CRITICAL ANALYSIS AND EFFECTIVENESS OF KEY PARAMETERS IN RESIDENTIAL PROPERTY VALUATIONS

by

CHUNG-CHUN LIN

December 2, 2010

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Civil, Structural, and Environmental Engineering

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGMENTS

I would like to express my profound indebtedness and sincere appreciation to Prof. Satish Mohan, my major professor, for his invaluable advice and insightful guidance throughout this endeavor. I would also like to thank my committee members, Prof. Alan Hutson and Prof. Shahid Ahmad, whose thoughtful comments and constructive contribution were important to the completion of this dissertation. I owe much to all of the faculty and staff in Department of Civil, Structural and Environmental Engineering for their enduring help and warm friendship.

My grateful thanks also are due to Mr. Brian Barnes of the Engineering Department of the Amherst Town for his assistance in providing the town's GIS data for this dissertation. Special thanks to all my friends at Buffalo who helped make my time unforgettable.

Finally, the greatest thanks go to my parents, Tzi Chi Lin and Ya Chiang. Without their endless love and care throughout my life, this lifetime milestone would not have been achieved.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF MAPS

ABSTRACT

All municipalities are required to re-assess their real estate periodically which they do manually spending large sums. This research has developed statistical and AI models for such mass appraisals. Three statistical and AI models: multiple regression (MR), additive nonparametric regression (ANR), and artificial neural network (ANN), were developed using the housing database of the Town of Amherst with 33,342 residential houses. Prediction accuracies of the three models were checked and found to be acceptable, and the results of each of the three models were linked to the GIS map layer of the municipality to draw various maps showing the distinct and wide variations in the prices of homes based on location or neighborhood. The research confirmed that statistical or artificial neural network models are reliable and cost effective methods for mass appraisal of residential property values.

The time variations of the housing prices and their interaction with the macroeconomic indicators: Oil Price (*OIL*), 30-year Mortgage Interest Rate (*IR*), Consumer Price Index (*CPI*), Dow Jones Industrial Average (*DJIA*), and Unemployment Rate (*UR*), were analyzed using Vector Autoregression (VAR) on the monthly housing sales data for the Town of Amherst, State of New York, for the period: 1999 – 2008. The various analyses concluded that the 30-year mortgage interest rate (*IR*) has the highest effect on the housing prices progressing from 4.97 percent in the first month to 8.51 percent in the twelfth month. The unemployment rate (*UR*) was next in order followed by Dow Jones Industrial Average (*DJIA*), and Consumer Price Index (*CPI*).

This research, also finalized a methodology for dividing the housing stock of a municipality into uniform zones or districts for value assessments and other planning purposes. The cluster

analysis was utilized to regroup the existing 67 neighborhoods of the Town of Amherst into 25 districts to simplify the priori classifications.

The methodologies formulated and tested in this dissertation will be useful for municipalities and consultants while performing mass appraisals. Town planners will also find the various methodologies and the resulting patterns useful for determining the development needs of one district in comparison to all other districts.

CHAPTER 1: Introduction

1.1 Introduction

All municipalities are required by law to assess the value of their real estate including: residential properties, commercial properties, and utilities etc; annually or at other intervals.

In the State of New York, the State Office of Real Property Services requires that all properties be assessed to their full value (DRPS, RPTL 301 and 305), and awards a $5 per property stipend to municipalities, for full value assessments. Some other States require full value assessments at intervals of more than one year. The State of California requires reassessment of the property when it changes ownership (California Property Taxes). This requirement, although necessary for an equitable and fair distribution of property tax levies, costs the municipalities. At this time, most municipalities and other local governments do their assessments manually, using data of actual sales of properties during the last one or more years. Most municipalities hire a consultant who charges a fee of $15 to $20 per residential property, and the large commercial properties are individually bid for assessment by the appraisers.

It is universally known that the value of a residential building is a function of its location within the town: the downtown, or suburb; the class of neighborhood in which it is situated; the size of the building: the lot area, frontage length, and the square feet living area; the quality and type of exterior: stone, brick, wood, or vinyl siding etc; and similar other measurable features. There could be some price variations within buildings in the same neighborhood, based on the interior layout, materials and fixtures used in the kitchen and lavatories etc. But it seems logical that given some minimum necessary information on a residential property, and given the previous 1-3

years data on the market sales of similar properties, a reasonably accurate statistical model can be developed and used to assess the values of the properties in a municipality.

This research attempts to establish the fact that it is possible to use statistical models and artificial intelligence (AI) based models for assessing the values of properties with reasonable accuracy and has developed such models. The focus of the work done in this research is on single-family residential properties. Three types of models have been developed:

(i) Multiple Regression (MR) Model,

(ii) Additive Nonparametric Regression (ANR) Model, and

(iii)Artificial Neural Network (ANN) Model.

All three models did well on validation tests, with reasonably high accuracy, and any of the models can be developed for a municipality using in house expertise or a consultant. This research has compared the effectiveness of each of the three models in terms of accuracy of the predicted values.

The statistical models used data for 33,342 residential properties in the Town of Amherst, State of New York, which was available in the electronic format. Assuming that the town decides to use the statistical models to assess their real estate, using in-house staff, a saving of 33,342 × $20 = $666,840 can be expected, every year.

Besides the substantial cost savings in the annual reassessments, the use of statistical or AI models eliminates the influences of any bias inherent in any manually done value assessments.

1.2 Dissertation Organization

This research has been organized into 11 chapters, as briefly describe below:

Chapter 2 reviews the existing literature on the mathematical models of residential property value assessments, and presents some previous studies on the time trends of the housing price markets.

Chapter 3 provides a description of the residential properties data used in this research in models development and several critical analyses. This Chapter begins with a brief description of the commonly used initial list of variables to include in the model development. This chapter outlines the processes used for cleaning up the data. Outlier data points have been identified and removed, and the variables with high covariance with other variables have also been removed. Chapter 3 finally presents the significant independent variables, both qualitative and quantitative, selected for the model development.

Chapter 4 describes the stepwise development of a multiple regression model, and examines the impact of the various housing attributes on the prices of single-family homes. This Chapter presents interpretation of the results from the multiple regression model. A spatial analysis using ArcGIS has also been described.

Chapter 5 presents the development of an additive nonparametric regression model for estimation of single-family residential property values. This Chapter shows the details on the estimation of housing prices, the model prediction performance, and the spatial analysis using ArcGIS.

Chapter 6 describes the development of an artificial neural network for estimation of property values. This Chapter presents the processes used to construct the architecture of artificial neural network. This Chapter also presents the details on the estimation of housing prices, the model prediction performance, and the spatial analysis using ArcGIS.

Chapter 7 compares the effectiveness of the three residential property mass appraisal methodologies, and gives the results of the comparative prediction performance accuracy and the prediction accuracy of each of the models by neighborhood.

Chapter 8 has attempted to quantify the effects of major macroeconomic indicators: Oil Price (*OIL*), 30-year Mortgage Interest Rate (*IR*), Consumer Price Index (*CPI*), Dow Jones Industrial Average (*DJIA*), and Unemployment Rate (*UR*), on the regional housing prices, over time. Vector Autoregression (VAR) was used to analyze the temporal variations of the housing prices, and their interaction with the macroeconomic indicators. This Chapter presents the impact of major macroeconomic indicators on the housing prices in the example municipality.

Chapter 9 investigates the ripple effects of housing prices between different housing sizes at regional level. This Chapter examined the interrelationship among housing price changes in different housing sizes and found that the two-bedroom housing market is more sensitive to the change in three and four-bedroom housing markets.

Chapter 10 lays out a new methodology for dividing the housing stock of a municipality into uniform zones or districts for averaging the value assessment and for other planning purposes. This Chapter presents the methodology development and presents the results from this methodology and spatial analyses using ArcGIS.

Finally, Chapter 11 summarizes the findings and contributions arrived in this research.

CHAPTER 2: Literature Review

2.1 Mass Appraisal Techniques

The residential housing pricing levels play a key role in our economy, therefore the pricing of houses and its relation to society's affordability has always been a concern of the real estate industry, and also the governments at every level: federal, state and the local municipalities. An accurate assessment of the value of a house is important to its current owner, prospective supplier of mortgage funds, property tax collecting municipality and potential buyer of the house. Academia and the real estate industry have attempted to efficiently and accurately estimate housing prices through various mass appraisal techniques. Multiple regression analysis has increasingly been used in the real estate industry for mass appraisal. Jack F. Eisenlauer (1968) addresses the concept of using multiple regression analysis as an appraisal method. Blettner (1969) extends the concepts introduced by Eisenlauer to estimate housing prices appraisal also by using multiple regression analysis. Kang and Reichert (1987) observe that the significance of regression coefficients and the prediction accuracy depend on the choice of estimating technique and the functional form of the regression equation;

Ma (2006) pointed out with respect to mass appraisal, the estimates yielded by multiple regression equation have been used as the basis for the taxation of properties, and to assess the value of properties for mortgage underwriting, and for the performance analysis of real estate portfolios. Basu and Thibodeau (1998) pointed out that spatial dependence exists in the multiple regression models because houses that are in closer proximity are more likely to have similar environmental and accessibility characteristics. Moreover, houses in the same area are more

likely to have been constructed at a similar point in time and hence to have similar structural characteristics.

But, according to one other study (Dubin, 1998), if the data are spatially correlated, multiple regression estimates will be unbiased but inefficient and inconsistent. Also, the estimate of the variance will be biased, which makes statistical inference difficult or impossible (Anselin, 1988).

The concepts of spatial dependence and housing submarkets are closely related. Goodman and Thibodeau (1998) observed that market segmentation is a key feature for the successful modeling of housing prices. The idea of housing submarkets provides a useful conceptual framework for modeling spatial dependence. In this research, a multiple regression model has been developed which incorporates a series of dummy variables for neighborhoods (housing sub-markets) defined by a tax assessor. All these neighborhood dummy variables have tried to capture the locational or spatial effect on the housing prices.

Alternatives to multiple regression model in the real estate price estimate modeling have increased in popularity in recent years due to the computational advances. The additive nonparametric regression (ANR) and artificial neural network (ANN) are two of the more popular examples of such alternatives technique, as illustrated in this dissertation. Several housing price studies have shown growing interest in additive nonparametric regression (ANR) as an estimation method. Anglin and Gencay, 1996; Gencay and Yang, 1996; Pace, 1995 found that the additive nonparametric estimation allows greater flexibility in the functional form of regression, avoiding the restrictive assumptions of its parametric counterparts. Meese and Wallace (1991) compare several common parametric specified models with nonparametric regression models. The result shows that the additive nonparametric model can avoid

assumptions on parametric functional forms. Also, they advocated the use of nonparametric regression techniques to construct housing price indices.

Coulson (1992) used a spline smoothing method to the additive nonparametric estimates of the housing price on floor space size. This paper also shows that the additive nonparametric method surpasses the parametric counterparts because the former lose the restrictive assumption of functional form. Their data are drawn from a sample of 402 residential real estate sales in the State College, PA, metropolitan area from the calendar year 1987. Pace (1993, 1995) applied the kernel nonparametric estimators to assess residential housing price. In his paper, both in-sample and out-of-sample prediction performances of the nonparametric regression and parametric regression are compared. The nonparametric estimator outperforms the parametric estimator in the measure of R^2, root mean squared error, and mean absolute error. Anglin and Gencay (1996) compared the prediction of the parametric model (multiple regression model) with that of the semiparametric model (additive nonparametric regression model). They estimated a benchmark parametric model which passes several common specification tests and then calculated the associated prediction values for both models. Their data was provided by the Windsor and Essex County Real Estate Board and describes residential houses sold during July, August, and September of 1987 through the local Multiple Listing Service. The total numbers of 546 records is included in the research. They concluded that the semiparametric model provides more accurate mean predictions than the benchmark parametric model.

Gencay and Yang (1996) examined the out-of-sample forecast comparison between the parametric and semiparametric regression model. They reported that the semiparametric model provides the smallest out-of-sample mean square prediction error in comparison with the parametric specifications such as the ordinary least squares regression, the Box-Cox and the

Wooldridge transformations. Also, they further suggested that semiparametric regression can be successfully used for prediction and assessment of residential housing prices.

Iwata et al. (2000) utilized a semiparametric regression model to examine the effects of land uses on residential property values in Lawrence, Kansas. They illustrated that the additive nonparametric model provides better prediction performance than parametric model based on the comparison of prediction accuracy. All these researches show that nonparametric estimation reveals nonlinear relationships that cannot be captured by traditional parametric regression, and results in much smaller prediction errors than parametric regression.

On the other hand, artificial neural network (ANN) can perform housing price prediction after learning the underlying relationships between the input variables and their corresponding outputs. In fact, a number of published studies (McCluskey and Borst, 1997; Nguyen and Cripps, 2001; Visit et al., 2004) investigated the application of artificial neural network technology to residential property appraisal. McCluskey and Borst (1997) applies three techniques of mass appraisal, namely, multiple regression analysis (MRA), comparable sales analysis, and artificial neural networks (ANNs), to a data set of residential sales from the suburbs of Londonderry, Northern Ireland. The objective is to analyze the performance of the models in terms of predictive ability and explain ability. They observed that the consistency of all three models in terms of predictive accuracy.

Nguyen and Cripps (2001) compared the predictive performance of artificial neural networks (ANN) and multiple regression analysis (MRA) for single family housing sales. They reports that as the functional model specification improves, the performance of multiple regression analysis model improves where as the performance of the artificial neural networks model improves as

the training size increases. Also, they pointed out that the ANN performs better based on the mean absolute percentage error and forecasting error than the MRA when a moderate to large data sample size is used. The MRA performs better in term of the mean absolute percentage error than the ANN when a small data sample size is used.

Visit et al. (2004) studied the predictive power of the multiple regression model and artificial neural network model on housing price prediction. A sample of 200 houses in Christchurch, New Zealand is randomly selected from the Harcourt website. Their empirical results found that artificial neural network model outperform the multiple regression model based on the in sample and out of sample prediction. They further considered the potential of artificial neural network on housing price prediction.

While Din et al. (2001) and McGreal et al. (1998) strongly support the use of ANN, the results of Worzala et al. (1995) and Lenk et al. (1997) had some reservations about this technique. Din et al. (2001) compared various real estate valuation models which take into account different definition of environmental variables. The benchmark model is taken to be a multiple regression model. The artificial neural network models which are non-linear were applied to compare the multiple regression models. He found that the correlation coefficients for ANN models are somewhat higher than for the regression models. He advanced a suggestion that the ANN models have a potential for more realistic pricing of individual properties.

McGreal et al. (1998) adopted a more skeptical approach to evaluate the prediction ability of artificial neural network. They found that there is a tendency towards better results with more homogeneous data. From their research, some very close predictions are possible, others can

deviate appreciably from the sale price. They suggested that the use of neural networks for mass appraisal purposes must remain problematic.

Worzala et al. (1995) adopt a contrary position and cast some doubt upon the role of artificial neural networks against the multiple regression analysis models, suggesting that caution is needed when working with neural networks. In undertaking analysis at three specific cases, the error magnitude for individual properties was found in some cases to be very significant (up to 70 per cent) and clearly not acceptable for a professional appraisal. Furthermore, the analysis showed that even when using the same data, results from models prepared by different neural network software packages could be inconsistent and do not always outperform regression models.

In the Lenk et al. (1997) research, three artificial neural network models and one multiple regression model were created and tested for their ability to perform mass valuation estimations for residential properties. They inferred that substantial value estimation errors are possibly occurred with all of the four mass valuation models. The empirical results indicated that artificial nueral network models do not necessarily outperform multiple regression model. Moreover, the choice of performance measure dictated which mass valuation technique was superior.

The literature shows that there is mixed success with the artificial neural network method, probably due to different variable inputs and market conditions. In this research, the multiple regression (MR) model, additive nonparametric regression (ANR), and artificial neural network (ANN) were used to estimate housing prices, respectively. So far, the relative accuracy of these three models has not been compared. This dissertation aims to compare the housing price prediction accuracies of these three models.

2.2 Impact of Macroeconomic Indicators

Housing prices are known to reflect local economic conditions. Also, the global economic conditions and the national business cycle both influence the local housing markets. In fact, a large number of economic variables affect variation in housing price over time. For instance, income, mortgage interest rates, construction costs, labor market variables, stock prices, industrial production, consumer confidence index, etc., act as potential predictors (Cho, 1996; Abraham and Hendershott, 1996; Johnes and Hyclak, 1999; and Rapach and Strauss, 2007, 2009). The highly volatile crude oil price causes large movements in global economic conditions. These shocks in turn feed into the regional economy.

History shows that over long periods, stock prices and house prices move together. Sutton (2002) presents evidence that a significant part of house price fluctuations can be explained by stock prices in six countries (USA, UK, Canada, Ireland, the Netherlands and Australia).

With regard to the impact of inflation on the housing sector, different views have been argued. Baffoe-Bonnie (1998) found that shocks to inflation may increase housing prices in the western region of USA. On the other hand, Kearl (1979) indicated that inflation causes nominal housing payments to rise, which results in a lower housing demand, and which in turn lowers housing prices. Baffoe-Bonnie (1998) further reported that recent innovations in the Welsh economy led to the view that the region's economy is likely to exhibit differential responses to financial and external shocks compared to the rest of the UK. He thus concluded that different regions have different responses to the same macroeconomic shock.

In addition to the Dow Jones Index Average (DJIA) and inflation, other economic variables, such as mortgage interest rates and unemployment rate, can affect both housing prices and the construction of new housing.

The mortgage interest rate is a very important variable influencing the decision of individuals to buy a house. Housing almost always involves mortgage borrowing because of the high purchase price. A rise in mortgage interest rate increases the cost of home ownership relative to other consumption items. Therefore, people are prevented from buying houses. Therefore, the demand for housing decreases.

The regional unemployment rate has also proven to be the predictor of fluctuations in the regional housing prices. The rate of unemployment is an indicator of the local economic conditions. Building activity is stimulated by higher employment growth (Smith and Tesarek, 1991; Sternlieb and Hughes, 1977), while Hartzel, et al. (1993) argued that certain regional employment characteristics play a significant role in investors' decisions, and thus, in the determination of housing prices. Abelson *et al.* (2005) also found that the high unemployment rate and the mortgage interest rate had a negative impact on housing prices.

The motivation of this dissertation is to quantify the effects of major macroeconomic indicators: Oil Price (*OIL*), 30-year Mortgage Interest Rate (*IR*), Consumer Price Index (*CPI*), Dow Jones Industrial Average (*DJIA*), and Unemployment Rate (*UR*), on the regional housing prices, over time. Vector Autoregression (VAR) was used to analyze the time variation of the housing prices, over time, and their interaction with the macroeconomic indicators.

2.3 Housing Price Diffusion Patterns at Local Level

The residential housing market plays a key role in our society as a whole. Recent research found that the ripple effect of housing prices or housing price diffusion does exist. That is, housing price shocks in one region are likely to cause subsequent shocks in other regions. This fact is confirmed by Alexander and Barrow, 1994, Ashworth and Parker, 1997, Cook, 2005, MacDonald and Taylor, 1993, Meen, 1999, Pollakowski and Ray, 1997, Stevenson, 2004, and Tu, 2000. Many researches (Alexander and Barrow, 1994; Ashworth and Parker, 1997; Pollakowski and Ray, 1997) have mentioned that the ripple effect or house price diffusion in the UK regional house prices. Stevenson (2004) examined the causal relationships between Irish regional housing markets and found Dublin has a lead effect with other markets. Tu (2000) applies the Granger causality test and finds two diffusion paths which formed a geographic diffusion pattern in the Australian housing market: starting from Brisbane via Sydney ending at Melbourne, and starting from Brisbane via a national path and ending at Melbourne.

Moreover, the so-called Housing price diffusion at the local level has been examined by Tirtiroglu (1992) and Clapp and Tirtiroglu (1994). They shows that for submarkets in Hartford, CT, housing price changes in a given submarket depend not only on their own lagged values, but also on the lagged values of price changes in contiguous submarkets. If there is a negative shock to a given market, potential home buyers gradually become aware of the information through both the news media and person-to-person contact and are reminded of the risk of owning real estate.

This ripple effect of housing prices can occur along a number of dimensions, including structure and neighborhood type, and distance. This dissertation separates the existing literature by

addressing housing price diffusion for different housing types at the regional level: Amherst Town, New York. Causality methodology within vector autoregression was used to determine whether causal relationships between house prices in three market segments exist. Then, the impulse responses and sensitivity among three housing markets were examined.

2.4 Dividing the Housing Stock into Uniform Districts

The housing market differs from that of many other goods in that it is fundamentally spatial in nature. Housing segmentations are typically defined as areas in which the quantities of different housing characteristics differ from those of another area (Goodman and Thibodeau 1998). In the USA, census tracks and zip codes are often used as proxies for housing segmentations. In the Town of Amherst, New York, the 67 neighborhood codes are used as housing segmentations by town's property tax assessment office. The variation in the quantities of different housing characteristics is analyzed in this dissertation. The issue of dividing a large housing market into segmentations has been researched in a number of papers.

Straszheim (1974) applied the multiple regression model to estimate the housing price in the Bay Area of San Francisco. The multiple regression equations were estimated for three broad geographic areas and then again when the data were pooled across areas. He argued that the housing market consists of a series of separate markets with different multiple regression functions in each.

Dale-Johnson (1982) applies factor analysis of housing variables and neighborhood characteristics to define a number of submarkets. He used 13 housing variables, and grouped the data into market segments by Q-factor analysis. He concluded that submarkets are critical consideration in the analysis of housing price characteristics data.

Goodman and Thibodeau (1998) observed that market segmentation is a key feature for the successful modeling of housing prices. Submarket identification is important because property prices in different submarkets are controlled by different functional relationship.

Goodman and Thibodeau (1998) propose to identify housing submarket boundaries by developing and estimating the parameters of a hierarchical model for house prices. Their overall objective was to identify areas in which the price of housing services is uniform and then to use this submarket information in conjunction with the multiple regression technique. His conclusion is that hierarchical models provide a useful framework for defining housing submarkets and could be adapted to other property types to describe submarket boundaries for multifamily, office, retail, and industrial/warehouse properties.

Bourassa *et al,* (1999) suggest that the classification derived from the clustering procedure was significantly better than other submarket classification methods. They did that by combining principal component and cluster analysis. The principal component analysis is applied to extract a number of factors from the original variables. The factor scores are then used in the clustering technique to assign individual properties to different housing submarkets. They found that all submarket classifications perform better than the overall market equation. This implies that the housing prices vary across submarkets.

Goodman and Thibodeau, 2003; and Chen, et al., 2009 utilized predefined or otherwise convenient geographical boundaries to identify submarkets.

Goodman and Thibodeau (2003) extended their earlier analysis by comparing the hedonic prediction accuracy for four different methods of delineating housing submarkets: (1) no spatial disaggregation; (2) by zip code districts; (3) by census tracts and (4) by Goodman–Thibodeau

(GT) technique: hierarchical modeling approach. His empirical results revealed that spatial disaggregation yields significant gains in prediction accuracy.

Chen *et al*, (2009) used three types of market segmentation strategies: (1) no segmentation; (2) by using statistical clustering methods; (3) by using *a priori* information, to estimate housing prices. Forecasting accuracies are examined by using housing sales data from Knox County, Tennessee, USA. The results found models with spatially disaggregated submarkets perform better in forecasting housing prices than a model without sub-markets. They further point out that sub-market segmentation models with predefined geographical areas and a method that performs statistical clustering within local government jurisdictions perform better than statistical clustering methods alone.

In this dissertation, a methodology for dividing the housing stock of a municipality into uniform zones or districts was presented. In the first place, ArcGIS software Version 9.3 was applied to analyze the spatial variations in the housing markets on the Town of Amherst, New York database of 33,342 houses. Then, the cluster statistical method was utilized on the housing characteristics such as: housing value per square feet of the living area, and the building age, etc. to regroup the 67 neighborhoods of an example municipality.

CHAPTER 3: Description of Residential Property Data Used in Models Development and Critical Analyses

3.1 Project Data

In this dissertation, the data provided for the residential properties in the Town of Amherst, State of New York was used. The data included assessment information for the year 2009. The level of collection for this data was at the town level. The information about each of the residences includes: the assessed value, address, name of owner, year built, zip code, neighborhood code, basement type, school district, square foot of living area, the number of bedrooms and bathrooms, building style, etc.

The variables provided in the original data set are listed and defined in Appendix A. The raw data set of for Amherst Town contained 42,433 observations for all property classes. After the data was cleaned up, the sample size for Amherst Town was 33,342 observations. The data set cleanup processes are discussed in detail in the following sections.

3.2 List of Variables

The main data came from the Amherst Town assessment database, representing a total of 33,342 single-family residential homes in the year 2009. These homes encompass the full spectrum of lot sizes, structural designs, and household amenities. The existing literature on this subject has included a number of variables in their multiple regression models. However, the property characteristics considered as potential regressors in the model used in this dissertation were limited to those variables that were available in the Amherst Town assessment database.

3.2.1 Dependent Variable

The dependent variable to be explored is the assessed housing price of single-family properties in Amherst Town, State of New York. Properties that are assessed in the year 2009 will be included in the estimation. The assessed prices range continuously from $4,000 to $1,600,000 in the database. The average housing price is 151,745. Table 3.1 shows the distribution of the assessed housing prices. Also, Figure 3.1 presents the histogram of the assessed prices. Inspection of the histogram reveals that the vast majority of the observations are located between $80,000 and $160,000. The number of residences with assessed prices between $ 80,000 and $160,000 is 19,120, 57.34% of the total. However, there is the dispersion of assessed prices, and some observed prices are quite distanced from the average assessed price.

TABLE 3.1: Distribution of the Assessed Prices

Range ($)	Number	%
4,000 ~ 80,000	3,321	9.96%
80,000 ~ 160,000	19,120	57.34%
160,000 ~ 240,000	8,055	24.16%
240,000 ~320,000	1,830	5.49%
320,000 ~ 400,000	563	1.69%
400,000 ~ 480,000	215	0.64%
480,000 ~ 560,000	113	0.34%
> 560,000	125	0.38%
Total	**33,342**	**100%**

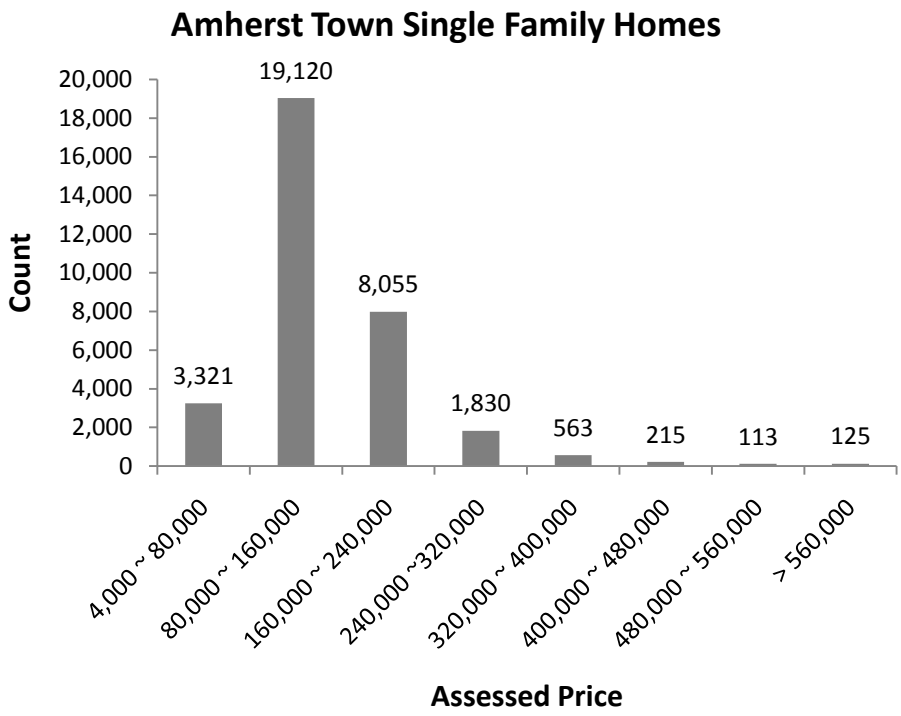


FIGURE 3.1: Distribution of the Assessed Prices

3.2.2 Independent Variables

Eighty five variables were used to describe factors such as location, important structural and physical characteristics and household amenities that could likely influence the assessed price of a home. Additional variables were included to measure the specific relationship between dependent variable and the independent variables.

Several housing characteristics expected to influence the housing price value can be classified into three categories: (i)physical structural attributes, (ii)household amenity features, and (iii)locational characteristics. The following section contains the description of each of the variables in detail by in the categories.

3.2.2.1 Physical Structural Attributes

For the regression model to accurately assess the price of the single-family home, housing physical structural attributes that are important to tax assessment officers must be carefully considered. In the effort to capture the unique structural characteristics of the property, the following eighteen distinct variables were employed in the model:

1) *age*: the age of the building, measured in years, was examined as an important condition. This variable is identified as continuous variable. To take depreciation into account, including this variable as the independent variable can easily capture the variation in the housing price. Furthermore, this variable may explain why older houses generally have less value than the newer ones. Table 3.2 and Figure 3.2 show the distribution of the age of the houses. As the table and diagram indicate, most housing ages are located within 20 years through 60 years. The number of the residences with housing age between 20 years and 60 years is 21,775, about 65% of the total number. Few houses are older than 100 years.

TABLE 3.2: Distribution of the Housing Age

Range (yr)	Number	%
1 ~ 20	6,518	19.55%
20 ~ 40	9,345	28.03%
40 ~ 60	12,430	37.28%
60 ~80	3,561	10.68%
80 ~ 100	1,102	3.31%
> 100	386	1.16%
Total	**33,342**	**100%**

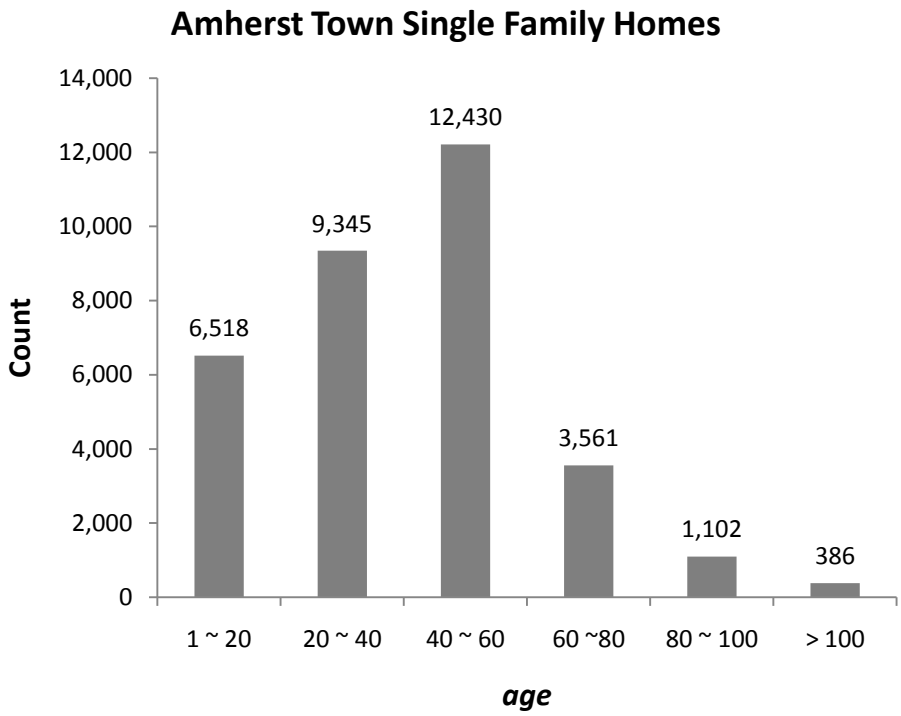


FIGURE 3.2: Distribution of the Housing Age

2) *frontage*: the width of the parcel, the physical dimensions of the parcel, measured in feet. Based on some prior expectation on price effects, the width of the parcel was expected to have the positive association with housing prices. Table 3.3 and Figure 3.3 indicate the distribution of the frontage of the parcel. These figure shows that a lot of the frontage of the observations locates between 40 feet and 100 feet.

TABLE 3.3: Distribution of the Frontage of the Parcel

Range (feet)	Number	%
0 ~ 40	3,497	10.49%
40 ~ 60	8,285	24.85%
60 ~ 80	13,022	39.06%
80 ~100	5,173	15.51%
100 ~ 120	1,753	5.26%
> 120	1612	4.83%
Total	**33,342**	**100%**

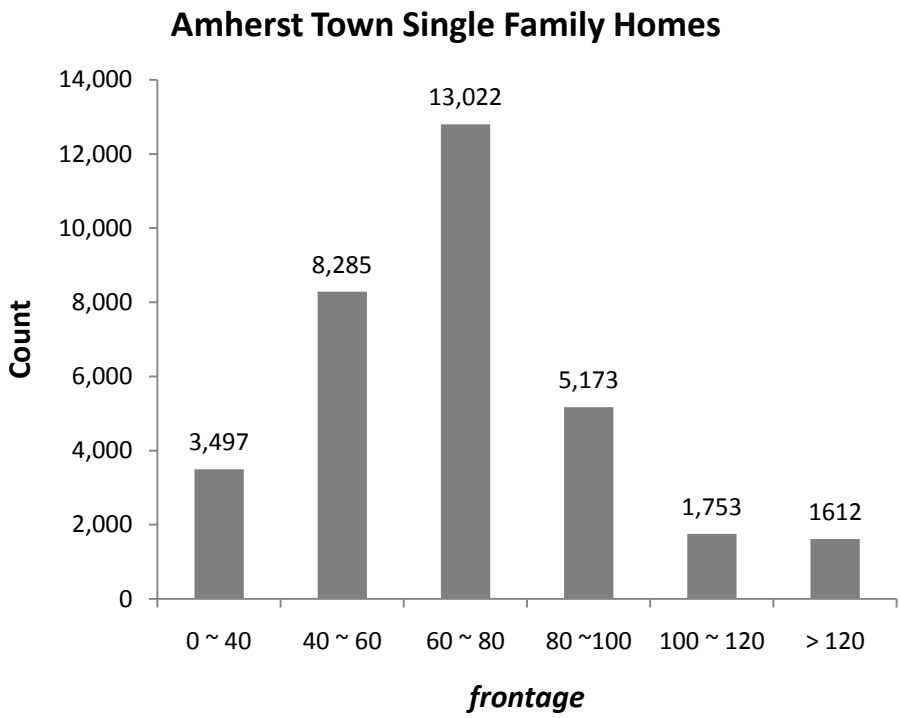


FIGURE 3.3: Distribution of the Frontage of the Parcel

3) *depth*: the depth of parcel, the physical dimension of the parcel, measured in feet.

The depth of the parcel also has the positive relationship with the housing prices. Figure 3.4 and Table 3.4 illustrate the distribution of the depth of the parcel. Most of the depths of the parcel are between 120 feet and 160 feet.

TABLE 3.4: Distribution of the Depth of the Parcel

Range (feet)	Number	%
0 ~ 80	2,551	7.65%
80 ~ 120	3,110	9.33%
120 ~ 160	18,377	55.12%
160 ~200	4,881	14.64%
200 ~ 240	1,716	5.15%
240 ~ 320	1,387	4.16%
> 320	1320	3.96%
Total	**33,342**	**100%**

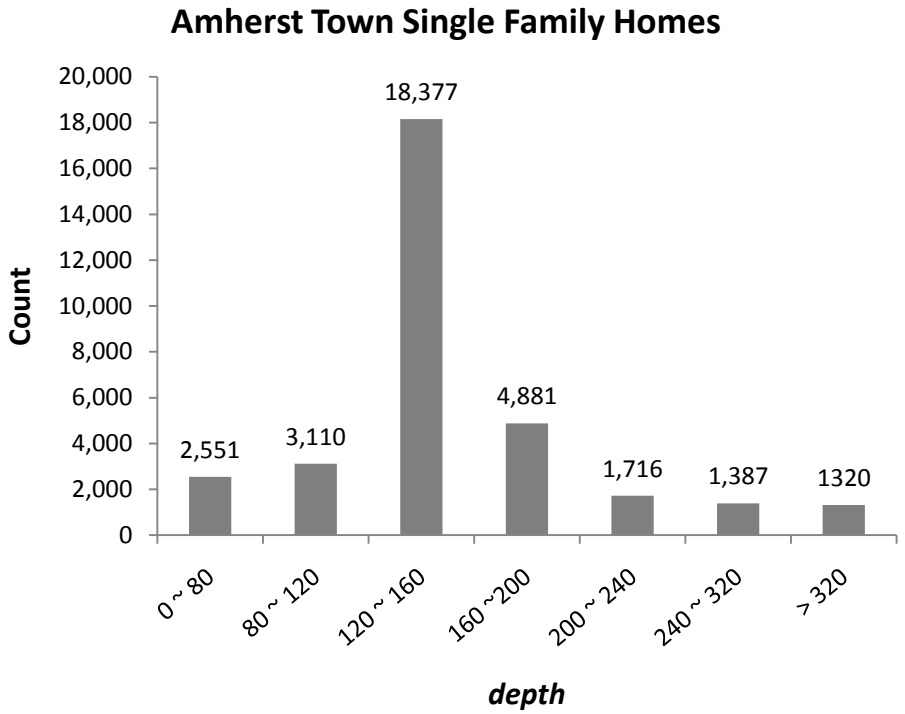


FIGURE 3.4: Distribution of the Depth of the Parcel

4) *sfla*: the square feet of living area, the physical dimension of the living area, was measured in square feet. This variable is involved in that the size of the house is a key indicator of the assessed prices. Generally speaking, holding all physical attributes as the same, the larger the house is, the higher the assessed housing prices will be. In this dissertation, this variable, *sfla*,

will be treated as the continuous variable. Figure 3.5 and Table 3.5 present the distribution of the living area of the houses. The range of the square feet of the living area is from 480 square feet to 10,163 square feet in this dataset. The histogram illustrates that most of the *sfla* of the residences fall between 1,200 square feet and 2,000 square feet.

Table 3.5: Distribution of the Living Area of the Houses

Range (square feet)	Number	%
400 ~ 1200	4,483	13.45%
1200 ~ 1600	7,856	23.56%
1600 ~ 2000	7,958	23.87%
2000 ~2400	5,975	17.92%
2400 ~ 2800	3,645	10.93%
2800 ~ 3600	2,534	7.6%
> 3600	891	2.67%
Total	**33,342**	**100%**

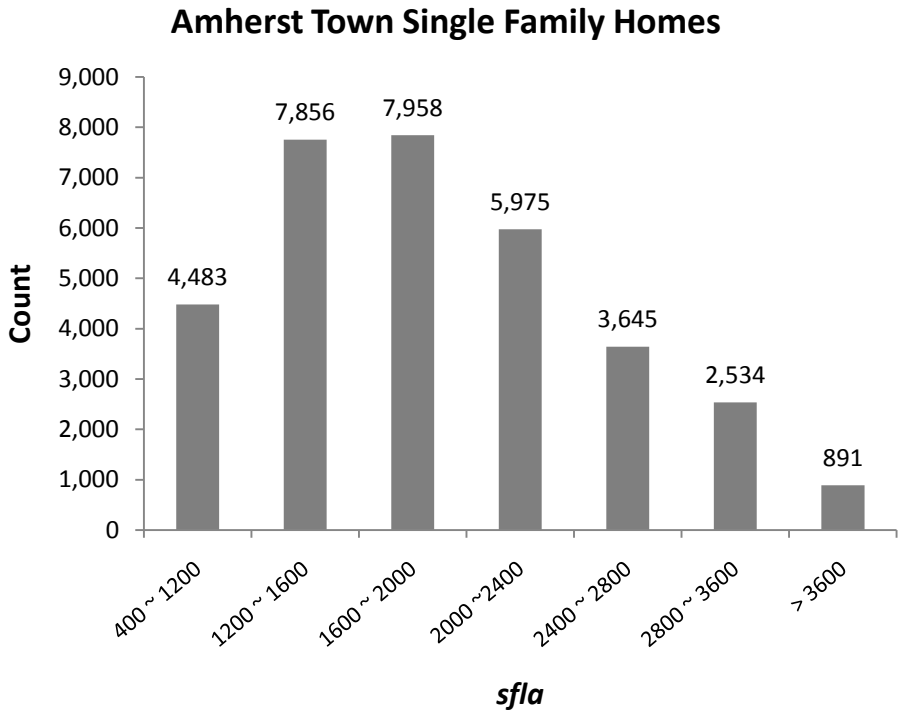


FIGURE 3.5: Distribution of the Living Area of the Houses

5) *nobedrms*: the total number of bedrooms.

This variable is designed to measure the allocation of living space within the home. Table 3.6 and Figure 3.6 represent the distribution of the total number of bedrooms in the house. Most homes have the number of bedroom from 2 to 5. It counts around 99% of the total house.

TABLE 3.6: Distribution of the Total Number of Bedrooms

Range	Number	%
1	221	0.66%
2	3,584	10.75%
3	16,039	48.10%
4	12,110	36.32%
5	1,246	3.74%
> 6	142	0.43%
Total	**33,342**	**100%**

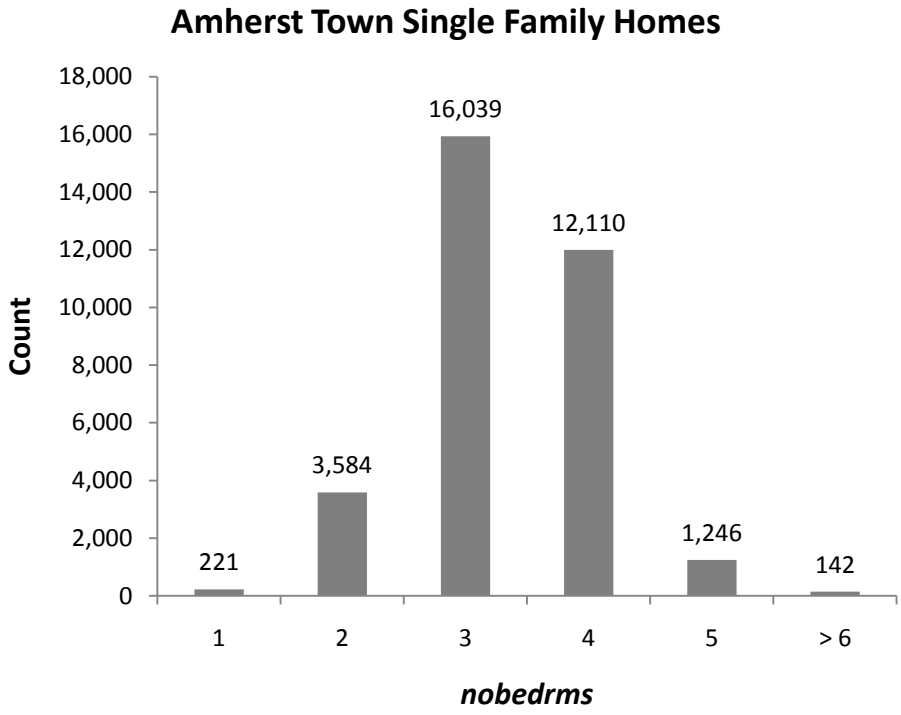


FIGURE 3.6: Distribution of the Total Number of Bedrooms

6) *nobaths*: the total number of bathrooms.

This variable is selected to measure the allocation of living space within the home. Table 3.7 and Figure 3.7 delineate the distribution of the total number of bathrooms in the house. It shows that a large number of the observations locate between 1 and 2.5.

