the lowest level of aggregated data whenever possible.
In Canada, the smallest geographic area available for census information is the Enumeration Area (EA). These areas are primarily defined by the territory a census representative can enumerate in a day. EAs are designed to be as compact as possible and are nested within the boundaries of a CT. Figure 9 outlines the boundaries of the EAs for the Municipality of Vancouver in 1996.

![Figure 9: Enumeration Area boundaries for the Municipality of Vancouver](image)

The size of each polygon, relative to the CT boundaries (Figure 8), clearly illustrates the difference in density between the two. The increased boundary resolution that is available when using EA data will undoubtedly improve the ability to represent heterogeneity inherent in socioeconomic data.

---

7. The number of dwellings in each of these areas varies between a maximum of 440 to a minimum of 125 (Statistics Canada, 1999). For the 2001 census, Statistics Canada is developing Dissemination Areas (DA) for data output. These new census units will supposedly address requested improvements to EAs such as increased temporal stability, reduced data suppression, intuitive boundaries, compactness and homogeneity (Puderer, 2000).
3.2.3 Suppression of Socioeconomic Data

Unfortunately for researchers, due to important issues of confidentiality, Statistics Canada applies procedures to prevent the possibility of associating data with any identifiable individual. Data are, therefore, routinely randomly rounded (to a multiple of 5 or 10) and are suppressed for certain census collection units (outlined above). Suppression results in the deletion of all characteristic data for units with populations below a specified threshold. Specifically, data for an area with a population of 40 individuals or less is suppressed. If the data contain information on income distribution, EAs or CTs with less than 250 persons have their data suppressed (Statistics Canada, 1997b; 1999).

3.3 Data Reduction Techniques—Factor Analysis

In order to introduce a socioeconomic dimension to the model, the entire census profile series for EAs—consisting of 1605 variables—needed to be condensed to a few variables. In particular, the database was initially reduced, then sharpened using factor analysis techniques for the extraction of variables that would explain or account for a substantial proportion of the interrelations (multivariate patterns) of a diverse range of urban and social phenomena.

8. Confidentiality is a major concern of census-takers around the world. For instance, the 1981 census in West Germany was abandoned due to widespread disbelief in the confidentiality of computerized census files (Coombes, 1995).
Factor analysis\textsuperscript{9} is often used in data reduction for the identification of a small number of factors that explain most of the variance observed in a much larger number of manifest variables. The procedure attempts to identify underlying variables or factors that explain the pattern of correlations within a set of observed variables. For an exhaustive account of the process, see Harman (1976) and Gorsuch (1983). Some see the analysis then as a systematic way of deriving generalisations—reducing the complexity of the real world into a set of basic dimensional traits by determining which variables are related, and which of them are not (Davies, 1984; Bryman and Cramer, 1997). Bailey and Gatrell (1995: 225-226) elaborate:

The analysis attempts to identify how many of these factors are significant in the data, in what relative order, and how each of these relate to the observed attributes... Factor analysis can thus be thought of as a technique by which a minimum number of hypothetical variables are specified in such a way that after controlling for these hypothetical variables all the remaining correlations between the original observed variables become zero...

When factor analysis techniques are applied to the problem of areal differentiation, the study has been identified as factorial ecology\textsuperscript{10} (see Rees, 1971; Berry, 1971; Meyer, 1971; and Davies, 1984). Essentially, factorial ecologies are descriptive summaries of

\begin{itemize}
  \item Factor analysis essentially distinguishes which variables are measuring the same concept and to what extent. The procedure differs from principal components in that the latter is simply concerned with accounting for the variance of each of the original variables by rotating axes. In factor analysis the attention is focused on explaining correlations in the original variables with respect to a model which proposes a certain number of common factors (Bailey and Gatrell, 1995). While these two methods are based on different mathematical models, they can produce similar results. The main distinction between cluster analysis and factor analysis is that former is hierarchical, driven by the strength of individual correlations. Factor analysis, on the other hand, considers the relations between all variables concurrently.
  \item The term was originally coined by Sweetser (1965) to denote a novel approach to the study of urban dimensions and the classification of social areas.
\end{itemize}

\footnotesize
\textsuperscript{9} Factor analysis essentially distinguishes which variables are measuring the same concept and to what extent. The procedure differs from principal components in that the latter is simply concerned with accounting for the variance of each of the original variables by rotating axes. In factor analysis the attention is focused on explaining correlations in the original variables with respect to a model which proposes a certain number of common factors (Bailey and Gatrell, 1995). While these two methods are based on different mathematical models, they can produce similar results. The main distinction between cluster analysis and factor analysis is that former is hierarchical, driven by the strength of individual correlations. Factor analysis, on the other hand, considers the relations between all variables concurrently.

\textsuperscript{10} The term was originally coined by Sweetser (1965) to denote a novel approach to the study of urban dimensions and the classification of social areas.
all the dimensions of residential variation that have been measured, which are amalgamated into a single model (Meyer, 1971).

The reliability of factors that emerge from these analyses, however, depends largely on the size of the sample. While there is no consensus in the literature as to what the ideal size should be, there is universal agreement that there should be more cases than attributes, with more cases providing greater validity of results. Various authors have recommended that the minimum number of cases per attribute range from five to ten. An absolute minimum of five cases per attribute and not less than one-hundred cases per analysis, for instance, has been proposed by Gorsuch (1983).

3.3.1 Variable Selection

An important set of challenges associated with factor analysis are the decisions regarding which variables should be included in the analysis. The selection of variables, at least initially, is guided by intuition. Primarily, significant variables were identified based upon relevancy to the problem domain, expected axes of differentiation and those used in previous studies that demonstrated good results in defining the separate structures, while taking care to have an appropriate balance between the available categories.
The 1605 original variable set was initially reduced through a subjective removal of variables that were redundant, severely suppressed and irrelevant. In addition, variable sets that add up to 100% of the sample were split,\textsuperscript{11} and certain variable pairs such as South and East European ethnic origins were combined. This procedure included a re-run of the analysis on successive reductions of the variable set in order to assess the loss of dimensions and descriptiveness.\textsuperscript{12} Eventually, the 768 case (EAs) data set was further reduced to 29 variables (transformed if not already expressed as a percent, rate or dollar value) and subjected to a final factor extraction. The variables selected for the analysis are listed in the table below.

\textsuperscript{11} Using the percentage of immigrants and non-immigrants, for example, will bias the correlations while appearing on opposite sides of a factor (see Davies, 1984).
\textsuperscript{12} Variables were also removed after examining their communality and loadings (i.e., if they had small correlations with the other variables).
<table>
<thead>
<tr>
<th>Variable #</th>
<th>Variable Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>% Ages 0 to 19</td>
</tr>
<tr>
<td>2</td>
<td>% Ages 25 to 44</td>
</tr>
<tr>
<td>3</td>
<td>% Ages 65 and over</td>
</tr>
<tr>
<td>4</td>
<td>Average number of persons in private households</td>
</tr>
<tr>
<td>5</td>
<td>% Female lone parent</td>
</tr>
<tr>
<td>6</td>
<td>% Living alone</td>
</tr>
<tr>
<td>7</td>
<td>% All immigrants</td>
</tr>
<tr>
<td>8</td>
<td>% immigrants arriving between 1981-1990</td>
</tr>
<tr>
<td>9</td>
<td>% immigrants arriving between 1991-1996</td>
</tr>
<tr>
<td>10</td>
<td>% Non-official languages</td>
</tr>
<tr>
<td>11</td>
<td>% British Isles origins</td>
</tr>
<tr>
<td>12</td>
<td>% South, East European origins</td>
</tr>
<tr>
<td>13</td>
<td>% South, East and Southeast Asian origins</td>
</tr>
<tr>
<td>14</td>
<td>% African/Pacific Islands/Latin American/Carribean/Aboriginal origins</td>
</tr>
<tr>
<td>15</td>
<td>% Visible minorities</td>
</tr>
<tr>
<td>16</td>
<td>Unemployment rate (Male)</td>
</tr>
<tr>
<td>17</td>
<td>Participation rate (Female)</td>
</tr>
<tr>
<td>18</td>
<td>% Management occupations</td>
</tr>
<tr>
<td>19</td>
<td>% Trades, transport and equipment operators and related occupations</td>
</tr>
<tr>
<td>20</td>
<td>% Occupations unique to processing, manufacturing and utilities</td>
</tr>
<tr>
<td>21</td>
<td>% Self-employed (unincorporated and incorporated)</td>
</tr>
<tr>
<td>22</td>
<td>% Public transit (mode of transportation)</td>
</tr>
<tr>
<td>23</td>
<td>% Less than grade 9</td>
</tr>
<tr>
<td>24</td>
<td>% With certificate or diploma (other non-university education)</td>
</tr>
<tr>
<td>25</td>
<td>% With bachelor's degree or higher (University)</td>
</tr>
<tr>
<td>26</td>
<td>% Movers 1 year ago (Mobility status)</td>
</tr>
<tr>
<td>27</td>
<td>Average family income $</td>
</tr>
<tr>
<td>28</td>
<td>% Owned Dwelling</td>
</tr>
<tr>
<td>29</td>
<td>Average gross rent $</td>
</tr>
</tbody>
</table>

Table 3: Variable Listing for Factor Analysis

Note: See Statistics Canada (1999) for variable definitions and how they are derived.

### 3.3.2 Factor Extraction

An important step in attempting to reduce the number of variables is actually determining the number factors to be retained.\(^\text{13}\) Essentially this is a question of how many of
the less significant factors should be eliminated. Plotting the eigenvalues against the corresponding factor numbers—known as a scree test (Cattell, 1966)—is a useful method that provides insight into the number of factors to extract. The graph simply illustrates the variance accounted for (in descending order) by the extracted factors. The plotted line of the graph typically depicts a break between the steep slope of the initial significant factors and the gentle slope of the smaller ones (Bryman and Cramer, 1997). Theory supporting the scree test advises to drop all factors that follow the point where the curve forms an elbow toward a less steep decline.

Given the scree test produced by SPSS (Figure 10), only the first four factors are worth retaining for the analysis, all of which, incidentally, have eigenvalues greater than one. After examining the results from a rotated factor solution, with four factors requested to be extracted, the analysis was subsequently performed withdrawing only three factors. The last factor was excluded from the procedure as its eigenvalue was close to one (the top three eigenvalues stand apart considerably) and, therefore, added little to the variance accounted for by the solution. Moreover, the loadings on the fourth factor were scattered with weak associations and thus became harder to interpret.

13. The optimal number of factors to be retained is not a straightforward exercise since the decision is ultimately subjective. Although several criteria exist for the number of factors to be extracted, they merely provide suggested guidelines.
14. Eigenvalues are simply the variance extracted by the factors.
15. Scree is a geological term denoting the debris that is typically found at the bottom of a rocky slope.
16. An alternative criteria for deciding how many factors to retain is known as the Kaiser rule where all components with eigenvalues under 1.0 are dropped.
3.3.3 Results of Factor Analysis

Performing an oblique\textsuperscript{17} factor analysis on the remaining data yielded loading results displayed in Table 4. Examining the structure matrix,\textsuperscript{18} it is noted that factor one has particularly high loadings on four of the variables: visible minorities; South, East and

\begin{itemize}
\item An oblique rotation produces factors which are allowed to be correlated with one another. Conversely, implementing an orthogonal rotation, the factors may be forced to be independent when in reality they may indeed be related. Comparing results from each rotation method, it was discovered that the extracted factors were identical and save for two sets of variables in opposing order (the second and third, and sixth and seventh ranked variables on the first component), the variation in returned coefficients was negligible.
\item The structure matrix, which is generally used to interpret the factors, is made up of weights which reflect the unique variance each variable contributes to a factor. This differs from a matrix consisting of correlations between the variables and the factors, known as a pattern (Bryman and Cramer, 1997). The values in the structure matrix are calculated by multiplying the pattern matrix by the factor correlation matrix.
\end{itemize}

\textsuperscript{17} An oblique rotation produces factors which are allowed to be correlated with one another. Conversely, implementing an orthogonal rotation, the factors may be forced to be independent when in reality they may indeed be related. Comparing results from each rotation method, it was discovered that the extracted factors were identical and save for two sets of variables in opposing order (the second and third, and sixth and seventh ranked variables on the first component), the variation in returned coefficients was negligible.

\textsuperscript{18} The structure matrix, which is generally used to interpret the factors, is made up of weights which reflect the unique variance each variable contributes to a factor. This differs from a matrix consisting of correlations between the variables and the factors, known as a pattern (Bryman and Cramer, 1997). The values in the structure matrix are calculated by multiplying the pattern matrix by the factor correlation matrix.
Southeast Asian origins; non-official languages; and all immigrants. This factor can be interpreted broadly as a component of *ethnicity*. The second factor has high positive loadings for average family income, university degree or higher, average gross rent and self employment variables. Generally speaking, this factor can be defined as an *economic* status dimension. Finally, two age components (25 to 44 and 65 +) with opposing signs, female participation rate and the mobility variable (#26) are strongly associated with the third factor. Conclusively, the last factor can be characterised as an indication of *family* structure; occasionally interpreted as a *life cycle* component.

By examining EA *choropleth* maps of the variables loading on the family factor and comparing them to a City of Vancouver zoning map, the factor may also be interpreted as a housing market indicator distinguishing between single and multiple family dwellings. In particular, the polarised pattern that emerged identified an inner city zone (of multiple family dwellings) extending outward from the Central Business District (CBD) and bounded by Alma, 16th Avenue, and Nanaimo (the edge of the surrounding single

19. Recent immigration has redefined the demographic, social, economic and cultural landscape of large Canadian cities. In particular, British Columbia and, of concern to this thesis, Vancouver has experienced high levels of immigration over the past decade, not only surpassing immigration levels of the post Second World War period, but since the first part of the century when the country was engaged in an intense process of nation building. Specifically, international annual arrivals to Canada have fluctuated around 200,000 while provincial counts (the vast majority of which settle in the vicinity of Vancouver) exceeded 50,000 in 1996. New immigration legislation in 1967, that redefined the terms of entry, favoured arrivals from non-traditional sources. By 1996, 80 percent of arrivals were originating from Asia, led by Hong Kong and Taiwan (Ley, 1999; Hiebert, 1999). Immigration to cities such as Vancouver has been coupled with very low and decreasing domestic in-migration and frequently with considerable out-migration. Consequently, immigration has become a major component of the demand for housing (Ley, Tutchner and Cunningham, 2001).
family dwelling districts). There also existed a spatially disaggregated, but similarly multiple dwelling zoned area, at the southern extent of Granville (see Figure 6 on page 42) that coincided as well with the distribution of the family factor variables.

20. Griffith (1992: 23) emphasizes a few contributing components of location on value of residential real estate in the context of the spatial distribution of income, “residential location... is a function of neighbourhood amenities, a preference for pleasant living environments, and socio-economic features of neighbourhoods. These elements lead to spatial externalities, which arise from specific sites conferring advantages in addition to their accessibility to centrally located places in a city. Thus, similar income groups tend to cluster together in neighbourhoods, with the higher-income households being willing to pay for these positive externalities. Such spillover effects embed spatial autocorrelation in the geographic distribution of income”.

21. A choropleth map, is defined by Chrisman (1997: 13) as, “a thematic mapping technique that displays a quantitative attribute using ordinal classes applied as uniform symbolism over a whole areal feature. Sometimes extended to include and thematic map based on symbolism applied to areal objects”.
<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Factor One</th>
<th>Factor Two</th>
<th>Factor Three</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Visible minorities</td>
<td>0.963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% South, East and Southeast Asian origins</td>
<td>0.940</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Non-official languages</td>
<td>0.939</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% All immigrants</td>
<td>0.900</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average number of persons in private households</td>
<td>0.858</td>
<td>0.352</td>
<td></td>
</tr>
<tr>
<td>% Ages 0 to 19</td>
<td>0.777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% immigrants arriving between 1981-1990</td>
<td>0.772</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% immigrants arriving between 1991-1996</td>
<td>0.634</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Owned Dwelling</td>
<td>0.593</td>
<td>0.576</td>
<td></td>
</tr>
<tr>
<td>% Occupations unique to processing, manufacturing and utilities</td>
<td>0.534</td>
<td>-0.335</td>
<td>0.486</td>
</tr>
<tr>
<td>% British Isles origins</td>
<td>-0.488</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Female lone parent</td>
<td>0.463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average family income $</td>
<td></td>
<td></td>
<td>0.779</td>
</tr>
<tr>
<td>% With bachelor's degree or higher (University)</td>
<td></td>
<td></td>
<td>0.744 0.412</td>
</tr>
<tr>
<td>Average gross rent $</td>
<td>0.323</td>
<td>0.696</td>
<td>0.325</td>
</tr>
<tr>
<td>% Self-employed (unincorporated and incorporated)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Less than grade 9</td>
<td>0.528</td>
<td>-0.558</td>
<td></td>
</tr>
<tr>
<td>% Management occupations</td>
<td></td>
<td>0.527</td>
<td></td>
</tr>
<tr>
<td>Unemployment rate (Male)</td>
<td></td>
<td>-0.519</td>
<td></td>
</tr>
<tr>
<td>% Living alone</td>
<td>-0.446</td>
<td>-0.455</td>
<td>0.441</td>
</tr>
<tr>
<td>% Trades, transport and equipment operators and related occupations</td>
<td></td>
<td></td>
<td>-0.425</td>
</tr>
<tr>
<td>% African/Pacific Islands/Latin American/Carribean/Aboriginal origins</td>
<td></td>
<td></td>
<td>-0.385</td>
</tr>
<tr>
<td>% Ages 25 to 44</td>
<td>-0.349</td>
<td></td>
<td>0.854</td>
</tr>
<tr>
<td>% Ages 65 and over</td>
<td></td>
<td>-0.785</td>
<td></td>
</tr>
<tr>
<td>Participation rate (Female)</td>
<td></td>
<td></td>
<td>0.404 0.703</td>
</tr>
<tr>
<td>% Movers 1 year ago (Mobility status)</td>
<td></td>
<td></td>
<td>0.691</td>
</tr>
<tr>
<td>% With certificate or diploma (other non-university education)</td>
<td></td>
<td></td>
<td>0.620</td>
</tr>
<tr>
<td>% South, East European origins</td>
<td></td>
<td></td>
<td>0.426</td>
</tr>
<tr>
<td>% Public transit (mode of transportation)</td>
<td></td>
<td>-0.315</td>
<td>0.383</td>
</tr>
</tbody>
</table>