TABLE 3.7: Distribution of the Total Number of Bathrooms

Range	Number	%
1	7,123	21.36%
1.5	10,048	30.14%
2	4,173	12.52%
2.5	10,522	31.56%
3	587	1.76%
3.5	571	1.71%
> 4	318	0.95%
Total	**33,342**	**100%**

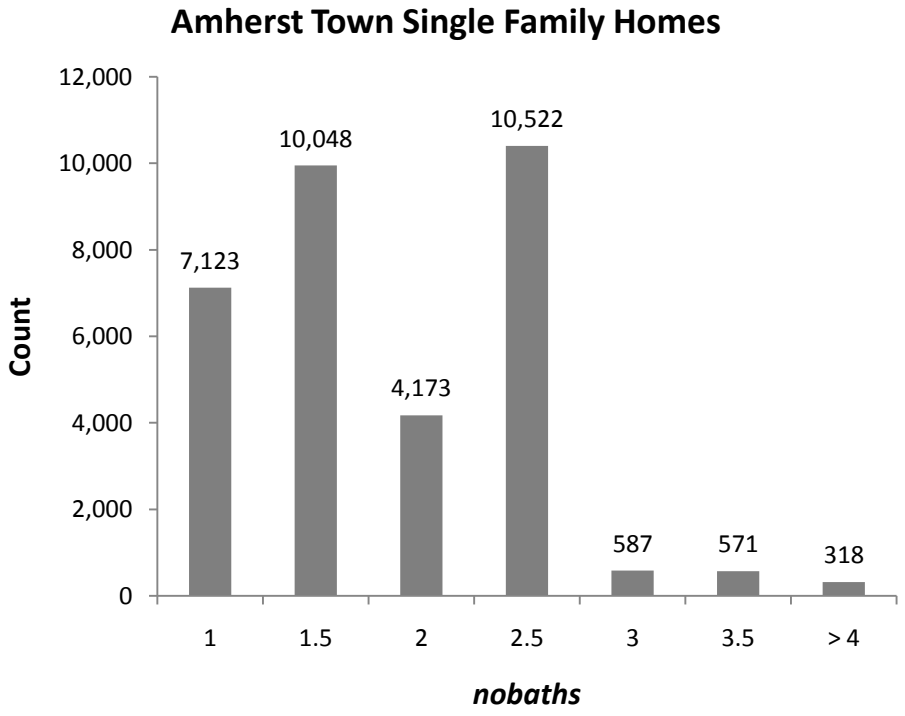


FIGURE 3.7: Distribution of the Total Number of Bathrooms

7) *bstyle1*, *bstyle2*, *bstyle3*,……, *bstyle12*: the dummy variables of external building style. Different exterior building styles have different appearances on the roof, garage, basement type, façade, and so on. In the Town of Amherst, exterior building style is utilized to classify the residence as to its architectural style. In the dataset, with regard to the building style, the dwellings were divided into thirteen types, which were: ranch, raised ranch, split level, cape cod, colonial, contemporary, mansion, old style, cottage, log cabin, duplex, town house, and others. Ranch building style was used as the benchmark group. The remaining building styles were identified by twelve dummy variables: *bstyle1*, *bstyle2*, *bstyle3*, *bstyle4*, *bstyle5*, *bstyle6*, *bstyle7*, *bstyle8*, *bstyle9*, *bstyle10*, *bstyle11*, and *bstyle12*, respectively. The brief description of the building style is in Appendix B. Also, the distribution of the building style is presented in Table 3.8 and Figure 3.8. From the table, the Colonial building style which is very popular style in the Town of Amherst counts 40.22% of the total single family homes. Totally, the style of cottage and log cabin which is usually not the common style of single family residence counts only 0.09% of the total number of houses.

TABLE 3.8: Distribution of Building Style

Style	Number	%
Ranch	7,152	21.45%
Raised Ranch	383	1.15%
Split Level	2,960	8.88%
Cape Cod	4,059	12.17%
Colonial	13,409	40.22%
Contemporary	355	1.06%
Mansion	166	0.50%
Old Style	3,039	9.11%
Cottage	26	0.08%
Log Home	3	0.01%
Duplex	1,031	3.10%
Town House	734	2.20%
Other	24	0.07%
Total	**33,342**	**100%**

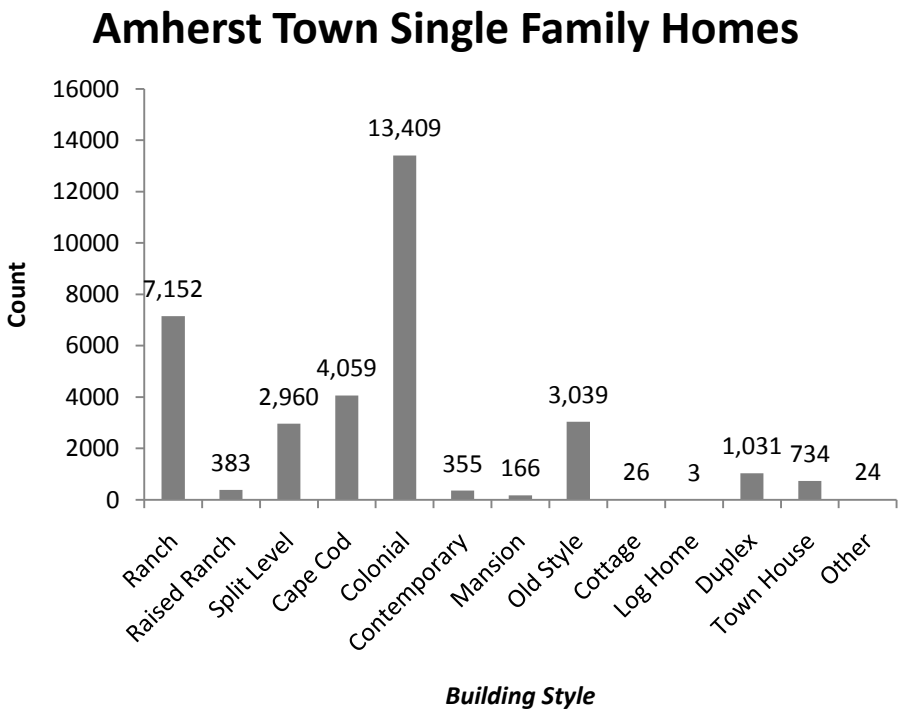


FIGURE 3.8: Distribution of the Building Style

3.2.2.2 Household Amenity Features

Amenities can be either positives or negatives. Fireplace is a positive amenity. The decision of the tax assessment officer may be influenced by fireplace. The following variables were employed in the model to capture the positive effect of fireplace on the home value:

1) *nofirepl*: indicated how many built-in fireplaces in the house. The existence of the fireplace is anticipated to have positive effects on the house prices. Table 3.9 and Figure 3.9 show the distribution of *nofirepl*. Most houses have either one fireplace or none.

TABLE 3.9: Distribution of the Number of the Fireplaces

Range	Number	%
0	11,825	35.47%
1	19,882	59.63%
2	1,474	4.42%
3	125	0.37%
> 4	36	0.11%
Total	**33,342**	**100%**

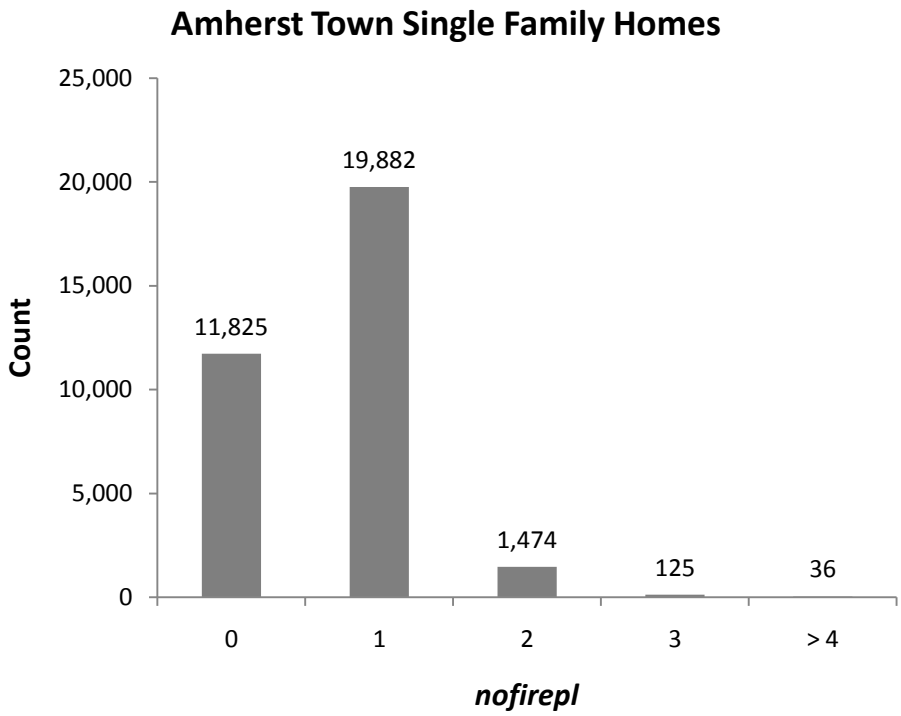


FIGURE 3.9: Distribution of the Number of the Fireplaces

3.2.2.3 Locational Characteristics

Location is important in determining the value of the house, since the values of identical houses are thought to vary significantly over space because of an unmeasured location difference. The variable, *nghdcode,* from the original dataset represents the different neighborhoods defined by the Amherst County property tax assessment office in the Town of Amherst. Different geographical areas are given different neighborhood codes by the assessor's office of the Town of Amherst. The number of the residences in each neighborhood code is presented in Table 3.10.

TABLE 3.10: *nghdcode* Distribution

#	*nghdcode*	Number of Residences	%	#	*nghdcode*	Number of Residences	%
1	100	496	1.49%	35	3300	13	0.04%
2	200	230	0.69%	36	3400	551	1.65%
3	300	127	0.38%	37	3500	649	1.95%
4	400	1532	4.59%	38	3600	1,317	3.95%
5	500	1,089	3.27%	39	3700	1,657	4.97%
6	600	176	0.53%	40	3800	1,748	5.24%
7	700	345	1.03%	41	3900	345	1.03%
8	701	311	0.93%	42	4000	386	1.16%
9	800	320	0.96%	43	4100	68	0.20%
10	900	279	0.84%	44	4200	133	0.40%
11	1000	556	1.67%	45	4300	609	1.83%
12	1100	472	1.42%	46	4400	594	1.78%
13	1200	1,115	3.34%	47	4500	84	0.25%
14	1300	102	0.31%	48	4600	46	0.14%
15	1400	172	0.52%	49	4700	155	0.46%
16	1401	134	0.40%	50	4800	117	0.35%
17	1500	341	1.02%	51	4900	560	1.68%
18	1600	175	0.52%	52	5000	452	1.36%
19	1700	176	0.53%	53	5100	378	1.13%
20	1800	415	1.24%	54	5200	1,933	5.80%
21	1900	277	0.83%	55	5300	359	1.08%
22	2000	572	1.72%	56	5400	694	2.08%
23	2100	301	0.90%	57	5500	87	0.26%
24	2200	241	0.72%	58	5600	712	2.14%
25	2300	496	1.49%	59	5700	13	0.04%
26	2400	155	0.46%	60	5800	1,078	3.23%
27	2500	195	0.58%	61	5900	432	1.30%
28	2600	427	1.28%	62	6000	540	1.62%
29	2700	255	0.76%	63	6100	672	2.02%
30	2800	1,312	3.93%	64	6200	4	0.01%
31	2900	438	1.31%	65	6300	239	0.72%
32	3000	254	0.76%	66	6400	629	1.89%
33	3100	454	1.36%	67	6500	399	1.20%
34	3200	1,749	5.25%				
				Total		33,342	100%

1) Locational variables: *nghd1*, *nghd2*, *nghd3*,……, and *nghd66*

In an attempt to quantify the impact of location on housing value, sixty six neighborhood dummy variables, *nghd1*, *nghd2*… and, *nghd66* were included. The neighborhood dummy variables appear to do a good job of capturing the effects of local public services and taxes. Each neighborhood has its own government, responsible for the school system, other public services and taxation. For the locational variables, neighborhood 100 was used as the base neighborhood. *nghd1* was used to designate whether or not the observed property was located within the neighborhood 200. *nghd2* was used to designate whether or not the observed property was located within the neighborhood 300. The remaining neighborhood dummy variables, *nghd3*,……, and *nghd66*, follow this rule.

In the previous literature, other variables capturing interior and exterior characteristics commonly found in the multiple regression model are excluded from this research in that those data are not available. These variables include the number of stories, the presence of a garage and the presence of the swimming pool. If the certain housing features important for property tax assessment officer to judge the housing price are not included in the model, the model could have "missing variable bias". While no regression model could take all influential factors into account, this dissertation attempts to be as thorough as possible. It is based on the assumption that the number of variables of the property could explain the variation of the property physical characteristics for the housing prices.

3.3 Development of Models

3.3.1 Project Data Cleanup and Eliminating Outlier Records

In building the data set used for this dissertation, a number of selection rules were followed to insure a valid data set. Prior to estimation, the data set included parcels that are assessed for the year 2009. All parcels that had missing data, or had unreasonable define data were eliminated. If any of the key variables: *noberdrms*, *nobaths*, *age*, *sfla*, *frontage*, or *depth* have missing data, these parcels were eliminated from the data set. Removal of these observed records prevents coding errors. Moreover, parcels with unreasonable data (both housing characteristics and/or assessed prices) are eliminated from the data set. See Table 3.11 for a complete definition of the data set cleanup process. Including properties in the data set that violate any of the cleanup process rules will cause a bias in the estimation of coefficients, which will lead to misunderstanding of results. After the data were cleaned, the sample size for Amherst Town is 33,342 observations.

TABLE 3.11: Data Set Cleanup Process

Data Cleanup

1. Delete *sfla* records if
    a. = 0 (or Null)
    b. Missing
2. Delete *nobedrms* records if
    a. = 0 (or Null)
    b. Missing
3. Delete *nobaths* records if
    a. = 0 (or Null)
    b. Missing
4. Delete *frontage* records if
    a. = 0 (or Null)
    b. Missing
5. Delete *depth* records if
    a. = 0 (or Null)
    b. Missing
6. Delete *age* records if
    a. < 0 (or Null)
    b. Missing
7. Delete *nofirepl* records if
    a. < 0
    b. Missing
8. Delete *bldstyle* records if
    a. < 0 (or Null)
    b. Missing
9. Delete *nghdcode* records if
    a. Null
    b. Missing
10. Delete records if no assess year

3.3.2 Dividing the Dataset into Training Set and Validation Set

For the purpose of the model validation, the 33,342 records were randomly divided into two parts using a split of 80% - 20%. The portion used for validation comprises of 6,668 records - 20 percent of the total, the remaining 26,674 records were applied for the training set. The root

mean squared error (RMSE), the mean absolute error (MAE), and the Theil's U statistic were used for validation.

3.3.3 Examination of the Covariance Structure across Model Variables

Multicollinearity between housing characteristics was anticipated and therefore examined. In order to examine which variables may pose multicollinearity issues, the correlations between variables were examined using Pearson Correlation method. Table 3.12 presents the results of the Pearson correlations, wherein highly significant correlations are highlighted in bold. As can be seen in the table, the variable *age* is highly negatively correlated with the *nobaths*. Also, the *sfla* has a high positive correlation to *nofirepl*, *nobedrms*, and *nobaths*, these three variables were therefore excluded from the benchmark multiple regression model.

TABLE 3.12: Pearson Correlation Coefficient

	frontage	*depth*	*age*	*sfla*	*nofirepl*	*nobedrms*	*nobaths*
frontage	1.00						
depth	0.28	1.00					
age	-0.17	0.07	1.00				
sfla	0.35	0.08	-0.33	1.00			
nofirepl	0.23	0.05	-0.21	**0.51**	1.00		
nobedrms	0.18	0.01	-0.22	**0.62**	0.28	1.00	
nobaths	0.30	0.01	**-0.43**	**0.77**	**0.47**	**0.58**	1.00

Note: all correlation coefficients are significant at the 0.05 level.

3.3.4 Determination of the Functional Form of the Dependent Variable: *hprice*

The linear, semi-logarithmic and root transformations of dependent variable *hprice* were evaluated. In order to examine the model fit, the natural log: $\log_n(hprice)$ and root transformation: $\sqrt{hprice}$ were compared to the straight linear form: *hprice*. The distribution of standardized residuals was used to examine the goodness-of-fit. All three models had reasonably good fit. The R^2 value for the linear, natural log, and root transformation was 0.90, 0.89, and 0.91, respectively. However, the linear fit is relatively easier to interpret and to compute by the users, therefore, the linear form of the dependent variable *hprice* was used in the model.

Figure 3.10 shows the distribution of standardized residuals for the linear functional form. As seen in this figure, the standardized residuals appear to clearly satisfy the normality assumption, which is important relative to the appropriateness of our inference.

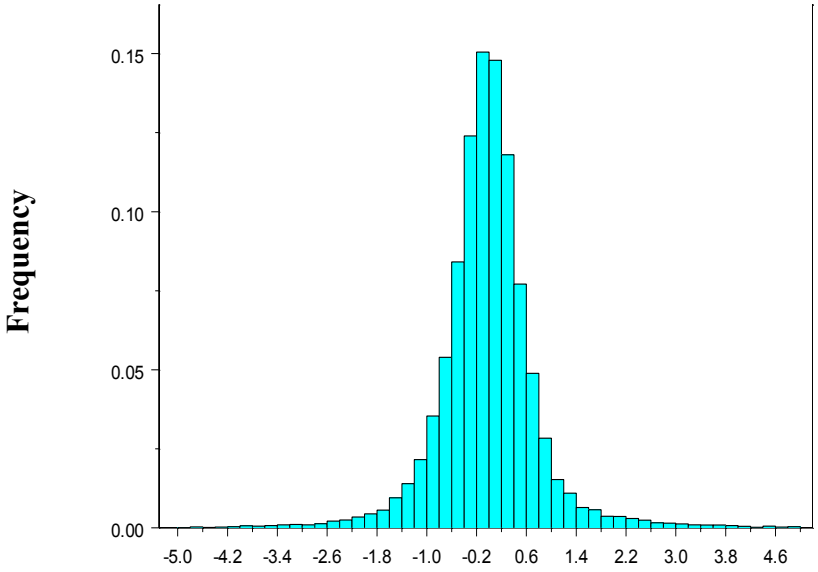


FIGURE 3.10: Distribution of the Standardized Residuals

3.3.5 Selection of Significant Independent Variables: Stepwise Multiple Regression

A stepwise regression approach was employed. Variables were included/excluded at each step in the model based at an $\alpha = 0.05$ significance level. One advantage of this approach is to minimize the impact of any potential multicollinearity. Variables offering significant explanatory variance conditional on all variables considered were retained and included in the final model.

The complete data set represents a total of 33,342 single-family residential homes for the year 2009. To capture the variation in housing prices, a total of 85 independent variables are applied to construct the mass appraisal models. Table 3.13 provides the definitions and descriptions of the housing attributes.

TABLE 3.13: Description of Variables Initially Considered for Model Development

Variable	Description	Unit of Measures
Dependent Variable:		
hprice	assessed housing prices	dollar
Independent Variables:		
(1) Property Physical Structural Attributes:		
frontage	Parcel frontage	feet
depth	Parcel depth	feet
age	Year house was built subtracted from 2008	year
sfla	Square footage of gross living area	square foot
nobedrms	Number of bedrooms	NA
nobaths	Number of bathrooms	NA
bstyle1	Dummy variable for building style (1 if raised ranch, 0 otherwise)	binary
bstyle2	Dummy variable for building style (1 if split level, 0 otherwise)	binary
bstyle3	Dummy variable for building style (1 if cape cod, 0 otherwise)	binary
bstyle4	Dummy variable for building style (1 if colonial, 0 otherwise)	binary
bstyle5	Dummy variable for building style (1 if contemporary, 0 otherwise)	binary
bstyle6	Dummy variable for building style (1 if mansion, 0 otherwise)	binary
bstyle7	Dummy variable for building style (1 if old style, 0 otherwise)	binary
bstyle8	Dummy variable for building style (1 if cottage, 0 otherwise)	binary
bstyle9	Dummy variable for building style (1 if log cabin, 0 otherwise)	binary
bstyle10	Dummy variable for building style (1 if duplex, 0 otherwise)	binary
bstyle11	Dummy variable for building style (1 if town house, 0 otherwise)	binary
bstyle12	Dummy variable for building style (1 if others, 0 otherwise)	
(2) Household Amenity Features:		
nofirepl	Number of fireplaces	NA
(3) Locational Characterisitcs:		
nghd1~nghd66	Dummy variables for neighborhood	binary

3.3.5.1 Description of Qualitative Variables

The final qualitative independent variables included in the multiple regression model are listed in Table 3.13. They are external building style variables: *bstyle1*, *bstyle2*, *bstyle3*,……, *bstyle12* and locational variables: *nghd1*, *nghd2*, *nghd3*,……, and *nghd66*.

3.3.5.2 Statistics of Quantitative Variables

The final quantitative independent variables included in the multiple regression model are also listed in Table 3.13. Table 3.14 gives the summary statistics for all quantitative variables used in the mass appraisal models. On average, a house in the Town of Amherst, New York was assessed for $151,745. The standard deviation from the mean is relatively high because of the large variation in housing prices. The smallest house that was assessed had only 480 ft^2 of living area and the largest had 10,163 ft^2. The mean size of living area was 1,927 ft^2 and the standard deviation was about 732 ft^2. The building age varies from 1 to 206 years with a mean age of 42 years.

TABLE 3.14: Summary Statistics of Quantitative Variables

Variable	Mean	Standard Deviation	Minimum	Maximum
Dependent Variable:				
hprice	151,745.44	77,267.59	4,000.00	1,600,000
Independent Variables:				
frontage	70.65	34.77	20	2,060
depth	156.5	96.37	30	4,138
age	42.05	23.19	1	206
sfla	1,927.03	732.36	480	10,163
nobedrms	3.33	0.77	1	12
nobaths	1.86	0.69	1	7.5
nofirepl	0.7	0.58	0	10

CHAPTER 4: Multiple Regression Model for Estimation of Property Values

4.1 Introduction

This chapter describes the development of a multiple regression model for housing price estimation, and links it to the GIS map layer of a municipality to draw various maps showing the wide variation in the prices of homes based on location or neighborhood. A statistical model was developed using the 2009 housing assessment data of 33,342 houses of a town in Western New York. The multiple regression model had high accuracy with a R^2 value of 0.9046, and tested well on validation. The price estimates of 33,342 houses using the model were tested for the number of overestimates and underestimates. Multiple regression model have been developed in the past, but the research presented in this chapter links the statistical model to GIS map layers to enable the appraising authority to visualize the distribution of housing based on price, or age by neighborhood. This can help in resolving any inequities with in a neighborhood.

This chapter takes a new approach and differs in a number of respects from the earlier studies. In the first place, ArcGIS software Version 9.3 was applied on the results from the multiple regression model to analyze and visualize the spatial variations in the housing sub-markets. The housing dataset for the Town of Amherst, State of New York that was segmented into 67 neighborhoods, was used to develop and clearly demonstrate the spatial patterns. Secondly, tables, graphs, and maps of prediction accuracy were created based on the standard error of housing price residuals for each neighborhood. These tables, graphs and maps can assist a town's property tax assessment officers in housing price estimation and comparisons within the sub-markets. The major focus of the research described in this chapter is to develop a comprehensive explanatory model of housing price determination using statistical methods and also visualize the

price variation patterns of different neighborhoods using GIS. The model developed in this research also gives several numerical summaries about the example housing market.

The format of the chapter is constructed as follows. In section 2, multiple regression (MR) model and ArcGIS Software is briefly introduced. Section 3 contains a description of the model development, and Section 4 describes the general geographic locations of the housing price estimates through GIS maps and reports spatial analysis. The final section provides the summary of multiple regression model analysis results.

4.2 Introduction to Multiple Regression (MR) Model and ArcGIS software

4.2.1 Introduction to Multiple Regression (MR) Model

Multiple regression analysis is the statistical approach for modeling the relationship between two sets of variables, and in its linear additive format can be written as:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon,} \quad \text{or}$$

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon, \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (1)$$

where: Y is the dependent variable: housing price in this research;
$X_1, \ldots, X_p$, is a set of independent variables; and
$E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.

A multiple regression model can have meaningful interpretation in relating the variable of interest. The superiority of multiple regression model is its simplicity, good statistical properties and computational advantages. However, the main concern of multiple regression model is the difficulty in choosing the right functional form of the dependent variable and secondly the other assumptions related to the error term in the regression model may not be satisfied.

4.2.2 Introduction to ArcGIS

ArcGIS is a geographic information system (GIS) software produced by ESRI Press. This software allows us to visualize, manage, create, and analyze the geographic data. Using ArcGIS, the geographic context of the data can be better understood. Moreover, it allows us to see the relationships and patterns in the geographic data in identifiable ways.

4.3 Model Development

Multiple regression was used for developing a model to estimate housing prices for mass appraisal. Figure 4.1 represents the flow chart for the model development process. The following steps were used in formulating the multiple regression model:

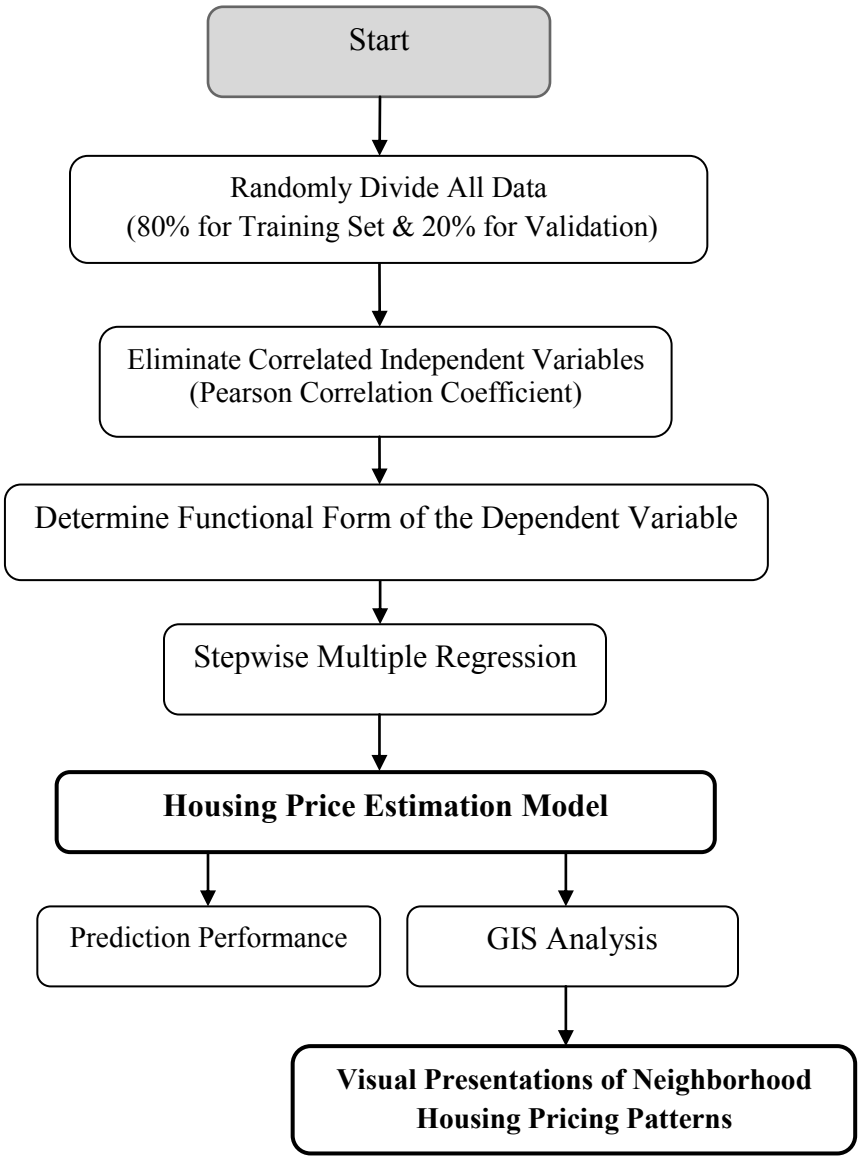


FIGURE 4.1: Model Development Process

4.3.1 Dividing the Dataset into Training Set and Validation Set

Refer to 3.3.2

4.3.2 Examination of the Covariance Structure across Model Variables

Refer to 3.3.3

4.3.3 Determination of the Functional Form of the Dependent Variable *hprice*

Refer to 3.3.4

4.3.4 Selection of Significant Independent Variables: Stepwise Multiple Regression

Refer to 3.3.5

4.3.5 Housing Price Estimation Model

The following parametric model with a linear form was proposed:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots\cdots + \beta_{82} X_{82i} + \varepsilon_i , \qquad (1)$$

where Y donates assessed housing price, β_0 is the intercept parameter. The variables: X_{1i} through X_{82i} correspond to the 82 independent variables considered in the model and β_1 through β_{82} are the corresponding regression coefficients. We assumed that the ε_i are normally distributed with mean 0 and variance σ^2. The factors we considered include measures such as *frontage*, *depth*, *age*, *sfla*, *bstyle1 ~ bstyle12*, and *nghd1 ~ nghd66*. The regression coefficients of the final model developed are given in Table 4.1.

TABLE 4.1: Multiple Regression Model Estimates

Variable	Regression Coefficient Estimate	Standard Error	t Value*	Pr > |t| **
Dependent Variable:				
hprice				
Independent Variables:				
intercept	47,049	1,761.17	**26.71**	**< 0.0001**
frontage	88.60	6.94	**12.77**	**< 0.0001**
depth	14.18	2.00	**7.09**	**< 0.0001**
age	-383.93	15.67	**-24.5**	**< 0.0001**
sfla	75.97	0.41	**183.29**	**< 0.0001**
bstyle1 (raised ranch)	-28,364	1,685.53	**-16.83**	**< 0.0001**
bstyle2 (split level)	-13,130	739.52	**-17.75**	**< 0.0001**
bstyle3 (cape cod)	-12,924	662.39	**-19.51**	**< 0.0001**
bstyle4 (colonial)	-6,594.26	575.67	**-11.45**	**< 0.0001**
bstyle5 (contemporary)	-9,659.54	1,835.08	**-5.26**	**< 0.0001**
bstyle6 (mansion)	105,207	3,127.53	**33.64**	**< 0.0001**
bstyle7 (old style)	-1,564.6	978.38	-1.6	0.1098
bstyle8 (cottage)	-8,610.58	9,604.72	-0.9	0.37
bstyle9 (log home)	-3,067.71	15,658	-0.2	0.8447
bstyle10 (duplex)	-62,279	13,557	**-4.59**	**< 0.0001**
bstyle11 (town house)	-15,270	1,354.04	**-11.28**	**< 0.0001**
bstyle12 (other)	45,915	7,569.53	**6.07**	**< 0.0001**
nghd1 (200)	-4,982.167	2,686.89	-1.85	0.0637
nghd2 (300)	-14,670	3,667.71	**-4.00**	**< 0.0001**
nghd3 (400)	-6,857.62	1,768.69	**-3.88**	**0.0001**
nghd4 (500)	-965.01	1,728.56	-0.56	0.5767
nghd5 (600)	13,181	2,857.76	**4.61**	**< 0.0001**
nghd6 (700)	23,411	2,198.55	**10.65**	**< 0.0001**
nghd7 (701)	77,727	2,620.43	**29.66**	**< 0.0001**
nghd8 (800)	18,861	2,279.93	**8.27**	**< 0.0001**
nghd9 (900)	8,607.37	2,376.47	**3.62**	**0.0003**
nghd10 (1000)	-2,511.45	1,992.03	-1.26	0.2074
nghd11 (1100)	19,058	2,091.87	**9.11**	**< 0.0001**
nghd12 (1200)	-415.66	1,748.36	-0.24	0.8121
nghd13 (1300)	31,939	3,361.03	**9.5**	**< 0.0001**
nghd14 (1400)	-5,016.49	2,814.89	-1.78	0.0747
nghd15 (1401)	-6,371.45	3,174.18	**-2.01**	**0.0447**
nghd16 (1500)	11,196	2,400.4	**4.66**	**< 0.0001**
nghd17 (1600)	24,112	2,776.76	**8.68**	**< 0.0001**
nghd18 (1700)	-3,483.57	2,764.19	-1.26	0.2076
nghd19 (1800)	2,756.09	2,111.47	1.31	0.1918
nghd20 (1900)	7,229.47	2,367.16	**3.05**	**0.0023**

TABLE 4.1: - continued

Variable	Regression Coefficient Estimate	Standard Error	t Value*	Pr > |t| **
nghd21 (2000)	36,757	2,035.90	**18.05**	**< 0.0001**
nghd22 (2100)	-2,779.54	2,368.01	-1.17	0.2405
nghd23 (2200)	77,275	2,551.82	**30.28**	**< 0.0001**
nghd24 (2300)	12,099	2,064.14	**5.86**	**< 0.0001**
nghd25 (2400)	84,237	2,829.65	**29.77**	**< 0.0001**
nghd26 (2500)	170,050	3,230.85	**52.63**	**< 0.0001**
nghd27 (2600)	4,201.24	2,124.93	**1.98**	**0.0480**
nghd28 (2700)	6,038.39	2,445.08	**2.47**	**0.0135**
nghd29 (2800)	-20,195	1,696.79	**-11.9**	**< 0.0001**
nghd30 (2900)	1,039.64	2,057.26	0.51	0.6133
nghd31 (3000)	43,748	2,546.62	**17.18**	**< 0.0001**
nghd32 (3100)	-29,292	2,068.2	**-14.16**	**< 0.0001**
nghd33 (3200)	-10,506	1,622.44	**-6.48**	**< 0.0001**
nghd34 (3300)	22,568	8,309.64	**2.72**	**0.0066**
nghd35 (3400)	-2,790.7	1,970.81	-1.42	0.1568
nghd36 (3500)	-1,616.62	1,892.32	-0.85	0.3929
nghd37 (3600)	-2,223.19	1,676.26	-1.33	0.1848
nghd38 (3700)	-28,434	1,669.39	**-17.03**	**< 0.0001**
nghd39 (3800)	-18,628	1,694.84	**-10.99**	**< 0.0001**
nghd40 (3900)	-29,006	2,248.41	**-12.90**	**< 0.0001**
nghd41 (4000)	6,405.03	2,134.23	**3.00**	**0.0027**
nghd42 (4100)	64,844	3,965.26	**16.35**	**< 0.0001**
nghd43 (4200)	-30,576	3,007.3	**-10.17**	**< 0.0001**
nghd44 (4300)	12,079	1,838.00	**6.57**	**< 0.0001**
nghd45 (4400)	2,992.88	1,943.66	1.54	0.1236
nghd46 (4500)	12,720	4,056.49	**3.14**	**0.0017**
nghd47 (4600)	177,940	4,794.58	**37.11**	**< 0.0001**
nghd48 (4700)	67,508	2,892.78	**23.34**	**< 0.0001**
nghd49 (4800)	108,115	3,696.43	**29.25**	**< 0.0001**
nghd50 (4900)	44,499	1,870.92	**23.78**	**< 0.0001**
nghd51 (5000)	-1,617.49	2,172.34	-0.74	0.4565
nghd52 (5100)	45,773	2,143.05	**21.36**	**< 0.0001**
nghd53 (5200)	-928.08	1,613.21	-0.58	0.5651
nghd54 (5300)	-9,804.03	2,169.01	**-4.52**	**< 0.0001**
nghd55 (5400)	-16,160	1,906.70	**-8.48**	**< 0.0001**
nghd56 (5500)	2,172.54	3,596.19	0.6	0.5458
nghd57 (5600)	-14,649	1,888.50	**-7.76**	**< 0.0001**
nghd58 (5700)	94,106	8,311.20	**11.32**	**< 0.0001**
nghd59 (5800)	-5,762.82	1,724.35	**-3.34**	**0.0008**
nghd60 (5900)	13,111	2,234.99	**5.87**	**< 0.0001**
nghd61 (6000)	12,987	2,096.56	**6.19**	**< 0.0001**
nghd62 (6100)	6,587.94	1,899.39	**3.47**	**0.0005**
nghd63 (6200)	215,249	13,643	**15.78**	**< 0.0001**
nghd64 (6300)	200,425	2,761.77	**72.57**	**< 0.0001**
nghd65 (6400)	-3,370.94	1,903.45	-1.77	0.0766
nghd66 (6500)	-173.93	2,100.51	-0.08	0.9340

* The 5% critical value for a two-tailed test with large dataset is 1.96.
** The critical p-value for the significance is 0.05.

The model can be expressed as below:

$$\hat{hprice} = 47{,}049 + 88.6\,frontage + 14.18\,depth - 383.93\,age + 75.97\,sfla - 28{,}364\,raisedranch$$
$$- 13{,}130\,splitlevel - 12{,}924\,capecod - 6{,}594.26\,colonial - 9{,}659.54\,contemporary + 105{,}207\,mansion$$
$$- 1{,}564.6\,oldstyle - 8{,}610.58\,cottage - 3{,}067.71\,loghome - 62{,}279\,duplex - 15{,}270\,townhouse$$
$$+ 45{,}915\,other - 4{,}982.17\,nghd1 - 14{,}670\,nghd2 - 6{,}857.62\,nghd3 - 965.01\,nghd4 + 13{,}181\,nghd5$$
$$+ 2{,}3411\,nghd6 + 77{,}727\,nghd7 + 1{,}8861\,nghd8 + 8{,}607.37\,nghd9 - 2{,}511.45\,nghd10 + 19{,}058\,nghd11$$
$$- 415.66\,nghd12 + 31{,}939\,nghd13 - 5{,}016.49\,nghd14 - 6{,}371.45\,nghd15 + 11{,}196\,nghd16 + 24{,}112\,nghd17$$
$$- 3{,}483.57\,nghd18 + 2{,}756.09\,nghd19 + 7{,}229.47\,nghd20 + 36{,}757\,nghd21 - 2{,}779.54\,nghd22 + 77{,}275\,nghd23$$
$$+ 12{,}099\,nghd24 + 84{,}237\,nghd25 + 170{,}050\,nghd26 + 4{,}201.24\,nghd27 + 6{,}038.39\,nghd28 - 20{,}195\,nghd29$$
$$+ 1{,}039.64\,nghd30 + 43{,}748\,nghd31 - 29{,}292\,nghd32 - 10{,}506\,nghd33 + 22{,}568\,nghd34 - 2{,}790.7\,nghd35$$
$$- 1{,}616.62\,nghd36 - 2{,}223.19\,nghd37 - 28{,}434\,nghd38 - 18{,}628\,nghd39 - 29{,}006\,nghd40 + 6{,}405.03\,nghd41$$
$$+ 64{,}844\,nghd42 - 30{,}576\,nghd43 + 12{,}079\,nghd44 + 2{,}992.88\,nghd45 + 12{,}720\,nghd46 + 177{,}940\,nghd47$$
$$+ 67{,}508\,nghd48 + 108{,}115\,nghd49 + 44{,}499\,nghd50 - 1{,}617.49\,nghd51 + 45{,}773\,nghd52 - 928.08\,nghd53$$
$$- 9{,}804.03\,nghd54 - 16{,}160\,nghd55 + 2{,}172.54\,nghd56 - 14{,}649\,nghd57 + 94{,}106\,nghd58 - 5{,}762.82\,nghd59$$
$$+ 13{,}111\,nghd60 + 12{,}987\,nghd61 + 6{,}587.94\,nghd62 + 215{,}249\,nghd63 + 200{,}425\,nghd64 - 3{,}370.94\,nghd65$$
$$- 173.93\,nghd66$$

$$\boldsymbol{hprice = 47{,}049 + 88.60(frontage) + 14.18(depth) - 383.93(age) + 75.97(sfla)}$$
$$\boldsymbol{\pm\, bstyle_i(X_i) \pm nghd_j(Y_j)}$$

where: $x_1 = -28{,}364$ for bstyle1……, and
$x_{12} = 45{,}915$ for bstyle12, and

$y_1 = -4987.167$ for nghd1,……, and
$y_{66} = -173.93$ for nghd66.

As we can see from the model statistics in Table 4.2, the p-value from the stepwise regression for the overall test that all β_i's = 0 was highly significant ($p < 0.0001$) and the overall model fit was high with the Adjusted $R^2 = 0.9046$. These results were encouraging and illustrate the potential utility for this regression model to predict the housing prices in the Town of Amherst.

TABLE 4.2: Analysis of Variance Model Statistics

	Sum of Squares	Mean Square	F Value	Pr > F*
Model	1.71E+14	2.08E+12	2838.80	< .0001
Error	1.79E+13	7.32E+08		
R-Square	0.9049			
Adj R-Sq	0.9046			

* The critical p-value for the significance is 0.05.

4.3.6 Estimation of the Housing Price

Table 4.1 gives the regression coefficients of the multiple regression model. The variables that were found to be statistically significant are highlighted with bold letters in Table 4.1.

The model establishes the relationships between the estimated housing price and each independent variable found significant. The influence of each of the independent variable on the housing price is briefly described below:

I) Property Physical Structural Attributes

1) *frontage*: the variable, *frontage*, is considered significant at the 5% level. The *frontage* has a positive effect on the price. The price increases by $89 for every additional one foot length of *frontage*.
2) *depth*: the coefficient for *depth* is significant at the 5% level. The *depth* has a positive association with the housing price. For every additional one foot length of depth of the parcel, the price goes up by $14.

3) *age*: the variable, *age*, is negatively related to house price as expected. The expected negative coefficient of *age* reveals that an older home is worth less than a newer home, by $384 per year.

4) *sfla*: the variable, *sfla,* coincides with prior expectation. The coefficient for *sfla* has a positive impact on housing prices. The model concluded that the value increases by $76 per square foot of living area.

5) *bstyle1*, *bstyle2*, *bstyle3*,……, *bstyle12*: for analyzing the effect of the building styles, *ranch* building style was considered as the base group. Any building style coefficients that were insignificant implies that building style's impact on housing price is not statistically different than a *ranch*; three building styles: *old style*, *cottage*, and *log home* fell into this category. Table 4.1 summarizes the regression coefficients of building style dummy variables. The housing prices for *raised ranch*, *split level*, *cape cod*, *colonial*, *contemporary*, *duplex*, and *town house* are $28,364, $13,130, $12,924, $6,594, $9,660, $62,279 and $15,270 lower than that of a *ranch*, respectively. And, the building style for *mansion* and *others* are predicted to assess for $105,207 and $45,915 more, respectively. The building style of *duplex* is the least desirable due to lack of privacy since the *duplex* house comprises two units. While the building style of *mansion* is the most expensive style, which typically is referred to luxury houses with many bedrooms and bathrooms.

II) Locational Characteristics

1) Locational variables: *nghd1*, *nghd2*, *nghd3*,……, and *nghd66*

For the locational variables, the base neighborhood code was 100. The coefficients of the locational variables are significant except for 18 of the 67 neighborhoods which were: *nghd1*, *nghd4*, *nghd10*, *nghd12*, *nghd14*, *nghd18*, *nghd19*, *nghd22*, *nghd30*, *nghd35*, *nghd36*, *nghd37*,

nghd45, *nghd51*, *nghd53*, *nghd56*, *nghd65*, and *nghd66*. Table 4 indicates that the estimated prices are significantly neighborhood sensitive. The range of coefficients ranges from $-29,292 to $215,249. There are large differences in the estimated prices from neighborhood to neighborhood due to the location effect. The neighborhood codes 2500, 4600, 6200, and 6300 have the highest housing prices. When looking carefully into these neighborhoods, the housing price increases for houses that have better structural attributes such as larger square foot of the living area and less age. On the other hand, the neighborhood with code 3100 has the lowest housing prices. This research provides evidence that the location of a house, or its neighborhood code, has the largest impact on its price.

4.3.7 Model Prediction Performance

The prediction accuracy of the model was an important goal of this research, therefore 20% of the data was utilized for model validation. Three widely accepted measures were utilized: (i) the root mean squared error (RMSE), (ii) mean absolute error (MAE), and (iii) Theil's U statistic were applied to the 80% model data and the 20% validation group of houses to analyze the prediction performance.

The root mean square error (RMSE) is the square root of the average of the squared values of the prediction errors and weights large errors more heavily than small errors. The root mean squared error (RMSE) is defined as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t}\left(y_t - \hat{y}_t\right)^2},$$

where y_t is the actual housing price, and $\hat{Y}_t$ is the fitted price from the regression model.

The mean absolute error (MAE) is the average of the absolute values of the prediction errors and is given by:

$$MAE = \frac{1}{n}\sum_{t}\left|y_t - \hat{y}_t\right|.$$

Theil's U statistic is the square root of the ratio of the mean squared error of the predicted change to the average squared actual change. It is defined as:

$$U = \sqrt{\frac{\sum_{t}\left(y_t - \hat{y}_t\right)^2}{\sum_{t}(y_t)^2}}.$$

For all three criteria values closer to 0 indicate better fitting models. In addition, model validity can be considered by examining differences between the respective goodness-of-fit statistics.

Table 4.3 gives the results of model prediction performance of the training set versus the validation set. Table 4.3 compares the three accuracy measures. The validation set's MAE is 3% higher than the MAE of the training set; the RMSE is 8% higher, and the Theil's U statistic is 6% higher. The relative difference between the model and validation set in terms of the respective goodness-of-fit statistics is relatively minimal, validating the accuracy of the model. Also, the percent data within 10% in the validation set computes to 2% less than the training set; and the percent data within 20% is almost equal in both sets. The above measures of predicting the performance of the model developed in this research indicate that the model is highly accurate and the model can be used for mass appraisal of the residential properties in the example town.

TABLE 4.3: Model Prediction Performances

Measure of Accuracy	Training Set	Validation Set	% Difference
MAE	15,504	15,997	+3%
RMSE	27,027	29,284	+8%
Theil's U Statisitc	0.132	0.141	+6%
Percent within 10%	70.8%	69.60%	-1.02%
Percent within 20%	91.9%	91.20%	-1.00%

4.4 Spatial Analysis Using ArcGIS

4.4.1 General Location Description

The Amherst Town is located on the northern part of Erie County in the western part of the New York State, USA (Map 4.1). The Town covers an area of 54 square miles and had a population of 116,510 in the 2000 census (USACE, 2005). Also, Amherst Town includes the village of Williamsville with a population of 5,573 in the 2000 census (USACE, 2005). Approximately, 45 percent of the total land area is developed for the purpose of residential use. In June 2009, there were an estimated 33,342 single-family residences in the Amherst Town.

4.4.2 Neighborhood Analysis

Since the number of disputes about the assessed housing prices is rising, it would be convenient for tax assessment officials to recognize and prove to complainants the accuracy of assessment by incorporating multiple regression model with the ArcGIS spatial analysis. Table 4.5 provides a distribution of the 26,674 house prices estimated by the model, and compares with their actual 2009 assessed values. The estimated values have been categorized as: (i) Over Estimate, (ii) Fair Estimate, and (iii) Under Estimate. Table 4.4 defines the criteria used for categorizing the error classification. Estimated values that are more than one (1) standard deviation from the mean

have been categorized as Over Estimate, and values below one (1) standard deviation have been categorized as Under Estimate, and all of the middle values have been categorized as Fair Estimate. Table 4.5 in columns 5, 6, and 7 gives the percent Over Estimate, Fair Estimate, and Under Estimate for each neighborhood with its average housing price per square foot of living area (*sfla*).

TABLE 4.4: Criteria for Categorizing Error Classification

Standard Error of Estimate	Classification
> 1 SE	Over Estimate (OE)
-1 SE – 1 SE	Fair Estimate (FE)
< -1 SE	Under Estimate (UE)

SE: Standard Error

TABLE 4.5: Proportion of Over Estimate, Fair Estimate, and Under Estimate Observations for Each Neighborhood

nghdcode	Number of Houses	Number of Sale Transactions (2008)	$^{hprice}/_{sfla}$ ($)	Over Estimate	Fair Estimate	Under Estimate
1	2	3	4	5	6	7
100	496	9	76	13.00%	72.87%	14.13%
200	230	7	78.17	14.61%	74.16%	11.24%
300	127	0	69.7	13.58%	71.60%	14.81%
400	1532	19	78.54	2.72%	94.01%	3.27%
500	1,089	14	79.58	7.88%	87.33%	4.79%
600	176	8	89.85	6.96%	89.24%	3.80%
700	345	11	76.43	8.92%	85.41%	5.68%
701	311	9	99.35	21.05%	50.24%	28.71%
800	320	4	85.63	3.15%	96.53%	0.32%
900	279	11	83.59	3.27%	95.27%	1.45%
1000	556	22	78.78	3.22%	94.32%	2.46%
1100	472	22	94.43	3.70%	94.55%	1.74%
1200	1,115	23	81.99	1.67%	95.78%	2.55%
1300	102	1	88.02	9.80%	82.35%	7.84%
1400	172	6	77.78	0.00%	98.14%	1.86%
1401	134	6	75.83	1.64%	93.44%	4.92%
1500	341	10	83.35	1.53%	96.17%	2.30%
1600	175	5	93.45	19.77%	65.70%	14.53%
1700	176	7	75.25	10.23%	76.14%	13.64%
1800	415	13	79.59	2.96%	93.84%	3.20%
1900	277	10	83.13	3.65%	92.70%	3.65%
2000	572	19	92.61	11.34%	82.90%	5.76%
2100	301	14	82	6.44%	89.77%	3.79%
2200	241	7	99.27	21.52%	59.49%	18.99%
2300	496	14	83.61	7.50%	83.75%	8.75%
2400	155	9	94.8	25.15%	47.95%	26.90%
2500	195	4	115.19	29.93%	29.93%	40.15%
2600	427	12	73.08	6.47%	87.31%	6.22%
2700	255	6	82.77	4.12%	93.00%	2.88%
2800	1,312	35	72	3.19%	93.55%	3.27%
2900	438	16	76.41	3.92%	90.55%	5.53%
3000	254	4	91.58	5.33%	87.56%	7.11%
3100	454	15	65.14	1.65%	94.81%	3.54%
3200	1,749	49	71.67	2.03%	93.39%	4.58%
3300	13	0	76.89	15.38%	84.62%	0.00%

TABLE 4.5: - continued

nghdcode	Number of Houses	Number of Sale Transactions (2008)	*hprice* / *sfla* ($)	Over Estimate	Fair Estimate	Under Estimate
1	2	3	4	5	6	7
3400	551	19	78.23	7.91%	84.18%	7.91%
3500	649	18	74.56	1.86%	97.52%	0.62%
3600	1,317	46	76.93	3.21%	93.20%	3.60%
3700	1,657	44	65.32	1.66%	94.65%	3.69%
3800	1,748	49	70.63	3.03%	93.40%	3.57%
3900	345	11	54.89	0.91%	94.83%	4.26%
4000	386	10	84.07	7.82%	87.33%	4.85%
4100	68	1	97.3	30.88%	39.71%	29.41%
4200	133	4	61.44	1.50%	92.48%	6.02%
4300	609	25	78.9	9.22%	78.77%	12.01%
4400	594	22	74.59	10.15%	78.66%	11.19%
4500	84	2	66.5	17.19%	62.50%	20.31%
4600	46	1	107.05	41.30%	23.91%	34.78%
4700	155	4	90.79	35.95%	30.07%	33.99%
4800	117	3	106.48	41.38%	20.69%	37.93%
4900	560	23	87.58	21.96%	48.95%	29.09%
5000	452	9	70.36	3.90%	89.42%	6.69%
5100	378	7	86.07	14.51%	68.87%	16.62%
5200	1,933	64	71.18	6.32%	86.89%	6.78%
5300	359	16	76.2	1.11%	97.21%	1.67%
5400	694	12	70.49	3.53%	90.25%	6.22%
5500	87	5	68.8	9.30%	82.56%	8.14%
5600	712	19	73.32	4.13%	90.79%	5.08%
5700	13	0	109.74	15.38%	76.92%	7.69%
5800	1,078	39	73.64	4.20%	90.57%	5.23%
5900	432	8	81.45	12.28%	79.64%	8.08%
6000	540	12	81.74	10.48%	79.52%	10.00%
6100	672	25	78.71	7.59%	85.95%	6.46%
6200	4	0	118.1	25.00%	0.00%	75.00%
6300	239	9	132.74	29.69%	11.98%	58.33%
6400	629	13	76.8	0.65%	96.60%	2.76%
6500	399	7	79.38	1.28%	93.88%	4.85%

Table 4.5 statistic revealed that neighborhoods: 2400, 2500, 4100, 4600, 4700, 4800, 6200, and 6300 have relatively higher proportion on both overestimates and underestimates. That indicates that the prediction accuracy of the multiple regression model is relatively weak for those neighborhoods. On further investigation, it was found that the average housing prices per square foot of living area for all of these neighborhoods have higher housing prices per square foot of living area. On sorting of the Table 3-6 data by $^{hprice}/_{sfla}$ showed that homes costing more than $88.00 per square foot of living area (*sfla*) had much higher percentage of Over Estimate and Under Estimate, while those costing below $88.00 per square foot of living area were mostly in the Fair Estimate region. Figures 4.2, 4.3, and 4.4 have illustrated the Table 4.5 data into three separate figures. Figures 4.2, 4.3, and 4.4 show the percentage of estimate error classifications against the average housing prices per square foot of living area for each neighborhood. These figures also pictorially show that the model estimates have higher accuracy for houses costing less than $88.00 per square foot of living area.

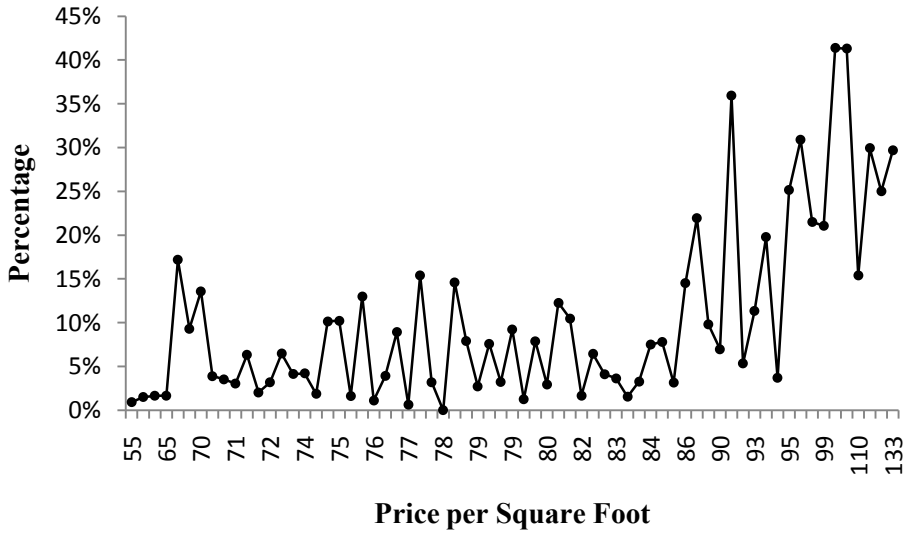


FIGURE 4.2: Over Estimate Error vs. Price per Square Foot

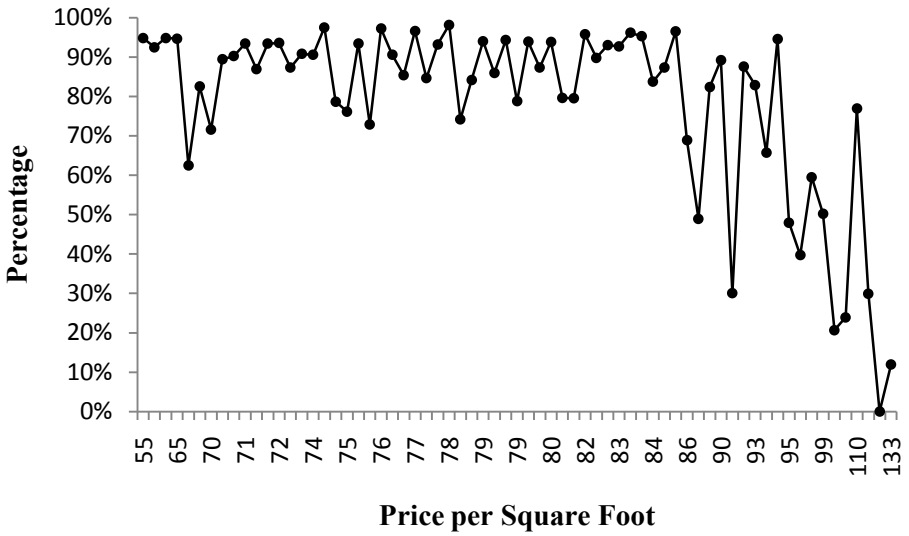


FIGURE 4.3: Predicted Fair Estimate vs. Price per Square Foot

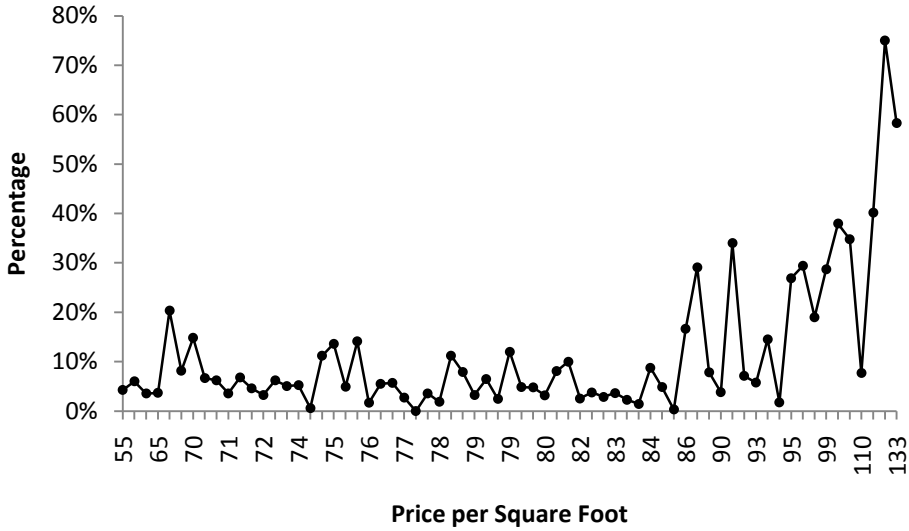


FIGURE 4.4: Under Estimate Error vs. Price per Square Foot

4.4.3 Spatial Housing Price Patterns

A series of maps showing the attributes of each neighborhood as found in the regression model were created by ArcGIS software. Map 4.2 shows the average assessed housing prices for each

neighborhood, the highest priced homes have a red color mark. The highest housing prices are indicated in the eastern, northern, and southwestern area of the Town.

Map 4.3 shows the proportion of fair estimate for each neighborhood. A large majority of the neighborhoods fall into the Fair Estimate category, leading to the conclusion that most of the assessed housing prices can be accurately forecasted from the multiple regression model. Map 4.4 illustrates the proportion of overestimates for each neighborhood. The neighborhoods of higher percentage of overestimates are indicated in the eastern, northern, and southwestern area of the Town, almost all in the high priced neighborhoods. Map 4.5 presents the proportion of underestimates for each neighborhood. The spatial result shows that the most underestimated houses are located on the eastern boundary of Amherst Town. In comparing Map 4.2, Map 4.4, and Map 4.5, the prediction accuracy for the higher-priced homes tends to be lower. This result supports the early findings from Table 4.5. The reason for this result is that the method of least squares used in multiple regression represents the "averaging" behavior or "central" tendency of a distribution, the tail behaviors of that distribution are therefore hard to recognize.

Another reason is the lack of enough number of sale transactions in some neighborhoods, especially for higher pricing neighborhoods. Usually, when assessing the housing prices, the tax assessment officials will compare the target house with the sale history of a particular neighborhood. Table 4.5, in column 3 gives the number of 2008 sale transactions for each neighborhood. The total number of sale transactions is 948. But, the total number of sales that are transacted in the neighborhood 2400, 2500, 4100, 4600, 4700, 4800, 6200, and 6300 is only 31. It counts only 3.3% of the total number of sale transactions.

Furthermore, the higher housing price homes have a personal social status factor. It is difficult to quantify this factor and include in the model development. Therefore, the prediction accuracy for higher priced homes is relatively lower.

4.5 Summary of Multiple Regression Model Analysis Results

The research presented in this dissertation has successfully developed a multiple regression model and linked it to the map layer for the Town of Amherst, New York. Utilizing GIS methodology, various maps showing the spatial distribution of housing price and estimation errors by neighborhood have been generated. Such visual patterns will enable the town assessors and appraisal companies to assess the value of homes in a neighborhood, on a uniform basis. This will also help the town planners in evaluating future development of their town. These methods are easily generalized to towns of a similar type.

The estimation model had a high accuracy with an R^2 value of 0.9046. The estimation errors of the validation set had minor differences with the model estimation errors. Also, a large majority of the houses were found to be within the *Fair Estimation* group, which was defined to be within ± 1 standard error of the residual distribution. On viewing the pattern of the *Over Estimate* and *Under Estimate* neighborhoods, it was found that higher priced homes had a lower estimation accuracy. In Town of Amherst used in this research, houses costing more than $88.00 per square foot of living area fell in this category.

The following conclusions can be drawn from the research presented in this paper.

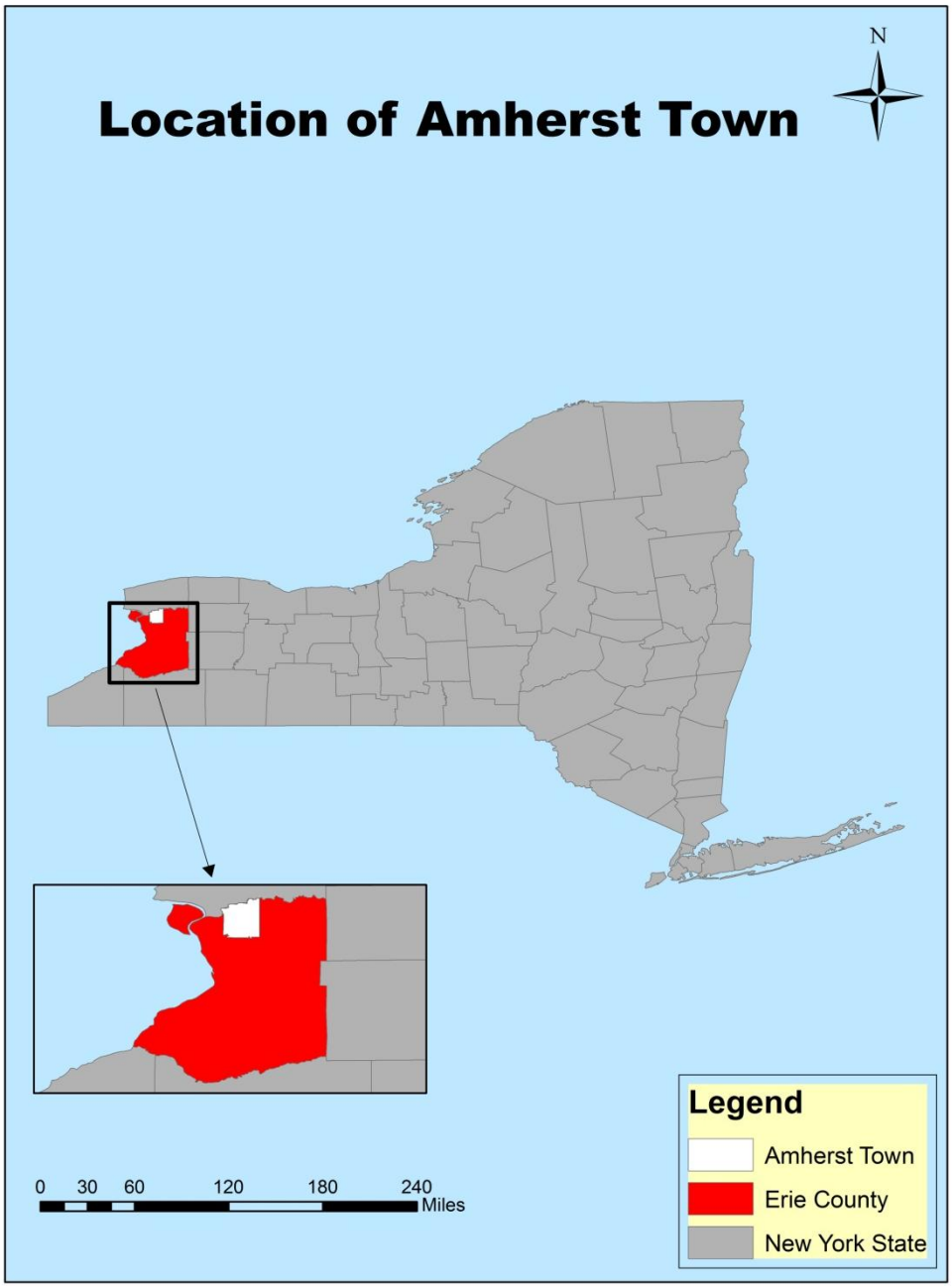
(i) A multiple regression model can be developed for a town, village, or a city with high accuracy and can be linked to the map layers of the town using GIS technology for viewing spatial patterns based on housing price, housing age, etc..

(ii) While determining the functional form of the dependent variable, it was found that a linear form gives as accurate results as a semi-logarithmic or a root transformation form.

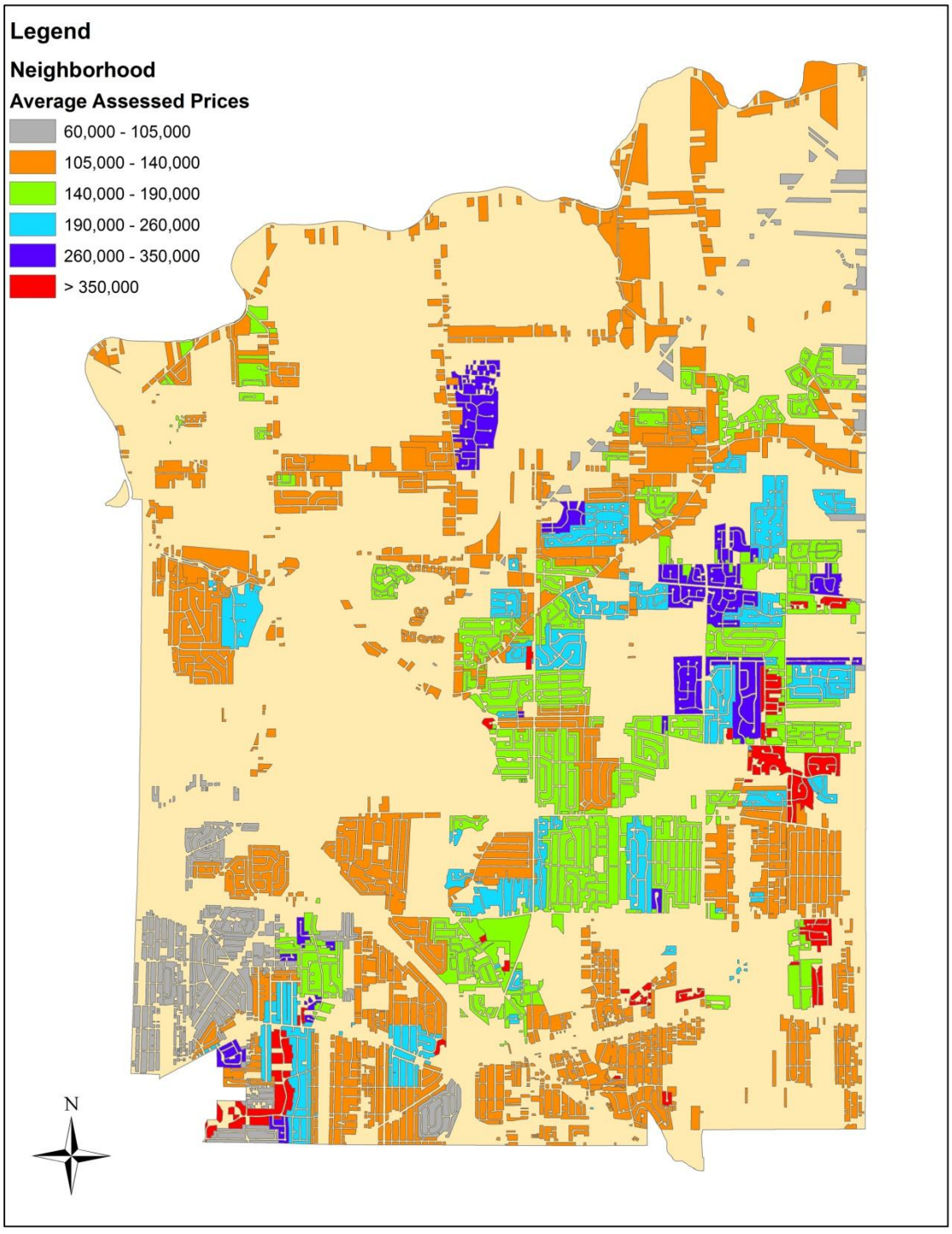
(iii) The factors that significantly effect the price of a house include: frontage width, depth of the parcel, age, square foot of living area, architectural building style, and the neighborhood. The neighborhood or location of a house has the maximum impact on its price.

(iv) The higher priced houses have a lower prediction accuracy within the context of our model. In this research, houses costing higher than $88.00 per square foot of living area were found to fall in the higher prices houses category.

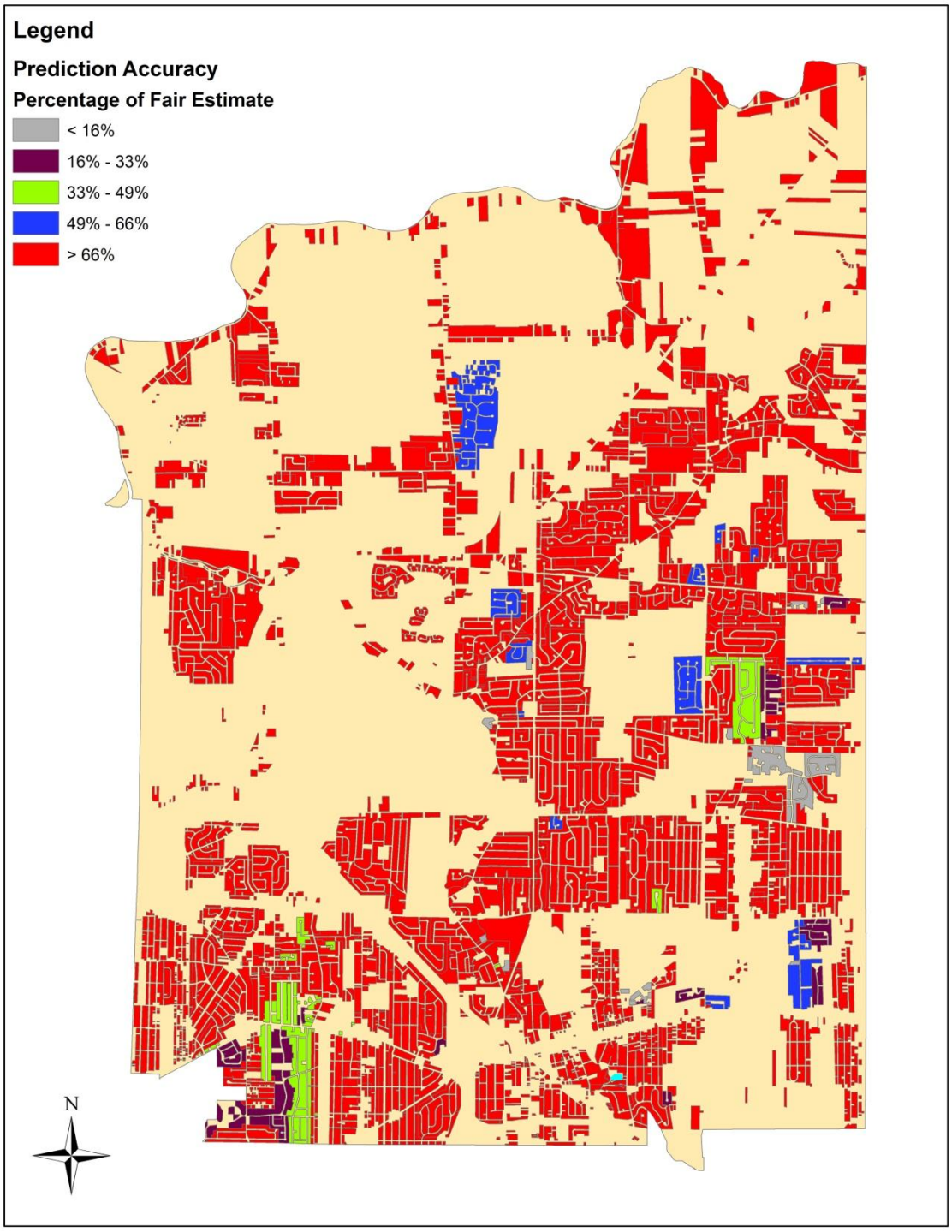
(v) This research has demonstrated the novel and beneficial uses of GIS technology in housing price estimations. Visual patterns of the distribution of prices of houses, or age of houses, or the square foot of living area, by neighborhood will facilitate in any dispute resolutions due to inequity in assessments.



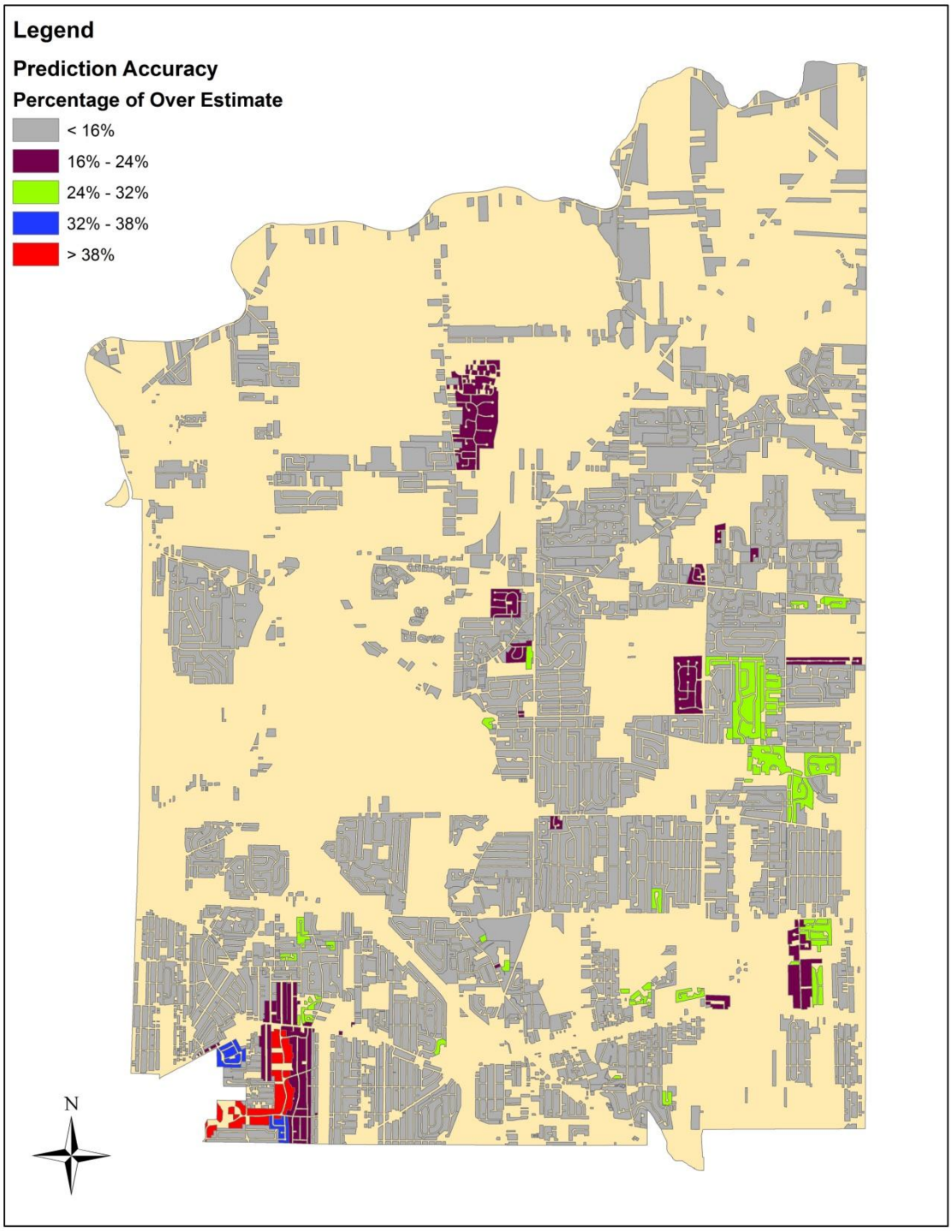
MAP 4.1: Location of Amherst Town in the State of New York



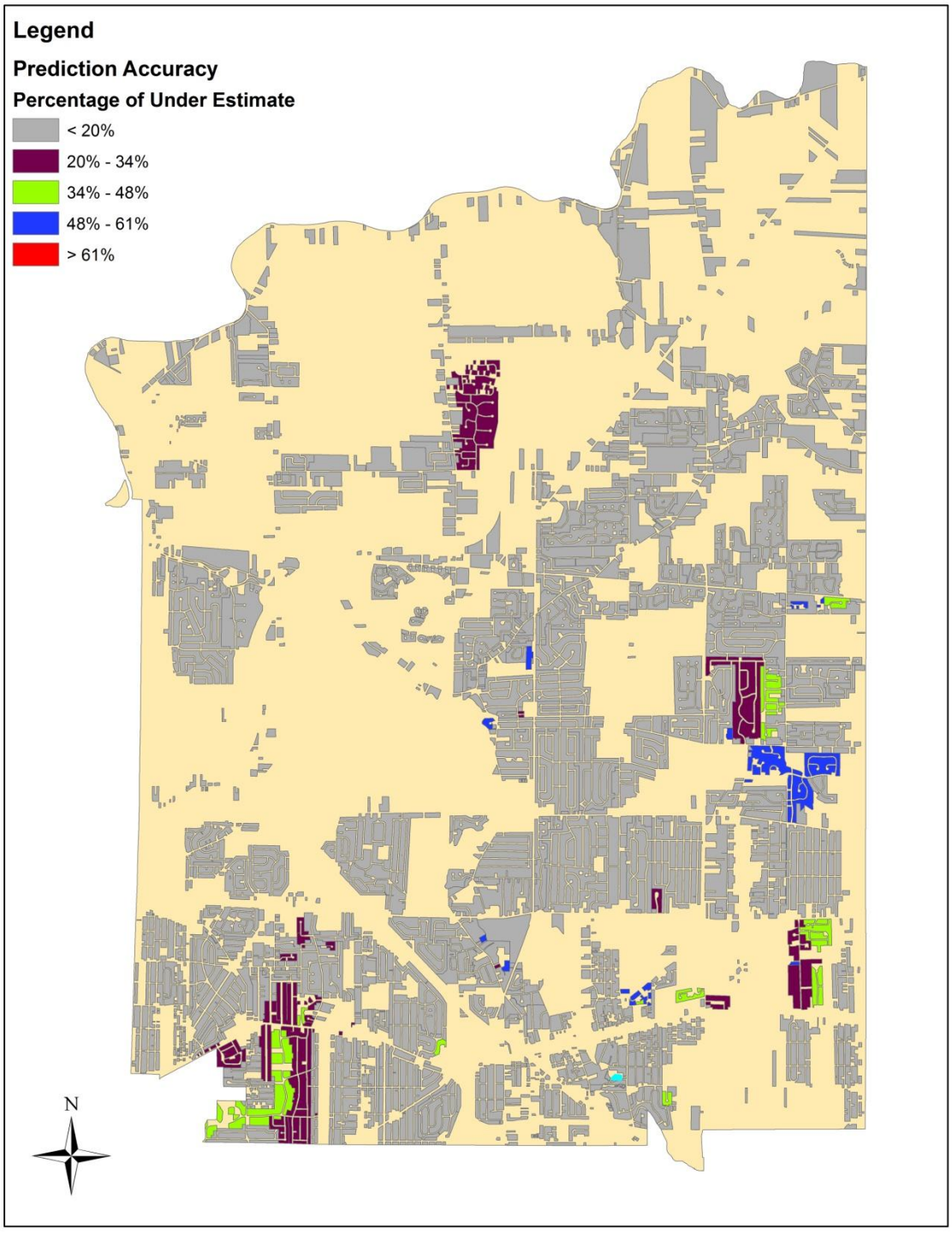
MAP 4.2: Spatial Distribution of Average Assessed Housing Price ($)



MAP 4.3: Distribution of Fair Estimation for Housing Prices



MAP 4.4: Distribution of Overestimation for Housing Prices



MAP 4.5: Distribution of Underestimation for Housing Prices

CHAPTER 5: Additive Nonparametric Regression (ANR) Model for Estimation of Property Values

5.1 Introduction

The chapter 5 presents the application of an additive nonparametric regression (ANR) model for housing price estimation, and also links it to the GIS map layer of a municipality to draw various maps showing the wide variation in the prices of homes based on location or neighborhood. The additive nonparametric regression (ANR) model was developed using the 2009 housing assessment data of 33,342 houses of a town in Western New York.

This research was motivated by the current trend towards the wide application of statistical technique in the mass appraise of residential real estate. This chapter takes a new approach and differs in a number of respects from the earlier studies. In the first place, ArcGIS software Version 9.3 was applied on the results from the additive nonparametric regression (ANR) to analyze and visualize the spatial variations in the housing sub-markets. The housing dataset for the Town of Amherst, State of New York, which was segmented into 67 neighborhoods, was used to develop and clearly demonstrate the spatial patterns. Secondly, tables, graphs, and maps of prediction accuracy were created based on the percentage of the estimated prices within 10% and 20% of the actual assessed values for each neighborhood. These tables, graphs and maps can assist a town's property tax assessment officers in housing price estimation and comparisons within the sub-markets. The major focus of the research described in this chapter is to develop a comprehensive explanatory model of housing price determination using additive nonparametric regression (ANR) and also visualize the price variation patterns of different neighborhoods using GIS. The model developed in this research also gives several numerical summaries about the example housing market.

The format of the chapter is constructed as follows. Section 2 introduces the additive nonparametric regression model. In section 3, the model development processes and the estimation using the additive nonparametric regression model are represented. Section 4 illustrates the empirical result of spatial analysis. The final section is to summarize the results of the additive nonparametric regression model analysis.

5.2 Introduction to Additive Nonparametric Regression (ANR) Model

Additive nonparametric regression models were originally suggested by Friedman and Stuetzle (1981) and popularized by Hastie and Tibshirani (1990). The additive nonparametric models are more flexible and utilize a data-driven approach that allows the data points themselves to determine the shape of the fitted curve. The central idea of the additive nonparametric models is to replace the linear function $\beta_j X_j$ by an unspecified non-linear smooth function, to get an equation of the form:

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \qquad (1)$$

where the f_j's denote arbitrary smooth functions whose shapes are unrestricted. In this paper, the f_j's are estimated by local polynomial regression. There are several advantages of an additive nonparametric regression model. First, there is little or no past experience, theory or other source of information available about the relationship between the dependent variable and its independent variables. Second, the additive nonparametric regression model is more flexible in exploring the data. In practice, there are many chances where there is little or no prior specific quantitative information available about the regression curve. In such conditions, the additive nonparametric regression can provide a wide range of functional forms of the regression curve in determining the unknown regression relationship.

This model can also be extended by incorporating linear terms, as in the model

$$Y = \beta_0 + \beta_1 X_1 + \sum_{j=2}^{p} f_j\left(X_j\right) + \varepsilon. \qquad (2)$$

Such semiparametric regression models are particularly useful for including dummy regressors or other contrasts derived from categorical predictors.

5.2.1 Fitting Additive Nonparametric Regression (ANR) Model

The main method used to estimate the additive nonparametric model is the backfitting algorithm.

The backfitting algorithm estimates the function f_j in the model $Y = \beta_0 + \sum_{j=1}^{p} f_j\left(X_j\right) + \varepsilon$, as follows:

I. Initialization: β_0= average (Y), $f_j = f_j^{(0)}$, $j = 1, \ldots, p$

II. Cycle: $f_j = S_j\left(Y - \beta_0 - \sum_{k \neq j} f_k / X_j\right), j = 1, \ldots, p, 1, \ldots, p$

III. Repeat step II until the individual function f_js converge

For each estimate of $f_j\left(x_j\right)$, the backfitting algorithm computes the residuals

$$Y - \beta_0 - \sum_{k \neq j}\left(f_k / X_j\right)$$

and smoothenes them against X_j's.

Here $S_j\left(Y / X_j\right)$ is a scatterplot smoother which generates the regression of the dependent variable Y on the predictors X_j. In this research, local polynomial regression is applied as the smoother. The linear terms in the model are fitted by the least squares method.

5.2.2 Smoothing Techniques: Local Polynomial Regression

A smoothing technique is the tool that smoothes out the rough edges of a dataset, and then constructs the regression function *E(Y/X)*, which is less variable than the observed *Y* itself.

In this paper, the local polynomial regression is utilized as our smoother. Local polynomial regression was proposed by Cleveland, Devlin, and Grosse (1988). Local polynomial regression (Cleveland, 1979) uses weighted least squares (WLS) regression to fit a D^{th} degree polynomial ($D \neq 0$) to data.

In local polynomial regression, weighted least square is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that it contains a specified percentage of the data points. The smoothing parameter *s*, in each local neighborhood controls the smoothness of the estimated data. When $s < 1$, the local neighborhood used at a point *x* contains the *s* fraction of the data points closest to the point *x*. When $s \geq 1$, all data points are used.

Generally, for the majority of cases, a first order fit (local linear regression) is an adequate choice for *D*. Local polynomial linear regression is suggested (Cleveland 1979) in order to balance computational ease with the flexibility to reproduce patterns that exists in the data. Nonetheless, local linear regression may fail to capture sharp curvature if present in the data structure (Mays 1995). In such cases, local quadratic regression (D=2) may be needed to provide an adequate fit. Most authors agree there is usually no need for polynomials of order $D>2$ (Mays 1995). As a result, we have only considered first order local polynomial regression in this research.