Table 4: Structure Matrix Results From an Oblique Factor Analysis

Note: Coefficients have been sorted according to size and suppressed if absolute value is less than 0.3. The absolute value of 0.3 has emerged as generally acceptable because it is the closest whole number to the situation in which 10% of the variance (the square of 0.317) in a variable is accounted for by the factor (Davies, 1984).

The results from the analysis are congruent with the literature on factor ecology (studies on Canadian, American and European cities) as the identification of three general factors: ethnic status, economic status and family status was confirmed (Shevky and
Williams, 1949; Bell, 1953; Shevky and Bell, 1955; Schwirian and Matre, 1974). From these factors, the three top ranking variables, visible minority, average family income and middle age, were extracted with the assumption that they constitute a meaningful combination that underlie theoretical dimensions and generally describe the socioeconomic characteristics of Vancouver.

3.4 Surface Modelling of Socioeconomic Variables

In this thesis, an enhanced surface model of census data has been explored. Specifically, the Inverse Distance Weighting (IDW) module in ArcInfo was employed to develop surface representations of the top three ranking socioeconomic variables from the above factor analysis. Statistics Canada's 1996 geography files for the Municipality of Vancouver were used to facilitate the creation of the models.

22. More recent studies have revealed that the Shevky and Bell's (1955) three axes of differentiation, although initially helpful, were an inadequate description of the social mosaic of cities (see Murdie, 1969; Davies, 1984; and Davies and Mudie, 1993). However, for the purpose of this thesis, the objective was simply to extract three dominant factors.

23. Given the strong level of correlation amongst the four highest variable loadings for each factor (for example, the four variables from the first factor yield bivariate Pearson correlations greater than 0.85), selection of lower ranking variables would presumably have little impact on the outcome of the model results.

24. ArcInfo is a GIS application developed by Environmental Systems Research Institute (ESRI).
3.4.1 Data Set Development

The challenge of matching residential transaction records with precise population characteristics presents a major obstacle to a spatially-aware analysis of housing markets. To elaborate, an aggregation of EA units (see Figure 9) poses a problem in that they encompass urban space up to the road segment where a dwelling's address is fixed. Therefore, attribute values could be misrepresented (i.e., assigned a value from an adjacent EA) for those houses that are situated on streets that straddle EAs. Furthermore, non-residential landuses are included in the census units, while polygons with missing data will simply propagate erroneous values to the model.

If GIS are to be used to extract socioeconomic characteristics, for use by separate models, it is essential that care is taken to ensure that a particular model is an acceptable reflection of the real world. Traditional models of urban population structure have assumed discontinuous patterns of population, as demographic landscapes are carved into discrete blocks. The alternative is to model the distribution of the phenomena by disregarding the areal units. If it is agreed that socioeconomic characteristics of an area are influenced by its surroundings (spatial changes are continuous), then population phenomena can theoretically be represented by three-dimensional surfaces. While a surface model of population will inevitably contain artifacts inherited from

25. Although Griffith (1992) has proposed a technique of dealing with missing values in urban census data, the dilemma of value extraction fails to be resolved. Alternatively, EA units with suppressed data could be assigned values that are statistically derived from units at a larger scale, such as CTs. However, with the confidentiality techniques (random rounding) employed by Statistics Canada, this will likely introduce more error to the model.
the interpolation procedure employed (see the surface swellings, or blisters and fault lines left from the original EA centroids in Figure 14), it is likely that in most cases they will be more accurate estimates of the underlying characteristic (Wood, Fisher, Dykes, Unwin and Stynes, 1999).

Socioeconomic data surfaces derived from census centroids, have recently been explored by Martin (1989) and Bracken and Martin (1989). However, the general technique can be traced back well before the widespread use of GIS (see Nordbeck and Rystedt, 1970). One of the advantages of modelling socioeconomic phenomenon as a surface is that analysis and manipulation may be performed independently of any fixed set of areal units. The resulting model incorporates a high degree of truly geographic information about the underlying population, therefore enabling analyses which are meaningful in terms of geographic concepts such as the variance of an attribute across space (Martin, 1996a).

For the nominated socioeconomic variables, masking (removal) of the non-residential areas in the city is critical for accuracy. To elaborate, when representing a population as a surface, discontinuity in residential landuse (for example, parks or golf courses) presents an interruption in the flow of settlement. Therefore, an interpolation routine

26. It is assumed that the population attribute peaks at the EA centroid location and decays with distance away from that point. Alternatively, a volume preserving smoothing filter could be applied to the surface so that idiosyncrasies are stretched across space, therefore removing sudden changes in the model. It can be argued, however, that these sporadic fluctuations more realistically represent the distribution of an underlying population characteristic.

27. A method that involves a process to estimate the value of a continuous variable at an unknown location between two or more known neighbouring locations.
should not be permitted to leak into or across non-residential space as adjacent, or contiguous, residences should have more association than those that are spatially detached.

The image in Figure 11 partially emphasizes the consequence of this challenge. For instance, the large hatched area in the top of the image constitutes one EA, containing 569 individuals. Inaccuracy arises from the fact that these individuals reside strictly in the southeast portion of the EA (delineated by the white hatched square shape) adjacent to the non-populated grey area known as Stanley Park. Nevertheless, the population is affiliated with the entire EA, grossly misrepresenting the actual location of the population and, thus, significantly impacting the eventual interpolation of the surrounding area.

Figure 11: Populated Versus Non-populated Portion of an Enumeration Area
As the purpose of representing socioeconomic data extends beyond simple visual exploration, accuracy is crucial. The imprecision illustrated above will drastically distort the numerical values being propagated through to the prediction models. Thus, attention has been directed towards the creation of an accurate non-residential geographic layer.

### 3.4.2 Non-residential Area Masking

In an attempt to develop a surface model of population characteristics, different objectives among data providers can introduce challenges when considering the masking (or removing) of non-residential areas from EAs. A single data source will be utilized in order to circumvent positional accuracy inconsistencies found among the various data suppliers.

Fortunately, Statistics Canada's 1996 Street Network File (SNF) database contains information on visible features such as streets, railways and hydrography, and information—albeit incomplete—on invisible (or abstract) features such as political and park boundaries that properly match the polygon coverage of the EAs. Feature categories included in the classification field of the SNF allowed for the separation of the major

28. While the Street Network Files contain many non-street features (for example, railways, hydrography and parks), the complete representation of these secondary features was neither intended nor guaranteed. In general, these were included if they appeared in base maps and update materials and were deemed to be of importance to the Street Network File (Statistics Canada, 1997a).
non-populated areas in the city, such as the Botanical Gardens, golf courses, parks and National Defence properties.

In addition, I developed boundaries to account for the relatively large undeveloped land surrounding the east end of False Creek and the rail lands located just to the east of the northern section of Main St. These polygons were created using a 10 metre buffer from the surrounding roads to enable the capture of peripheral dwellings. Furthermore, for model simplicity, Granville and Deering Islands and the separate University Endowment Area, a Subdivision of the Regional District (SRD), have been removed from the locational data sets. The non-residential masking polygons (denoted by the dark-grey areas) are illustrated in Figure 12 (these polygons only represent those areas that are large and contiguous).

![Figure 12: Polygons Masking Non-residential Areas](image-url)
3.4.3 Enumeration Area Centroid Correction

Figure 13 demonstrates the repositioning of an EA centroid, from the white to the black circular symbol, to properly represent the actual populated portion of the polygon (hatched area). In particular, after the grey area has been designated as non-populated parkland, the data point must be relocated so that it is not eliminated (masked) from the interpolation procedure. Furthermore, use of ortho-photography allowed for visual estimation of centroid placement across the city to facilitate the eradication of arbitrary positioning—reflecting the geographically correct, population weighted, centre of the population cluster.

Figure 13: Correction of Enumeration Area Centroid Placement
In addition, all EAs with suppressed data (this imposes yet another constraint as gaps or holes in the spatial database hinder attempts to generalize attribute surfaces), duplicate polygon identifiers and institutional collective dwellings containing no data were withdrawn from the sample, so that the values would not erroneously interfere with the interpolation of the surface.

### 3.4.4 Surface Creation

The interpolation method examined in this thesis is referred to as *Inverse Distance Weighting* (IDW). This procedure assumes that each data point, or EA centroid, has a local influence that decreases with distance. Using a specified number of points, or all points within a specified radius, a surface is created by filling in the gaps of void cells in a grid by linearly weighting the value of each point by the distance that point is from the cell being analysed, and then averaging the values. The surface coverage of the

29. These include: children's group homes (orphanages), nursing homes, chronic care hospitals, residences for senior citizens, hospitals, psychiatric institutions, treatment centres and institutions for the physically handicapped, correctional and penal institutions, young offenders' facilities and jails. Only basic data (age, sex, marital status and mother tongue) were collected for institutional residents (Statistics Canada, 1999).

30. The ArcInfo Grid options used for the IDW interpolation were: exponent of power = 2 (this relatively high power produces a detailed surface by placing more emphasis on closer points and giving lesser influence to those of greater distance); number of nearest points = 12 (delimiting the number of points used in the interpolation procedure, again keeping the influence localized); maximum radius = 1000 metres (this establishes the maximum distance that the procedure will search for the minimum number of points); grid cell size = 25 metres (a resolution small enough to adequately cover the confined areas of the city while retaining sufficient detail relative to the density of points).

31. Note that account of areal boundary location is not considered in these models. Consequently, attribute information is potentially lost into or gained from adjacent units (see Martin, 1996b for a boundary-constrained, data preserving surface model).
grid can be confined to areas defined by a barrier layer (note how the interpolated surface is restricted to the light-grey area depicted in Figure 12) while the specification of cell dimension delineates the resolution of the output.\textsuperscript{32}

![Figure 14: Surface Representation for the Percentage of Visible Minority Population](image)

The outcome of the process is a raster layer containing counts (heights or z values) of the specified characteristic, showing the population at a resolution that is much finer—more appropriately communicating the heterogeneous nature—than that of conventional vector based choropleth mapping (see Figure 14 above; the darker areas delineate the higher concentration). The IDW surface considerably improves upon the accuracy of variable representation by progressing from a static, irrelevant object de-

\textsuperscript{32} Due to the necessity of covering the entire street network of the City, the cell dimensions were reduced from an initial 50 to 25 metres. At 50 metres, tight areas (i.e., less than 50 metres in dimension) defined by the barrier layer were left uninterpolated or void.
lineation to a flexible field depiction. It is, however, important to recognize that in calculating unknown values from surrounding points that the data is expected to behave in a spatially predictable manner over the surface area (regardless of the fact that data collected from the census is not continuous). Furthermore, the size and shape of the original areal units will have a profound influence on the form of the constructed surface. If these predefined areas are particularly large, meaningful variations in the data will be dismissed (Schmid and MacCannell, 1955 in Martin, 1996a). Nevertheless, in reality, all representations of socioeconomic characteristics are to some degree arbitrarily imposed constructs.

Following the above methodology, three surface models were developed from irregularly spaced, sufficiently dense point files of the three top ranking variables of each factor that were outlined in the data reduction section. Values from these three surface representations of visible minority, average family income and middle age were subsequently appended to the residential property transaction database (described below).

---

33. The underlying assumption of this procedure is that the distribution of the EA centroids is a summary of the distribution of the characteristic to be modelled.

34. Population surfaces do not resemble physical terrains; they tend to be far less spatially autocorrelated and actually consist of 'spikey' features (Wood, Fisher, Dykes, Unwin and Stynes, 1999).
3.5 Extrinsic Characteristics—Spatial Attribute Data

Beyond the complex complexion of socioeconomic space exists an additional spatial dimension that incorporates an elaborate mesh of locational externalities. This diverse collection of characteristics is typically difficult to specify and often more laborious to measure and capture. Furthermore, neighbourhood externalities become very difficult to model because of the subjectivity of an individual's utility structure. For example, accessibility to amenities such as greenspace, recreational facilities, shopping, public transit, population density and residential investment may differ from distance from pollution, noise, commercial and industrial landuse, street quality and zoning, traffic and crime. These elements can effect property values either positively or negatively, depending on preferences, while their impact is conceivably rather localised with regard to influence. It is anticipated, however, that spatial variance can be explained through one or few variables when combined with the socioeconomic dimension. A discussion of variable relevance and development ensues.

The Central Business District (CBD) distance function commonly used in the hedonic literature seems unreasonable to include in the proposed model. Perhaps if the study encompassed a greater geographical expanse, including the greater metropolitan area, the CBD function would be warranted as it is clear that land values decay,

35. That is, influence from externalities depreciates significantly, often in a complex non-liner manner, with distance (see Orford, URL; Ding, Simons and Baku, 2000 and Castle, 1998 for the analysis of influence for selected externalities using GIS; and O'Sullivan, 1996 for discussions on locational and neighbourhood effects).

36. The straight-line distance variable, from dwelling to the CBD, was an artefact from the monocentric models of urban spatial structure (Alonzo, 1964; Muth, 1969).
albeit in a complex spatial manner, as distance increases from a city centre. Beyond the fact that the City of Vancouver is geographically complex, with regard to its physical landscape, there exist residential pockets and east / west divisions that remain isolated from this influence (due to historical significance, declining significance of central employment location, peripheral multicentricity—edge cities and telecommuting, consumer behaviour and lifestyle patterns).  

A seldom used, but spatially significant, characteristic is crime data. Publicly available crime data for the City of Vancouver is divided up into four districts, which are, firstly, at too gross a scale and, secondly, appear to be aggregated based upon equalizing overall incidence rates to present any detail to the model. For a cost, data can be aggregated down to a minimum collection of a five-block radius. This procedure precludes an extracted data set for the entire city, and to date, remains prohibitively expensive. Access to crime data was initially highly desirable, however, it is anticipated that the included socioeconomic variables will indirectly speak to the majority of the idiosyncrasies existing in the urban network.

Several studies have explored and quantitatively attempted to confirm the magnitude of public school quality's contribution to house prices. Many have suggested that public school quality is one of the most important determinants of house price.  

37. In her paper, Rena Sivitanidou (1997) explores changes in office-commercial land value gradients within the multicentric urban area of Los Angeles, California. Quantitative results indicate a substantial flattening in these gradients in recent years. The findings are consistent with speculations that the information revolution is contributing to the weakening of spatial links between office-commercial activities and large business centers—resulting in the progressively dispersed pattern of business location.
assortment of researchers have used expenditure data to measure the impact of schools, Rosen and Fullerton (1977) criticized the use of expenditures figures. Instead, they suggested that proficiency test scores, an outcome of the schooling process, may better reflect school quality than expenditures, an input to education. Most of the subsequent research has followed Rosen and Fullerton's advice (Jud and Watts, 1981; Jud, 1985; Walden, 1990; Haurin and Brasington, 1996; Goodman and Thibodeau, 1998; Brasington, 1999).

The average percentage achieved by secondary school students on the uniform examinations in all provincially examinable courses was introduced to the proposed model as an indicator of effective teaching and general school district performance. Associating this data with a discrete object layer of school district boundaries (see the district boundaries for the city of Vancouver below) facilitated the transfer of this attribute to the property database.

38. For examples of hedonic studies highlighting the significance of education variables see Grether and Mieszkowski (1974), Kain and Quigley (1970), Dale-Johnson (1982), Follain and Jimenez (1985) and for schools and districts in particular see Brasington (1999), Bogart and Cromwell (2000).

39. The percentages, organized by the Fraser Institute, are derived from student results data provided by the BC Ministry of Education. For each secondary school in the province, an average is calculated by taking the mean scores achieved by the school's students on the standardized final examinations in all of the provincially examinable courses weighted by the relative number of students who participated in the examination (Fraser Institute, URL).