5.3 Additive Nonparametric Regression (ANR) Model Development

The additive nonparametric regression model (ANR) was used for developing a model to estimate housing prices for mass appraisal. Figure 5.1 represents the flow chart for the model development process. The following steps were used in formulating the additive nonparametric regression model (ANR):

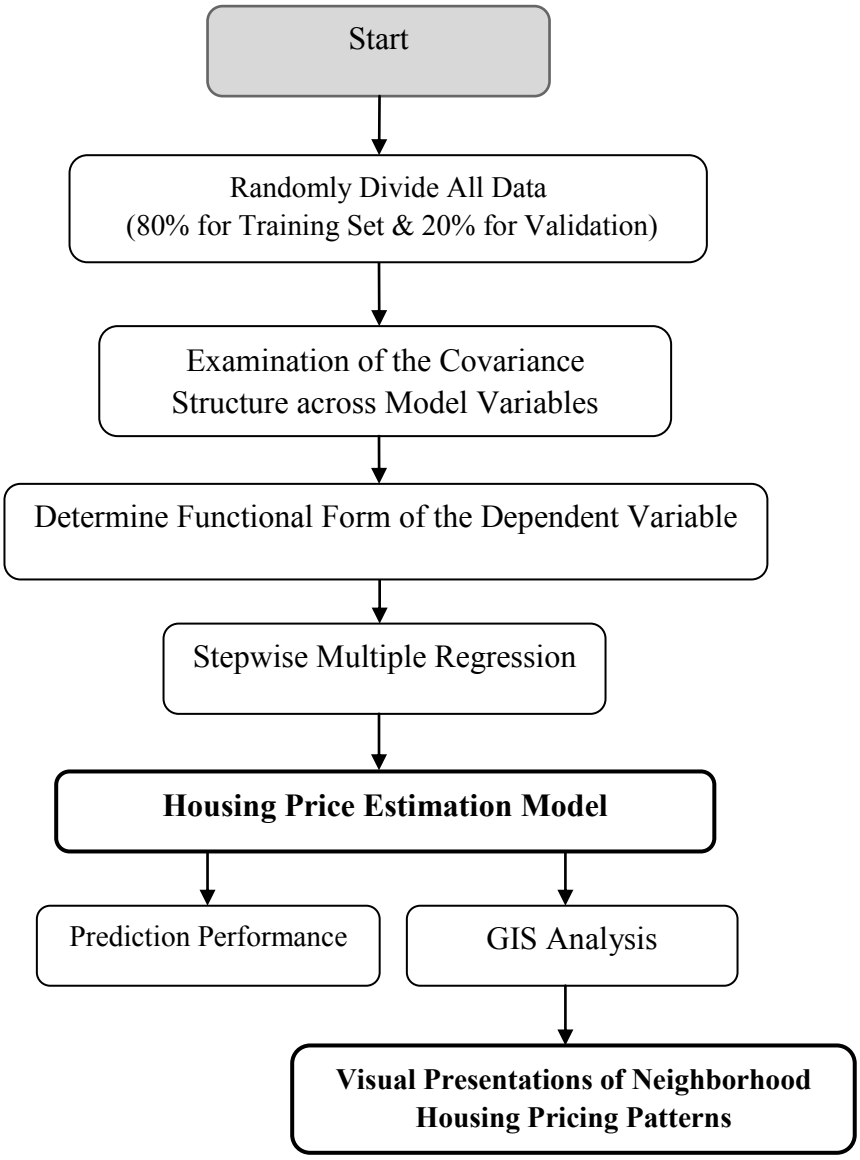


FIGURE 5.1: Model Development Process

5.3.1 Dividing the Dataset into Training Set and Validation Set

Refer to 3.3.2.

5.3.2 Examination of the Covariance Structure across Model Variables

Refer to 3.3.3.

5.3.3 Determination of the Functional Form of the Dependent Variable *hprice*

Refer to 3.3.4.

5.3.4 Stepwise Multiple Regression

Refer to 3.3.5.

5.3.5 Additive Nonparametric Regression (ANR) Model

The following form of the additive nonparametric regression model popularized by Hastie and Tibshirani (1990) was used:

$$\boldsymbol{hprice} = \beta_0 + f_1(\boldsymbol{frontage}) + f_2(\boldsymbol{depth}) + f_3(\boldsymbol{age}) + f_4(\boldsymbol{sfla}) + \sum_{i=1}^{12} \boldsymbol{bstylei} + \sum_{j=1}^{66} \boldsymbol{nghdj} + \varepsilon, \quad (1)$$

where, $E(\varepsilon) = 0$, and $Var(\varepsilon) = \sigma^2$.

This additive nonparametric regression model utilized the same variables as in the stepwise multiple regression model. Note that some linear functions are replaced by the unspecified functions. The unknown functions are traditionally composed of continuous variables. So, the variables: *frontage*, *depth*, *age*, and *sfla* are in the form of the unspecified functions: *f(frontage)*, *f(depth)*, *f(age)* and *f(sfla)*, respectively.

The key idea of the additive nonparametric regression model (ANR) is to replace the usual linear function with an unknown function. The specification assumes that the effect of individual attribute is additively separable. The functions *f(x)* that appears in the above equation were

estimated using the iterative procedure known as the backfitting algorithm in combination with the local polynomial regression.

5.3.6 Estimation of the Housing Price

In this research, the functions *m(x)* that appear in the above equation are estimated using the iterative procedure known as the backfitting algorithm in combination with the local polynomial regression.

Figure 5.2 through 5.5 illustrates the graphical representation of the estimated regression functions: *f(frontage)*, *f(depth)*, *f(age)* and *f(sfla)*. These estimated functions can be considered as the analogues of the coefficients in the multiple regression model. These figures describe the contribution of various physical structure characteristics to the housing prices for the data range where most of the observations fall.

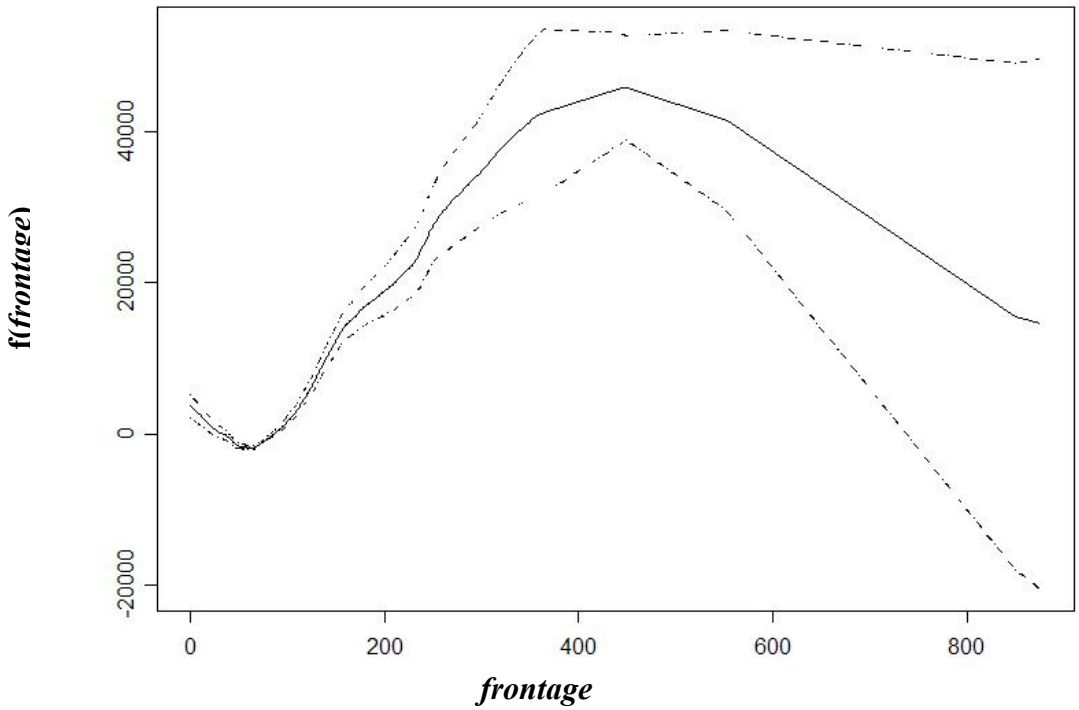


FIGURE 5.2: Estimated Regression Function for *frontage*

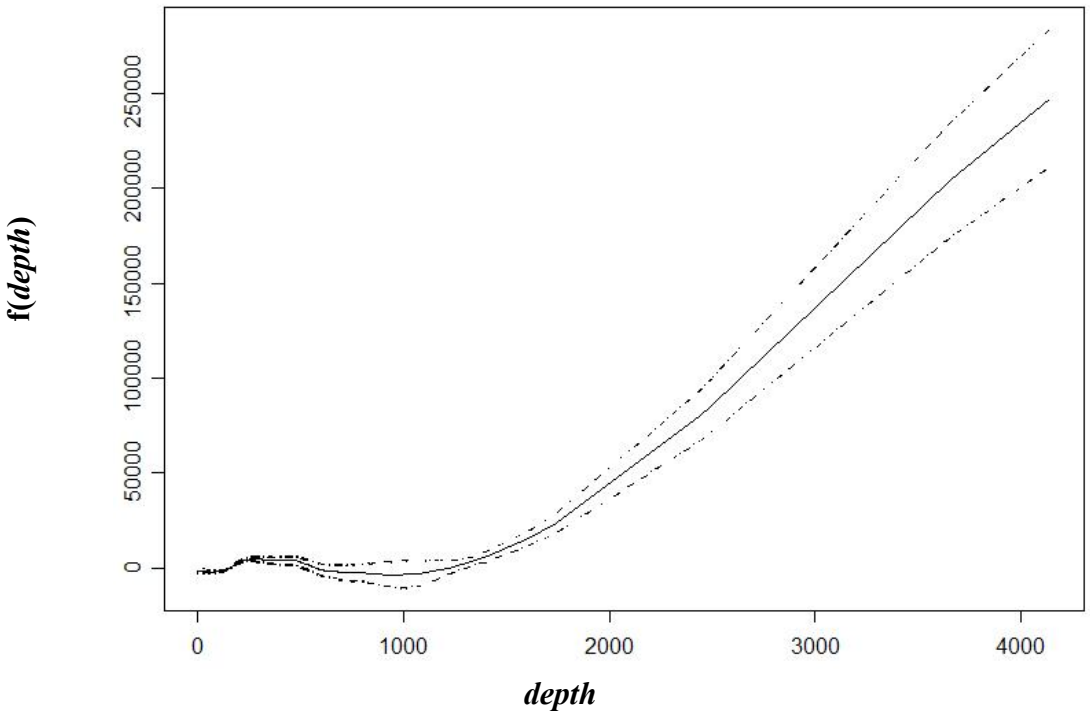


FIGURE 5.3: Estimated Regression Function for *depth*

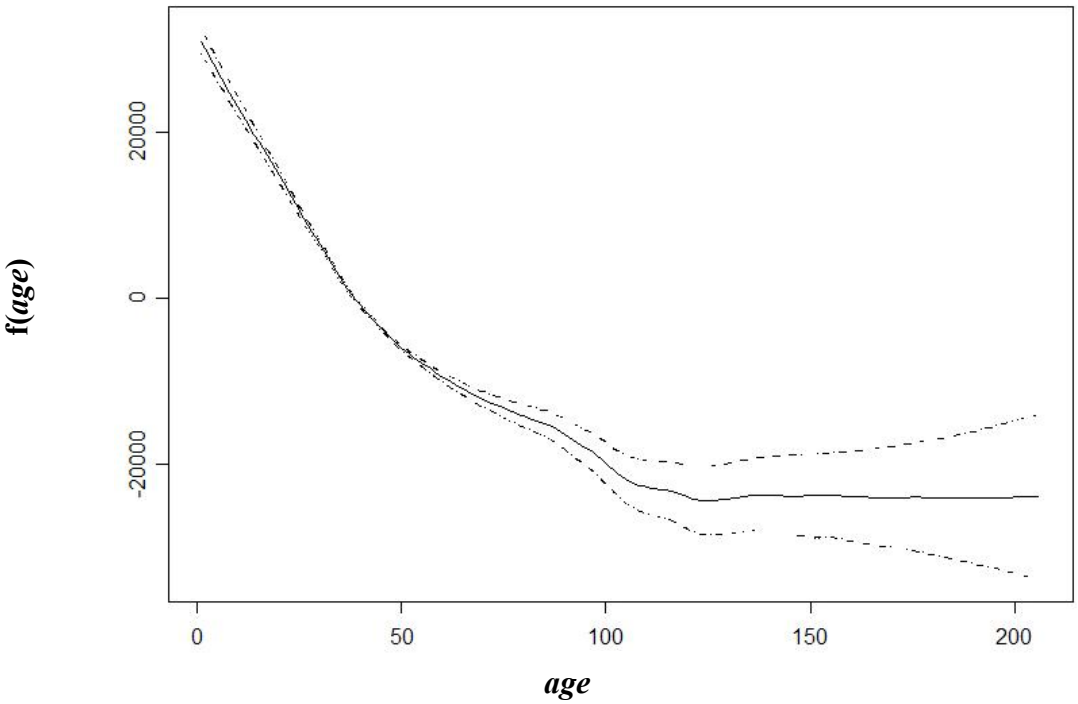


FIGURE 5.4: Estimated Regression Function for *age*

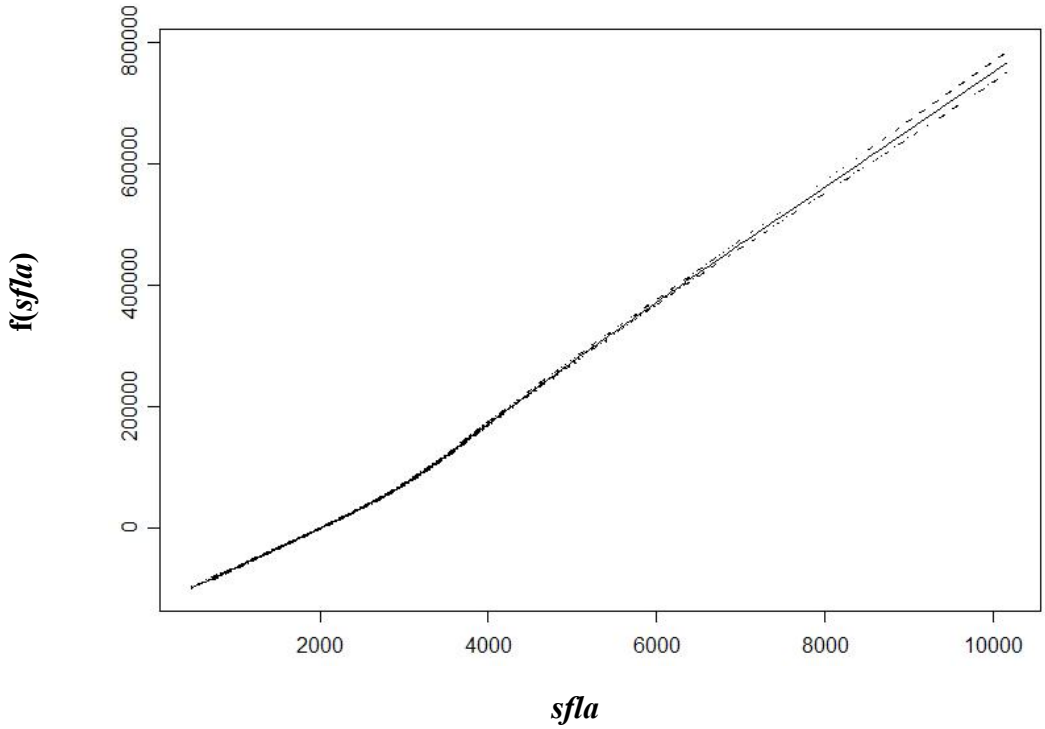


FIGURE 5.5: Estimated Regression Function for *sfla*

1) *frontage*: the plot of *frontage* from Figure 5.2 shows the relationship between frontage and housing prices. The dotted lines represent the 95% confidence intervals. The relationship between frontage and housing prices is not linear. The housing prices decreases first as the length of frontage increases in the range (20 - 60). This length of frontage is considered as the standard length for the parcel of many U.S. residences. So, when estimating the additive nonparametric regression model, this group of houses is regarded as the base group. Then, the housing prices start increasing. The rate of increase declines after above 400*ft*, and continues upon 500*ft*. After the length of frontage is 500*ft*, there is a drop in the housing prices. Obviously, after the frontage becomes larger than 500*ft*, the housing falls in the farming or rural category which has lower price ranges.
2) *depth*: the estimated function for the depth of parcel from Figure 5.3 also clearly illustrates the advantage of using the additive nonparametric regression model. In the beginning, the depth has no impact, especially in the range between 20 *ft* and 1000*ft*. For values above 1000 *ft,* the depth has a positive influence in determining the housing prices. It seems reasonable to expect the positive association between the length of depth and the housing prices due to the more privacy and quieter environment. .
3) *age*: Figure 5.4 shows that the housing age has a negative relationship with the property value. The downward-slope fitted function for the housing age implies that the house value declines as the house gets older. There is depreciation in the house values until the housing age reached 100 years. After 100 years, the effect of age on the housing prices levels out.
4) *sfla*: the estimated regression for the influence of the total living area on the housing price has very expected shape. Figure 5.5 indicates that the square footage of living area has a positive impact on the property value. The upward-slope fitted function implies that property

value increases as the square footage of living area increases. In the range 500 to 6000, where most of the data point fall. The fit function is increasing and nearly linear. The estimated regression is slightly convex in the beginning, implying a varying stronger price effect as the house becomes larger.

Table 5.1 gives the regression coefficients of the additive nonparametric regression model for the parametric portion. The model establishes the relationships between the estimated housing price and each independent variable found significant. The influence of each of the independent variables on the housing price is briefly described below:

TABLE 5.1: Estimate of the Additive Nonparametric Regression Model for the Parametric Part

Variable	Regression Coefficient Estimate	Variable	Regression Coefficient Estimate
Dependent Variable: *hprice*			
Independent Variables:			
intercept	180,179.90	*nghd28* (2700)	11,097.40
bstyle1 (raised ranch)	-26,121.74	*nghd29* (2800)	-8,151.44
bstyle2 (split level)	-9,225.02	*nghd30* (2900)	11,535.77
bstyle3 (cape cod)	-11,656.11	*nghd31* (3000)	45,383.59
bstyle4 (colonial)	-4,458.50	*nghd32* (3100)	-23,879.95
bstyle5 (contemporary)	-8,417.76	*nghd33* (3200)	-1,494.65
bstyle6 (mansion)	61,316.28	*nghd34* (3300)	25,603.11
bstyle7 (old style)	-2,564.17	*nghd35* (3400)	7,027.91
bstyle8 (cottage)	-17,295.79	*nghd36* (3500)	8,746.04
bstyle9 (log home)	-15,459.93	*nghd37* (3600)	5,318.65
bstyle10 (duplex)	-55,983.46	*nghd38* (3700)	-22,065.61
bstyle11 (town house)	-13,089.71	*nghd39* (3800)	-12,909.59
bstyle12 (other)	40,752.81	*nghd40* (3900)	-24,930.60
nghd1 (200)	-2,806.77	*nghd41* (4000)	15,721.87
nghd2 (300)	-13,144.53	*nghd42* (4100)	67,352.52
nghd3 (400)	-1,218.58	*nghd43* (4200)	-23,574.84
nghd4 (500)	1,531.15	*nghd44* (4300)	20,790.84
nghd5 (600)	10,313.81	*nghd45* (4400)	11,638.50
nghd6 (700)	28,246.91	*nghd46* (4500)	19,183.29
nghd7 (701)	76,623.60	*nghd47* (4600)	155,589.60
nghd8 (800)	22,796.87	*nghd48* (4700)	74,978.58
nghd9 (900)	10,637.03	*nghd49* (4800)	103,321.90
nghd10 (1000)	5,254.30	*nghd50* (4900)	55,029.85
nghd11 (1100)	21,569.98	*nghd51* (5000)	6,522.11
nghd12 (1200)	6,691.21	*nghd52* (5100)	57,525.55
nghd13 (1300)	3,536.64	*nghd53* (5200)	7,545.40
nghd14 (1400)	-1,708.82	*nghd54* (5300)	-3,100.19
nghd15 (1401)	-1,636.79	*nghd55* (5400)	-8,044.32
nghd16 (1500)	15,368.78	*nghd56* (5500)	10,118.35
nghd17 (1600)	28,587.60	*nghd57* (5600)	-9,016.34
nghd18 (1700)	750.87	*nghd58* (5700)	96,319.92
nghd19 (1800)	10,686.80	*nghd59* (5800)	4,975.51
nghd20 (1900)	14,585.56	*nghd60* (5900)	18,908.63
nghd21 (2000)	39,588.76	*nghd61* (6000)	18,098.37
nghd22 (2100)	2,508.82	*nghd62* (6100)	15,608.23
nghd23 (2200)	71,454.67	*nghd63* (6200)	203,782.20
nghd24 (2300)	19,605.34	*nghd64* (6300)	190,552.10
nghd25 (2400)	83,021.82	*nghd65* (6400)	6,102.50
nghd26 (2500)	157,298.70	*nghd66* (6500)	7,454.85
nghd27 (2600)	13,931.73		

1) *bstyle1*, *bstyle2*, *bstyle3*,……, *bstyle12*: for analyzing the effect of the building styles, *ranch* building style was considered as the base group. Table 5-1 summarizes the regression coefficients of building style dummy variables. The housing prices for *raised ranch*, *split level*, *cape cod*, *colonial*, *contemporary*, *duplex*, and *town house* are $26,122, $9,225, $11,656, $4,459, $8,418, $55,983 and $13,090 lower than that of a *ranch*, respectively. And, the building style for *mansion* and *others* are predicted to assess for $61,316 and $40,753 more, respectively. The building style of *duplex* is the least desirable due to lack of privacy since the *duplex* house comprises two units. While the building style of *mansion* is the most expensive style, which typically is referred to luxury houses with many bedrooms and bathrooms.

2) Locational variables: *nghd1*, *nghd2*, *nghd3*,……, and *nghd66*

   For the locational variables, the base neighborhood code was 100. Table 5-1 indicates that the estimated prices are significantly neighborhood sensitive. The range of coefficients ranges from $-24,931 to $203,782. There are large differences in the estimated prices from neighborhood to neighborhood due to the location effect. The neighborhood codes 2500, 4600, 6200, and 6300 have the highest housing prices. When looking carefully into these neighborhoods, the housing price increases for houses that have better structural attributes such as larger square foot of the living area and less age. On the other hand, the neighborhood with code 3900 has the lowest housing prices. This research provides evidence that the location of a house, or its neighborhood code, has the largest impact on the price of a home.

5.3.7 Model Prediction Performance

The prediction accuracy of the model was an important goal of this research; therefore 20% of the data was utilized for model validation. Three widely accepted measures were utilized to measure the prediction accuracy of the model:

(i) the root mean squared error (RMSE);
(ii) mean absolute error (MAE); and
(iii) Theil's U statistic.

These there statistics were applied to the 80% model data and the 20% validation group of houses to analyze the prediction performance.

The root mean square error (RMSE) is the square root of the average of the squared values of the prediction errors and weights large errors more heavily than small errors. The root mean squared error (RMSE) is defined as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t}\left(y_t - \hat{y}_t\right)^2}$$

where y_t is the actual housing price, and $\hat{Y}_t$ is the fitted price from the regression model.

The mean absolute error (MAE) is the average of the absolute values of the prediction errors and is given by:

$$MAE = \frac{1}{n}\sum_{t}\left|y_t - \hat{y}_t\right|$$

Theil's U statistic is the square root of the ratio of the mean squared error of the predicted change to the average squared actual change. It is defined as:

$$U = \sqrt{\frac{\sum_t \left(y_t - \hat{y}_t \right)^2}{\sum_t (y_t)^2}}$$

For all three criteria values closer to 0 indicate better fitting models. In addition, model validity can be considered by examining differences between the respective goodness-of-fit statistics.

Table 5.2 compares the results of model prediction performance of the training set versus the validation set. The validation set's MAE is 2% higher than the MAE of the training set; the RMSE is 6% higher, and the Theil's U statistic is 5% higher. The relative differences between the model and validation set in terms of the respective goodness-of-fit statistics are relatively small, validating the accuracy of the model. Also, the percent data within 10% in the validation set computes to 2% less than the training set; and the percent data within 20% is almost equal in both sets. The above measures of predicting the performance of the model developed in this research indicate that the model is highly accurate and the model can be used for mass appraisal of the residential properties in the example town.

TABLE 5.2: Model Prediction Performances

Measure of Accuracy	Training Set	Validation Set	% Difference
MAE	14,580	14,934	+2%
RMSE	25,844	27,385	+6%
Theil's U Statisitc	0.126	0.132	+5%
Percent within 10%	73.9%	72.7%	-2%
Percent within 20%	92.6%	92.7%	+0.1%

5.4 Spatial Analysis Using ArcGIS

5.4.1 Neighborhood Analysis

Since the number of disputes about the assessed housing prices is rising, it would be convenient for tax assessment officials to show to the complainants the accuracy of assessment by incorporating the additive nonparametric regression (ANR) model with the ArcGIS spatial analysis. Table 5.3 provides a distribution of the 33,342 house prices estimated by the model, and compares with their actual 2009 assessed values. The estimated values have been categorized as: (i) within 10% and (ii) within 20% range of the actual assessed values. Table 5.3 in columns 4 and 5 gives the percentage of houses within 10% and within 20% of the actual assessed prices for each neighborhood with its average housing price per square foot of living area ($^{hprice}/_{sfla}$).

TABLE 5.3: Proportion of Estimated Price within 10% and 20% of Assessed Price for Each Neighborhood

nghdcode	Number of Houses	$^{hprice}/_{sfla}$ ($)	Estimated Price within 10%	Estimated Price within 20%
1	2	3	4	5
100	496	76	57.85%	86.77%
200	230	78.17	53.93%	87.64%
300	127	69.7	41.98%	76.54%
400	1532	78.54	84.64%	95.86%
500	1,089	79.58	75.55%	94.71%
600	176	89.85	67.72%	96.84%
700	345	76.43	86.22%	95.41%
701	311	99.35	64.11%	96.17%
800	320	85.63	96.21%	99.68%
900	279	83.59	86.55%	99.27%
1000	556	78.78	91.67%	98.48%
1100	472	94.43	95.21%	99.56%
1200	1,115	81.99	89.79%	98.53%
1300	102	88.02	89.22%	99.02%
1400	172	77.78	87.58%	98.14%
1401	134	75.83	86.89%	96.72%
1500	341	83.35	92.72%	99.23%
1600	175	93.45	66.86%	94.19%
1700	176	75.25	73.30%	93.75%
1800	415	79.59	90.64%	98.28%
1900	277	83.13	90.88%	98.54%
2000	572	92.61	86.06%	98.51%
2100	301	82	77.27%	95.83%
2200	241	99.27	71.73%	93.67%
2300	496	83.61	82.92%	96.25%
2400	155	94.8	57.31%	83.04%
2500	195	115.19	56.93%	79.56%
2600	427	73.08	83.83%	97.01%
2700	255	82.77	82.72%	96.30%
2800	1,312	72	86.08%	98.37%
2900	438	76.41	83.18%	97.47%
3000	254	91.58	92.00%	98.67%
3100	454	65.14	59.91%	87.74%
3200	1,749	71.67	75.41%	95.53%
3300	13	76.89	100.00%	100.00%

TABLE 5.3: - continued

nghdcode	Number of Houses	$^{hprice}/_{sfla}$ ($)	Estimated Price within 10%	Estimated Price within 20%
1	2	3	4	5
3400	551	78.23	73.30%	94.11%
3500	649	74.56	93.49%	99.38%
3600	1317	76.93	85.93%	96.87%
3700	1657	65.32	57.89%	88.28%
3800	1,748	70.63	66.43%	89.20%
3900	345	54.89	42.25%	87.23%
4000	386	84.07	74.39%	94.88%
4100	68	97.3	54.41%	80.88%
4200	133	61.44	53.38%	85.71%
4300	609	78.9	56.56%	81.28%
4400	594	74.59	65.58%	90.19%
4500	84	66.5	62.50%	84.38%
4600	46	107.05	65.22%	93.48%
4700	155	90.79	41.18%	65.36%
4800	117	106.48	40.23%	67.82%
4900	560	87.58	43.78%	75.10%
5000	452	70.36	67.69%	89.97%
5100	378	86.07	59.37%	87.60%
5200	1933	71.18	59.61%	87.00%
5300	359	76.2	73.26%	97.77%
5400	694	70.49	75.97%	92.77%
5500	87	68.8	65.12%	84.88%
5600	712	73.32	70.32%	92.70%
5700	13	109.74	84.62%	100.00%
5800	1078	73.64	80.58%	95.61%
5900	432	81.45	64.97%	86.23%
6000	540	81.74	59.52%	83.57%
6100	672	78.71	63.17%	88.05%
6200	4	118.1	0.00%	25.00%
6300	239	132.74	34.90%	76.04%
6400	629	76.8	90.11%	99.35%
6500	399	79.38	86.73%	96.68%

Table 5.3 statistics revealed that neighborhoods: 300, 2500, 4700, 4800, 4900, 6200, and 6300 have relatively lower percentage in the category of within 20% range of the actual assessed values. That indicates that the prediction accuracy of the additive nonparametric regression (ANR) model is relatively weak for those neighborhoods. On further investigation, it was found that the average housing prices per square foot of living area for all of these neighborhoods are either in the very low range, or very high range. Figure 5.6 illustrates the Table 5.3 data by sorting $^{hprice}/_{sfla}$ from low to high. Figure 5.6 shows that the percentage of houses within 10% and 20% against the average housing prices per square foot of living area for each neighborhood. This figure pictorially shows that the additive nonparametric regression (ANR) model have higher accuracy for houses costing from $72.00 to $88.00 per square foot of living area. The majority of houses in the example town fell in this range.

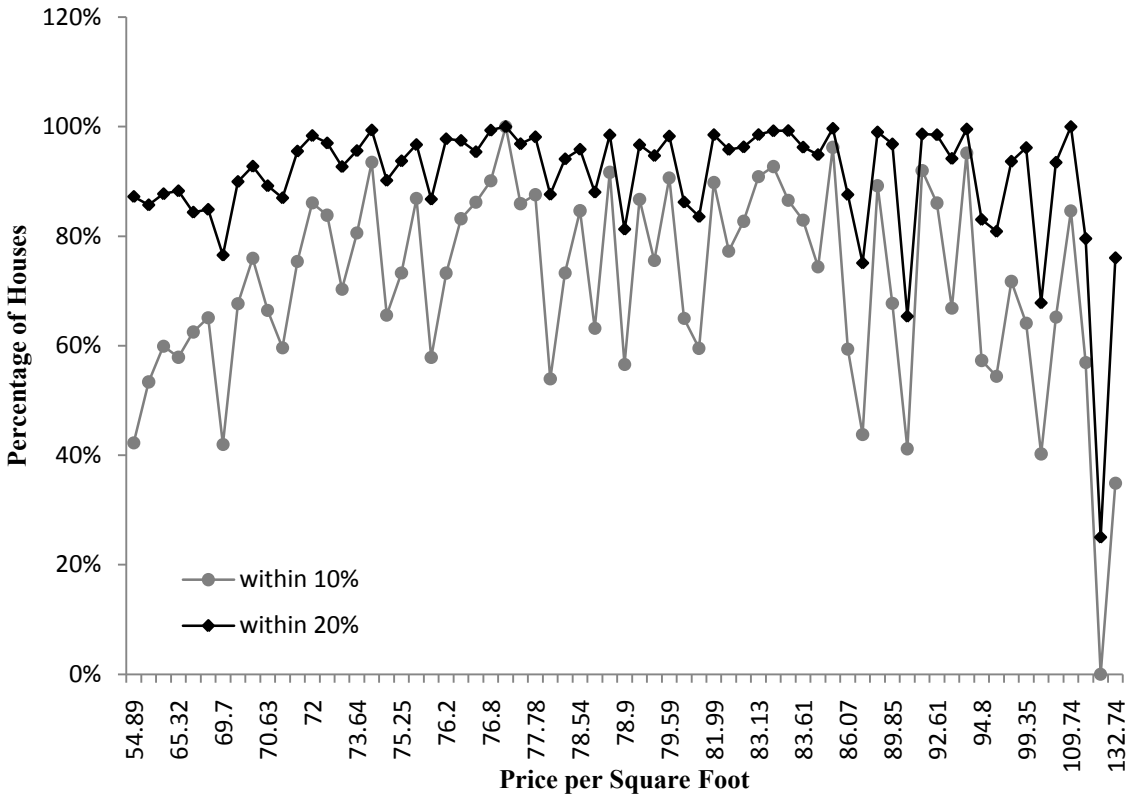


FIGURE 5.6: Predicted House Values within 10% and 20% of the Assessed Values

5.4.2 Spatial Housing Price Patterns

A series of maps showing the attributes of each neighborhood as found in the additive nonparametric regression (ANR) model were created by ArcGIS software. Map 5.1 shows the average assessed housing prices for each neighborhood, the highest priced homes have a red color mark. The highest housing prices are indicated in the eastern, northern, and southwestern areas of the Town.

Map 5.2 shows the distribution of proportion of estimated price within 10% of assessed price. A large majority of the neighborhoods fall into the > 80% category, leading to the conclusion that most of the assessed housing prices can be accurately forecasted from the additive nonparametric regression (ANR) model. The neighborhoods of lower percentage of within 10% are indicated in

the southwestern and northeastern of the Town, almost all in the lower priced neighborhoods. Map 5.3 presents the proportion of estimated price within 20% of assessed price. In comparing Map 5.1 and Map 5.3, the results support the early findings in Map 5.2 that the prediction accuracy for the lower priced homes tends to be lower.

5.5 Summary of Additive Nonparametric Regression Model Analysis Results

The research presented in this paper has successfully demonstrated the application of an additive nonparametric regression (ANR) model and linked it to the map layers of the Town of Amherst, New York which supplied assessment data of 33,342 houses. Utilizing GIS methodology for establishing the prices of houses in a municipality, various maps showing the spatial distribution of housing price and estimation accuracy by neighborhood have been generated. Such visual patterns will enable the town assessors and appraisal companies to assess the value of homes in a neighborhood, on a uniform basis. This will also help the town planners in evaluating future development of their town.

The estimation model had the ability to predict the housing prices with relatively high accuracy. The estimation errors of the validation set had minor differences with the model estimation errors. Also, a large majority of the houses were found to be within 20% of the actual assessed values. On viewing the pattern of the within 10% and within 20% neighborhoods, it was found that very low and very high priced homes had lower estimation prediction accuracy. The reason for this result is that the additive nonparametric regression (ANR) model represents the “averaging” behavior or “central” tendency of a distribution, the tail behaviors of that distribution are therefore hard to recognize.

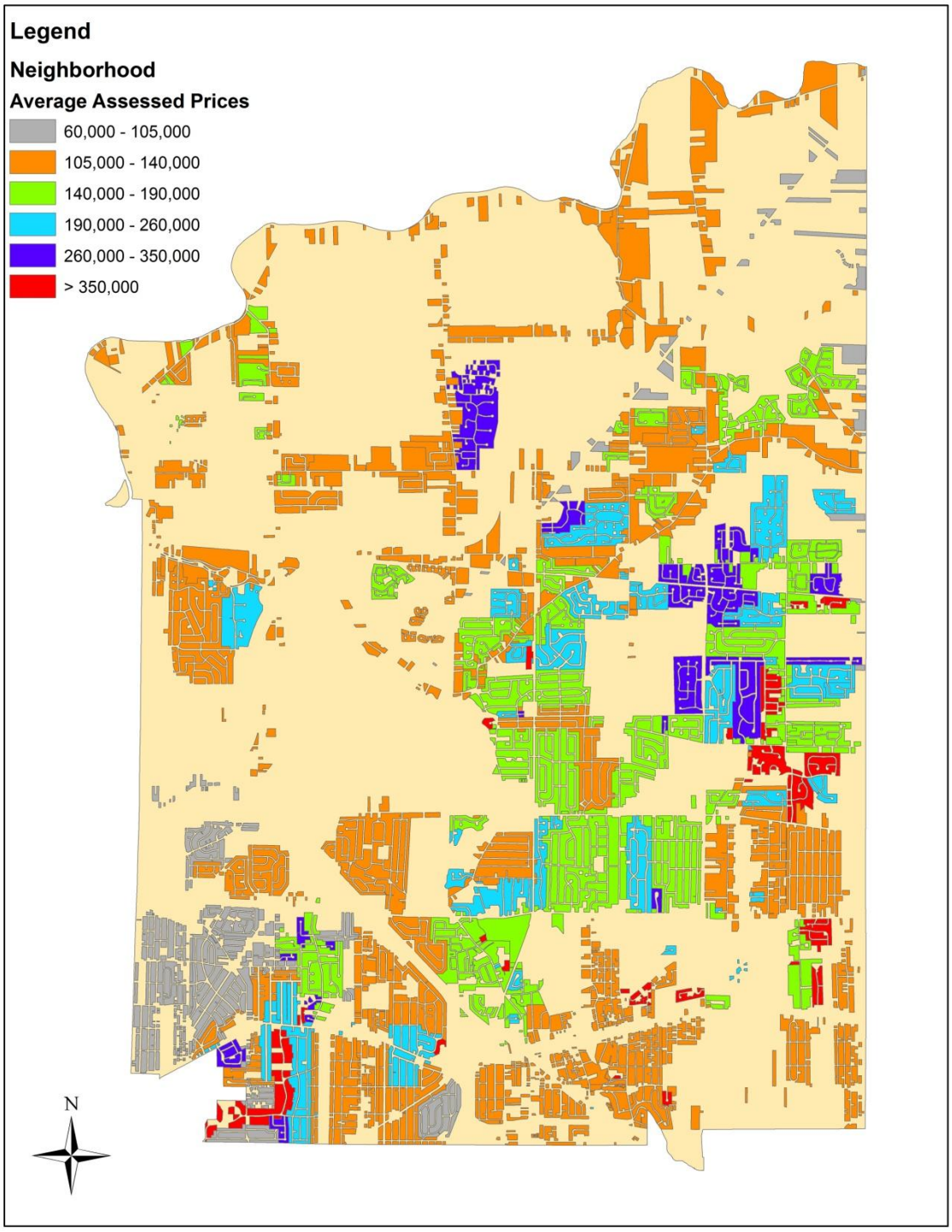
The following conclusions can be drawn from the research presented in this paper.

(i) An additive nonparametric regression (ANR) model can be developed and applied for a town, village, or a city with high accuracy and can be linked to the map layers of the town using GIS technology for viewing spatial patterns based on housing price, housing age, etc..

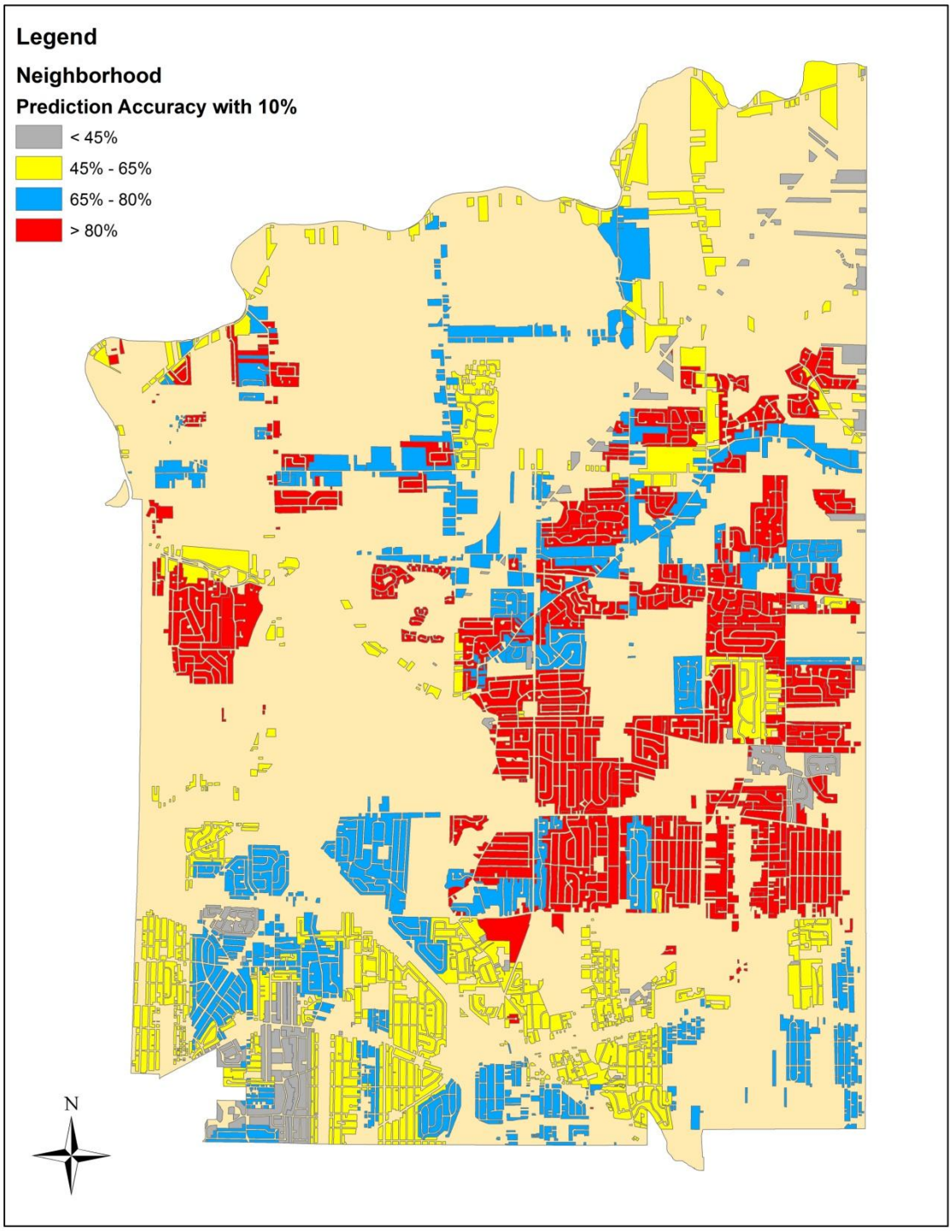
(ii) The factors that significantly affect the price of a house include: frontage width, depth of the parcel, age, square foot of living area, architectural building style, and the neighborhood.

(iii) The very low and high priced houses have lower prediction accuracy within the context of the additive nonparametric regression model.

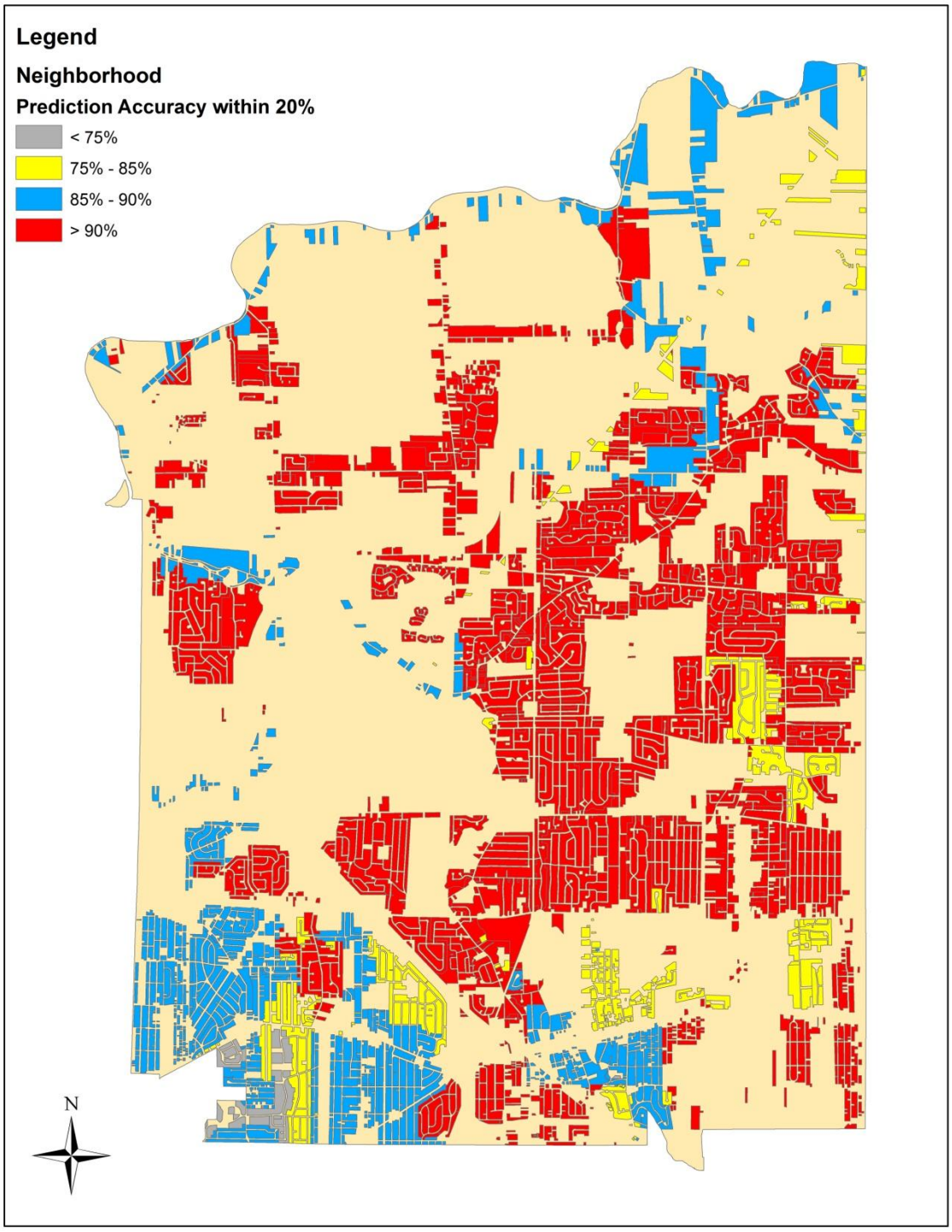
(iv) Using GIS technology, visual patterns of the distribution of prices of houses, or age of houses, or the square foot of living area, by neighborhood can facilitate in any dispute resolutions due to inequity in assessments.