40. The catchment polygons were developed from Statistics Canada's 1996 SNF using information on street boundaries provided by the Vancouver School Board (VSB).
3.6 Address Matching

In order to incorporate the unique field and object values into the model, knowledge of absolute location is compulsory. As the MLS property listings contain an address field, the entire database can be mapped to the Street Network File (SNF) illustrated below.
The light grey lines displayed in the image outline the street network of the Municipality of Vancouver and the University Endowment Area. Following the geocoding of these lines in ArcView, the relative geographic locations of residential properties can be interpreted by accessing the address field supplied in the MLS database. By augmenting the functionality of ArcView, with the assistance of extensions, cell values from field and object attribute layers registered in a GIS can be extracted. As a result, the four spatial variables (listed in Table 5) can be appended to the transaction database. To elaborate, each house (represented by a point) is passed through spatial attribute layers collecting information associated with its position in space, therefore, supplement-

41. Geocoding, in a GIS, simply adds locational intelligence to every line in a street network. That is, each segment is provided with a start and finish value associated with its address range in the real world.

42. ArcView is a GIS application developed by Environmental Systems Research Institute (ESRI).
ing the intrinsic attributes. Their correlations—along with those for the intrinsic characteristics—with dwelling value, are presented in Table 6.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Rating</td>
<td>61.20%</td>
<td>75.20%</td>
<td>68.75%</td>
<td>70.00%</td>
</tr>
<tr>
<td>Ethnic status - % Visible Minority</td>
<td>1.89%</td>
<td>90.31%</td>
<td>51.95%</td>
<td>56.98%</td>
</tr>
<tr>
<td>Economic status - Average Income</td>
<td>$26,962</td>
<td>$172,263</td>
<td>$60,781</td>
<td>$48,592</td>
</tr>
<tr>
<td>Family status - % Middle Age (25-44)</td>
<td>12.35%</td>
<td>66.68%</td>
<td>33.57%</td>
<td>33.48%</td>
</tr>
</tbody>
</table>

Table 5: Descriptives of Extrinsic Characteristics

Note: The above statistics were derived from those cases incorporated in the analysis (n=1908).

<table>
<thead>
<tr>
<th>Variables</th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
<th>(E)</th>
<th>(F)</th>
<th>(G)</th>
<th>(H)</th>
<th>(I)</th>
<th>(J)</th>
<th>(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Sale Price</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(B) Lot Dimension</td>
<td>0.699</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(C) Square Footage</td>
<td>0.782</td>
<td>0.536</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D) Dwelling Age</td>
<td>-0.385</td>
<td>0.027</td>
<td>-0.436</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(E) Bedrooms</td>
<td>0.285</td>
<td>0.121</td>
<td>0.572</td>
<td>-0.374</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F) Fireplaces</td>
<td>0.464</td>
<td>0.344</td>
<td>0.528</td>
<td>-0.387</td>
<td>0.272</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(G) Bathrooms</td>
<td>0.585</td>
<td>0.234</td>
<td>0.716</td>
<td>-0.771</td>
<td>0.565</td>
<td>0.495</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H) Visible Minority</td>
<td>-0.309</td>
<td>-0.172</td>
<td>-0.171</td>
<td>-0.038</td>
<td>0.026</td>
<td>-0.097</td>
<td>-0.089</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(I) Average Income</td>
<td>0.687</td>
<td>0.574</td>
<td>0.452</td>
<td>-0.049</td>
<td>0.050</td>
<td>0.310</td>
<td>0.271</td>
<td>-0.598</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(J) Middle Age</td>
<td>-0.461</td>
<td>-0.456</td>
<td>-0.339</td>
<td>0.149</td>
<td>-0.069</td>
<td>-0.287</td>
<td>-0.259</td>
<td>-0.177</td>
<td>-0.452</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>(K) School Rating</td>
<td>0.569</td>
<td>0.429</td>
<td>0.370</td>
<td>-0.103</td>
<td>0.066</td>
<td>0.293</td>
<td>0.261</td>
<td>-0.561</td>
<td>0.636</td>
<td>-0.305</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 6: Pearson Correlations of the Intrinsic and Extrinsic Characteristics

Note: The above statistics were derived from those cases incorporated in the analysis (n=1908).

3.7 Data Set Division, Spatial Stratification and Data Preprocessing

Following Gopal and Fischer (1996), the data set for the ANN model was randomly divided into a 70, 20 and 10% split for the training, cross validation and test sets respectively. In an attempt to avoid model bias, the cross validation set was also eliminated.
from the regression modelling procedure. Therefore, identical data sets were retained to train and test the competing models so as not to introduce error.

To test for the spatial representativeness of the originally obtained samples, and to ensure that the entire range of data is adequately captured in each sample, a second analysis of the data was performed by using data extracted from a spatially stratified random sample. To accomplish this, a three by four grid of 3800 metre\(^2\) cells was imposed on the City of Vancouver (see map below) and subsequently used to delimit sub-areas from which the three individual data sets were randomly constructed (equal percentages were taken from each strata).

![Three by Four Grid for Spatial Stratified Random Sample](image)

**Figure 17: Three by Four Grid for Spatial Stratified Random Sample**

43. Stratified random sampling is often applied, and is most successful, when the sampling population can be divided into strata that individually have less variance than the overall variance of the population. Where there exist no readily apparent divisions, arbitrary grid cells can be used from which random samples are extracted and then combined.
In order to promote better model fitting, transformed variables have been used in this analysis rather than the original raw data values. In hedonic applications, it is customary to transform the data by using the natural log of the variables (see, for instance, Hamilton and Hobden, 1992; and Hoyt and Rosenthal, 1997). Alternatively, it is common practice to linearly scale the data from 0.0 to 1.0, when implementing ANN models, so that the backpropagation algorithm is not biased by variable magnitudes and variance (Garson, 1998; Sarle, 1997).

To extend current research, considerable efforts have been made to explore and incorporate detailed extrinsic data into property prediction models. The preceding sections have discussed how GIS can be used to link locational data across spatial scales. In particular, three socioeconomic variables were extracted from the 1996 census profile using factor analysis and subsequently modelled into surfaces that more accurately depicted the underlying characteristics. A school quality indicator was also developed and appended, along with the socioeconomic characteristics, to the intrinsic variables associated with the dwellings. These procedures enriched the housing transaction database and, therefore, resolved many technical and conceptual problems existing in previous studies.

44. Log transformations perform well for data where residuals get larger for larger values of the dependent variable. Logs essentially reduce the residuals resulting from greater dependent values. For example, after the data in this study was transformed, correlations fluctuated slightly and $R^2$ results improved 0.059 for the best performing regression model.
Chapter IV: Model Development

4.1 Multiple Linear Regression Models

The simple linear regression method used in this analysis is termed stepwise. This technique determines the priority of variable entry into the multiple regression equation. In particular, as each variable is entered into the model the contribution of the variable to $R^2$ is calculated, thus the statistical significance of the variable can be assessed. The order in which the variables are presented to the model is a function of the magnitude of their contribution to $R^2$, starting with the variable that exhibits the largest partial correlation with the dependent variable. Any variables that fail to meet statistical tolerance criteria set out in the routine are excluded from the analysis (SPSS, 1996; Bryman and Cramer, 1997).

Despite the fact that all of the independent variables will be related in some way to the dependent variable, it is important to ensure that each of them are not too highly related to one another. As a guideline, the Pearson's $r$ between any pair of variables should not exceed 0.80. Variables that do reveal a relation of 0.80 or above may be suspected of exhibiting multicollinearity. The phenomenon of multicollinearity is generally considered to be problematic, as it suggests that resulting coefficients are likely to be unstable. Regardless, when two variables express a high level of correlation, there appears

---

1. The regression equations were modelled using Statistical Package for the Social Sciences (SPSS).
2. Although stepwise is the most commonly used approach, it is no less controversial than alternative methods as it gives priority to statistical criterion for inclusion rather than theoretical ones (Bryman and Cramer, 1997). Conversely, the enter method for variable selection is where all variables in a block are entered in a single step, without exclusion.
little justification for considering them as separate entities (Bryman and Cramer, 1997). According to the output in Table 6 there exist no correlations above 0.80. Once the model specifications have been determined, coefficients—stored in a separate file during the routine—are then used to calculate predictions on the remaining test data set.

4.2 Artificial Neural Network Models—Network Configuration

Developing an ANN, through input variable selection and optimizing the network architecture by heuristic search, can be a time consuming and tedious procedure. However, an automatic network designer included in the interface of the application\(^3\) allows the user to easily select the data and design an optimal network architecture. Experimenting with different numbers of hidden units, the automated designer performs a number of training runs with unique network architectures and selects the best configuration on the basis of verification error, all without the requirement of human interaction. The challenge of architecture determination is regarded as an optimization problem: a superior network architecture is searched for in a highly non-linear, noisy and multi-optima search-space. However, there is no guarantee that the configuration it allocates will be the best available (StatSoft, URL).

\(^3\) Statsoft's STATISTICA Neural Networks package was used to model neural networks.
Running the network designer\textsuperscript{4} on all intrinsic and extrinsic dwelling attributes (for a total of 10 independent variables—see Table 7), the advisor suggested a 10:88:1 three-layer model. When a four-layer model was requested, the advisor recommended a 10:19:22:1 formation.\textsuperscript{5} Monitoring a real-time training graph,\textsuperscript{6} it was apparent that in both model formations errors for both the training and testing data sets demonstrated persistently wild fluctuations, even after several hundred epochs, when it was expected that the network would start to stabilize. Although a more complex network will often reach a lower error eventually, it is typically overfitting the data as networks with more weights are prone to overfitting.

Following the principle that—all else being equal—a simple model is always preferable to one that is complex, the robustness of the models from the network designer was measured against several alternative network configurations. The standard, practical approach is to simply experiment with a large number of architectures, and decide heuristically which model(s) are favourable. A good starting point is to use one hidden layer, with the number of units equal to half the sum of the number of input and output units (Stanley, 1988 and Ripley, 1993 in Garson 1998; StatSoft, URL).

\textsuperscript{4} It was determined that more theoretically sound variable reduction procedures, as explored in chapter three, would be implemented instead of relying on techniques in the network advisor.

\textsuperscript{5} ANN models with five or more layers were not investigated as they seldom out-perform three or four layer networks (StatSoft, URL).

\textsuperscript{6} The training error graph simply plots the Root Mean Squared (RMS) error of the network against epochs.
The task of selecting a competitive network involves multiple trial and error runs—a natural selection procedure. As the backpropagation algorithm is sensitive to unique initializations, or starting values in the weight matrix, each network is trained four separate times with a consistent number of iterations (and unique initializations) while the performance is recorded. A key indicator of the potential of a specific network is the cumulative training and verification errors. Therefore, a model is selected based upon a minimisation of the associated errors terms. To add confidence to the promise of a particular configuration, the test data set is run after training is complete to ensure that the results on the verification and training set are real, and not artifacts of the training process. Unfortunately, the learning procedure in the program does not minimize the error term most important for prediction—unseen data from the test data set.

After considerable effort, through trial and error, three logistic network architectures were retained for modelling the individual problem sets. Specifically, five neurons in one hidden layer, were integrated in the simple intrinsic model (variable sequence 1 in Table 7), six hidden neurons were employed to model the second through fifth variable sequences, while only three hidden neurons were used for the extrinsic-only network.

<table>
<thead>
<tr>
<th>Model</th>
<th>Network Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No Extrinsic Dwelling Attributes</td>
<td>6:5:1</td>
</tr>
<tr>
<td>2. Average Income Included</td>
<td>7:6:1</td>
</tr>
<tr>
<td>3. Average Income and School Rating Included</td>
<td>8:6:1</td>
</tr>
<tr>
<td>4. Average Income, School Rating and Visible Minority Included</td>
<td>9:6:1</td>
</tr>
<tr>
<td>5. All Extrinsic Dwelling Attributes (simple random sample)</td>
<td>10:6:1</td>
</tr>
<tr>
<td>6. All Extrinsic Dwelling Attributes (random stratified sample)</td>
<td>10:6:1</td>
</tr>
<tr>
<td>7. No Intrinsic Dwelling Attributes (random stratified sample)</td>
<td>4:3:1</td>
</tr>
</tbody>
</table>

Table 7: ANN model Network Configurations
Once the ANN models have been generated, they are trained with observed data so that the network can shape itself by constructing an internal representation rules that ensure the model can recognize, adjust for and generalize unseen data. Training of the networks can be set to terminate when a specific target error level is obtained, or when the verification error deteriorates over a given number of iterations—indicating overtraining.

4.3 Framework for the Proposed Model

The dimensions of the model, established in the preceding chapters, are assembled in a detailed graphic framework below. A summary of the operations in the illustration—outlining the flow of data from the spatial layers in a GIS to the ANN model—follow.
• The elongated arrows originating from the grey elliptical nodes denote the locational dimension supplied through raster and vector representations of data. Firstly, the address attribute included in the property listing is georeferenced using ArcView and the SNF for Vancouver (white elliptical node). Subsequently, values from equivalent geographic positions in the object (school rating) and field (family income example provided) GIS overlays are extracted and fed into individual neurons in the input layer of the error backpropagation Multi-Layer Perceptron.

Figure 18: Framework for the Model of Prediction
The ANN at the bottom represents the nonlinear dimension of the proposed model. Triangles in the right layer receive the input characteristics (intrinsic and extrinsic) of the residential property parcel. These attributes are propagated through to the hidden layer where incoming values are multiplied by a weight matrix, summed and passed to an activation function which generates an output. The results are advanced to the sole grey circle in the output layer which represents the predicted value of the residential property.
Chapter V: Model Results

Modelling with ANNs is experimental and does not necessarily provide an improvement over traditional methods. The network can fail miserably, or perform efficiently and precisely. The black-box nature of ANN models require tests to confirm the integrity of the results—the proof of performance rests with the user (Openshaw, URL).

Upon completion of the model development stage, separate experiments are conducted on the linear regression and ANN models. Predictions are produced for the quasi-independent samples in the test data set (10% of the entire data set). In this section, various quantitative and visual measures are investigated in order to gauge model performance and to determine which of the methodologies exhibit the most effective prediction model for each of the variable sequences.

5.1 Model Evaluation Statistics

In order to provide a quantitative comparison of the proposed models, three measures of linear association were evaluated. First, Pearson’s $r$, was calculated. The resulting correlation coefficient, ranging from -1.0 to 1.0, can be interpreted as the proportion of the variance in the observed value attributable to the variance in the predicted value.
The square of Pearson's product moment, the multiple coefficient of determination $R^2$, indicates the proportion of the total variation explained by the model and is defined as:

$$R^2(S) = \frac{\Sigma(p_i - \bar{p}_i)^2}{\Sigma(y_i - \bar{y}_i)^2},$$

(3)

where $p_i$ and $y_i$ signify the predicted and observed values respectively, and $\bar{p}_i$, $\bar{y}_i$ denote the average of the particular variable over the data set ($S$).

Although the correlation coefficient is extensively used as a measure of similarity between predicted and observed values, it is not the most reliable indicator of correspondence. While the measure of correlation describes consistent proportional increases and decreases about the respective means of two variables, this measure makes too few distinctions between the type or magnitudes of possible covariations (Wilmott, 1981). Consistent over or under-prediction by the model, for instance, will still result in a high degree of association (Gould, 1994). To circumvent the potential of an unstable evaluation, Wilmott (1981, 1984) advocates the Index of Agreement ($d$), which measures the extent to which a model's predictions are error free. The value of $d$, evaluated for a supporting statistic, is expressed as:

$$d(S) = 1 - \frac{\Sigma(p_i - y_i)^2}{\Sigma(|p_i' + y_i'|)^2},$$

(4)
where $p'_i = p_i - \tilde{y}_i$ and $y'_i = y_i - \tilde{y}_i$. The results of this statistic range from 1.0 to 0.0 where a value of 1.0 indicates perfect agreement between $p_i$ and $y_i$, and 0.0 denotes complete disagreement. Not simply a measure of correlation or association in the formal sense, $d$ reflects the degree to which predicted deviations about the $\tilde{y}_i$ differ from observed deviations about $\tilde{y}_i$, while taking into consideration both sign and magnitude (Wilmott, 1981, 1984).

The third statistical measure used in this study is the Average Relative Variance (ARV). This alternative performance evaluation is commonly used in ANN literature (see Weigend, Rumelhart and Huberman, 1991; Gopal and Fischer, 1996; and Weigel, Horton, Tajima and Detman, 1999). The equation for $ARV$ is as follows:

$$ARV(S) = \frac{\sum(y_i - p_i)^2}{\sum(y_i - \tilde{y}_i)^2}. \quad (5)$$

The result communicates the fraction of the variance of the data which fails to be predicted by the model. If the $ARV$ is equal to 1, then the predicted values from the model are no better than an estimate that simply extracts the mean from the observed values for the predictions (Weigel, Horton, Tajima and Detman, 1999). Improved predictability is indicated by $ARV$ values that approach 0.
5.2 Statistical Results

As indicated by the statistics below, the ANN models consistently outperform the regression models, save for the $d$ and ARV results from the second variable sequence. Overall, the variance explained by the ANNs does not differ greatly from the regression techniques. In particular, however, two instances do yield significantly higher and lower $R^2$ and ARV statistics respectively (see the first and sixth variable sequences in Tables 8 and 9). As illustrated in the tables, inclusion of extrinsic characteristics (locational attributes) significantly improves model predictiveness. For the regression model, the $R^2$ values increased to 0.8828 from 0.7833 with the $d$ and ARV values following a similar and inverse trend respectively. $R^2$ values increased from 0.8388 to 0.9169 for the ANN model, while the other statistics followed a proportionate pattern to the regression models. A stratification of the sample also increased the prediction accuracy for both the regression and ANN models with $R^2$ results of 0.9011 and 0.9322 respectively (variable sequence 6 in the tables below).