MAP 5.1: Spatial Distribution of Average Assessed Housing Price ($)



MAP 5.2: Distribution of Estimated Price within 10% of Assessed Price



MAP 5.3: Distribution of Estimated Price within 20% of Assessed Price

CHAPTER 6: Artificial Neural Network Model for Estimation of Property Values

6.1 Introduction

The chapter 6 presents the application of an artificial neural network (ANN) model for housing price estimation, and also links it to the GIS map layer of a municipality to draw various maps showing the wide variation in the prices of homes based on location or neighborhood. The artificial neural network (ANN) model was developed using the 2009 housing assessment data of 33,342 houses of a town in Western New York. The artificial neural network (ANN) model had high accuracy and tested well on validation. Based on the validation test, using artificial neural network (ANN) model to predict real estate values seems promising. The research presented in this chapter links the artificial neural network (ANN) model to the GIS map layers of the town to enable the appraising authority to visualize the distribution of housing based on price on each of the neighborhoods.

To further investigate the ANN applications for housing price estimation, the back-propagation neural network was applied in this research on the 33,342 residential properties of a large town in the western New York with the results presented in this paper.

This paper takes a new approach and differs in a number of respects from the earlier studies. In the first place, ArcGIS software Version 9.3 was applied on the results from the artificial neural network (ANN) to analyze and visualize the spatial variations in the housing sub-markets. The housing dataset for the Town of Amherst, State of New York, which was segmented into 67 neighborhoods, was used to develop and clearly demonstrate the spatial patterns. Secondly, tables, graphs, and maps of prediction accuracy were created based on the percentage of the

estimated prices within 10% and 20% of the actual assessed values for each neighborhood. These tables, graphs and maps can assist a town's property tax assessment officers in housing price estimation and comparisons within the sub-markets.

The major focus of the research described in this chapter is to develop a comprehensive explanatory model of housing price determination using artificial neural network and also visualize the price variation patterns of different neighborhoods using GIS. The model developed in this research also gives several numerical summaries about the example housing market.

The format of the paper is constructed as follows. In section 2, the artificial neural network (ANN) model is reviewed. Section 3 contains a description of the model development. The results of spatial analysis are reported in section 4. The final section provides the summary of artificial neural network model analysis results arrived in this research.

6.2 Introduction to Artificial Neural Network (ANN) Model

Recently, artificial neural network, based on the neural architecture of the brain have been developed and successfully applied across a variety of disciplines including psychology, genetics, linguistics, engineering, computer science, and economics. Neural networks have been shown to be particularly well suited to solving problems involving pattern recognition, classification, and nonlinear feature detection.

The artificial neural network is an artificial intelligence model originally designed to replicate the human brain's learning processes, and thus is made up of a complex network of artificial neurons. Each artificial neuron can send a signal to the other artificial neurons in the network to which it is immediately connected. In essence, neural networks are densely interconnected networks of artificial neurons with a rule to adjust the strength or weight (analogous to

regression coefficient in multiple regression equation) of the connections between the neurons in response to externally supplied data.

The artificial neuron performs three basic functions like a neuron in the human brain. Figure 6.1 shows an example of an artificial neuron performing the three basic functions. Firstly, it receives inputs from other artificial neurons through weighted links; secondly, it sums and processes these inputs; and finally, it outputs the results to other immediately connected artificial neurons in the network. The processing stage involves the computation of a weighted sum of the inputs and then passing the sum through a mathematical transfer function which acts as a nonlinear threshold.

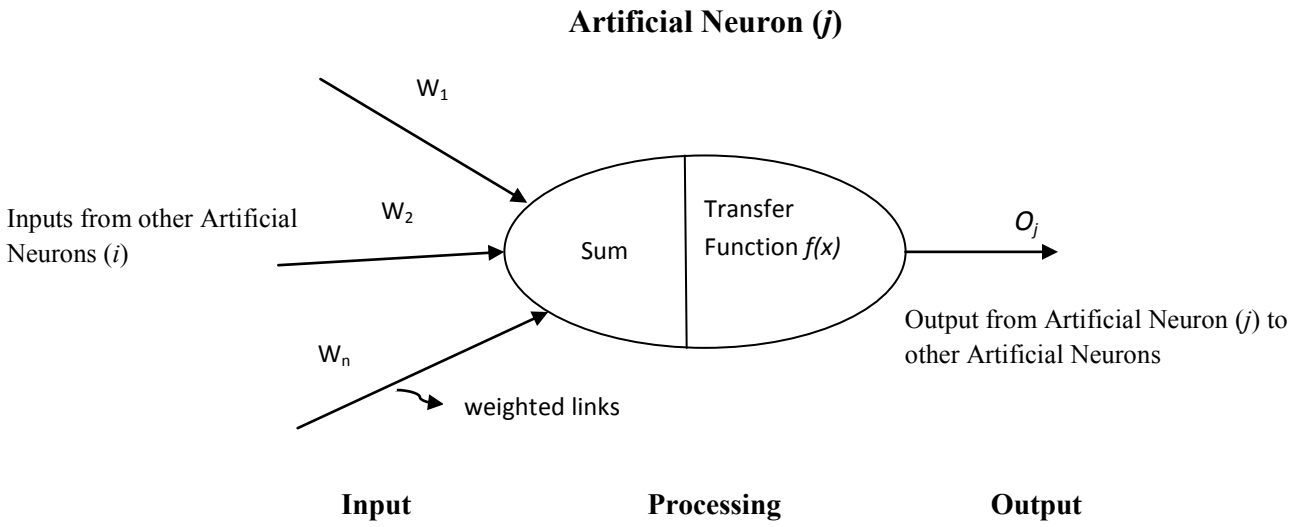


FIGURE 6.1: Three Basic Functions of an Artificial Neuron

The relationship between the input signals and the output to other artificial neuron is as in the following equation (Rumelhart et al., 1986), and thus the artificial neural network can perform prediction after learning the underlying relationship between the input variables and outputs.

$$X_j = \sum_{i=1}^{n} W_{ji} Y_i \qquad \text{…….....} \quad (1)$$

where: X_j is the net input to artificial neuron (j);
Y_i is the value of input signal from artificial neuron (i);
W_{ji} is the weight from artificial neuron (i) to artificial neuron (j);
n is the number of input signals to artificial neuron (j).

The output from artificial neuron (j) is a function of the following transfer function:

$$O_j = f(X_j) \qquad \text{……} \quad (2)$$

where: O_j is the output signal from artificial neuron (j);
$f(X_j)$ is the transfer function of artificial neuron (j).

6.2.1 Training Algorithm

Out of several training algorithms, back-propagation is one of the most widely used training algorithm and has been used in this research. During the first stage, the input is presented and propagated forward to produce the output for each neuron. This output is then compared with the desired output (in our case the actual assessed value) to produce the error signal. The second stage involves a backward pass through the networks, during which the error signal is passed to each neuron in the network. This backward pass allows recursive computation of the error signal and causes the network to adapt its weights. These stages are repeated with each set of training data to find an optimum set of weights so as to mirrors its target value. During the training process, the artificial neural network learns the relationship between a given set of inputs and outputs.

6.2.2 Artificial Neural Network Architecture

In this dissertation, the artificial neural network used to estimate the housing prices is shown in Figure 6.2. Each neuron is represented by a circle, and each neuron interconnection, with its

associated weight, by a line terminated by an arrow. The artificial neural network depicted in Figure 6.2 has an input layer, a hidden layer, and an output layer. Signal in the artificial neural network is fed forward from left to right.

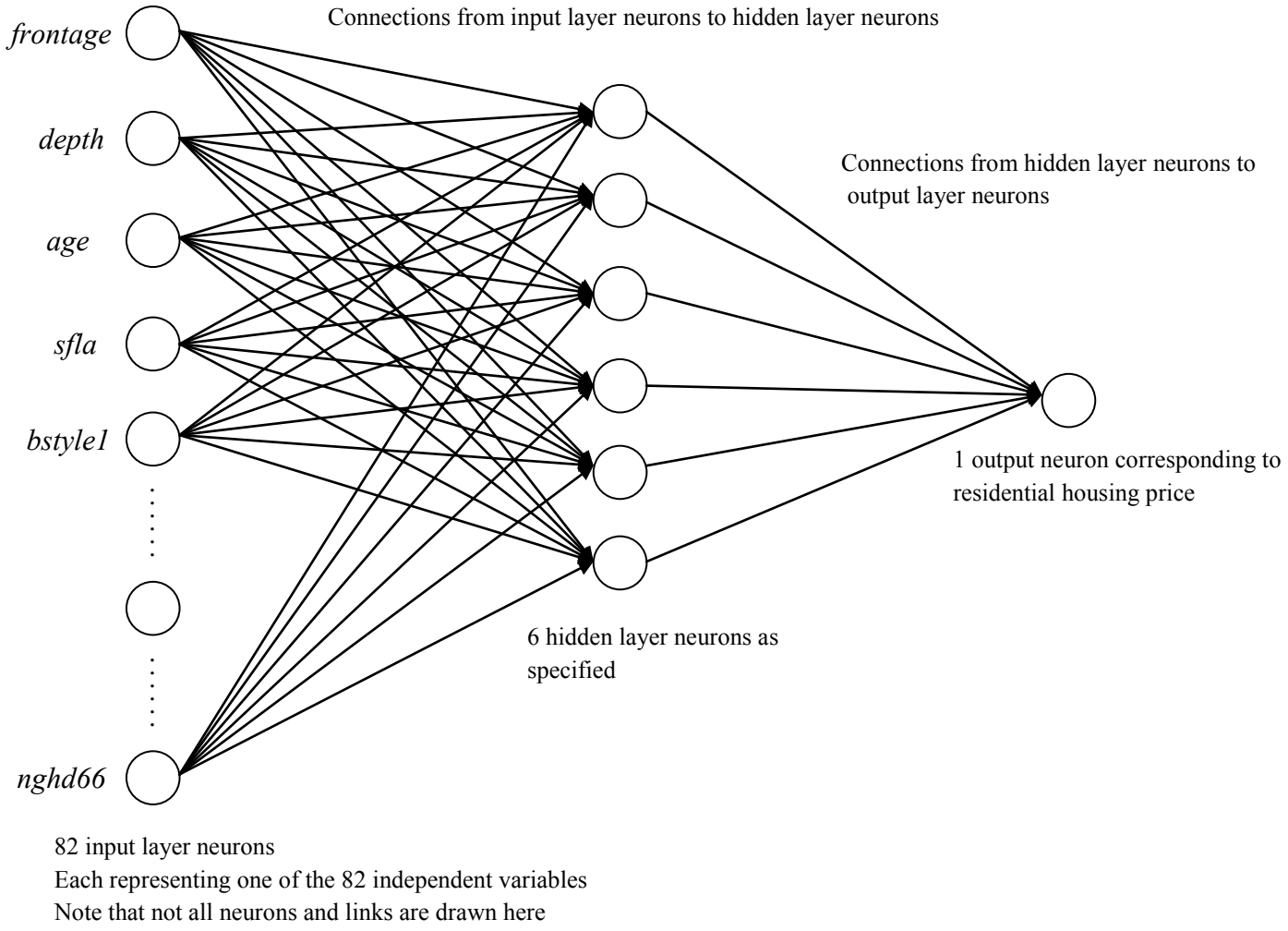


FIGURE 6.2: Artificial Neural Network Architecture

Artificial neural network technique has become more easily applicable due to computational advances, and has the ability to minimize**.** the consequences of several methodological problems such as functional form misspecification, multicollinearity, and heteroskedasticity associated with multiple regression models. The reason for the success of neural network over multiple regression models stems from the difference in how their functional form is specified. While traditional estimation methods must prespecify a functional form before fitting the data, neural network self determines their functional form based on the tuning of their weight parameters to best fit the data (Do and Grudnitski, 1993).

On their negative side, neural networks suffer from some small but annoying shortcomings. It is nearly impossible to specify an effective architecture given the specifications of a problem that must be done by experimentation to find the optimal artificial neural network. Moreover, the approach is notoriously "black-box" in nature. It has no explaining ability for cause and effect.

6.3 Artificial Neural Network (ANN) Model Development

The artificial neural network model (ANN) was used for developing a model to estimate housing prices for mass appraisal. Figure 6.3 represents the flow chart for the model development process. The following steps were used in formulating the artificial neural network model (ANN):

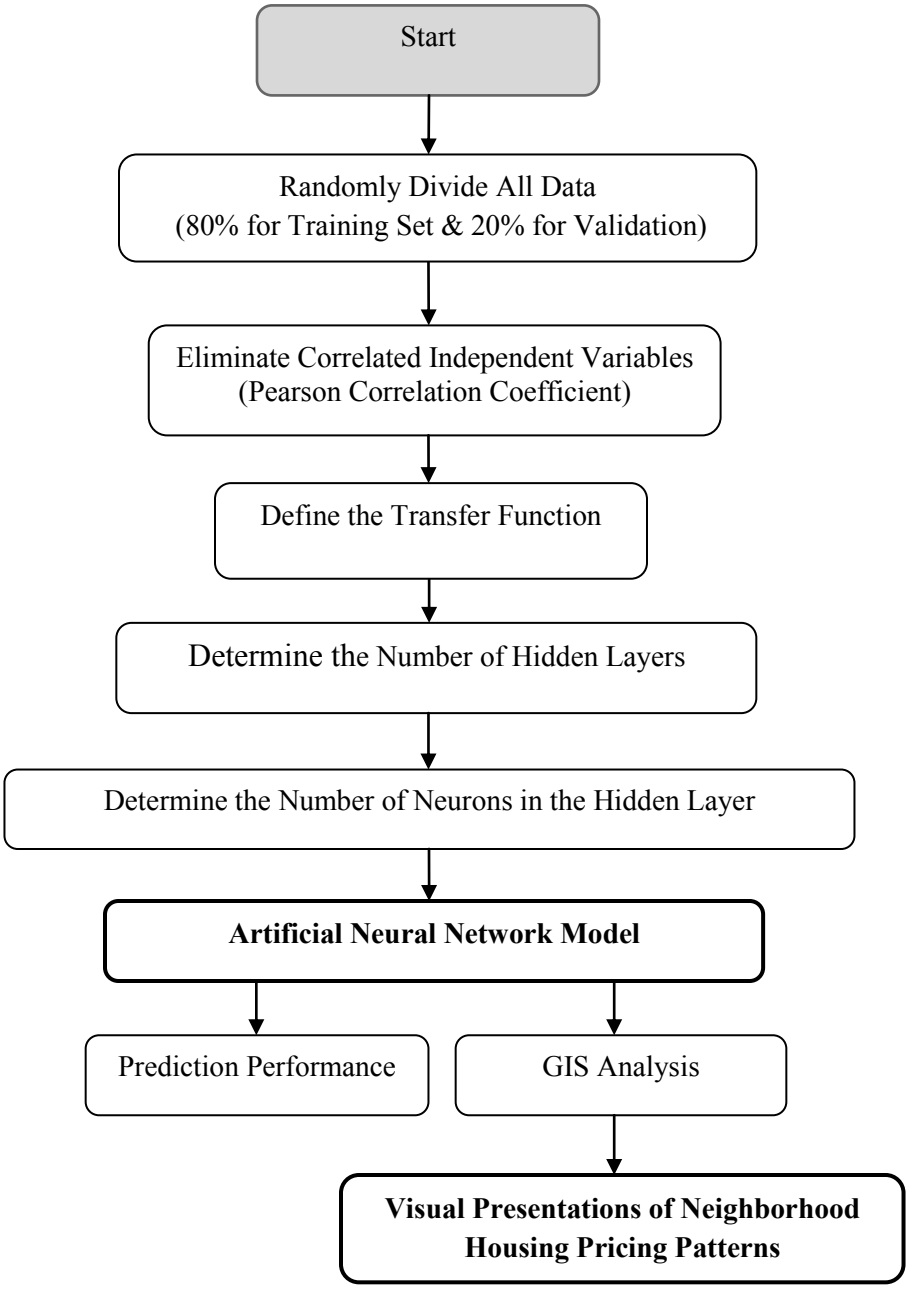


FIGURE 6.3: Model Development Process

6.3.1 Dividing the Dataset into Training Set and Validation Set:

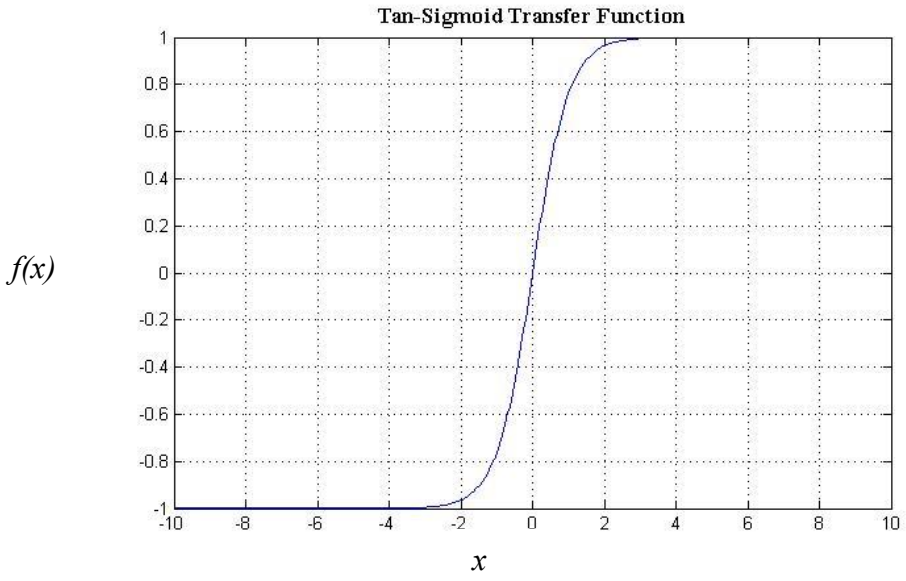
Refer to 3.4.1

6.3.2 Examination of the Covariance Structure across Model Variables:

Refer to 3.4.2

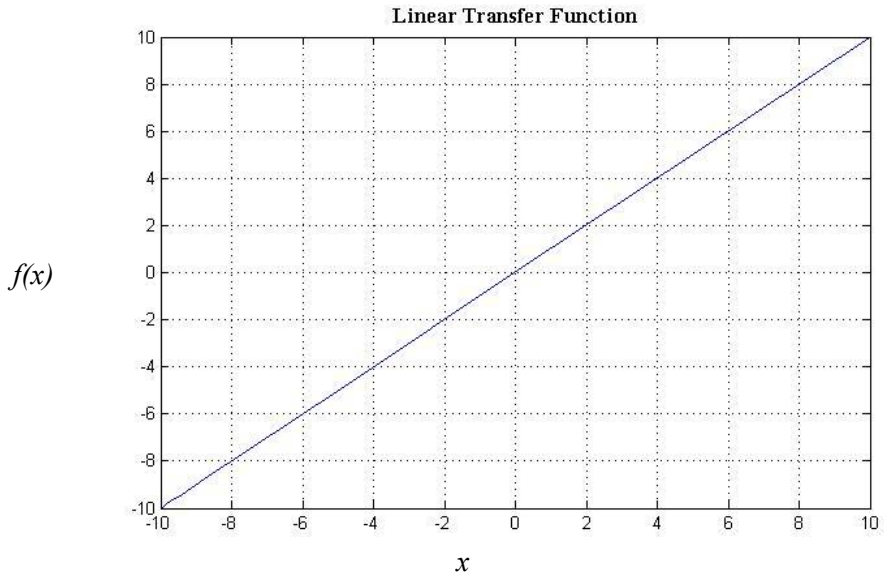
6.3.3 Define the Transfer Function

The transfer function determines the relationship between inputs and outputs of a neuron and its network. The most commonly used tan-sigmoid transfer function was used in the neurons of the hidden layer and the linear transfer function was utilized in the neurons of the output layer. These two transfer functions are pictorially presented in Figures 6.4 and 6.5.



$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

FIGURE 6.4: Tan-Sigmoid Transfer Function



$$f(x) = x$$

FIGURE 6.5: Linear Transfer Function

6.3.4 Determining the Number of Hidden Layers

Masters (1993) strongly recommended that one hidden layer be the first choice for any practical artificial neural network design. For the vast majority of practical problems, there is no reason to use more than one hidden layer. Hornik (1991) has shown that a single hidden layer is sufficient for an artificial neural network to approximate any complex non-linear function with any desired accuracy. Therefore, one hidden layer was applied to construct the architecture of artificial neural network in the research presented in this dissertation.

6.3.5 Determining the Number of Neurons in the Hidden Layer

The number of neurons in the hidden layer was determined from the data by trial and error method. The method to finding the optimal number of hidden neurons is the constructive algorithm, proposed by Kwok and Yeung, 1997. The basic idea of constructive algorithm is to start with a small number of neurons, then add hidden neurons incrementally until a satisfactory solution is found.

80% of the data was used to train the network, and 20% of the data was used for validation to see how well the network was generalized. Training on the training set continues as long as the training reduces the network's error on the validation set. After the network memorizes the training set, training is stopped. Bayesian regularization technique was used to improve network generalization. This technique automatically avoids the problem of overfitting, which plagues many optimization and learning algorithms. The performance function evaluated the sum of squares of the network errors (SSE) on the training set to achieve the best generalization.

For each number of neurons, the same 80% of the data was trained repeatedly for ten times, and the same remaining 20% of the data was used for the validation set. The variable used to decide the number of neurons was the correlation coefficient between the actual prices and predicted prices. The correlation coefficient is a measure of how well the variation in the output is explained by the targets. If this number is equal to 1, then there is perfect correlation between the actual prices and predicted prices.

After training ten times, the average of correlation coefficient was calculated. Table 6.1 shows the calculated results for each number of neurons and Figure 6.6 gives the relationship between the correlation coefficient and the number of neurons. The correlation coefficient gradually

increases at a declining rate as the number of neurons increases, with slighter improvement beyond 6 neurons. Therefore, 6 neurons were decided for constructing the architecture of artificial neural network. From Table 6.1, the correlation coefficient for 6 neurons in the hidden layer is 0.9723, that is, very close to 1, which indicates a good fit.

TABLE 6.1: Correlation Coefficient for Each Number of Neurons

Number of Neurons	Correlation Coefficient
1	0.9554
2	0.9628
3	0.9664
4	0.9691
5	0.9712
6	0.9723
7	0.9729
8	0.9731
9	0.9735
10	0.9739
15	0.9744
20	0.9745

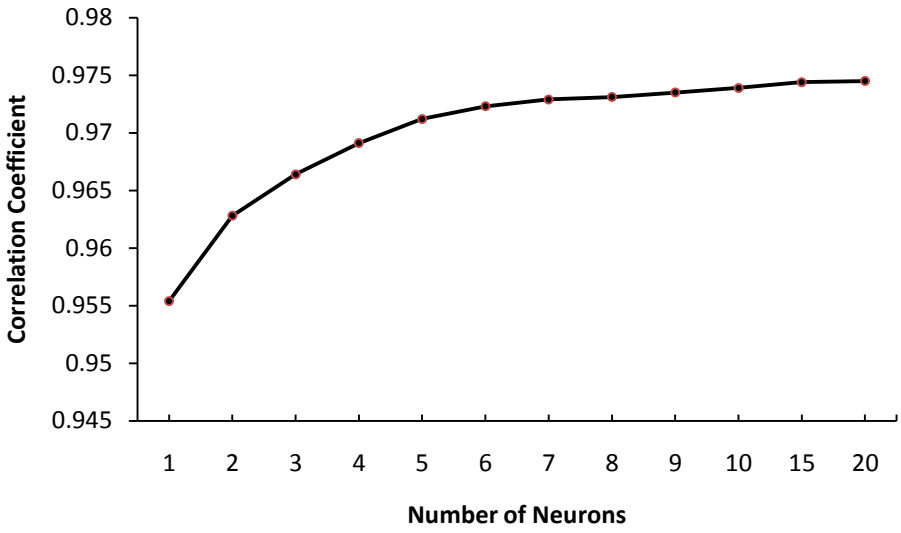


FIGURE 6.6: Relationship between Correlation Coefficient and the Number of Neurons

Figure 6.2 shows the final architecture of artificial neural network model (ANN) adopted in this chapter. It is composed of three layers, namely, an input layer, one hidden layer, and an output layer. An input layer consists of 82 input nodes as in Table 6.2. A hidden layer is with 6 neurons, as determined above in this section. An output layer is with 1 node corresponding to estimated housing prices for residential houses.

TABLE 6.2: Description of Variables Initially Considered for Model Development

Variable	Description	Unit of Measure
Dependent Variable:		
hprice	Assessed housing prices	dollar
Independent Variables:		
Physical Structural Attributes:*		
frontage	Parcel frontage	feet
depth	Parcel depth	feet
age	Year house was built subtracted from 2008	year
sfla	Square footage of living area	square foot
*nobaths***	Number of bathrooms	
*nobedrms***	Number of bedrooms	
*nofirepl***	Number of fireplaces	
bstyle1	Dummy variable for building style (1 if raised ranch, 0 otherwise)	binary
bstyle2	Dummy variable for building style (1 if split level, 0 otherwise)	binary
bstyle3	Dummy variable for building style (1 if cape cod, 0 otherwise)	binary
bstyle4	Dummy variable for building style (1 if colonial, 0 otherwise)	binary
bstyle5	Dummy variable for building style (1 if contemporary, 0 otherwise)	binary
bstyle6	Dummy variable for building style (1 if mansion, 0 otherwise)	binary
bstyle7	Dummy variable for building style (1 if old style, 0 otherwise)	binary
bstyle8	Dummy variable for building style (1 if cottage, 0 otherwise)	binary
bstyle9	Dummy variable for building style (1 if log cabin, 0 otherwise)	binary
bstyle10	Dummy variable for building style (1 if duplex, 0 otherwise)	binary
bstyle11	Dummy variable for building style (1 if town house, 0 otherwise)	binary
bstyle12	Dummy variable for building style (1 if others, 0 otherwise)	binary
Location Characteristics:		
nghd1~nghd66	Dummy variables for neighborhood	binary

*The total number of independent variables total to 82, as below

i. *frontage*, *depth*, *age*, *sfla* - 4
ii. *bstyle1 ~ 12* - 12
iii. *nghd1 ~ 66* - 66
Total 82

**These three variables were later eliminated from the model (see Sec 6.3.2.).

6.3.6 Model Prediction Performance

The prediction accuracy of the model was an important goal of this research; therefore 20% of the data was utilized for model validation. Three widely accepted measures were utilized to measure the prediction accuracy of the model: (i) the root mean squared error (RMSE), (ii) mean absolute error (MAE), and (iii) Theil's U statistic. These three statistics were applied to the 80% model data and the 20% validation group of houses to analyze the prediction performance.

The root mean square error (RMSE) is the square root of the average of the squared values of the prediction errors and weights large errors more heavily than small errors. The root mean squared error (RMSE) is defined as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t}\left(y_t - \hat{y}_t\right)^2}$$

where y_t is the actual housing price, and $\hat{Y}_t$ is the fitted price from the regression model.

The mean absolute error (MAE) is the average of the absolute values of the prediction errors and is given by:

$$MAE = \frac{1}{n}\sum_{t}\left|y_t - \hat{y}_t\right|$$

Theil's U statistic is the square root of the ratio of the mean squared error of the predicted change to the average squared actual change. It is defined as:

$$U = \sqrt{\frac{\sum_{t}\left(y_t - \hat{y}_t\right)^2}{\sum_{t}\left(y_t\right)^2}}$$

For all three criteria values closer to 0 indicate better fitting models. In addition, model validity can be considered by examining differences between the respective goodness-of-fit statistics.

Table 6.3 compares the results of model prediction performance of the training set versus the validation set. The validation set's MAE is 7% higher than the MAE of the training set; the RMSE is 26% higher, and the Theil's U statistic is 25% higher. The relative difference between the model and validation set in terms of the respective goodness-of-fit statistics is not too large, validating the accuracy of the model. Also, the percent data within 10% in the validation set computes to 2% less than the training set; and the percent data within 20% is almost equal in both sets. The above measures of predicting the performance of the model developed in this research indicate that the model is accurate enough and the model can be used for mass appraisal of the residential properties in the example town.

TABLE 6.3: Model Prediction Performances

Measure of Accuracy	Training Set	Validation Set	% Difference
MAE	12,213	13,107	+7%
RMSE	20,687	26,182	+26%
Theil's U Statisitc	0.101	0.126	+25%
Percent within 10%	79.3%	77.7%	-2%
Percent within 20%	94.6%	94.1%	-0.5%

6.4 Spatial Analysis Using ArcGIS

6.4.1 Neighborhood Analysis

Since the number of disputes about the assessed housing prices is rising, it would be convenient for tax assessment officials to show to the complainants the accuracy of assessment by incorporating the artificial neural network (ANN) model with the ArcGIS spatial analysis. Table 6.4 provides a distribution of the 33,342 house prices estimated by the model, and compares with their actual 2009 assessed values. The estimated values have been categorized as: (i) within 10% and (ii) within 20% range of the actual assessed values. Table 6.4 in columns 4 and 5 gives the percentage within 10% and within 20% of the actual assessed prices for each neighborhood with its average housing price per square foot of living area (*sfla*).

TABLE 6.4: Proportion of Estimated Price within 10% and 20% of Assessed Price for Each Neighborhood

nghdcode	Number of Houses	$^{hprice}/_{sfla}$ ($)	Estimated Price within 10%	Estimated Price within 20%
1	2	3	4	5
100	496	76	57.17%	87.00%
200	230	78.17	53.93%	87.64%
300	127	69.7	60.49%	85.19%
400	1532	78.54	85.51%	96.84%
500	1,089	79.58	79.84%	95.91%
600	176	89.85	91.14%	98.10%
700	345	76.43	92.16%	99.46%
701	311	99.35	88.52%	97.61%
800	320	85.63	96.53%	99.68%
900	279	83.59	93.09%	99.64%
1000	556	78.78	93.37%	98.48%
1100	472	94.43	96.95%	99.56%
1200	1,115	81.99	92.93%	99.12%
1300	102	88.02	90.20%	99.02%
1400	172	77.78	87.58%	98.76%
1401	134	75.83	91.80%	96.72%
1500	341	83.35	93.49%	99.23%
1600	175	93.45	80.81%	98.84%
1700	176	75.25	81.82%	96.02%
1800	415	79.59	93.35%	99.01%
1900	277	83.13	93.07%	99.27%
2000	572	92.61	90.15%	98.88%
2100	301	82	86.36%	97.35%
2200	241	99.27	74.68%	96.20%
2300	496	83.61	86.25%	98.96%
2400	155	94.8	60.82%	87.13%
2500	195	115.19	66.42%	91.97%
2600	427	73.08	87.06%	96.52%
2700	255	82.77	83.13%	97.53%
2800	1,312	72	93.55%	98.91%
2900	438	76.41	83.41%	97.24%
3000	254	91.58	94.22%	99.11%
3100	454	65.14	72.41%	94.10%
3200	1,749	71.67	80.80%	96.29%
3300	13	76.89	92.31%	100.00%

TABLE 6.4: - continued

nghdcode	Number of Houses	$hprice/sfla$ ($)	Estimated Price within 10%	Estimated Price within 20%
1	2	3	4	5
3400	551	78.23	83.79%	97.42%
3500	649	74.56	93.80%	99.38%
3600	1317	76.93	85.61%	96.40%
3700	1657	65.32	63.39%	88.86%
3800	1,748	70.63	69.70%	93.40%
3900	345	54.89	79.33%	96.05%
4000	386	84.07	78.17%	95.42%
4100	68	97.3	57.35%	89.71%
4200	133	61.44	51.13%	90.98%
4300	609	78.9	62.71%	85.75%
4400	594	74.59	69.54%	90.71%
4500	84	66.5	60.94%	82.81%
4600	46	107.05	71.74%	93.48%
4700	155	90.79	57.52%	81.70%
4800	117	106.48	40.23%	73.56%
4900	560	87.58	56.08%	85.45%
5000	452	70.36	74.65%	92.20%
5100	378	86.07	72.56%	91.03%
5200	1933	71.18	63.39%	88.07%
5300	359	76.2	81.06%	97.77%
5400	694	70.49	81.51%	94.45%
5500	87	68.8	63.95%	84.88%
5600	712	73.32	75.71%	94.29%
5700	13	109.74	92.31%	100.00%
5800	1078	73.64	84.03%	95.52%
5900	432	81.45	67.96%	84.73%
6000	540	81.74	66.43%	88.81%
6100	672	78.71	69.14%	92.25%
6200	4	118.1	100.00%	100.00%
6300	239	132.74	72.40%	95.31%
6400	629	76.8	94.49%	99.84%
6500	399	79.38	88.78%	98.72%

Table 6.4 statistics revealed that the prediction accuracy of the artificial neural network (ANN) model is relatively higher for each of the neighborhoods. Figure 6.7 illustrates the Table 6.4 data by sorting $^{hprice}/_{sfla}$ from low to high. Figure 6.7 shows that the percentage of houses within 10% and 20% against the average housing prices per square foot of living area for each neighborhood. This figure pictorially shows that the artificial neural network (ANN) model has high accuracy for houses costing from $54.00 to $132.00 per square foot of living area.

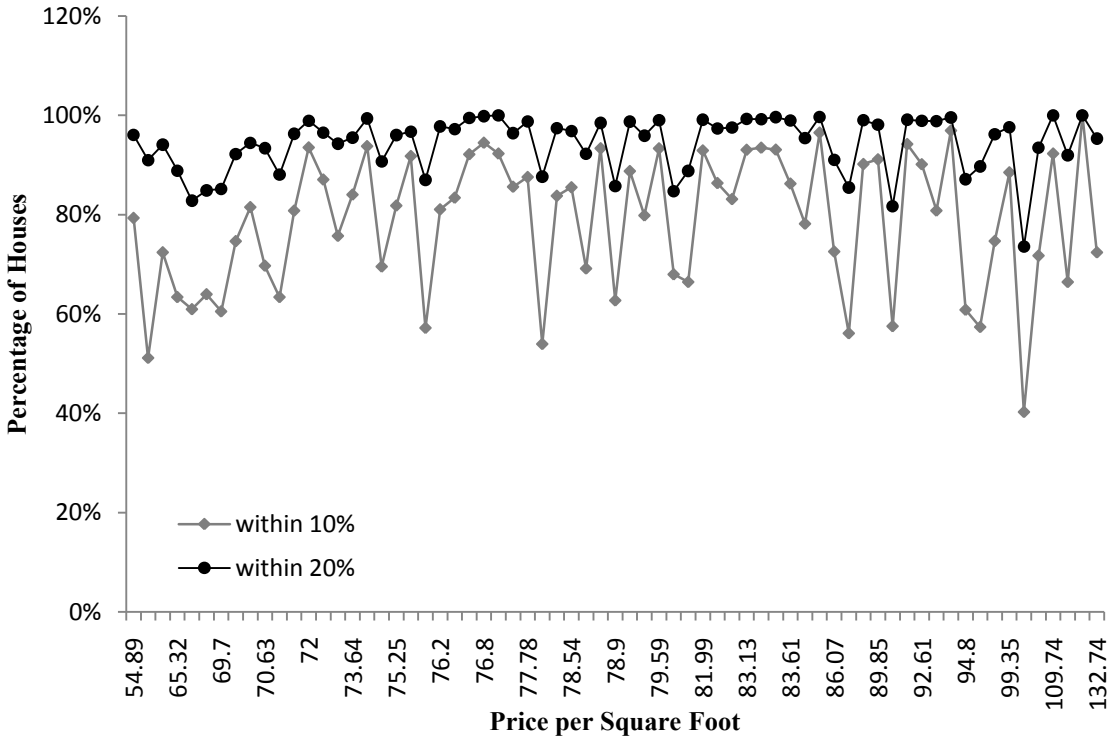


FIGURE 6.7: Distribution of the Model Prediction Accuracy for Price per Square Foot

6.4.2 Spatial Housing Price Patterns

A series of maps showing the attributes of each neighborhood as found in the artificial neural network (ANN) model were created by ArcGIS software. Map 6.1 shows the average assessed housing prices for each neighborhood, the highest priced homes have a red color mark. The highest housing prices are indicated in the eastern, northern, and southwestern areas of the Town.

Map 6.2 shows the distribution of proportion of estimated price within 10% of assessed price. A large majority of the neighborhoods fall into the > 85% category, leading to the conclusion that most of the assessed housing prices can be accurately forecast using the artificial neural network (ANN) model. The neighborhoods of lower percentage of within 10% are indicated in the southwestern and northeastern of the Town, almost all in the lower priced neighborhoods. Map 6.3 presents the proportion of estimated price within 20% of assessed price. In comparing Map 6.1 and Map 6.3, the results support the early findings in Map 6.2 that the prediction accuracy for the lower priced homes tends to be lower.

6.5 Summary of Artificial Neural Network Model Analysis Results

The research presented in this paper has successfully developed an artificial neural network (ANN) model and linked it to the map layer of the Town of Amherst, New York. Utilizing GIS methodology, various maps showing the spatial distribution of housing price and estimation accuracy by neighborhood have been generated. Such visual patterns will enable the town assessors and appraisal companies to assess the value of homes in a neighborhood, on a uniform basis. This will also help the town planners in evaluating future development of their town. These methods are easily generalized to towns of a similar type.

The estimation model had the ability to predict the housing prices with relatively high accuracy. The estimation errors of the validation set had minor differences with the model estimation errors. Also, a large majority of the houses were found to be within 20% of the actual assessed values. On viewing the pattern of the within 10% and within 20% neighborhoods, it was found that lower priced homes had lower estimation prediction accuracy.

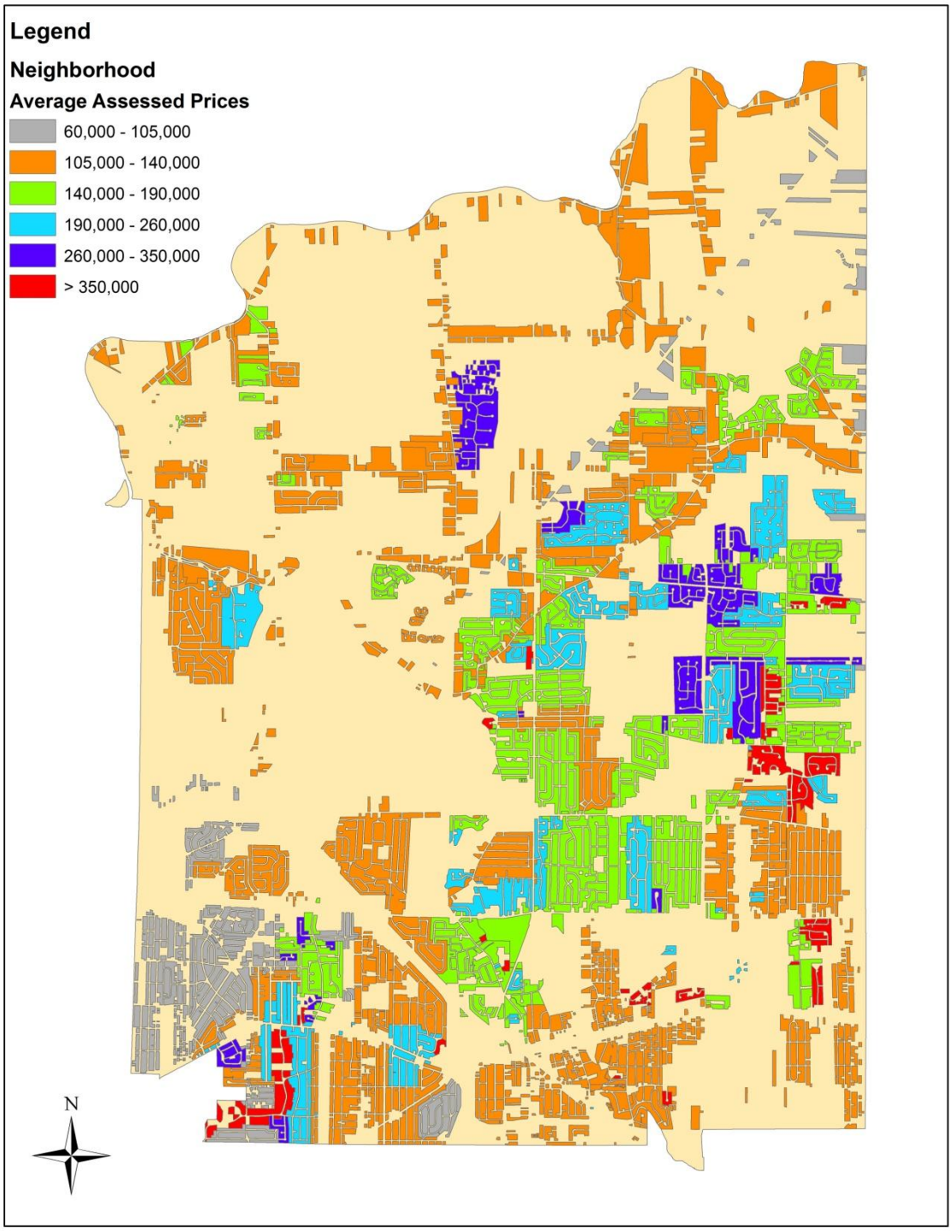
The following conclusions can be drawn from the research presented in this paper.

(i) An artificial neural network (ANN) model can be developed for a town, village, or a city with high accuracy and can be linked to the map layers of the town using GIS technology for viewing spatial patterns based on housing price, housing age, etc..