<table>
<thead>
<tr>
<th>Regression model</th>
<th>$R^2$</th>
<th>$d$</th>
<th>ARV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No Extrinsic Dwelling Attributes</td>
<td>0.7833</td>
<td>0.9351</td>
<td>0.2175</td>
</tr>
<tr>
<td>2. Average Income Included</td>
<td>0.8573</td>
<td>0.9596</td>
<td>0.1440</td>
</tr>
<tr>
<td>3. Average Income and School Rating Included</td>
<td>0.8828</td>
<td>0.9685</td>
<td>0.1178</td>
</tr>
<tr>
<td>4. Average Income, School Rating and Visible Minority Included</td>
<td>0.8828</td>
<td>0.9685</td>
<td>0.1178</td>
</tr>
<tr>
<td>5. All Extrinsic Dwelling Attributes (simple random sample)</td>
<td>0.8828</td>
<td>0.9685</td>
<td>0.1178</td>
</tr>
<tr>
<td>6. All Extrinsic Dwelling Attributes (random stratified sample)</td>
<td>0.9011</td>
<td>0.9723</td>
<td>0.0998</td>
</tr>
</tbody>
</table>

Table 8: Performance of the Multivariate Linear Regression Models

Note: The stepwise method was selected for the regression procedure and, therefore, produced similar results for three of the models (3, 4 and 5) due to the statistical programs' removal of variables that are not statistically related. In comparison, an enter method of variable selection produced unsubstantial differences in $R^2$ statistics of 0.000131 and 0.00081 for the forth and fifth variable sequences respectively.
<table>
<thead>
<tr>
<th>ANN model</th>
<th>$R^2$</th>
<th>$d$</th>
<th>ARV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No Extrinsic Dwelling Attributes</td>
<td>0.8388</td>
<td>0.9503</td>
<td>0.1652</td>
</tr>
<tr>
<td>2. Average Income Included</td>
<td>0.8606</td>
<td>0.9576</td>
<td>0.1495</td>
</tr>
<tr>
<td>3. Average Income and School Rating Included</td>
<td>0.8915</td>
<td>0.9702</td>
<td>0.1122</td>
</tr>
<tr>
<td>4. Average Income, School Rating and Visible Minority Included</td>
<td>0.9077</td>
<td>0.9754</td>
<td>0.0946</td>
</tr>
<tr>
<td>5. All Extrinsic Dwelling Attributes (simple random sample)</td>
<td>0.9169</td>
<td>0.9784</td>
<td>0.0852</td>
</tr>
<tr>
<td>6. All Extrinsic Dwelling Attributes (random stratified sample)</td>
<td>0.9322</td>
<td>0.9796</td>
<td>0.0724</td>
</tr>
</tbody>
</table>

Table 9: Performance of the Artificial Neural Network Models

Examining the results from the best performing regression model (variable sequence 6—model training data set), individual variable influences can be identified. Given stepwise selection, the top five variables entered into the equation were: average family income, square footage, lot dimension, dwelling age and school rating. The table below shows the correlation with the dependent variable and a bivariate $R^2$ model measure. The results for the remaining extrinsic characteristics were: correlation 0.512 and $R^2$ 0.262 for middle age (25 to 44), and correlation and $R^2$ values of -0.343 and 0.118 for visible minority.¹

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation with Price</th>
<th>Bivariate $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Income</td>
<td>0.757</td>
<td>0.573</td>
</tr>
<tr>
<td>Square Footage</td>
<td>0.708</td>
<td>0.501</td>
</tr>
<tr>
<td>Lot Dimension</td>
<td>0.709</td>
<td>0.502</td>
</tr>
<tr>
<td>Dwelling Age</td>
<td>-0.399</td>
<td>0.159</td>
</tr>
<tr>
<td>School Rating</td>
<td>0.661</td>
<td>0.437</td>
</tr>
</tbody>
</table>

Table 10: Bivariate Statistics for Variable Sequence 6—Model Training Data Set

¹ Strikingly, the correlation (see table 6 for variable correlations of the full data set), from the entire data set, between average family income and school rating was 0.636.
5.3 Graphical Results

Analysis of graphs showing predicted against observed dwelling values visually highlights the differences between the competing models (the test set cases for the graphs have been sorted in ascending observed value while the physical dimensions of the axes are held constant for comparability).\(^2\) The regression model results (Figures 19, 21 and 23) display wild oscillations compared to the relatively tight fluctuations around the observed values for the ANN models. A collective improvement of model 'fitness' can also be detected for the contrasted variable sequences. Furthermore, the ANN models (Figures 20, 22 and 24) predict the lower half of the observed dwelling values remarkably well. Note that several of the models (in particular, Figures 20, 21, 23 and 24), to varying degrees, have difficulty in capturing the upper and lower extents of the dwelling values.

---

2. The x and y axes of the graphs differ due to the random nature of the sample drawn and the separate scaling techniques implemented for each of the models.
Figure 19: Regression Model 1—Intrinsic Attributes Only

Figure 20: ANN Model 1—Intrinsic Attributes Only
Figure 21: Regression Model 5—With Extrinsic Attributes (simple random sample)

Figure 22: ANN Model 5—With Extrinsic Attributes (simple random sample)
Figure 23: Regression Model 6—With Extrinsic Attributes (stratified random sample)

Figure 24: ANN Model 6—With Extrinsic Attributes (stratified random sample)
5.4 Residual Mapping

In order to assess the spatial ability of the individual models to predict property values, the distribution of the models' residuals (equal interval categorization of absolute errors) are mapped in Figures 26 through 31. While the residual values themselves do not vary greatly amongst the variable sequences for each of the models, the geographic distribution of the residuals for the separate models do display pronounced differences. In particular, the regression model residuals clearly do not display any spatial dependency, even with regard to an east / west\(^3\) comparison that is abundantly evident in the surface representation of the property values rendered in Figure 25. An increase in spatial correlations (dwelling value and error residual) for each of the variable sequences, between the regression and neural network models, can be detected (e.g., comparing Figures 30 and 31).

As for the stratified ANN model, larger residuals are strongly spatially correlated with the more affluent neighbourhoods in the city (areas with higher property values). This implies that the ANNs benefited more from the geographic component, from a spatial perspective, while the regression models appear to simply generalize the error terms

---

3. Within the city of Vancouver a pronounced east / west pattern is generally observable; the east side of the city (with Main Street as the divide—see Figure 6) has typically been dominated by low to middle income households while the western half of the city is associated with middle to high income groups (Hiebert, 1999). With respect to dwelling values, the dark, erratic left hand side of the surface (Figure 25) represents the middle to upper middle-class inner city districts on the west side, particularly the traditional elite neighbourhood of Shaughnessy and the relatively high income neighbourhoods of Kerrisdale and Point Grey. Note that the upper and lower bounds of the 1996 MLS property values were $3,000,000 and $148,624 respectively, with a mean of $514,565 and median of $411,500.
across the entire surface. These images echo the results from the tight weave of the lines in the lower-half of the graphs in Figures 20, 22 and 24.

Figure 25: Surface Representation of MLS Property Values—Orthographic Perspective
Note: Surface created using IDW interpolation on all 1908 cases masking non-residential areas. Darker areas indicate higher dwelling values.
Figure 26: Regression Model 1 Residuals—Intrinsic Attributes Only

Figure 27: ANN Model 1 Residuals—Intrinsic Attributes Only
Figure 28: Regression Model 5 Residuals—With Extrinsic Attributes (simple random sample)

Figure 29: ANN Model 5 Residuals—With Extrinsic Attributes (simple random sample)
Figure 30: Regression Model 6 Residuals—With Extrinsic Attributes (stratified random sample)

Figure 31: ANN Model 6 Residuals—With Extrinsic Attributes (stratified random sample)
5.5 Extrinsic Data Models

The final analysis tested for the independent impact of the extrinsic characteristics. That is, the influence of intrinsic characteristics was removed by analysing only the four spatial variables. The outcomes are presented in the tables below. In two out of the three statistical measures the ANN model outperformed, albeit negligibly, the regression model. In particular, the $R^2$ value for the ANN model differed by only 0.0005 over the regression model.