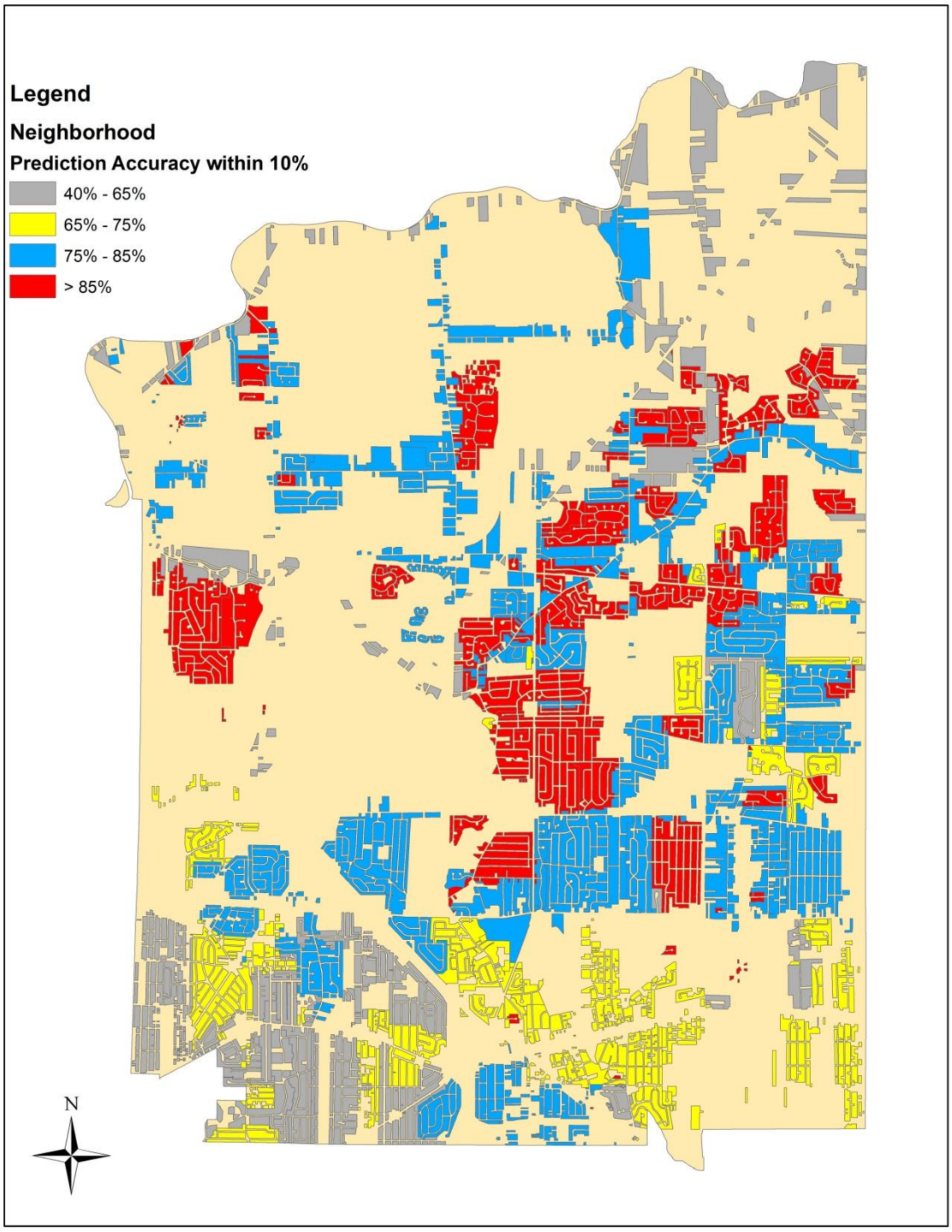
(ii) The factors that significantly affect the price of a house include: frontage width, depth of the parcel, age, square foot of living area, architectural building style, and the neighborhood.

(iii) The lower priced houses have lower prediction accuracy within the context of the artificial neural network model.

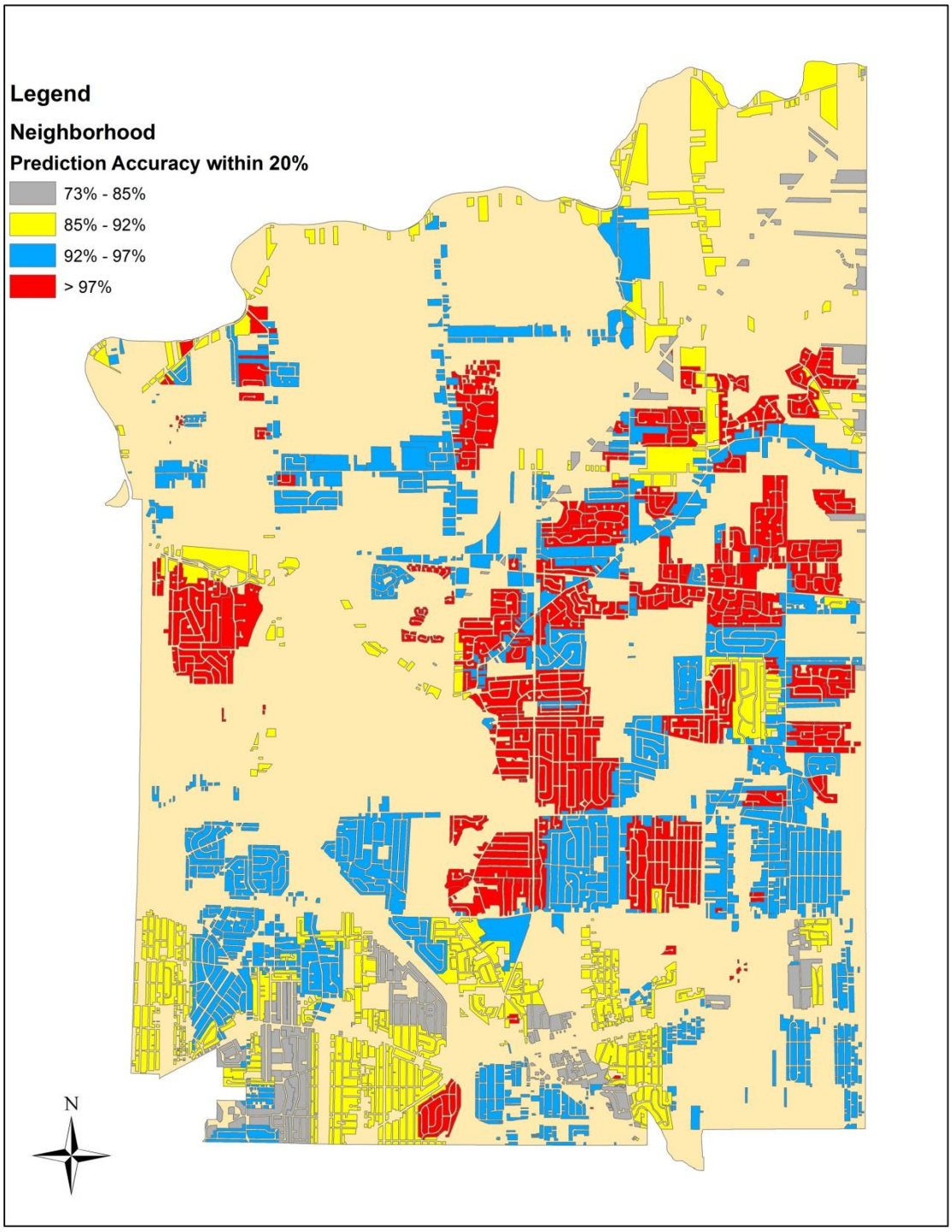
(iv) This research has demonstrated the novel and beneficial uses of GIS technology in housing price estimations. Visual patterns of the distribution of prices of houses, or age of houses, or the square foot of living area, by neighborhood can facilitate in any dispute resolutions due to inequity in assessments.



MAP 6.1: Spatial Distribution of Average Assessed Housing Price ($)



MAP 6.2: Distribution of Estimated Price within 10% of Assessed Price



MAP 6.3: Distribution of Estimated Price within 20% of Assessed Price

CHAPTER 7: Effectiveness Comparison of the Residential Property Mass Appraisal Methodologies

7.1 Introduction

Quite a few statistical and artificial neural network models have been developed for the mass appraisal of the real estate by the municipalities. This chapter reports the results of a research conducted to compare the prediction accuracy of the previous three most used models: (i) multiple regression model (MRM), (ii) additive nonparametric regression (ANR), and (iii) artificial neural network (ANN). The three models were developed using the housing database of a town with 33,342 residential houses. In this database, the cutoff point for higher priced homes was $88 per square foot of living area.

The research confirmed that using statistical and artificial neural network models are reliable and cost effective methods for mass appraisal of residential housing. It was found that any of the three models can be used, with similar accuracy, for lower and medium priced houses, but the artificial neural network (ANN) is considerably more accurate for higher priced houses.

In the research presented in this paper, back-propagation neural network is used for housing price prediction, and for comparing its accuracy with the traditional Multiple Regression Model and the Additive Nonparametric Regression. This research was motivated by the current trend towards the wide application of statistical technique in the mass appraise of residential real estate. Artificial neural network and additive nonparametric regression are now competing with the more traditional multiple regression model for adoption as a preferred mass valuation technique. The purpose of this research was to measure the estimation errors associated with these three

methods of mass valuation by applying the three techniques to the housing data of the Town of Amherst, State of New York. This town has 33,342 residential properties.

The format of the chapter is constructed as follows. Section 2 contains a description of the models development process and the results of empirical analysis. The final section 3 provides the concluding remarks arrived in this research on the estimation accuracy of the three models.

7.2 Prediction Accuracy Comparison of the Three Models

Multiple regression model was first developed as the benchmark model to estimate housing prices. Figure 7.1 represents the flow chart for the three models accuracy comparison development process.

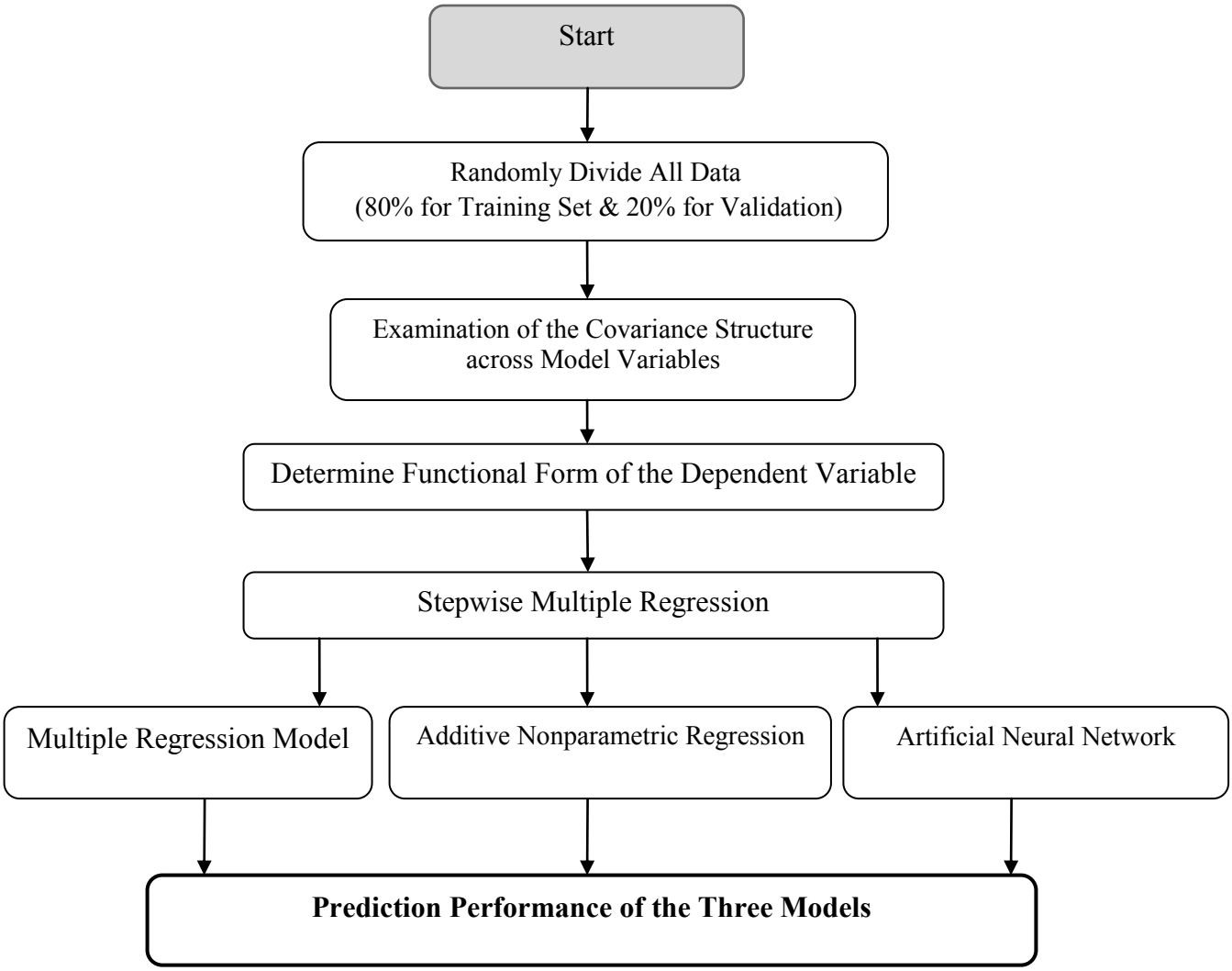


FIGURE 7.1: Model Comparison Development Process

7.2.1 Comparison of the Prediction Performance Accuracy of the Three Models

The comparison of the prediction accuracy of the three models was the goal of this research; therefore 20% of the data was utilized for models validation. For the comparison of the prediction accuracy, three widely accepted measures were utilized:

(i) the root mean squared error (RMSE);

(ii) the mean absolute error (MAE); and

(iii) the Theil's U statistic.

These three statistics were applied to the 80% data and the 20% validation data.

The root mean square error (RMSE) is the square root of the average of the squared values of the prediction errors and weights large errors more heavily than small errors. The root mean squared error (RMSE) is defined as below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t}\left(y_t - \hat{y}_t\right)^2} \qquad (7)$$

where y_t is the actual housing price, and $\hat{Y}_t$ is the fitted price from the regression model.

The mean absolute error (MAE) is the average of the absolute values of the prediction errors and is given by:

$$MAE = \frac{1}{n}\sum_{t}\left|y_t - \hat{y}_t\right| \qquad (8)$$

Theil's U statistic is the square root of the ratio of the mean squared error of the predicted change to the average squared actual change. It is defined as:

$$U = \sqrt{\frac{\sum_{t}\left(y_t - \hat{y}_t\right)^2}{\sum_{t}(y_t)^2}} \qquad (9)$$

For all of the three criteria, a value closer to zero indicates a better fitting model.

Table 7.1 shows the comparison of the training set and the validation set prediction accuracies, in terms of the above three errors, for the three models.

TABLE 7.1: Comparative Accuracies of the Three Models

	Measure of Accuracy	Multiple Regression Model	Additive Nonparametric Regression	Artificial Neural Network
	Training Set			
(a)	MAE	15,504	14,580	***12,213***
	RMSE	27,027	25,844	***20,687***
	Theil's U Statisitc	0.132	0.126	***0.101***
(b)	Percent within 10%	70.8%	73.9%	***79.3%***
	Percent within 20%	91.9%	92.6%	***94.6%***
	Validation Set			
(c)	MAE	15,997	14,934	***13,107***
	RMSE	29,284	27,385	***26,182***
	Theil's U Statisitc	0.141	0.132	***0.126***
(d)	Percent within 10%	69.6%	72.7%	***77.7%***
	Percent within 20%	91.2%	92.7%	***94.1%***

The upper part (a) of Table 7.1 illustrates the comparison of the training set prediction accuracies. The artificial neural network (ANN) results in significantly smaller prediction errors than the multiple regression model (MRM) and the additive nonparametric regression model (ANR) in the MAE, RMSE, and Theil's U statistic comparisons. The artificial neural network has about 21% lower MAE, 23% lower RMSE, and 23% lower Theil's U statistic than those of the multiple regression model.

Part (b) of Table 7.1 shows the comparison using another criterion, commonly used to evaluate housing price forecasting models. This criterion measures percent predicted values within 10% or 20% of the observed prices (Thibodeau, 2003). For the multiple regression model (MRM), 70.8% of predicted values are within 10% of the assessed prices and 91.9% are within 20%. For the additive nonparametric regression model (ANR), 73.9% of predicted values are within 10% of the assessed prices and 92.6% are within 20%.Using the artificial neural network (ANN) increases the within 10% predicted VALUES from 70.8% to 79.3% and within 20% predicted values from 91.9% to 94.6%. Again, the results clearly demonstrate the highest accuracy of the artificial neural network in predicting housing prices.

The second half of Table 7.1 compares the validation set price prediction accuracies. Similar to the results obtained from the training set, the MAE, RMSE, and Theil's U statistic are lower in the additive nonparametric model compared to the multiple regression model, and markedly decrease in the artificial neural network. Also, the number of houses within 10% and 20% of the observed assessed prices are notably higher in the artificial neural network model.

To sum up, the results show that artificial neural network model (ANN) outperforms the multiple regression model (MRM) and the additive nonparametric regression model (ANR) in both the training set and the validation set price predictions. Moreover, while the additive nonparametric regression slightly outperformed the multiple regression model, the difference between the two models was minimal. The artificial neural network model (ANN) outperforms both of the models by a significant margin due to the many interactive terms between the independent variables underline the ANN architecture. These results lead to the conclusion that the artificial neural network can be used for accurately predicting the housing prices in a municipality.

7.2.2 Prediction Accuracy by Neighborhood

The above section focused on the model prediction accuracies for the total population of houses. This section further investigates each model's prediction accuracy for each of the 67 neighborhoods. In this section, the 33,342 housing prices estimated by the three models are compared with their actual 2009 assessed values for each neighborhood code. Table 7.2 gives the mean absolute error (MAE) of the three models for each neighborhood with its average housing price per square foot of living area (*sfla*). Table 7.2 statistic revealed that: when applying the artificial neural network, the mean absolute error (MAE) of neighborhoods: 701, 2500, 4600, 4700, 4800, 6200, and 6300 is remarkably lower as compared to the results obtained from the multiple regression model and the additive nonparametric regression model. On further investigation, it was found that the average housing price per square foot of living area for all of these neighborhoods have higher housing prices of $88.00 per square foot of living area.

TABLE 7.2: Mean Absolute Error (MAE) of Three Models for Each Neighborhood

			Mean Absolute Error (MAE)			
nghdcode	Number of Houses	$^{hprice}/_{sfla}$ ($)	Multiple Regression Model	Additive Nonparametric Regression	Artificial Neural Network	% Difference in MAE between Model 1 and 3
1	2	3	4	5	6	7
3900	345	54.89	11,445	9,655	5,026	-56%
4200	133	61.44	11,322	10,639	10,502	-7%
3100	454	65.14	10,063	9,507	7,415	-26%
3700	1,657	65.32	10,987	10,343	9,415	-14%
4500	84	66.5	34,737	33,263	32,783	-6%
5500	87	68.8	17,328	17,698	17,218	-1%
300	127	69.7	22,998	22,104	18,190	-21%
5000	452	70.36	13,194	12,842	10,763	-18%
5400	694	70.49	11,878	10,081	8,815	-26%
3800	1,748	70.63	10,195	9,978	8,837	-13%
5200	1,933	71.18	15,111	15,236	14,534	-4%
3200	1,749	71.67	11,158	10,398	8,807	-21%
2800	1,312	72	11,450	10,356	7,027	-39%
2600	427	73.08	14,139	13,827	12,840	-9%
5600	712	73.32	13,130	11,436	10,272	-22%
5800	1,078	73.64	13,085	12,754	10,916	-17%
3500	649	74.56	9,720	8,282	7,548	-22%
4400	594	74.59	23,125	20,349	17,428	-25%
1700	176	75.25	20,292	21,197	18,742	-8%
1401	134	75.83	11,012	10,864	9,134	-17%
100	496	76	23,089	18,296	18,487	-20%
5300	359	76.2	9,089	8,126	6,246	-31%
2900	438	76.41	12,341	11,558	10,834	-12%
700	345	76.43	21,205	18,892	11,618	-45%
6400	629	76.8	8,748	7,572	5,908	-32%
3300	13	76.89	20,448	14,157	12,705	-38%
3600	1,317	76.93	10,288	9,416	8,965	-13%
1400	172	77.78	8,887	8,523	7,959	-10%
200	230	78.17	20,806	18,016	16,894	-19%
3400	551	78.23	17,990	18,833	14,221	-21%
400	1,532	78.54	9,695	8,957	8,114	-16%
6100	672	78.71	16,924	16,347	14,933	-12%
1000	556	78.78	9,379	8,520	8,527	-9%
4300	609	78.9	18,410	19,208	16,175	-12%
6500	399	79.38	9,769	9,334	8,076	-17%

TABLE 7.2: - continued

nghdcode	Number of Houses	$^{hprice}/_{sfla}$ ($)	Mean Absolute Error (MAE)			
			Multiple Regression Model	Additive Nonparametric Regression	Artificial Neural Network	% Difference in MAE between Model 1 and 3
1	2	3	4	5	6	7
500	1,089	79.58	14,734	11,887	10,314	-30%
1800	415	79.59	9,342	9,363	8,415	-10%
5900	432	81.45	21,252	19,806	18,931	-11%
6000	540	81.74	19,988	17,943	15,858	-21%
1200	1,115	81.99	9,002	8,169	6,681	-26%
2100	301	82	14,302	12,171	9,481	-34%
2700	255	82.77	12,853	11,700	10,320	-20%
1900	277	83.13	10,483	11,058	9,779	-7%
1500	341	83.35	8,492	8,040	7,056	-17%
900	279	83.59	9,449	9,623	7,629	-19%
2300	496	83.61	16,909	16,764	14,527	-14%
4000	386	84.07	13,436	13,260	11,724	-13%
800	320	85.63	9,474	8,060	6,890	-27%
5100	378	86.07	26,259	25,944	21,715	-17%
4900	560	87.58	35,243	34,735	28,148	-20%
1300	102	88.02	16,956	15,066	14,275	-16%
600	176	89.85	14,494	15,432	8,807	-39%
4700	155	90.79	60,308	56,289	41,652	-31%
3000	254	91.58	14,105	12,991	10,396	-26%
2000	572	92.61	18,343	16,680	14,694	-20%
1600	175	93.45	25,049	24,159	16,267	-35%
1100	472	94.43	9,493	8,518	8,309	-12%
2400	155	94.8	43,080	44,176	39,302	-9%
4100	68	97.3	39,015	37,346	32,642	-16%
2200	241	99.27	29,695	28,057	27,861	-6%
701	311	99.35	28,416	25,741	15,806	-44%
4800	117	106.48	92,143	87,105	74,795	-19%
4600	46	107.05	56,252	52,582	44,795	-20%
5700	13	109.74	19,531	20,730	16,038	-18%
2500	195	115.19	72,557	70,854	55,571	-23%
6200	4	118.1	132,704	119,459	21,625	-84%
6300	239	132.74	85,898	77,673	43,441	-49%

Figure 7.2 shows the mean absolute error (MAE) of the three models against the average housing price per square foot of living area. This figure shows that the estimates of multiple regression model have almost similar prediction accuracy as additive nonparametric regression and the artificial neural network for houses costing less than $88.00 per square foot of living area.

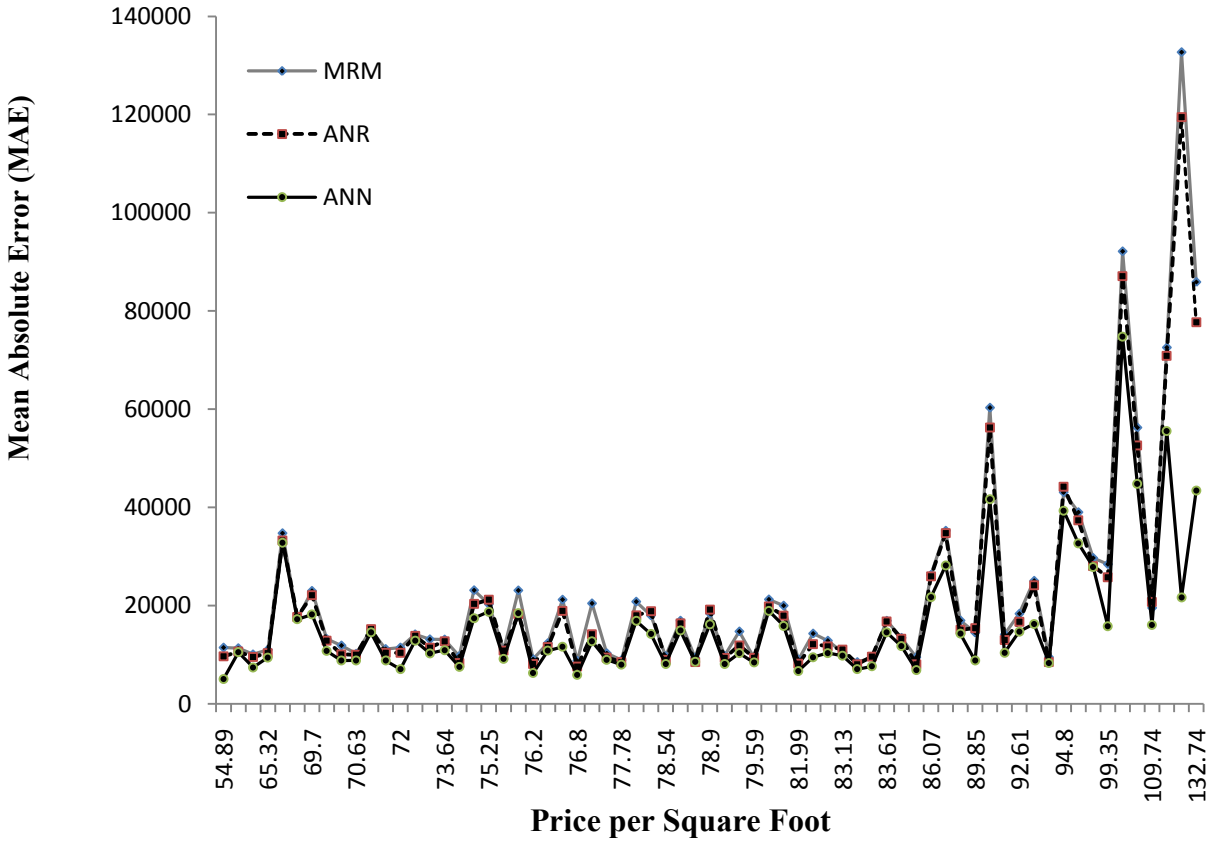


FIGURE 7.2: Model Prediction Accuracy in MAE versus Price per Square Foot for the Three Models Compared

Table 7.3 gives the recommendation about the optimal mass appraisal methodology for each neighborhood. It shows that:

(i) For houses costing more than $88.00 per square foot of living area, the artificial neural network should be considered, and

(ii) For houses costing less than $88.00 per square foot of living area, all three mass appraisal methodologies provide almost the same prediction accuracy.

TABLE 7.3: Recommendation of Mass Appraisal Methodology

Any one of the three Models				Artificial Neural Network	
nghdcode	$^{hprice}/_{sfla}$ **($)**	***nghdcode***	$^{hprice}/_{sfla}$ **($)**	***nghdcode***	$^{hprice}/_{sfla}$ **($)**
100	76	3400	78.23	600	89.85
200	78.17	3500	74.56	701	99.35
300	69.7	3600	76.93	1100	94.43
400	78.54	3700	65.32	1600	93.45
500	79.58	3800	70.63	2000	92.61
700	76.43	3900	54.89	2200	99.27
800	85.63	4000	84.07	2400	94.8
900	83.59	4200	61.44	2500	115.19
1000	78.78	4300	78.9	3000	91.58
1200	81.99	4400	74.59	4100	97.3
1300	88.02	4500	66.5	4600	107.05
1400	77.78	4900	87.58	4700	90.79
1401	75.83	5000	70.36	4800	106.48
1500	83.35	5100	86.07	5700	109.74
1700	75.25	5200	71.18	6200	118.1
1800	79.59	5300	76.2	6300	132.74
1900	83.13	5400	70.49		
2100	82	5500	68.8		
2300	83.61	5600	73.32		
2600	73.08	5800	73.64		
2700	82.77	5900	81.45		
2800	72	6000	81.74		
2900	76.41	6100	78.71		
3100	65.14	6400	76.8		
3200	71.67	6500	79.38		
3300	76.89				

Note: This table relates to Figure 6-7 where it can be more clearly seen that those neighborhood with average price per square foot of living area less than $88 show almost equal MAE error, as those more than $88 per square foot of living area end up in less MAE by using Artificial Neural Network Methodology.

7.3 Summary of Comparative Statistics

All municipalities have divided their residential real estate into several neighborhoods based on either the prices of homes or its zoning. They periodically conduct mass appraisal of their real estate, mostly by hiring appraisers who estimate the prices of properties in each neighborhood using the sales data of the last few years. This process is expensive, and is subject to human errors, and is also subject to biased appraisals. With a view to avoid these subjectivities and to confirm the reliability of the computer models for cost-effective appraisals, a research was conducted to compare the prediction accuracy of the computer models, and artificial neural network.

The research developed three models: (i) multiple regression model (MRM), (ii) additive nonparametric regression (ANR), and (iii) artificial neural network (ANN) for comparing the prediction accuracy of housing prices using the housing stock of 33,342 residential houses. The results obtained from this research clearly showed that:

1) The variables that significantly influence the housing price include: frontage width, parcel depth, age, square foot of living area, building style, and neighborhood.
2) The statistical models and an artificial neural network model using computers can be used for mass appraisal with accurate results.
3) Any of the three models can be used with similar prediction accuracy for lower and medium priced houses. For the municipality that provided the data of their 33,342 houses, this limit was the houses costing less than $88 per square foot of living area, and

4) For higher priced houses, the artificial neural network (ANN) model gives higher accuracy than the two statistical models due to the many interactive terms between the independent variables underline the ANN architecture.

CHAPTER 8: The Impact of Macroeconomic Indicators on Regional Housing Prices

8.1 Introduction

The aim of this chapter is to quantify the effects of major macroeconomic indicators: Oil Price (*OIL*), 30-year Mortgage Interest Rate (*IR*), Consumer Price Index (*CPI*), Dow Jones Industrial Average (*DJIA*), and Unemployment Rate (*UR*), on the regional housing prices, over time. Vector Autoregression (VAR) was used to analyze the time variation of the housing prices, over time, and their interaction with the macroeconomic indicators. The analyses used monthly housing sales data for the Town of Amherst, State of New York, for the period: 1999 – 2008. The various analyses concluded that the 30-year mortgage interest rate has the highest effect on the housing price ranging from 4.97 percent in the first month to 8.51 percent in the twelfth month. The unemployment rate was next in order followed by Dow Jones Industrial Average, and Consumer Price Index. The total effect of these five macroeconomic indicators ranged from 7.3% in the first month to 25.5% in the twelfth month. The conclusions arrived in this chapter, along with several related tables and figures will be useful to the housing community and the real estate companies in the region in planning their business for the next years.

The goal of this paper is to empirically investigate the changes in housing prices at the regional level with changes in: (i) the global macroeconomic indicator: Oil Price; (ii) national indicators: 30-year Mortgage Interest Rate, Consumer Price Index, and Dow Jones Industrial Average; and (iii) the local factor: Unemployment Rate.

Prices of houses sold in the Town of Amherst, State of New York during 1999 to 2008 have been used for developing the regional model. In modeling the impact of macroeconomic variables on

housing prices, a regional vector autoregressive (VAR) model is used to capture the full interaction of the housing sector with the rest of the economy. This statistical tool, VAR, has proven useful in analyzing time-series data. The global economic variables act as driving variables for the national and regional economic variables while the national economic variables drive the regional variables.

This chapter has been organized into five major sections. Section 2 introduces several major property market data resources, and outlines the data used in this dissertation. Section 3 discusses the methodology of the Unit Root Test and the Vector Autoregression (VAR) model to test the stationary and the optimal lag length order of the data series. It also presents the use of the Cointegration Test to test the long run equilibrium. Section 4 represents the empirical results, and the last section 5 provides the summary and conclusions.

8.2 Data Description

The model was formulated using a mix of regional economic variables, national economic variables, and global economic variable. The empirical analysis was carried out using monthly data from 1999 to 2008 for each economic variable. Figure 8.1 gives a schematic representation of the flow of causation from global economic variable to national economic variables and regional economic variables. Then, all national economic variables in turn feed into the regional economic variables.

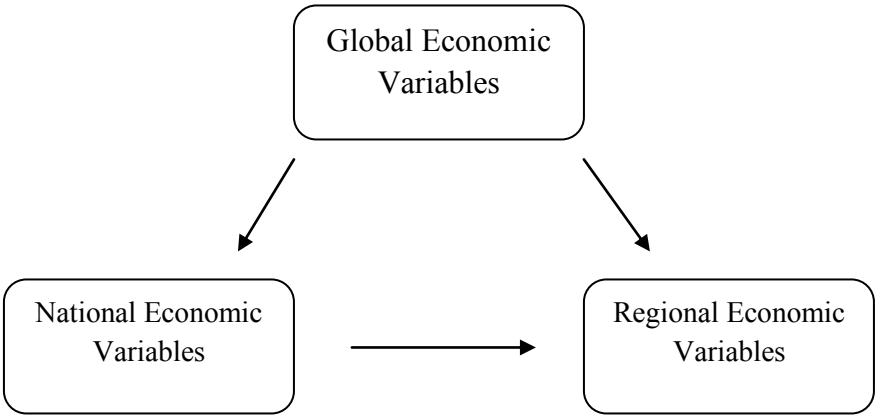


FIGURE 8.1: Influences on Regional Economic Variables

The definition of variables used is described as follows. Figures 8.2 ~ 8.7 show time series data of all variables, used in this research, in their original form, for the period 1999 to 2008.

1. Global Economic Variable:
    a. *OIL*: Crude Oil Price is measured in the US dollars per barrel. The monthly price data are calculated by taking the natural log of the crude oil price. The data were obtained from the U.S. Energy Information Administration.

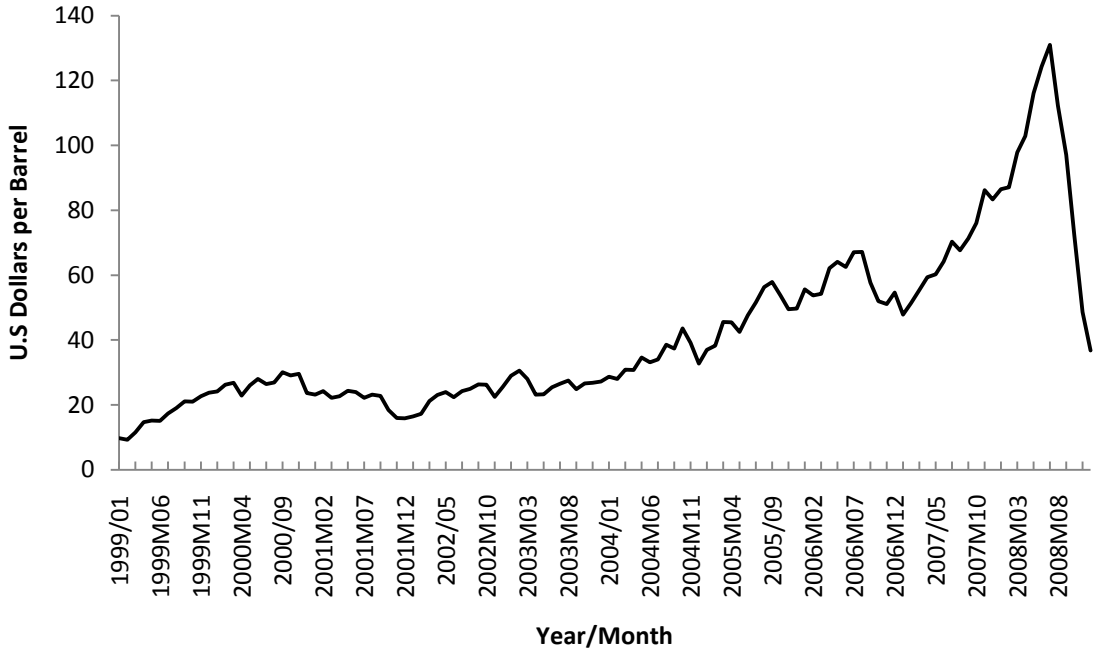


FIGURE 8.2: Time Series Data for U.S. Crude Oil Price per Barrel, 1999 – 2008

2. National Economic Variables:

    a. *IR*: 30-Year Fixed Mortgage Annual Interest Rate (%). The data were provided by the Federal Reserve. This variable captures the cost of borrowing to the household for house purchase. The annual mortgage interest rate is taken as a percent.

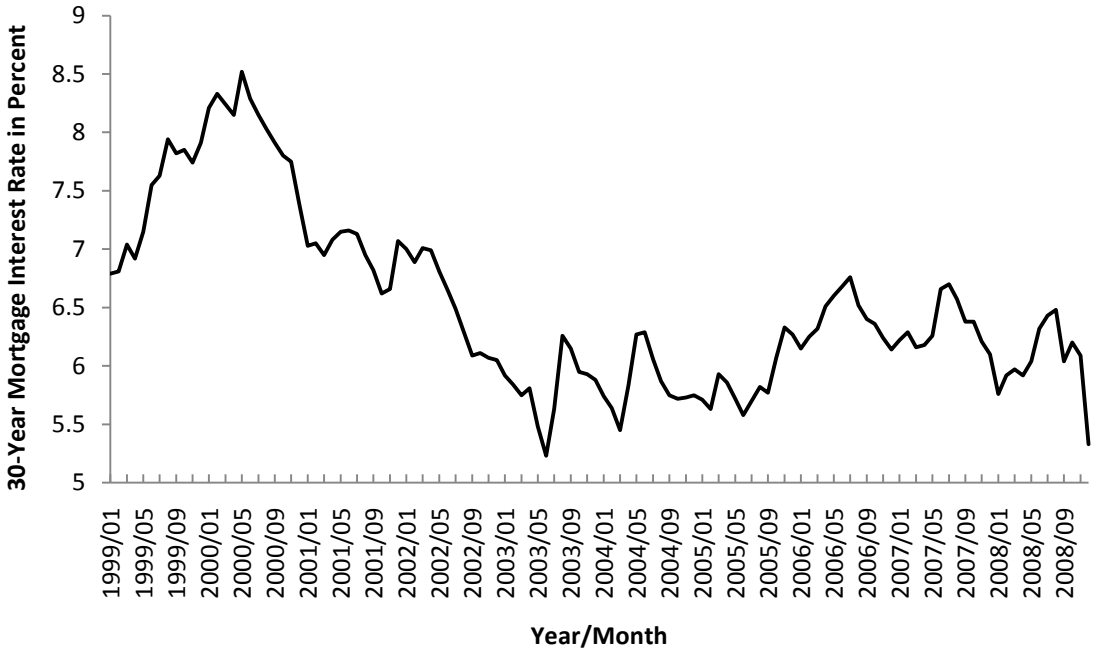


FIGURE 8.3: Time Series Data for 30-Year Fixed Mortgage Interest Rate (%), 1999 -2008

b. *CPI*: Consumer Price Index. These data were obtained from the Bureau of Labor Statistics. This variable was transformed into logarithms for the data analysis. The number 100 is assigned to the base period 1982 – 84.

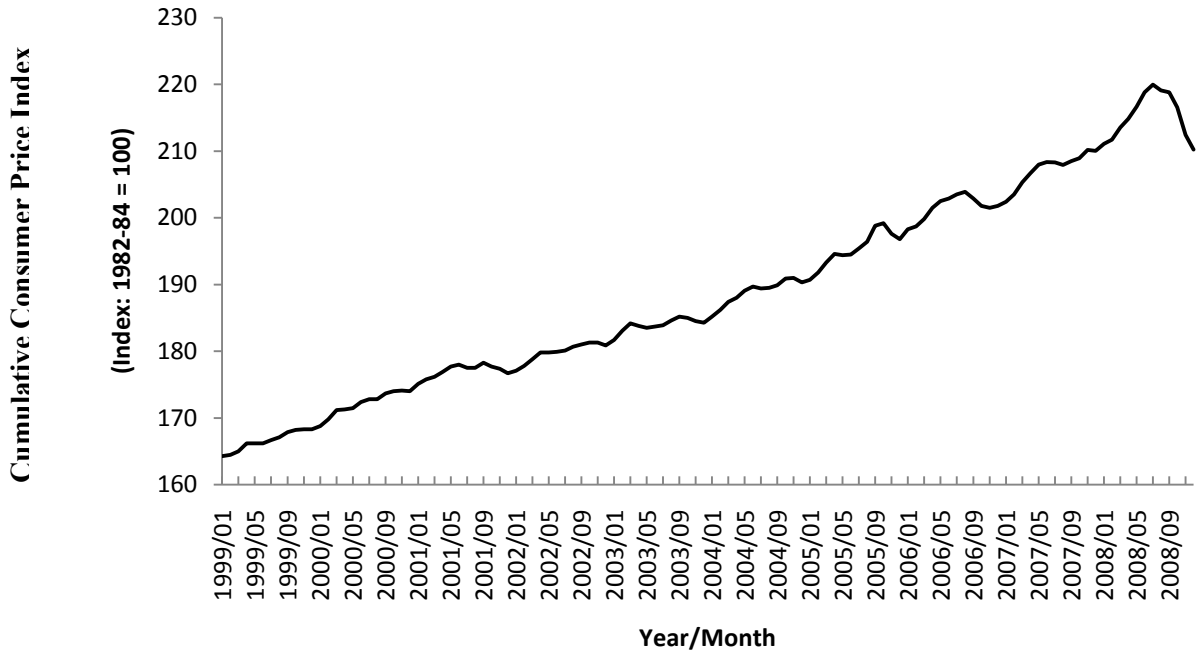


FIGURE 8.4: Time Series Data for Consumer Price Index, 1999 - 2008

c. *DJIA*: Dow Jones Industrial Average. It represents the stock price index. The closed prices are taken to represent this index. This variable was used in logarithm format in the data analysis.

FIGURE 8.5: Time Series Data for Dow Jones Industrial Average, 1999 – 2008

3. Regional Economic Variables:

    a. *UR*: Unemployment Rate. These data were provided by the Bureau of Labor Statistics. The unemployment rate is not taken in logarithm format as it is expressed in percentages.

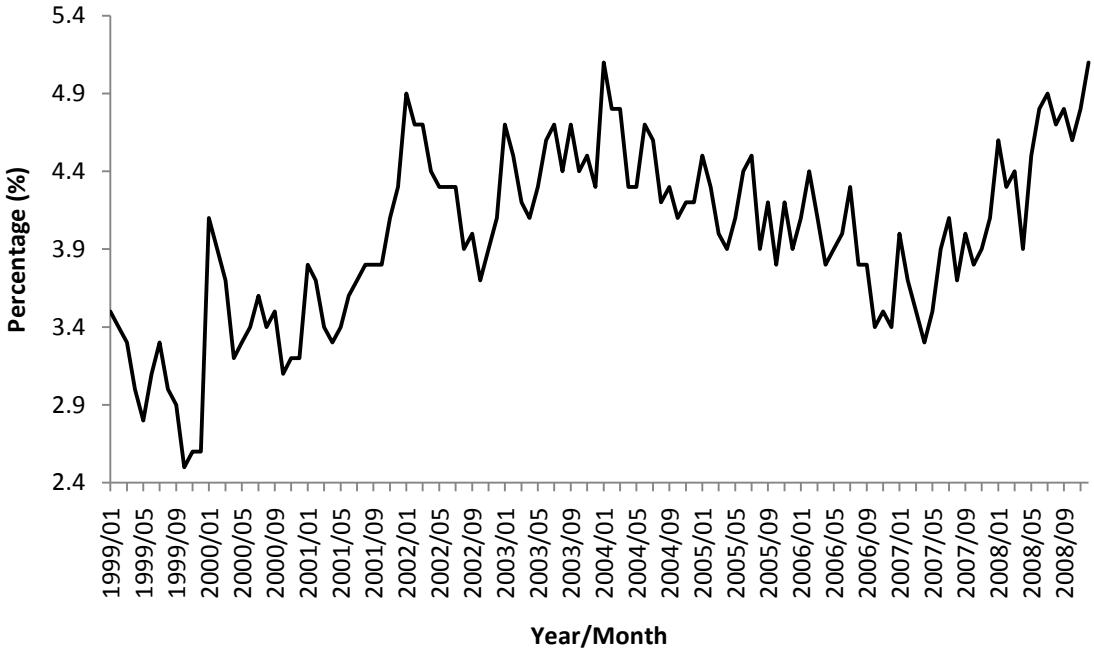


FIGURE 8.6: Time Series Data for Unemployment Rate, 1999 – 2008

b. *HPI*: Nominal Housing Price Index. The housing price indices are the median sale prices of single-family houses sold in the Town of Amherst, State of New York. The data set used in estimation and testing consists of monthly observations from the Amherst Town, and the sample period spans from January of 1999 to December of 2008. In Figure 8.7, the housing price indices have increased considerably during the sample period and only just begun to decline after 2007. The housing price index has been used in the logarithmic form in the analysis at data.

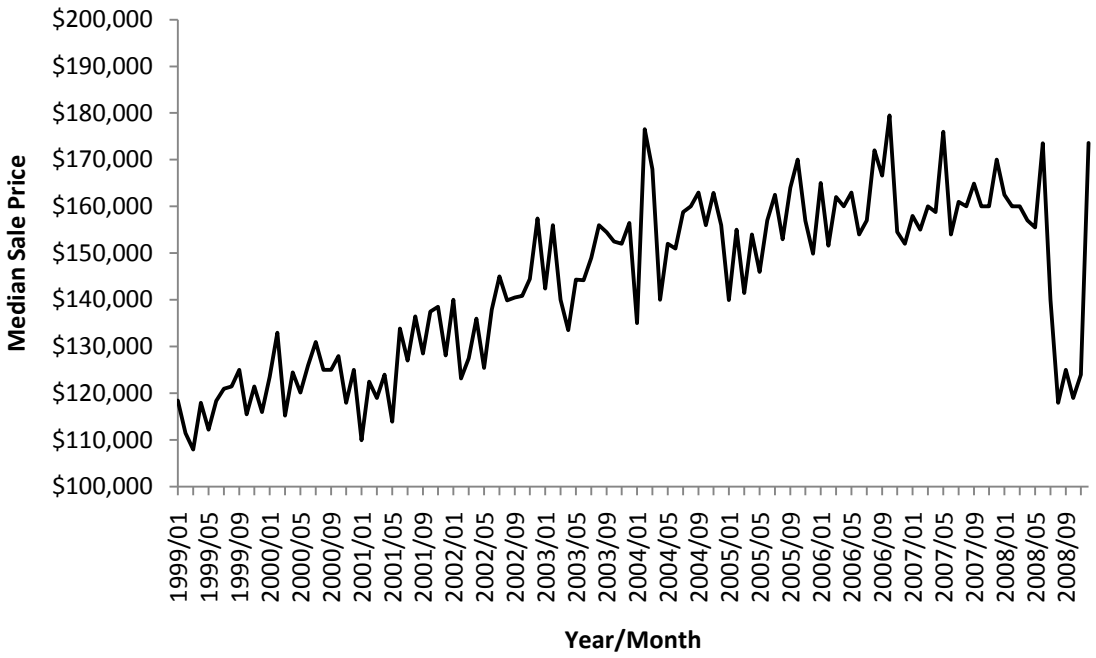


FIGURE 8.7: Time Series Data for Nominal Housing Price Index, 1999 - 2008

8.3 The Vector Autoregression (VAR) Method

This section is divided into two subsections. The first subsection provides a general overview of the vector autoregression (VAR) approach to time series analysis. The second subsection discusses the preliminary diagnostic steps needed to formulate the VAR estimation.

8.3.1 Overview of Vector Autoregression (VAR) Model

Traditionally, researchers have used models which impose a priori restrictions on the coefficients to analyze the impact of macroeconomic variables on housing prices. However, the Vector Autoregression (VAR) is a dynamic model of time series that allows the data, rather than the researcher, to specify the dynamic structure of the model.

VAR is a set of symmetric equations in which each variable is described by a set of its own lagged values, and the current and past lagged values of all other variables in the system. The

lagged value of a variable is simply the value that the variable took during a previous period. For example, the value of y_t lagged one period is written as y_{t-1}. Following Baffoe-Bonnie (1998), an n variable VAR system can be written as:

$$\Gamma(\phi)X_t = \Gamma + V_t \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (1)$$

and

$$\Gamma(\phi) = I - \Gamma_1\phi - \Gamma_2\phi^2 - \cdots - \Gamma_m\phi^m \qquad \ldots\ldots\ldots\ldots\ldots\ldots \qquad (2)$$

where: X_t is an $n \times 1$ vector of variables;
Γ is an $n \times 1$ vector of constants;
V_t is an $n \times 1$ vector of random variables, each of which is serially uncorrelated with constant variance and zero mean.

Equation (2) is an $n \times n$ matrix of a normalized polynomial in the lag operator $\phi\left(\phi^k X_t = X_{t-1}\right)$ with the first entry of each polynomial on Γ's being unity.

The appropriate estimation technique is the OLS (Ordinary Least Squares). However, before estimating using the model, the number of preliminary testing are applied such that each section has a number of lagged values in each equation.

8.3.2 Preliminary Testing

Before the results of the empirical analysis are presented, it is first important to introduce three preliminary diagnostic tests that are commonly employed in VAR analyses—namely, unit root tests, lag length specification, and cointegration tests. All of these issues are discussed in the following sections.

8.3.2.1 Integration Analysis*:* Unit Root Tests for Stationary

Testing for non-stationary time-series in the form of unit root has become a staple of time series econometrics. A stationary time series is significant to a regression analysis because useful information or characteristics are difficult to identify in a non-stationary time series. Moreover, it is well established that OLS produces spurious results when applied to non-stationary data. What OLS is really estimating in such situations are common trends and not the underlying relationships between two or more variables.

A time series is said to be stationary if its mean and variance are constant and the covariances depend on the distance of two time periods. In practice, most economic time series are non-stationary. By differentiating the data, time series can be induced to be stationary. Useful information can still be recognized in the data after differentiating. The first difference of y, also known as the change in y, and denoted Δy_t, is calculated as the difference between the values of y in this period and its value in the previous period. This is calculated as:

$$\Delta y_t = y_t - y_{t-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

where: the symbol, Δ, in tables, figures, or text stands for the first difference series.

All variables were tested for unit root non-stationary by using Augmented Dicky-Fuller (ADF) Test (Dickey and Fuller, 1979) at the level form and their first differential of data series. Two forms of the Augmented Dickey-Fuller (ADF) test were conducted:

Intercept and no trend: $\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta y_{t-i} + \varepsilon_t$ …….. (4)

Intercept and with trend: $\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=1}^{m} \beta_i \Delta y_{t-i} + a_2 t + \varepsilon_t$ …… (5)

where: Δy_t is the first-difference of the variable of interest y_t;
a_0 is the intercept;
and t is a deterministic trend.

The augmented ADF test determines whether to accept or reject the null hypothesis. If the null hypothesis of $\gamma = 0$ is accepted, the time series has the unit root. That is, the time series is nonstationary. Equation 4 comprises an intercept but no trend; this means that y is a stationary time series with a non zero mean. Equation 5 includes an intercept and a trend; this means that y is a stationary time series around a deterministic trend.

Table 8.1 reports the result of unit root tests with and without trend. The null hypothesis of non-stationary is performed at the 1% and the 5% significant levels. The result of the ADF test illustrates that all variables could not be rejected at the 5% significance level. When first differentials were used, non-stationarity was rejected at the 1% and 5% significance level, strongly supporting that all data series were stationary after the first difference.

TABLE 8.1: Six Time Series Data Unit Root Tests from 1999 M1 to 2008 M12

		ADF test at level			ADF test in first difference		
		t-statistic	Sig. level	lag	t-statistic	Sig. level	lag
No trend	***OIL***	-2.57	na	1	-7.26	**	0
	IR	-1.09	na	2	-7.65	**	1
	CPI	-0.97	na	2	-6.09	**	1
	DJI	-1.74	na	0	-10.3	**	0
	UR	-1.25	na	12	-3.72	*	11
	HPI	-2.62	na	1	-10.5	**	1
With trend	***OIL***	-2.77	na	1	-7.47	**	0
	IR	-1.99	na	2	-7.63	**	1
	CPI	-4.01	*	1	-6.07	**	1
	DJI	-1.36	na	0	-10.4	**	0
	UR	-1.74	na	12	-4.25	*	11
	HPI	-3.72	na	1	-10.5	**	1

Note: * and ** denote the rejection of the null hypothesis of unit root at 5% and 1% significant levels, respectively.

8.3.2.2 Selecting Optimal Lag Length Using Vector Autoregression Model

After testing for unit roots, the next step in the process involves specifying an appropriate lag length for the VAR estimation model. The lag length is time in months. The common problem using the VAR model is to select the optimal lag length. Often, economic theory has little information about what is an appropriate lag length for a VAR and how long changes in the variables should take to work through the VAR model.

When using VAR estimates, the decision of optimal lag length can be determined by comparing the Akaike Information Criterion (AIC) and the Schwarz Criterion (SC) (Grasa, 1989). Also, two other criteria can assist in judging the optimal lag length. These two criteria are: Final Prediction Error (FPE) and Hannan-Quinn information criterion (HQ). All of the 4 criteria are used in this study. When applying these criteria, the smallest value of these 4 criteria points to the optimal lag length.

Table 8.2 represents the results of the VAR lag length selection. The first left hand column shows the lag lengths from 0 to 3 months. The numbers with asterisks are the smallest value for each criterion. According to the Schwarz Criterion (SC), a model with no lagged month is appropriate; the Hannan-Quinn information criterion (HQ) indicates that a 1 month lag is optimal. However, as a further check with Final Prediction Error (FPE) and Akaike Information Criterion (AIC), the lag specification is 3 months. Most researchers have selected the maximum number of minimum lag length values by the various criterions. Therefore, based on the above results, the optimal lag length is considered as 3 months in the VAR model.

TABLE 8.2: Lag-Length Selection Tests

Lagged month	FPE	AIC	SC	HQ
0	5.50E-15	-15.80767	-15.66121*	-15.74825
1	2.56E-15	-16.57404	-15.54881	-16.15813*
2	2.40E-15	-16.64432	-14.74032	-15.87192
3	2.29E-15*	-16.70532*	-13.92256	-15.57644

Note: The asterisk indicates lag order selected by the criterion. The criterions include the final prediction error (FPE), the Akaike information criterion (AIC), the Schwarz information criterion (SIC), and the Hannan-Quinn information criterion (HQIC). For all four information criteria, smallest values indicate better model fit.

Then, VAR equations could be written as:

$$\Delta OIL_t = \alpha_{10} + \sum_{i=1}^{3} \beta_{11i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{12i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{13i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{14} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{15} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{16} \Delta HPI_{t-i} + e_t$$

$$\Delta IR_t = \alpha_{20} + \sum_{i=1}^{3} \beta_{21i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{22i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{23i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{24} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{25} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{26} \Delta HPI_{t-i} + e_t$$

$$\Delta CPI_t = \alpha_{30} + \sum_{i=1}^{3} \beta_{31i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{32i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{33i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{34} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{35} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{36} \Delta HPI_{t-i} + e_t$$

$$\Delta DJIA_t = \alpha_{40} + \sum_{i=1}^{3} \beta_{41i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{42i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{43i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{44} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{45} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{46} \Delta HPI_{t-i} + e_t$$

$$\Delta UR_t = \alpha_{50} + \sum_{i=1}^{3} \beta_{51i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{52i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{53i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{54} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{55} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{56} \Delta HPI_{t-i} + e_t$$

$$\Delta HPI_t = \alpha_{60} + \sum_{i=1}^{3} \beta_{61i} \Delta OIL_{t-i} + \sum_{i=1}^{3} \beta_{62i} \Delta IR_{t-i} + \sum_{i=1}^{3} \beta_{63i} \Delta CPI_{t-i} + \sum_{i=1}^{3} \beta_{64} \Delta DJIA_{t-i} + \sum_{i=1}^{3} \beta_{65} \Delta UR_{t-i} + \sum_{i=1}^{3} \beta_{66} \Delta HPI_{t-i} + e_t$$

ΔOIL_t is the first difference of natural log of the crude oil price at month t; α_{10} is an intercept; ΔOIL_{t-i} is the first difference of natural log of the crude oil price *i* months ago; finally, all Greek symbols, β, are parameters.

8.3.2.3 Cointegration Tests

Once having identified that these variables were stationary after the first difference, the possibility of cointegration among these variables was examined. Cointegration means that the economic variables share the same stochastic trend so that they are combined together in the long run. Even though economic variables deviate from each other in the short run, they tend to come back to a similar trend in the long run. In practice, most economic time series are nonstationary. If two or more variables are nonstationary and have the same order of integration, they can be constructed in a cointegration model. Therefore, the unit root test should be launched before the cointegration test. The results from the unit root test show that all variables are integrated at the first difference. The possibility of cointegration among these variables was examined. Thus, a VAR model was postulated to obtain a long run relationship. If variables are cointegrated and the

corresponding cointegration vector is not used in the VAR model, the model with only first differenced data will be misspecified. The Johansen cointegration test will reveal the evidence of cointegration. The methodology of Johansen cointegration test suggested by Johansen (1991 and 1995) and Johansen and Juselius (1990) is used to determine whether or not a stationary linear combination of a set of nonstationary series exists.

Table 8.3 is based on the Johansen cointegration test. The results report the hypothesized number of cointegration equations in the first left column, the eigenvalue, the likelihood ratio statistics and 5% critical value. The trace test statistic indicates there is no long run relationship in the VAR model. In other words, a simple VAR model is accepted instead of the vector error correction model (VECM).

TABLE 8.3: Johansen Cointegration Test

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.
None	0.295938	90.4909	95.75366	0.109
At most 1	0.15284	49.78773	69.81889	0.647
At most 2	0.135191	30.54725	47.85613	0.6906
At most 3	0.094684	13.69869	29.79707	0.8571
At most 4	0.016107	2.159977	15.49471	0.9925
At most 5	0.002379	0.276344	3.841466	0.5991

Note: Trace test indicates no integration at the 0.05 level. Trace statistic are the test statistics used to determine the existence of cointegration and, specifically, the number of cointegrating vectors.

8.4 Empirical Results

Based on the unit root and cointegration test results, the first differences of the data series were utilized in the VAR model. The six equations as specified in section 8.3.2.2 were used to examine two aspects of the data: i) Impulse Response Functions: the dynamic impact of changes in one variable on others in the model; and ii) Forecast Variance Decomposition: the amount of variance in the forecast error of one variable that can be attributed to changes in the other. These two components of the data analysis are briefly discussed as below:

8.4.1 Impulse Response Functions

Impulse responses trace out the effect of current and future values of each of the variables to one standard deviation increase in the current value of one of the VAR errors, assuming that this error returns to zero in subsequent periods and that all other errors are equal to zero. There is a corresponding impulse response for a shock to each variable in the system, which means that a system with **m** variables will have $\mathbf{m^2}$ impulse responses (including the response of each variable to its own shock). It can indicate whether the impact is positive or negative, or whether it is a temporary jump or long-run persistence. Thus, the impulse response figures provide information to analyze the dynamic behavior of a variable due to a random shock in other variables. Also, the impulse response functions can be used to predict the responses of housing price to a shock in the housing price determinants from one period to another. In the present case, this means that the response of housing price in Amherst Town to a shock in local unemployment rate can be traced over time. Specifically, one can investigate how housing price changes after they have been affected by a shock in local unemployment rate. The response of a stationary variable to a one-time shock should eventually dampen out to zero over time. Whether a response to a shock

is statistically different from zero is assessed by constructing 95% confidence intervals around a variable's time path. In line with classical hypothesis testing, a variable's response is considered statistically insignificant if the confidence interval includes zero for a given time horizon.

The ordering of the variables was made in such a way that variables that are not expected to have predictive value for other variables are put last. The estimates of the parameters are not reported since the goal of VAR analysis is to determine the dynamic interrelations among variables, not the parameter estimates.

Figures 8.8 to 8.13 depict the impulse response functions for housing price index (*HPI*) in response to changes in (1) crude oil price (*OIL*), (2) mortgage rate (*IR*), (3) consumer price index (*CPI*), (4) stock price index (*DJIA*), and (5) local unemployment rate (*UR*), respectively. The time horizon for the impulse response function is 12 months and is recorded on the X axis of each individual graph. The Y axis shows the movement trend of housing price index. The impulse response of housing price index is assumed as zero at time t = 0. And these macroeconomic variables are assumed to receive a one positive unit standard deviation shock at time t = 0. Then, the Figures 8.8 to 8.13 trace out the response of housing price index to those shocks at time t = 1, 2,…, 12, respectively. The thinner lines shown in Figures 8.8 to 8.13 represent ± 2 times the standard error bands, which yield an approximate 95% confidence interval. According to Sims and Zha (1999), the standard error bands give a more accurate summary of the central tendency of the response. The positive symbol means a favorable effect on house prices growth and a negative symbol means an adverse effect. In addition, the values shown in the figure indicate a change on the house prices movement trend.

For example, Figure 8.8 shows the responses of *HPI* to one standard deviation shock in crude oil price. In Figure 8.8, an increase in the crude oil price pushes slightly housing prices up. However, the impact of the shock is not significantly different from zero because the 95% confidence intervals for the response cross over the zero.

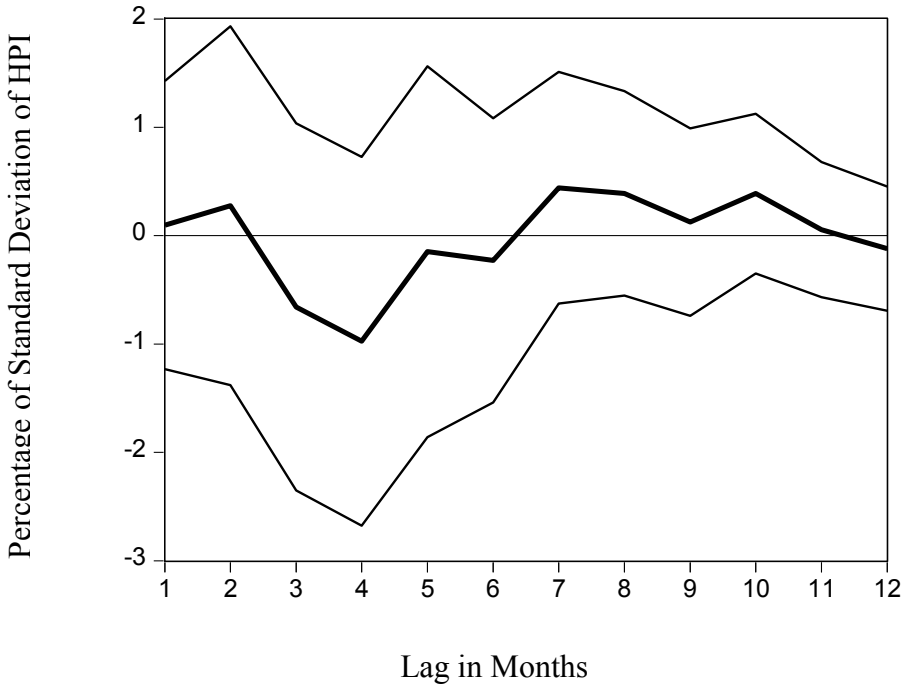


FIGURE 8.8: Response of *HPI* to *OIL*

*Think line is impulse response; thin lines represent 95% confidence intervals.

In Figure 8.9, a positive (higher) shock to the mortgage rate produces sharp cycles in housing price index. This confirms that housing prices are influenced by mortgage rates. A plausible explanation of this relationship may be that as the cost of financing a house (mortgage rate) rises, the demand for housing declines, and housing prices are likely to fall. The profound dynamic changes indicate that housing prices are sensitive to the mortgage market. A shock to the mortgage rate generates an immediate response in housing prices. Moreover, the confidence

interval never encompasses zero, which indicates that the response, while small in magnitude, is statistically significant.

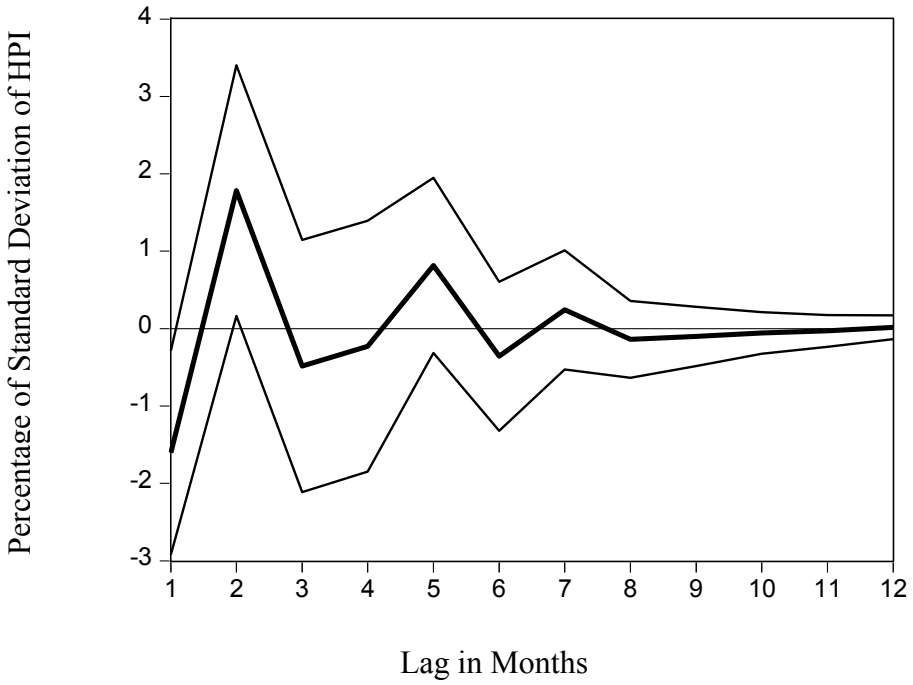


FIGURE 8.9: Response of *HPI* to *IR*

In figure 8.10, a shock to inflation does cause dynamic responses in housing prices. A shock to inflation seems to have positive effect and produce the increase in the housing prices immediately. But, it is not statistically different from zero after roughly twelve months. Because the lower confidence interval always encompasses zero, it appears that Housing price index does not significantly respond to consumer price index.

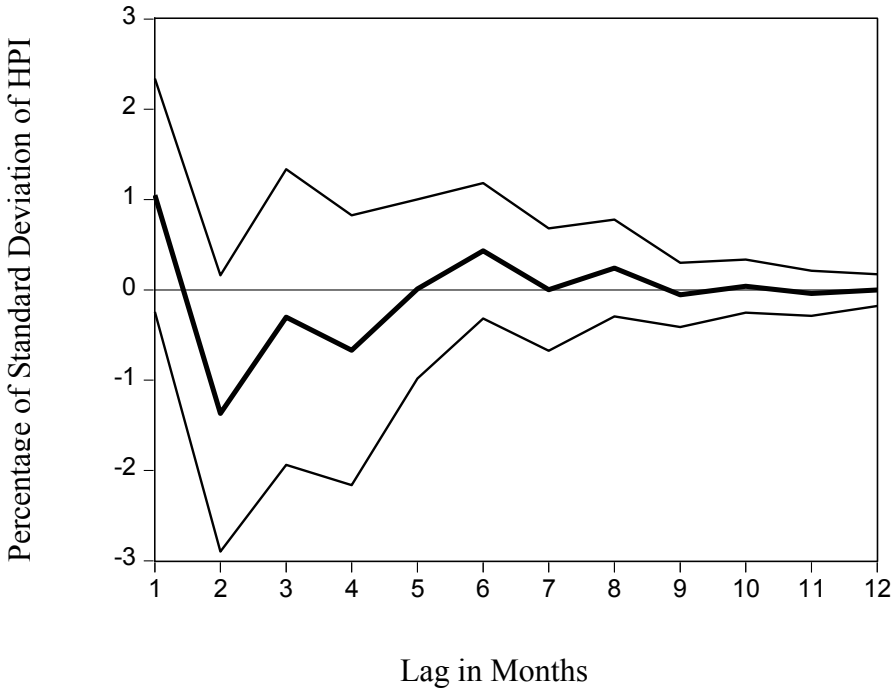


FIGURE 8.10: Response of *HPI* to *CPI*

An inspection of Figure 8.11 showed that a positive (increase) shock to stock price leads to higher housing prices as expected. The speed of adjustment is quick and they reach their steady-state level after 6 months from the occurrence of the shock.

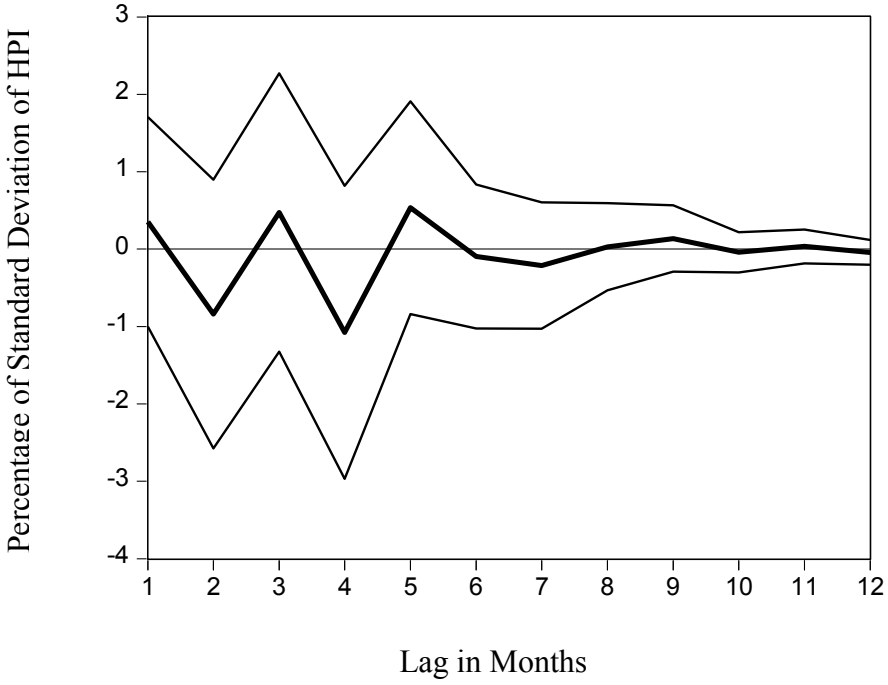


FIGURE 8.11: Response of *HPI* to *DJIA*

Inspection of Figure 8.12 reveals that a positive (increase) shock to unemployment rate tends to move housing prices in the opposite direction. In other words, a growth in the unemployment level may discourage individuals to purchase houses, and this decrease in demand seems to slightly decrease housing prices. The response quickly dies off, however, and is not significantly different from zero for most of the forecast horizon. The housing prices reached their steady state after 5 months from the occurrence of the shock.

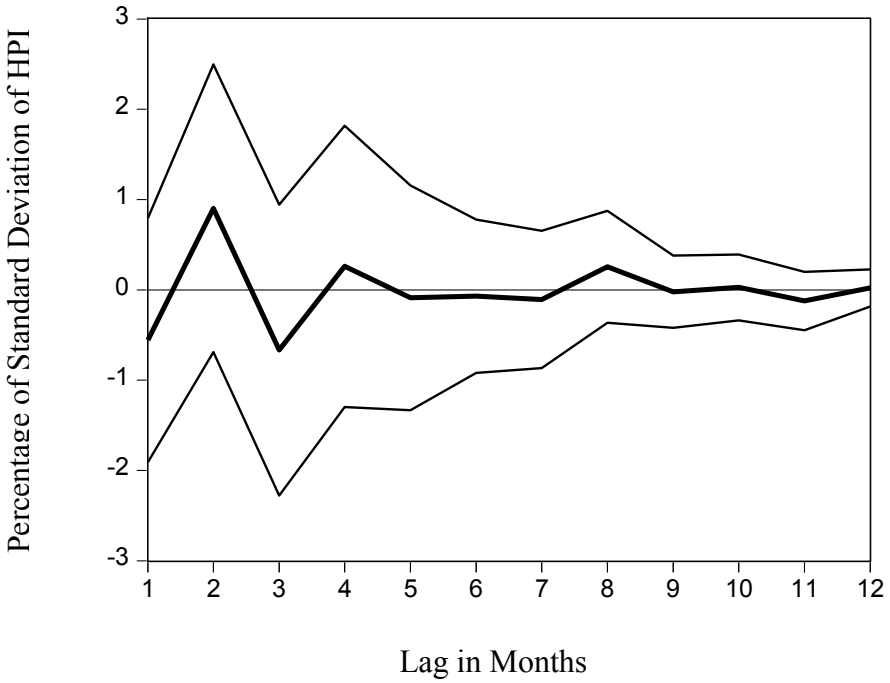


FIGURE 8.12: Response of *HPI* to *UR*

In response to a one standard deviation disturbance in current house price itself (Figure 8.13), future house prices increase in the first months. This appears to die out after 6 months, implying that the current price change has a greater impact on the next month's housing price rather than over longer term horizon.

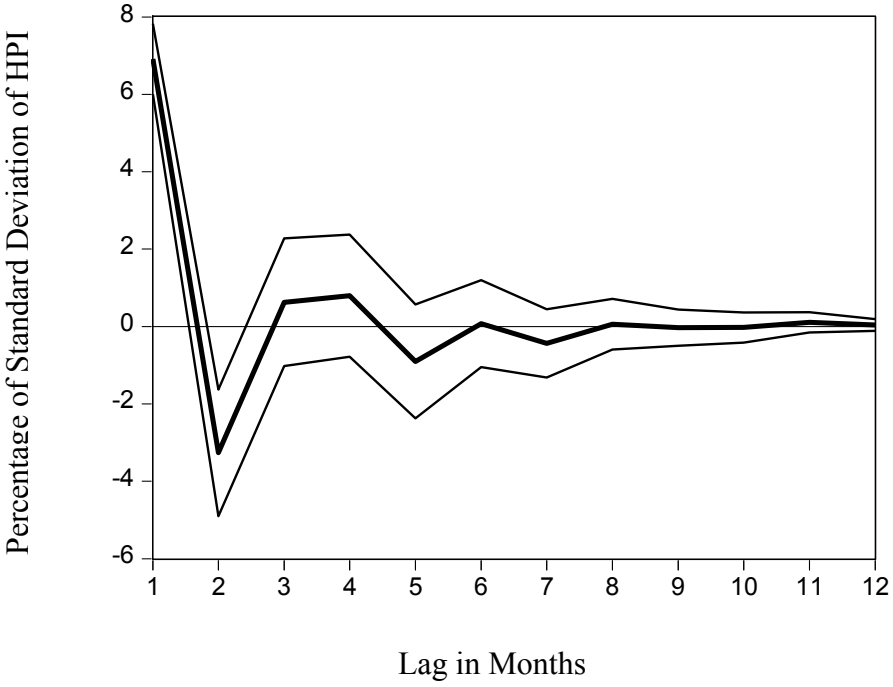


FIGURE 8.13: Response of *HPI* to *HPI*

8.4.2 Forecast Variance Decomposition

To indicate the relative importance of the shocks requires variance decomposition of the forecast errors for each variable. The variance decomposition of the forecast errors is the percentage of the variance of the error made in forecasting for each variable that is attributable to its own shocks and to shocks to other variables in the system. Thus, the forecast variance decomposition is like a partial R^2 for the forecast error. The larger the percentage of forecast error attributed to the one variable, the more important that variable is in explaining or predicting the target variable. Of key interest here is what percentage of the deviation in housing price index is due to changes in macroeconomic indicators. If, for example, local unemployment rate accounts for a high percentage of the forecast variance in housing price index, but not vice versa, this could be interpreted as evidence that local unemployment rate helps to predict the housing price index.

Given that the main focus of this paper is to investigate the influence of macroeconomic indicators on housing prices, Table 8.4 presents the variations of *HPI* which are explained by random shocks in the five macroeconomic variables at horizons up to 12 months. In practice, this decomposition is computed and reported in percentage terms so that the total amount of variation in one variable due to another can be expressed in absolute terms. The first column in Table 8.4 lists the steps in the forecast with each step corresponding to one month. Thus, the first step represents the first month of the forecast while the twelfth step represents the twelfth month. The total forecast horizon covers one year. The column 2, labeled "S.E." is the forecast error of the variable at the given forecast horizon. The next five columns in the table report the percentage of forecast variance in the *HPI* series explained by *OIL*, *IR*, *CPI*, *DJIA*, and *UR*, respectively. Because the VAR accounts for all forecast variance, each row sums to 100%.

TABLE 8.4: Forecast Variance Decomposition of *HPI*

Step	S.E.	*OIL*	*IR*	*CPI*	*DJIA*	*UR*	*HPI*
1	0.0718	0.0253	4.9741	2.1302	0.1110	0.0992	92.6602
2	0.0830	0.0263	8.3370	4.3142	0.5601	1.8951	84.8673
3	0.0856	0.2920	8.1602	4.1813	1.0233	5.9923	80.3508
4	0.0885	1.6670	7.6913	4.4793	4.5885	5.6841	75.8898
5	0.0894	1.6344	8.3738	4.3917	4.5127	5.6652	75.4222
6	0.0898	1.6622	8.4581	4.5847	4.7731	5.7622	74.7596
7	0.0900	1.6675	8.5018	4.5690	4.7627	5.7588	74.7402
8	0.0901	1.6814	8.5072	4.6312	4.7931	5.8016	74.5855
9	0.0901	1.7051	8.5110	4.6302	4.8225	5.8217	74.5096
10	0.0901	1.7105	8.5130	4.6313	4.8284	5.8212	74.4956
11	0.0901	1.7104	8.5113	4.6316	4.8323	5.8292	74.4851
12	0.0901	1.7108	8.5104	4.6309	4.83393	5.8386	74.4754

Note: presented are forecast variance decompositions after 1 to 12 months (percentage points). To calculate these variance decompositions, shocks are identified by imposing a ordering in which the *OIL* is placed first, followed by *IR*, *CPI*, *DJIA*, *UR*, and *HPI*.

The results indicate that disturbance originating from house price itself cause the greatest variability to future prices: it contributes up to 92.6 percent variability one month ahead, and approximately 74.5 percent 12 months ahead. The proportion of variance remains high (74.5 percent) even until one year (12 months). This result indicates that current change in house prices heavily influence people's expectation of future prices changes.

Despite a 74.5 percent variability contributed by current price changes, there remains 25.5 percent of the variability which is explained by other factors. The results suggest that shocks to the mortgage rate and unemployment rate account for more variation in housing prices than variations produced by shocks to the crude oil price, consumer price index, and stock price index.

Mortgage rate prevails over all other five house price determinants in influencing house prices. A shock to the mortgage rate explains about 8.51% of the variation in housing prices over 12 months. It accounts for approximately 33 percent of the total variance contributed by the five determinants.

The third largest source of house price variance appears to be from unemployment rate. A shock to unemployment rate explains 5.84% of the variation in housing prices, which accounts for approximately 23 percent of the total variance contributed by the five determinants.

Apart from these three determinants, the two remaining variables account for less than 10 percent of housing price variance. The corresponding variations in housing prices due to a shock in consumer price index and stock price index are 4.63% and 4.83%, respectively.

The final variable in the model, crude oil price, contributes very little to house price variance (1.71 percent of the total variance). The results suggest that the global economic variable doesn't impact the regional housing prices directly. But, it goes through the national economic variable to indirectly influence the housing prices.

8.5 Results of Analyses

The main objective of this study was to quantify the impact of global, national, and local variables on housing prices at the regional level. The global macroeconomic indicator was: Oil Price; national indicators were: 30-year Mortgage Interest Rate, Consumer Price Index, and Dow Jones Industrial Average; and local factor was: Unemployment Rate. The analyses used monthly housing sales data for the Town of Amherst, State of New York, for the period: 1999 – 2008. And, to analyze the time-series data of housing sales, the VAR methodologies were employed

including unit root tests, Johansen's cointegration test, impulse response functions and variance decompositions. These techniques appear to be effective.

The various analyses concluded that the 30-year mortgage interest rate has the highest effect on the housing price ranging from 4.97 percent in the first month to 8.51 percent in the twelfth month. The Unemployment Rate was next in order followed by Dow Jones Industrial Average, and Consumer Price Index. The total effect of these five macroeconomic indicators ranged from 7.3% in the first month to 25.5% in the twelfth month. The conclusions arrived in this paper, along with several related tables and figures will be useful to the housing community and the real estate companies in the region in planning their business for the next years.

CHAPTER 9: Housing Price Diffusion Patterns at Local Level: an Examination of Housing Markets at Amherst Town, New York

9.1 Introduction

The ripple effect of housing prices within metropolitan areas has been recognized by recent research. That is, housing price change in one area would reflect in housing prices change in other areas that have some certain relationship among them. This study explores the ripple effect of housing prices between different housing sizes at the regional level. The question of interest is whether housing price changes in one size of housing in terms of the number of bedrooms measured can be predicted not only by their own history, but also by housing price changes in other housing sizes.

This chapter examines the interrelationship among housing price changes in different housing sizes. Therefore, causality methodology within vector autoregression was used. The results confirm that four bedroom and three bedroom housing markets have bi-directioned influence, where two bedroom housing market is exclusive. The evidence of housing price diffusion between different housing sizes was found. Moreover, from the impulse response function, the sensitivity of one housing size to the shocks of others is determined. Finally, this study may shed light on predicting house price movement trends in three different housing size markets.

Generally speaking, three market segments, determined by housing price, can be called starter or entry-level homes, trade-up, and premium housing. Two bed-room homes, often considered as an entry-level or starter home, and three bedroom houses are regarded as the trade-up homes. The houses that have the number of bedrooms more than four are thought of as premium houses. In this study, stratification techniques are used to derive these three market segments. Then,

regional housing price indices are constructed for each housing size market from the housing sale transactions from the year 1999 through 2008.

Different price segments of markets have been known to display different price behavior. Granger causality methodology (Granger, 1969, 1981) is employed to determine whether causal relationships between house prices in three market segments exist and, if yes, what is the direction and magnitude of these relationships.

The organization of this paper is as follows. In Section 2, the movement of house prices for each housing size market is examined. Section 3 presents the results of unit root tests, the discussion of methodology for Granger causality tests, and reports the results of Granger causality tests. The following section shows the impulse responses and sensitivity among three housing markets. Section 5 provides some concluding comments.

9.2 Data Description

This article examines the interrelationship between housing price changes by using the lagged values of price changes in different size houses to estimate housing price changes in a given size of houses. Table 9.1 presents the number of sale transactions from 1999 through 2008, in the Town of Amherst, State of New York. The housing markets are divided into four categories:

TABLE 9.1: Number of Sale Transactions from 1999 through 2008

Year	Number of Bedrooms inside the House				Total
	<= 2	3	4	>=5	
1999	71	580	448	45	1,144
2000	80	568	460	41	1,149
2001	70	598	448	32	1,148
2002	77	603	471	50	1,201
2003	77	667	545	58	1,347
2004	76	664	537	58	1,335
2005	126	712	533	98	1,469
2006	151	625	509	94	1,379
2007	108	673	464	91	1,336
2008	67	457	378	46	948
Total	903	6,147	4,793	613	12,456

1. Houses with number of bedrooms equal or less than two: this categorical house belongs to starter homes. The housing prices are below $120,000. In this research, these size houses are to be known as two bedroom houses.
2. Three bedroom houses: the trade-up homes range between $120,000 and $200,000.
3. Four bedroom houses: the premium homes sell above $200,000.
4. Houses with number of bedrooms equal to or larger than five: these type houses also are categorized in the premium homes.

Because of the infrequency of transactions for fourth market category, this housing type is excluded from this research. The housing price indices are derived from the median sales price of single-family residential property for each different market. The housing indices were constructed within housing transaction data supplied by the local government. Housing price indices serve as proxies to identify price behavior. The price index is based on the median sale house prices as a better indicator of general house price movements especially in cases where

data distribution is significantly skewed. Next, the price index values are transformed by taking the natural logarithm of all values. Moreover, the housing price series are nominal housing prices and not adjusted for inflation. The following time series are selected for the empirical modeling. The selection is limited by the data availability.

Figure 9.1 shows the housing price index movements of the three different housing sizes. The period is from the January 1999 to the December 2008. The indices are based on the median prices of monthly sale at the Amherst Town, New York. From Figure 9.1, the three housing types represent a similar propensity during the investigated period. They all have a slow increasing trend at first which is followed by a slight decrease after December 2006. Moreover, the variances of two bedroom housing price are more volatile than three and four bedroom houses. The following analyses are performed at the regional level, using price index values for three housing categories.

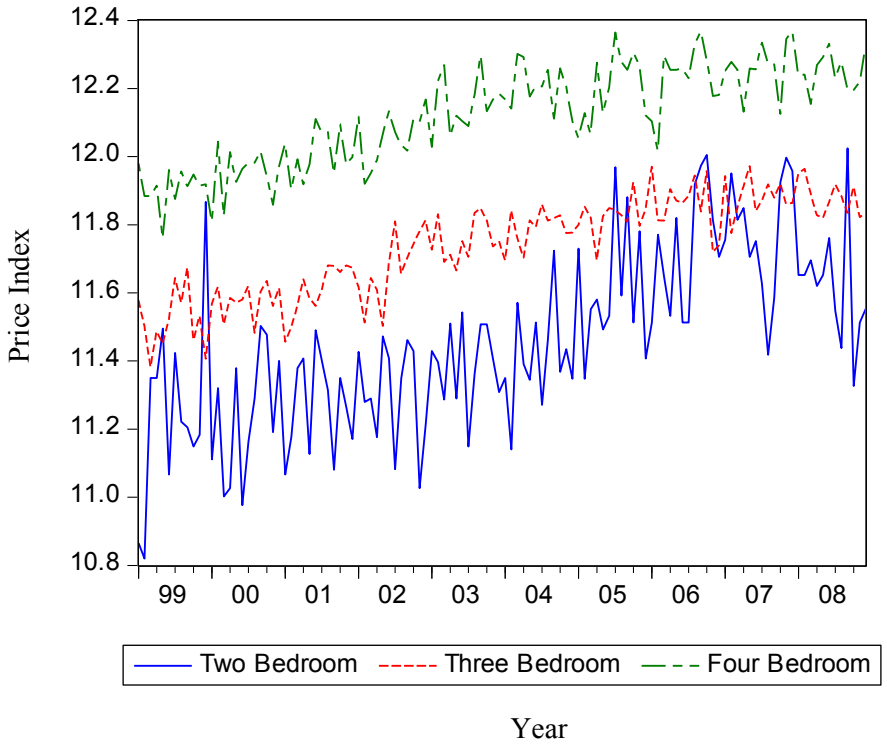


FIGURE 9.1: House Price Indices at Three Different Housing Markets

9.3 Empirical Analysis

9.3.1 Integration Analysis*:* Unit root tests for Stationary of House Price Indices

Unit root tests are applied to all housing price time series to investigate the stability. All time series were tested for unit root nonstationary by using Augmented Dicky-Fuller (ADF) unit root test at the level form. Table 9.2 reports the result of unit root tests with and without trend. The null hypothesis of non-stationary is performed at the 1% and the 5% significant levels. The result of the ADF test at level illustrates that all variables could not be rejected at the 5% significance level without deterministic trend. However, housing price index series can be rejected at the 1% significant level with deterministic trend. It indicates that all housing price index series are integrated at the level form. Furthermore, there are no cointegration relationships between all housing price index series. After the time trend is removed from the housing price index series,

the series become stationary. That is, they are a stationary time series around the time trend. Moreover, this means that the housing prices for each market segment increases from one month to another moved around a time trend during the investigation period.

TABLE 9.2: House Price Index Series Unit Root Tests from January 1999 to December 2008

	ADF test at level					
	without a deterministic trend			with a deterministic trend		
	t-statistic	Sig. level	Lag	t-statistic	Sig. level	Lag
2Br Houses	-1.284	na	6	-9.901	**	0
3Br Houses	-2.004	na	4	-4.373	**	2
4Br Houses	-2.125	na	12	-8.659	**	0

Notes: 1. Constant is included in the unit root test model.
2. The considered maximum lag is 12.
3. 1% test critical value: -4.037, 5% test critical value: -3.449
4. ** denotes the significance at 1%
5. H_0: Non-stationary, H_1: Stationary
6. na denotes: not applicable

9.3.2 Granger Causality Tests of Three Housing Market Segments

From the results of unit root test, there are no cointegration relationships between three housing markets. Then, Granger causality test must be conducted using VAR methodology. Granger causality tests are conducted within a vector autoregressive (VAR) context, pioneered by Sims (1980). As the VAR estimates are sensitive to the model specification, the selection of optimal lag length is important. When using VAR estimates, the decision of optimal lag length can be determined by comparing the Akaike information criterion (AIC) and the Schwarz criterion (SC) (Grasa, 1989). Moreover, other two criteria can assist to judge the optimal lag length. These two criteria are: Final prediction error (FPE) and Hannan-Quinn information criterion (HQ). All 4 criteria are

investigated in this study. When applying these criteria, the smallest value of these 4 criteria points to the optimal lag length.