<table>
<thead>
<tr>
<th>Regression model</th>
<th>$R^2$</th>
<th>$d$</th>
<th>ARV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No Intrinsic Dwelling Attributes</td>
<td>0.6854</td>
<td>0.9011</td>
<td>0.3156</td>
</tr>
</tbody>
</table>

Table 11: Performance of the Extrinsic Regression Model (stratified random sample)

<table>
<thead>
<tr>
<th>ANN model</th>
<th>$R^2$</th>
<th>$d$</th>
<th>ARV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No Intrinsic Dwelling Attributes</td>
<td>0.6859</td>
<td>0.9055</td>
<td>0.3367</td>
</tr>
</tbody>
</table>

Table 12: Performance of the Extrinsic ANN Model (stratified random sample)
Chapter VI: Conclusion

6.1 Discussion

This thesis investigates the ability of ANN models to predict residential real estate values in the City of Vancouver, British Columbia. Multivariate linear regression models have typically been used for this purpose and have provided reasonable results. However, the relation between the characteristics of a dwelling and its value are generally complex and nonlinear—conditions that might be better captured with ANN models. A comparison of ANN and linear regression modelling techniques has been presented in the preceding chapters. Several models are developed and then tested using MLS sales transaction records for a six month period in 1996. The evidence based on this investigation suggests that, for the problem of predicting property values, the proposed multi-level technique presents practical and theoretical advantages over conventional approaches.

ANN models showed a significant improvement over the classical hedonic approach to valuation, regardless of the variable combination applied. In particular, the ANN models consistently outperform the regression models, save for the d and ARV results from one of the six variable sequences. Notably, a random, spatial stratification of the sample also increased the accuracy of prediction for both the regression and ANN models.
A recurrent deficiency identified during the testing of the ANN models is their trouble with predicting the upper ranges of the testing data set. For instance, of the five largest residuals, from the sixth variable sequence, four had dwelling values greater than $1,400,000, all of which were extraordinarily underpredicted. This is often a common challenge associated with ANN modelling as they are customarily developed to generalize functions by establishing an internal representation of rules that are extracted from residential properties in the training data set. Thus, in order to generalise well for unseen events, the training set must contain an adequate balance and encompass a full range of cases. Given the minimum and maximum dwelling values of $148,624 and $3,000,000 respectively with a mean of $514,565 and median $411,500, there are comparatively few cases in the upper ranges of the data set. Consequently, the model tends to perform poorly when presented with extreme cases.

This challenge could potentially be eliminated by stratifying the data set into value ranges (i.e., $148,624–$300,000; $300,000–$450,000;... $900,000–$1,050,000; etc.) and then training multiple models for each variable sequence. However, this weakens the overall potential of the modelling technique—to handle complex, noisy data—and deters the spatial and quantitative portability of the model.

During the training and testing modelling phases, it appeared as though the performance of a specific network depended greatly on the random selection of the training, 

1. Although ANNs are relatively noise tolerant, there is a limit to this tolerance; if there are outliers outside the range of normal values for a variable they may bias training.
2. There are only 134 cases in the data set over $1,050,000, and 10 cases over $2,000,000.
verification and testing data sets.\(^3\) This suggests that the results of an individual model are biased by the artifacts intrinsic to particular divisions of the data (likely impacted by the number of outliers included). It is, however, important to note that when the stratified data samples were integrated, the random selection of the data sets were already established, prior to modelling, and failed to noticeably improve the results.

6.2 Uniqueness of Findings

When reviewing the results presented in the ANN literature, several methodological issues were commonly neglected. Of particular note, except for Lewis and Ware (1998), there was no inclusion of significant extrinsic (i.e., spatial) characteristics. In an attempt to extend current research efforts, a 'spatially-aware' model was proposed and tested in this study.

The preliminary results exhibited in this thesis are consistent with several studies (Borst, 1991; Evans, James and Collins, 1992; Do and Grudnitski, 1992; Lu and Lu, 1992; Tay and Ho, 1992; Lewis and Ware, 1998) investigating the potential of ANN models to predict residential parcels. Of significance, however, is that although ANNs consistently outperform traditional hedonic techniques, inclusion of a detailed spatial component added remarkable predictive capability. The spatial variables were quan-

---

3. With sufficient trial runs on various random data set selections it is conceivable that a particular network may encounter a satisfying fit which happens to perform well on the test data set.
titatively proved not to be simply redundant (merely a function intrinsic to the structural attributes), as the significance of the extrinsic components is pronounced in both prediction models, yielding surprisingly high $R^2$ measures. This was evident in both the improved performance of the variable sequences—as extrinsic characteristics were added—and the reverse analysis, where intrinsic characteristics were removed from the models altogether.

The removal of the intrinsic attributes produced striking $R^2$ results of 0.6854 and 0.6859 for the regression and ANN models respectively. Overall, the statistical measures are remarkably strong even though none of the intrinsic dwelling attributes were even considered. This clearly reinforces the significance of the spatial attributes, as the model compared favourably with the strictly intrinsic regression model, with the $R^2$ falling short by only 0.0979.

It is curious why the regression model performed similarly to the ANN for the last variable sequence. Perhaps the socioeconomic variables, modelled as continuous surfaces, exhibit less of a nonlinear nature than the intrinsic attributes. In particular, the extrinsic data generally consisted of relatively less range in the variables (refer to Tables 2 and 5) with fewer extreme outliers compared to intrinsic data. Furthermore, there is less variance within a block, or confined area (i.e., two adjacent houses may have extremely different intrinsic attributes, while the socioeconomic characteristics are similar as they have been generalised or smoothed).
This accomplishes the objectives set out in the introduction. Given the results, it is concluded that ANN models present a favourable alternative to the traditional regression method of appraisal. More importantly, the addition of extrinsic characteristics substantially improved the ability of both models to predict

6.3 Recommendations for Further Research

Having completed this study, it is believed that further improvements could likely be realized through the inclusion of additional extrinsic characteristics. Incorporation of crime statistics, population density and distance calculations for positive and negative amenities, such as open spaces, noise pollution and industrial landuse, could help increase the explanation of the predicted values. For example, buffer zones could be developed in a GIS with fuzzy (nonlinear) distance relations which model different levels of sensitivity to localised externalities.

Considerable effort was dedicated to modelling socioeconomic characteristics in this study. It would be beneficial to test the significance of the surface interpretation of population distribution against simple extractions from the original polygon assembly.

---

4. The analysis of sensitivity is regarded as the study of the correlation and calibration between the input parameters, configuration of components and the expected outcome of the model.
While extrinsic characteristics undoubtedly impact the value of a dwelling, inclusion of less significant intrinsic variables from the transaction database, such as view or water frontage, could possibly improve precision. Alternatively, an entire year or more could be modelled, while including the transaction date, so that a temporal element could internally be realized. Furthermore, extracting multiple dwelling types (apartments, townhouses, etc.) could help demonstrate the power of ANN modelling while truncation of the transaction records to exclude dwellings with a value greater than $1,000,000 could tighten the range and, therefore, the accuracy of the separate models.

Finally, given the breadth of ANN algorithms and the range of programs that implement them, it would be valuable to test uniform data sets across various programs to ensure stability and precision.

6.4 Summary

This research explored the ability of ANN techniques to predict residential property values when compared to the traditional hedonic approach using simple linear regression methodologies. Current research efforts were extended by incorporating detailed locational factors in ANN models. Through integrating ANN techniques and GIS, the ex-

5. Given the arbitrary nature of this variable, for example, any further additions to the data set may simply add noise instead of intelligence.
traction, transfer and recognition of spatial attributes, such as average family income and secondary school provincial examination performance, was facilitated.

Results primarily indicate that ANN techniques compare favourably with multiple regression models. Of particular geographic concern, the inclusion of extrinsic dwelling characteristics substantially improved the explanatory power of the competing models. In addition, spatial stratification of the samples further improved model fit in both circumstances.

While ANN models have been proven to outperform linear regression models in an operational setting, where accurate predictions of dwelling values must be supported, it may be argued that the relatively simple regression model is a more practical choice. Working with ANNs is not typically a straightforward process. The modelling procedures can be cumbersome and often require considerable effort and knowledge—for minor gain—which is exasperated by the fact that existing theory is diverse and largely inconsistent. Furthermore, given the black-box nature of the model, a considerable obstacle is presented when attempting to explain training progress and how results were received. Nevertheless, perhaps advances in ANN modelling software and theory in the near future will provide the researcher with a more straightforward and sound process.
Works Cited

*(URL) stands for Uniform Resource Locator and indicates that the material was obtained from the Internet. These web sites were still accessible at time of publication, October 12, 2001.