Table 9.3 represents the results of the VAR lag length selection criteria. The first left hand column shows the lag lengths from 0 to 3. The numbers with asterisks are the smallest value in each criterion. Therefore, based on the results, the optimal lag length is considered as 1 in the VAR model.

TABLE 9.3: Lag-Length Selection Tests

Lag	FPE	AIC	SC	HQ
0	9.66E-07	-5.3367	-5.2639*	-5.3071*
1	9.51E-07*	-5.3524*	-5.0611	-5.2341
2	1.02E-06	-5.2856	-4.7759	-5.0788
3	1.01E-06	-5.289	-4.5609	-4.9936

Note: The star indicates lag order selected by the criterion. The criterions include the final prediction error (FPE), the Akaike information criterion (AIC), the Schwarz information criterion (SIC), and the Hannan-Quinn information criterion (HQIC).

The foundation of the Granger causality technique is the idea that variable X causes variable Y if Y is predicted better by using lagged values of X together with lagged values of variable Y rather than by using just lagged values of variable Y. This similarly applies to the case when variable Y causes variable X. The test for Granger causality involves testing the joint significance of lagged values of X in the model

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p Y_{t-p} + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_p X_{t-p} + \varepsilon_t$$
$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + u_t$$

where: p is the number of lags included in the model; and

ε_t, u_t are random errors.

The Granger causality test is conducted using above equation under the null hypothesis H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ against the alternative H_1: at least one $\beta_i \neq 0$ to test whether X causes Y and H_0: $\varphi_1 = \varphi_2 = \ldots = \varphi_p = 0$ against the alternative H_1: at least one $\varphi_i \neq 0$ to test whether Y causes X.

If X Granger causes Y and Y Granger causes X, then the relationship exists in both directions. If X Granger causes Y but Y doesn't Granger cause X and vice versa, the unidirectional relationship exists. If neither of the variables Granger causes the other, then X and Y are statistically independent. It is important to note that the statement "X granger causes Y" does not mean that Y is the effect or result of X. Granger causality measures precedence and information content and does not relate to the meaning of causality in common use.

Granger causality method is applied to explore the short run relationships. The tests will show which housing price segmentation dynamic leads the changes of other segmentations. The results of Granger causality tests for the pairs of three different housing segments are presented in Table 9.4. All pairs for Granger causality tests were applied within VAR framework.

TABLE 9.4: Granger Causality Tests of Three Housing Market Segments, 1999–2008

	***F*-statistic**	***p*-value**	**Causal**
2BR to 3BR	0.3664	0.545	No
2BR to 4BR	0.0064	0.9361	No
3BR to 2BR	0.1483	0.7001	No
3BR to 4BR	5.3017	0.0213	Yes
4BR to 2BR	0.09334	0.7599	No
4BR to 3BR	3.6504	0.0461	Yes

Notes: 2BR, 3BR, 4BR indicates two bedroom, three bedroom, and four bedroom houses, respectively.

Analysis of results of the Granger causality tests presented in Table 9.4 reveals one bi-directional relationship between three bedroom and four bedroom houses. While three bedroom housing prices are influenced by housing prices in four bedroom houses, the four bedroom housing market appears to be affected by prices in three bedroom houses. The prices of two bedroom houses seem to be independent with no influence from any other housing market. Figure 9.2 is generated from the results of Table 9.4 and summarizes the results of Granger causality tests among three housing type markets. Moreover, Figure 9.2 shows the price diffusion pattern in Amherst Town, State of New York.

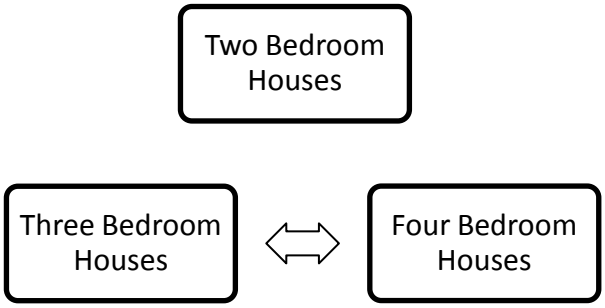


FIGURE 9.2 Causal Relationships among Three Housing Type Markets

9.4 Impulse responses among different housing markets

The impulse response analysis presents the dynamic effect of each variable response to the individual impulse from other variables. It determines the sensitivity of one variable to the change in another. Figure 9.3 ~ 9.5 represent the impulse response results of the three housing size markets individually. There are three curved lines in each figure. Each curved lines explains the response of one market to its past shock. The X axis shows the lag in months and Y axis shows the impulse response. The positive symbol means an increase in housing prices. In the

same way, the negative symbol means a decrease in housing prices. All of the three figures (9.3, 9.4, and 9.5) show that all of the three markets are impacted from the past performance of the market itself. Both two and three bedroom houses received a positive impact from the four bedroom house market. Two bedroom houses had less influence on both three and four bedroom house markets. The affect appears to gradually die out after four months. Comparing Figure 9.3 to Figure 9.4, some interesting facts were found. When the housing prices of three bedroom houses boomed, the prices of two bedroom houses dropped. That means that as the demand for three bedroom houses increases, the demand for tow bedroom houses declines, and therefore the two bedroom housing prices fall. But, this relationship is not reversible. It seems that three bedroom house market is not influenced by the boom of two bedroom housing prices.

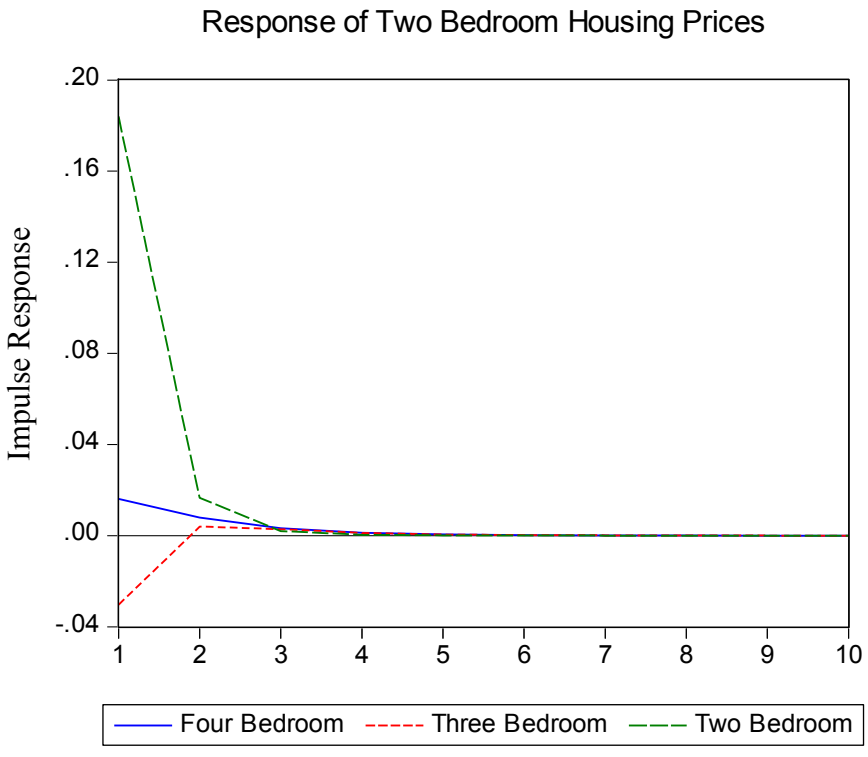


FIGURE 9.3: Impulse Response of Two Bedroom Housing Market to Shocks from other Markets

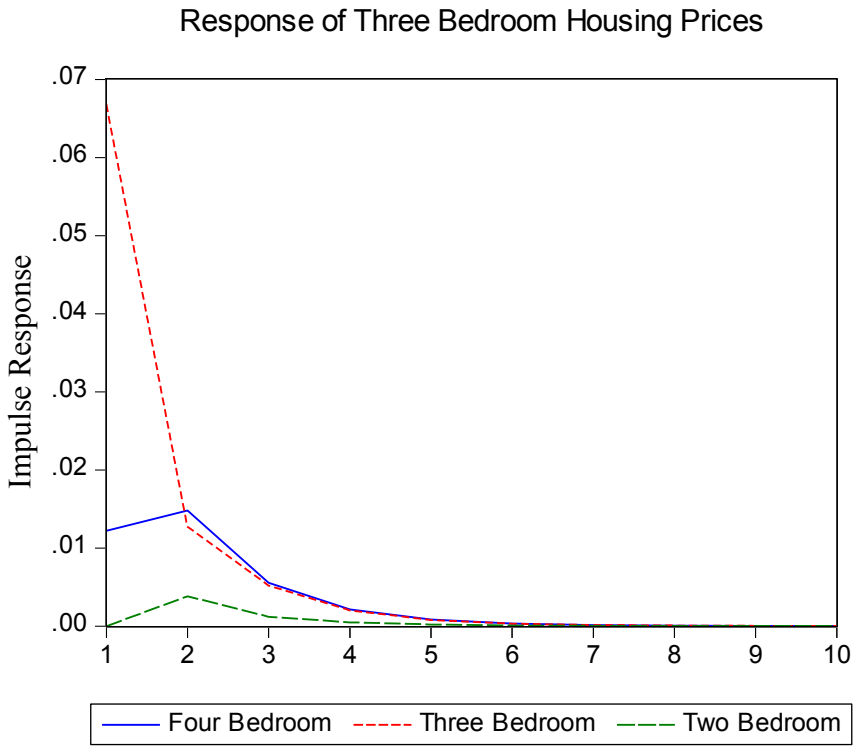


FIGURE 9.4: Impulse Response of Three Bedroom Housing Market to Shocks from other Markets

FIGURE 9.5: Impulse Response of Four Bedroom Housing Market to Shocks from other Markets

The speed of convergence to zero can scale the sensitivity of one market to the influence from other markets. In this study, 0.01 of absolute value are set up as the standard to measure this speed. Table 9.5 presents the number of lagged months when first reaching a value of impulse response of less than 0.01. The influence from three bedroom houses on four bedroom ones will persist for three months, and vice versa. The smaller number of the lagged months indicates the speed of convergence to zero is faster which means less sensitive to other markets. This suggests that both the three bedroom and four bedroom housing markets are less sensitive to the impacts from the two bedroom houses and the influence cannot persist for a long time. However, the market of two bedroom houses is more sensitive to the changes in the three bedroom and four bedroom housing markets. Furthermore, the numbers of lagged months shown on the diagonal line in Table 9-5 explains the duration of the time interval by each market is affected itself. It shows that the impacts of three and four bedroom houses influence shorter than that from itself on two bedroom housing market.

TABLE 9.5: Lagged Months when First Reaching a Value of Impulse Response of less than 0.01

	2BR house	3BR house	4BR house
response of 2BR house to	3	2	2
response of 3BR house to	0	3	3
response of 4BR house to	0	3	3

9.5 Sensitivity Analysis and Predictions

This study investigated linkages between different housing sizes, utilizing data from the Town of Amherst Town using Granger causality tests within VAR frameworks from the January 1999 to December 2008. It was found that a causal relationship in both directions between housing

prices in three and four bedroom houses. Moreover, the impulse response function is used to examine the response of one market to other markets and determined the sensitivity of each market.

The findings highlight a number of issues which are set out below. The speeds of convergence to zero were found with various numbers. The results suggest that two bedroom housing market is more sensitive to the change in three and four bedroom housing markets. From the Granger causality test, the housing price fluctuations in three bedroom should be considered as the explanatory variables in explaining four bedroom housing price dynamics and vice versa. Finally, this research shed light on some predictions for the developments in one market based on the developments in the other housing markets. It gives some useful information for the policy makers with respect to the changes in the demand for housing.

CHAPTER 10: A GIS-Based Methodology for Dividing the Housing Stock of a Municipality into Uniform Districts

10.1. Introduction

This chapter presents a methodology for dividing the housing stock of a municipality into uniform zones or districts for value assessment and other planning purposes. The 33,342 housing records from the Town of Amherst in the State of New York are utilized to demonstrate the methodology. The town's property tax assessment office has currently divided the total housing into 67 neighborhoods which were used as an initial set for this research. The cluster analysis is utilized to regroup the 67 neighborhoods into 25 districts to simplify the priori classifications. Then, the spatial analysis is done on these 25 districts. To perform this task, the multiple regression analysis and the geographic information system (GIS) methodology using ArcGIS software are applied. The valuable contributions of this research are to explore and establish the spatial patterns and relationships within the housing stock of towns. Also, the results of the GIS analyses and the cluster analysis used can be used in the land use planning and for the mass appraisal of houses.

This chapter takes a new approach and differs in a number of respects from the earlier studies. In the first place, ArcGIS software Version 9.3 was applied to analyze the spatial variations in the housing markets on the Town of Amherst, New York database of 33,342 houses. Then, the cluster statistical method was utilized on the housing characteristics such as: housing value per square feet of the living area, and the building age, etc.. Appendix A includes all of the seven housing characteristics. The model describes several numerical summaries about the housing market. In this chapter, the multiple regression model has incorporated a series of dummy

variables for neighborhoods (housing submarkets) defined by the local assessor, and the cluster analysis statistical method is used to regroup the 67 neighborhoods of an example municipality. The format of the chapter is constructed as follows. Section 2 contains a description of model development, and Section 3 describes the results of empirical spatial analyses. The final section 4 provides the summary of optimal methodology arrived in this research.

10.2. GIS Model Development

Figure 10.1 represents the sequence of steps in establishing the GIS model described in this chapter. The following steps were used in formulating the GIS model:

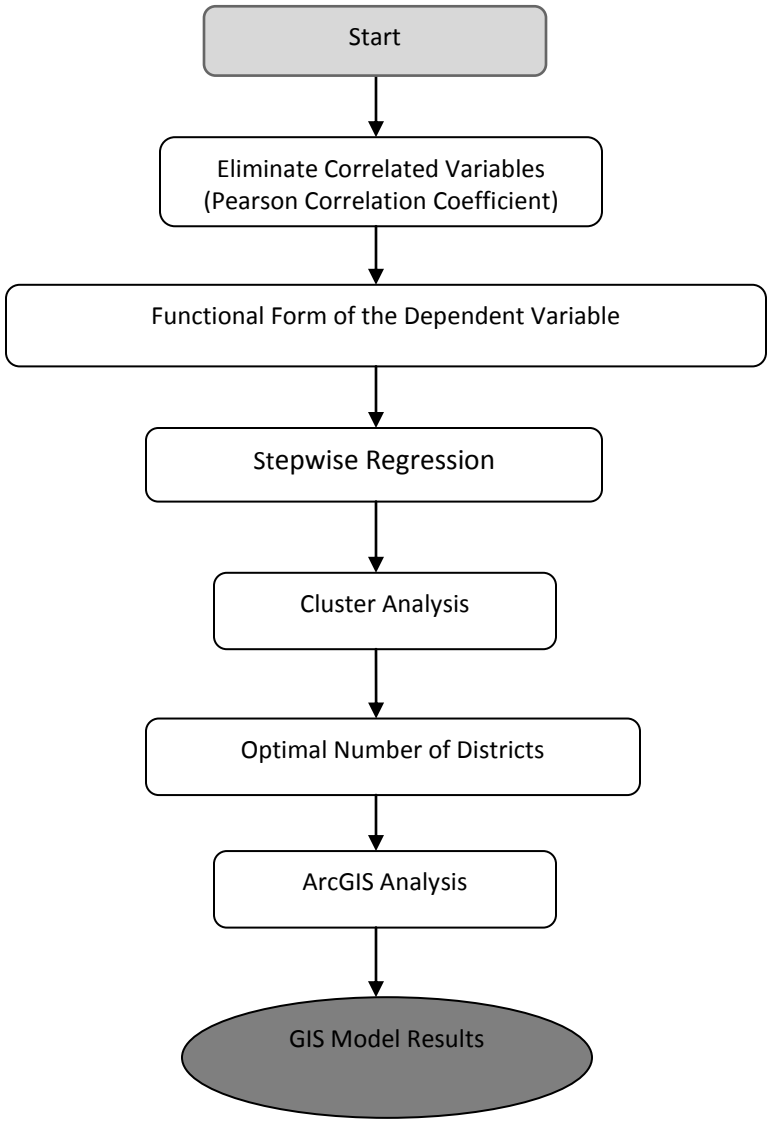


FIGURE 10.1: GIS Model Development Process

10.2.1. Eliminate Correlated Variables using Pearson Correlation Coefficient

Refer to 3.3.3.

10.2.2. Determination of Functional Form of Dependent Variable: *hprice*

Refer to 3.3.4.

10.2.3. Stepwise Regression

Refer to 3.3.5.

10.2.4. Cluster Analysis

The categorical variable, *nghdcode,* represents 67 different neighborhoods defined by the property tax assessment office of the Town of Amherst. Historically, the Town of Amherst was divided into neighborhoods based on geographical area and similar housing characteristics. The number of houses in each neighborhood code is presented in column 3 of Table 10.1. We utilized cluster analysis method to group houses that were relatively homogeneous in order to simplify the *a priori* classifications used groupings. Such grouping divides the total number of houses into uniform neighborhoods. For the cluster analysis, the Procedure FASTCLUS from the SAS statistical software is applied. Summary statistics for each of the 67 neighborhoods within the research area were calculated. These data include: mean values for the housing price per square foot and other characteristics such as *frontage*, *depth*, and *age*. Table 10.2 indicates the statistical summary of the means of variables used as the basis for cluster analysis. In Table 10.2, the variable *frequency* represents the number of houses in a given neighborhood.

TABLE 10.1: Existing Neighborhood (*nghdcode*) Districts

No.	*nghdcode*	Frequency	Percent	No.	*nghdcode*	Frequency	Percent
1	2	3	4	1	2	3	4
1	100	496	1.49%	35	3300	13	0.04%
2	200	230	0.69%	36	3400	551	1.65%
3	300	127	0.38%	37	3500	649	1.95%
4	400	1532	4.59%	38	3600	1,317	3.95%
5	500	1,089	3.27%	39	3700	1,657	4.97%
6	600	176	0.53%	40	3800	1,748	5.24%
7	700	345	1.03%	41	3900	345	1.03%
8	701	311	0.93%	42	4000	386	1.16%
9	800	320	0.96%	43	4100	68	0.20%
10	900	279	0.84%	44	4200	133	0.40%
11	1000	556	1.67%	45	4300	609	1.83%
12	1100	472	1.42%	46	4400	594	1.78%
13	1200	1,115	3.34%	47	4500	84	0.25%
14	1300	102	0.31%	48	4600	46	0.14%
15	1400	172	0.52%	49	4700	155	0.46%
16	1401	134	0.40%	50	4800	117	0.35%
17	1500	341	1.02%	51	4900	560	1.68%
18	1600	175	0.52%	52	5000	452	1.36%
19	1700	176	0.53%	53	5100	378	1.13%
20	1800	415	1.24%	54	5200	1,933	5.80%
21	1900	277	0.83%	55	5300	359	1.08%
22	2000	572	1.72%	56	5400	694	2.08%
23	2100	301	0.90%	57	5500	87	0.26%
24	2200	241	0.72%	58	5600	712	2.14%
25	2300	496	1.49%	59	5700	13	0.04%
26	2400	155	0.46%	60	5800	1,078	3.23%
27	2500	195	0.58%	61	5900	432	1.30%
28	2600	427	1.28%	62	6000	540	1.62%
29	2700	255	0.76%	63	6100	672	2.02%
30	2800	1,312	3.93%	64	6200	4	0.01%
31	2900	438	1.31%	65	6300	239	0.72%
32	3000	254	0.76%	66	6400	629	1.89%
33	3100	454	1.36%	67	6500	399	1.20%
34	3200	1,749	5.25%	**Total**		**33,342**	**100%**

TABLE 10.2: Mean Values of the Variables of the Existing 67 Neighborhoods

nghdcode	Frequency	$^{hprice}/_{sfla}$	*age*	*frontage*	*depth*	*nofirepl*	*nobedrms*	*nobaths*
100	496	76	46.72	63.97	193.68	0.58	3.2	1.6
200	230	78.17	56.45	88.09	246.49	0.6	3.14	1.55
300	127	69.7	61.18	95.27	259.78	0.4	2.98	1.35
400	1,532	78.54	36.74	69.41	157.48	0.34	3.29	1.51
500	1,089	79.58	47.16	74.56	259.82	0.51	3.16	1.51
600	176	89.85	7.84	74.47	199.13	0.72	3.34	2.07
700	345	76.43	15.99	75.54	162.37	1.02	3.89	2.52
701	311	99.35	6.89	94.02	195.47	1.06	3.76	2.52
800	320	85.63	15.3	75.48	157.35	0.96	3.7	2.41
900	279	83.59	21.73	64.51	167.18	0.77	3.15	1.85
1000	556	78.78	25.81	83.3	153.87	0.98	3.75	2.37
1100	472	94.43	11.42	81.02	179.08	0.95	3.79	2.5
1200	1,115	81.99	34.19	70.42	136.48	0.54	3.36	1.75
1300	102	88.02	16.18	80.11	227.84	1.02	3.73	2.53
1400	172	77.78	25.83	65.04	124.52	0.7	3.08	1.7
1401	134	75.83	22.25	83.59	145.08	0.99	3.22	2.1
1500	341	83.35	21.49	75.83	152.4	0.95	3.54	2.31
1600	175	93.45	14.36	81.16	162.28	1.04	3.79	2.48
1700	176	75.25	30.05	97.96	168.71	0.99	3.97	2.67
1800	415	79.59	27.25	74.08	152	0.91	3.68	2.2
1900	277	83.13	23.14	82.09	147.81	1.03	3.82	2.44
2000	572	92.61	15.54	89.01	173.07	1.05	3.95	2.58
2100	301	82	29.64	74.68	180.9	0.91	3.39	1.99
2200	241	99.27	16.34	98.54	184.55	1.23	4.11	2.85
2300	496	83.61	28.33	99.25	161.1	1.06	4.06	2.59
2400	155	94.8	34.57	142.05	188.01	1.41	4.16	2.94
2500	195	115.19	16.09	124.68	171.24	1.77	4.33	3.71
2600	427	73.08	35.6	99.13	188.55	0.97	3.86	2.33
2700	255	82.77	24.02	74.05	157.67	0.85	3.43	2.15
2800	1,312	72	37.31	77.21	148.52	0.89	3.88	2.29
2900	438	76.41	37.1	79.49	169.54	0.83	3.71	2.14
3000	254	91.58	9.62	83.32	168.06	0.99	3.68	2.49
3100	454	65.14	53.7	59.84	141.64	0.19	2.99	1.23
3200	1,749	71.67	46.97	72.84	144.76	0.41	3.23	1.52
3300	13	76.89	16.77	88.38	172.69	1.08	3.77	2.62
3400	551	78.23	42.55	91.57	179.2	0.99	3.67	2.41
3500	649	74.56	33.04	75.52	131.21	0.93	3.64	2.16

TABLE 10.2: continued

nghdcode	Frequency	$^{hprice}/_{sfla}$	*age*	*frontage*	*depth*	*nofirepl*	*nobedrms*	*nobaths*
3600	1,317	76.93	37.89	71.11	175.32	0.54	3.40	1.69
3700	1,657	65.32	56.80	53.67	126.37	0.28	3.04	1.27
3800	1,748	70.63	60.20	52.00	149.03	0.32	3.11	1.32
3900	345	54.89	61.31	56.08	136.05	0.04	2.85	1.06
4000	386	84.07	49.41	80.81	179.71	0.87	3.35	1.89
4100	68	97.30	24.25	96.71	155.90	1.15	3.72	2.63
4200	133	61.44	53.11	76.31	134.74	0.16	2.61	1.18
4300	609	78.90	58.43	65.17	183.19	0.74	3.23	1.70
4400	594	74.59	52.91	86.30	189.47	0.95	3.33	1.89
4500	84	66.50	52.58	110.79	316.90	1.30	3.79	2.33
4600	46	107.05	12.26	128.49	186.15	1.48	4.13	3.37
4700	155	90.79	56.29	95.56	161.95	1.33	3.95	2.79
4800	117	106.48	57.84	149.33	313.34	1.71	4.29	3.61
4900	560	87.58	65.33	70.67	160.77	1.14	3.63	2.14
5000	452	70.36	66.00	52.93	138.22	0.40	3.19	1.46
5100	378	86.07	61.33	72.55	146.73	1.14	3.63	2.12
5200	1,933	71.18	63.15	57.33	150.37	0.75	3.21	1.53
5300	359	76.20	55.82	59.94	138.29	0.38	2.89	1.31
5400	694	70.49	57.38	71.98	171.28	0.51	3.07	1.46
5500	87	68.80	62.08	67.49	223.69	0.65	3.27	1.56
5600	712	73.32	50.68	90.30	165.89	0.50	3.03	1.50
5700	13	109.74	9.46	76.54	166.08	1.00	2.23	2.31
5800	1,078	73.64	44.01	82.30	170.26	0.86	3.56	2.09
5900	432	81.45	79.05	61.05	186.48	0.54	3.28	1.55
6000	540	81.74	77.10	66.72	152.46	0.49	3.15	1.47
6100	672	78.71	65.45	67.17	167.18	0.65	3.16	1.54
6200	4	118.10	4.00	101.00	181.50	2.00	4.00	3.25
6300	239	132.74	5.00	111.34	180.43	1.39	3.97	3.21
6400	629	76.80	41.33	69.12	150.32	0.53	3.51	1.72
6500	399	79.38	44.33	69.73	168.42	0.50	3.20	1.62

10.2.5. Optimal Number of Districts

To obtain the optimal number of districts, entire dataset of 33,342 observations was used to determine the estimates of the Root Mean Squared Error (RMSE) obtained from the model (formula 1) against the number of districts (Figure 10.2). The multiple regression model was used with increasing number of the clustered districts. The root mean square error (RMSE) for each model is used as an indicator for choosing the optimal number of districts. Figure 10.3 illustrates how RMSE changes for incremental changes in the number of districts in the multiple regression model. There is a relatively large improvement found in RMSE when using the variables corresponding to five districts. The RMSE gradually decreases at a declining rate as the number of districts increases, with slighter improvements beyond 25 districts. For the purpose of simplifying the *a priori* classification of submarkets, the hypothesis is made that prediction accuracy with 25 districts should be nearly as good as with 67 districts (neighborhoods). Therefore, 25 districts can be used to map and do spatial analysis on the distribution of housing characteristics. Table 10.3 includes the grouping of the 33,342 houses into 25 districts from the earlier set of 67 neighborhoods. Table 10.4 provides the statistical summary for each of the 25 districts.

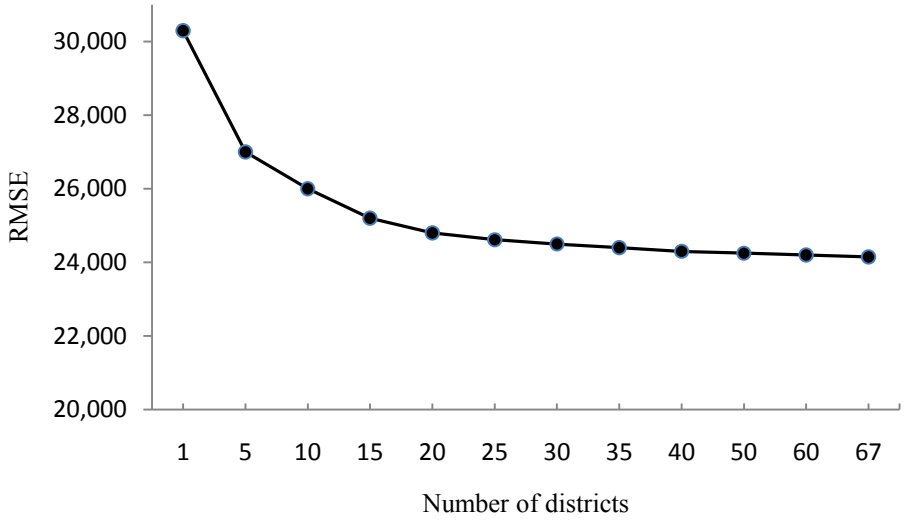


FIGURE 10.2: Relationship between RMSE and the Number of Districts

(RMSE: Root Mean Squared Error)

TABLE 10.3: Distribution of the 25 Districts after Merging

District	*nghdcodes*									
1	600									
2	701									
3	2200									
4	2300									
5	2500									
6	3900									
7	4500									
8	4600									
9	4800									
10	5200									
11	5700									
12	6200									
13	6300									
14	1100	3000								
15	1700	3300								
16	2400	4100								
17	3800	5300								
18	700	2600	3400							
19	1600	2000	4700							
20	3100	3700	4200							
21	800	1300	1900	4900	5100					
22	300	3200	5000	5400	5500	5600				
23	900	1200	1500	2100	2700	4000	5900	6000		
24	1000	1401	1800	2800	2900	3500	4400	5800		
25	100	200	400	500	1400	3600	4300	6100	6400	6500

TABLE 10.4: Summary Statistics of the Merged Districts

District	$hprice/sfla$	*age*	*frontage*	*depth*	*nofirepl*	*nobedrms*	*nobaths*
1	89.85	7.84	74.47	199.13	0.72	3.34	2.07
2	99.35	6.89	94.02	195.47	1.06	3.76	2.52
3	99.27	16.34	98.54	184.55	1.23	4.11	2.85
4	83.61	28.33	99.25	161.10	1.06	4.06	2.59
5	115.19	16.09	124.68	171.24	1.77	4.33	3.71
6	54.89	61.31	56.08	136.05	0.04	2.85	1.06
7	66.50	52.58	110.79	316.90	1.30	3.79	2.33
8	107.05	12.26	128.49	186.15	1.48	4.13	3.37
9	106.48	57.84	149.33	313.34	1.71	4.29	3.61
10	71.18	63.15	57.33	150.37	0.75	3.21	1.53
11	109.74	9.46	76.54	166.08	1.00	2.23	2.31
12	118.10	4.00	101.00	181.50	2.00	4.00	3.25
13	132.74	5.00	111.34	180.43	1.39	3.97	3.21
14	93.49	10.83	81.78	175.46	0.96	3.75	2.49
15	75.36	23.34	94.53	170.60	1.04	3.86	2.64
16	95.49	29.34	123.35	175.00	1.29	3.96	2.79
17	71.84	59.25	53.73	146.69	0.34	3.06	1.32
18	76.11	31.41	91.71	177.91	0.99	3.80	2.42
19	92.46	28.50	91.98	165.77	1.14	3.90	2.62
20	65.27	45.34	42.79	101.26	0.21	2.32	0.97
21	86.25	36.38	74.01	168.73	1.05	3.70	2.33
22	71.40	58.12	71.82	184.27	0.52	3.15	1.49
23	82.47	41.95	69.50	164.09	0.74	3.35	1.88
24	74.77	34.79	80.60	158.03	0.92	3.59	2.16
25	78.11	46.08	70.72	185.21	0.57	3.23	1.62

10.2.6. ArcGIS Analysis:

After linking the clustered districts to the map layer file obtained from the Amherst Town, the housing prices and housing attributes are installed in each district which presents the polygon in the ArcGIS. Later, this information can be used for mapping and analyzing the spatial pattern.

10.3. Results of Spatial Analyses

After the 25 districts are constructed, a series of maps showing the attributes of districts are created by ArcGIS software.

The Map 10.1 shows the average assessed housing prices for each district to see how different districts of Amherst Town are compared. The highest housing prices are indicated in the eastern, northern, and southwestern area of Amherst Town.

The Map 10.2, created by calculating the housing age for each district and marking the districts accordingly, shows the oldest house in red areas. The Map 10.2 presents most of the older houses are located on the southern boundary of Amherst Town.

The Map 10.3 shows the spatial distribution of housing parcel frontage. By comparing Map 10.3 to Map 10.1, the location of largest housing frontage is located on the same spot as the highest housing prices.

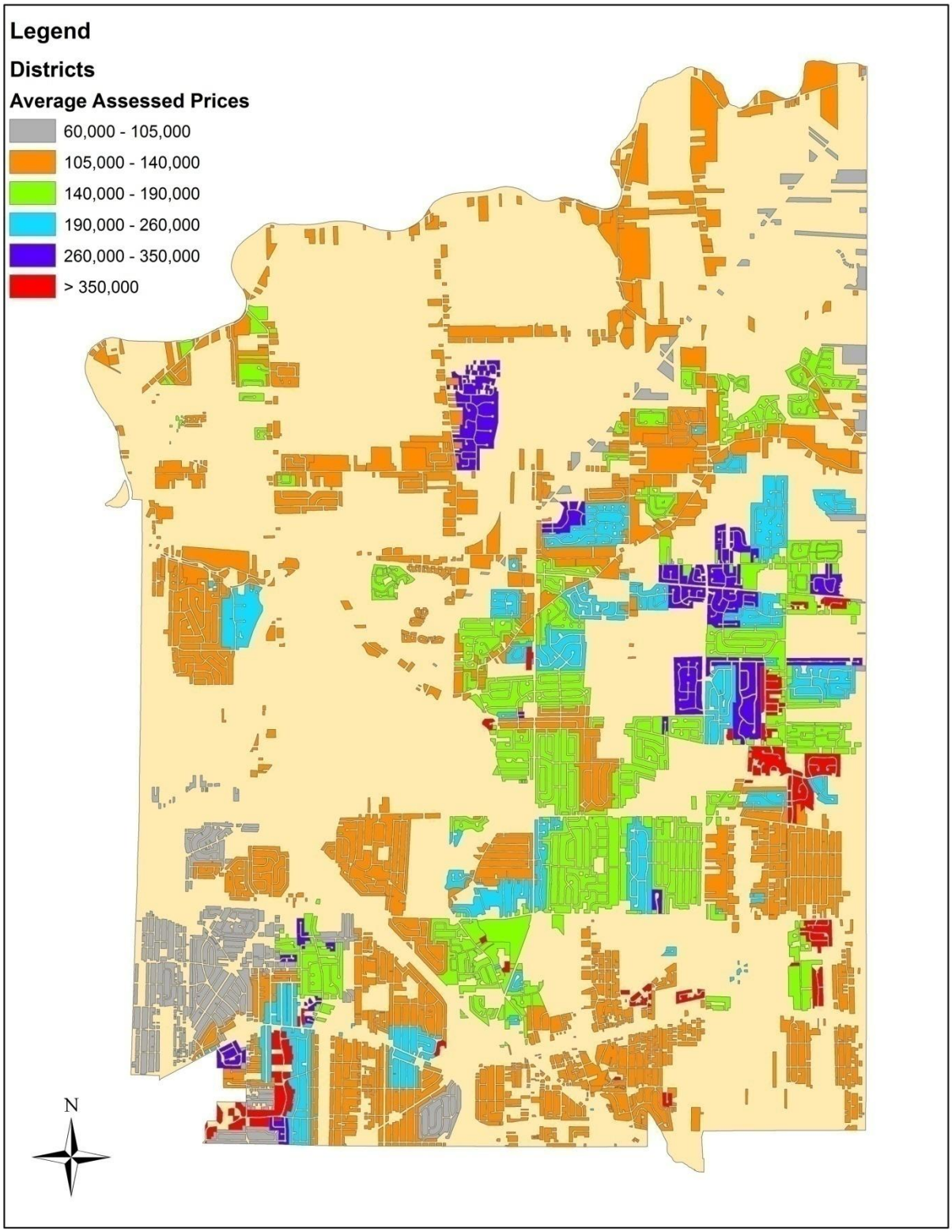
The Map 10.4 presents the spatial distribution of housing parcel depth. The deeper housing parcels are located on the Northern area of Amherst Town. The parcel depth is not correlated to housing price.

10.4. Summary of Optimal Methodology

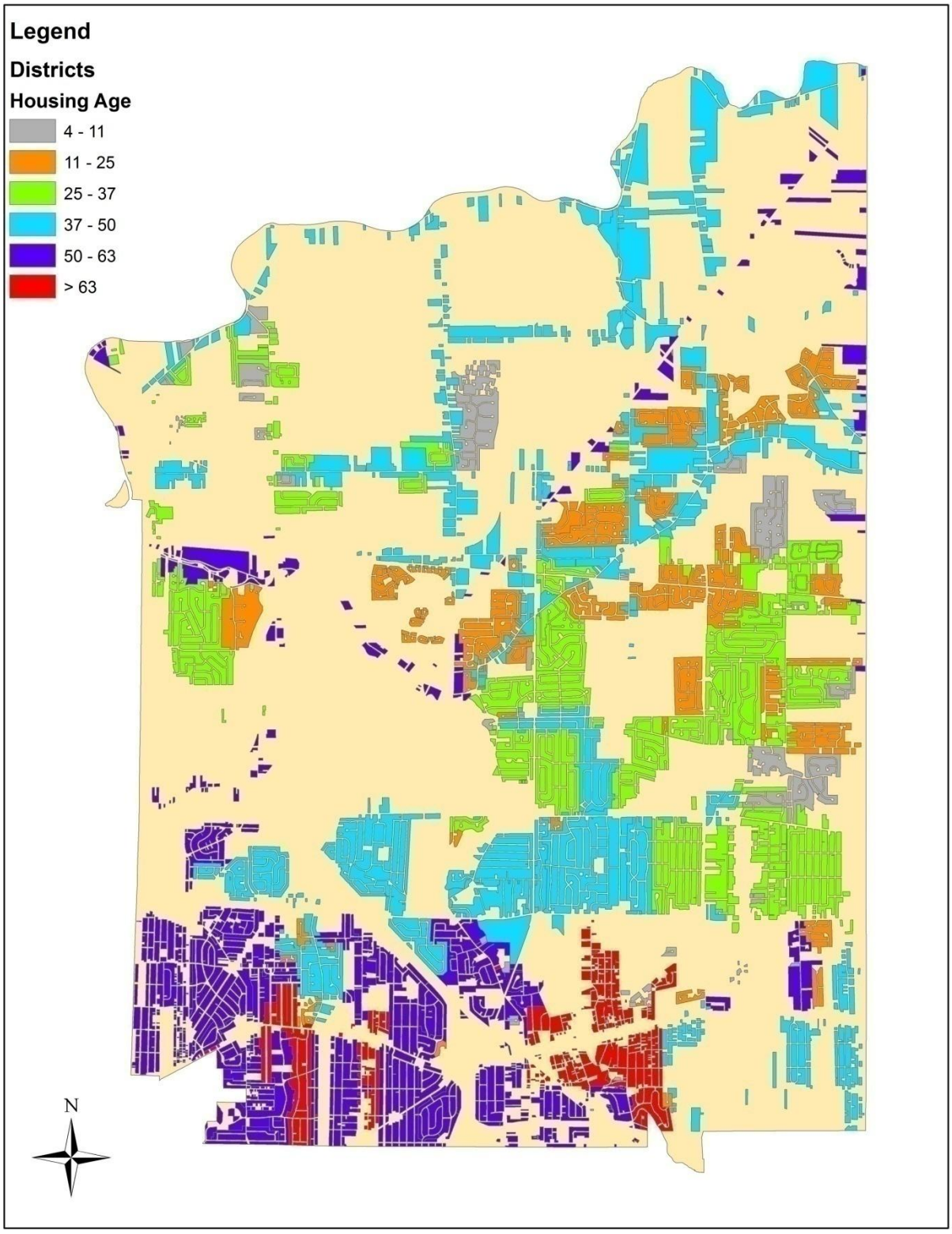
Dividing the housing stock of a municipality into uniform districts is the first important task in the mass appraisal of houses by a municipality. All statistics are then computed for the district and applied to each house uniformly within the district.

This chapter has outlined a methodology using GIS and advanced statistical tools for dividing the housing stock into uniform districts. The methodology outlined in this chapter will be useful for

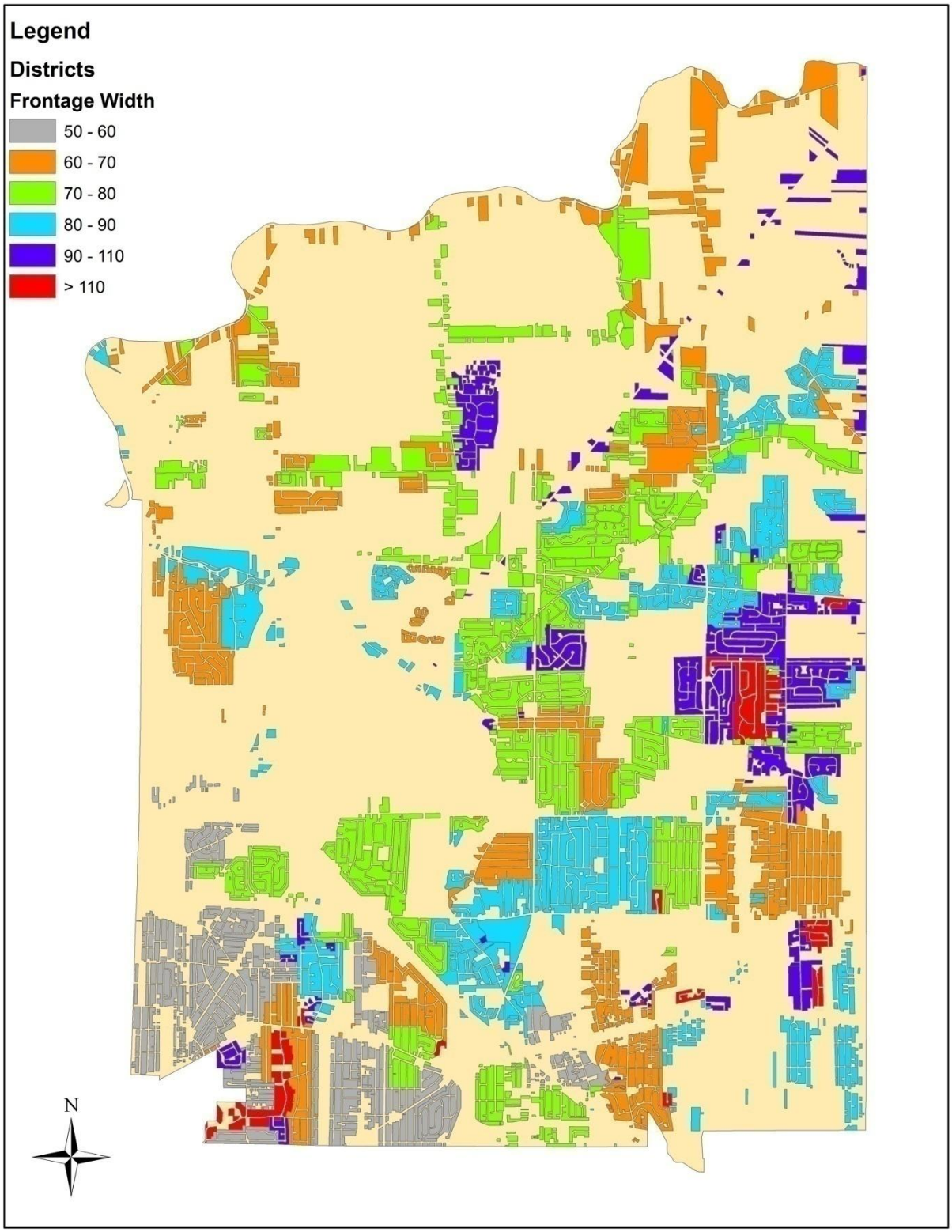
municipalities and consultants while performing mass appraisals. Town planners will also find the methodology useful for determining development needs of one district in comparison to all other districts.



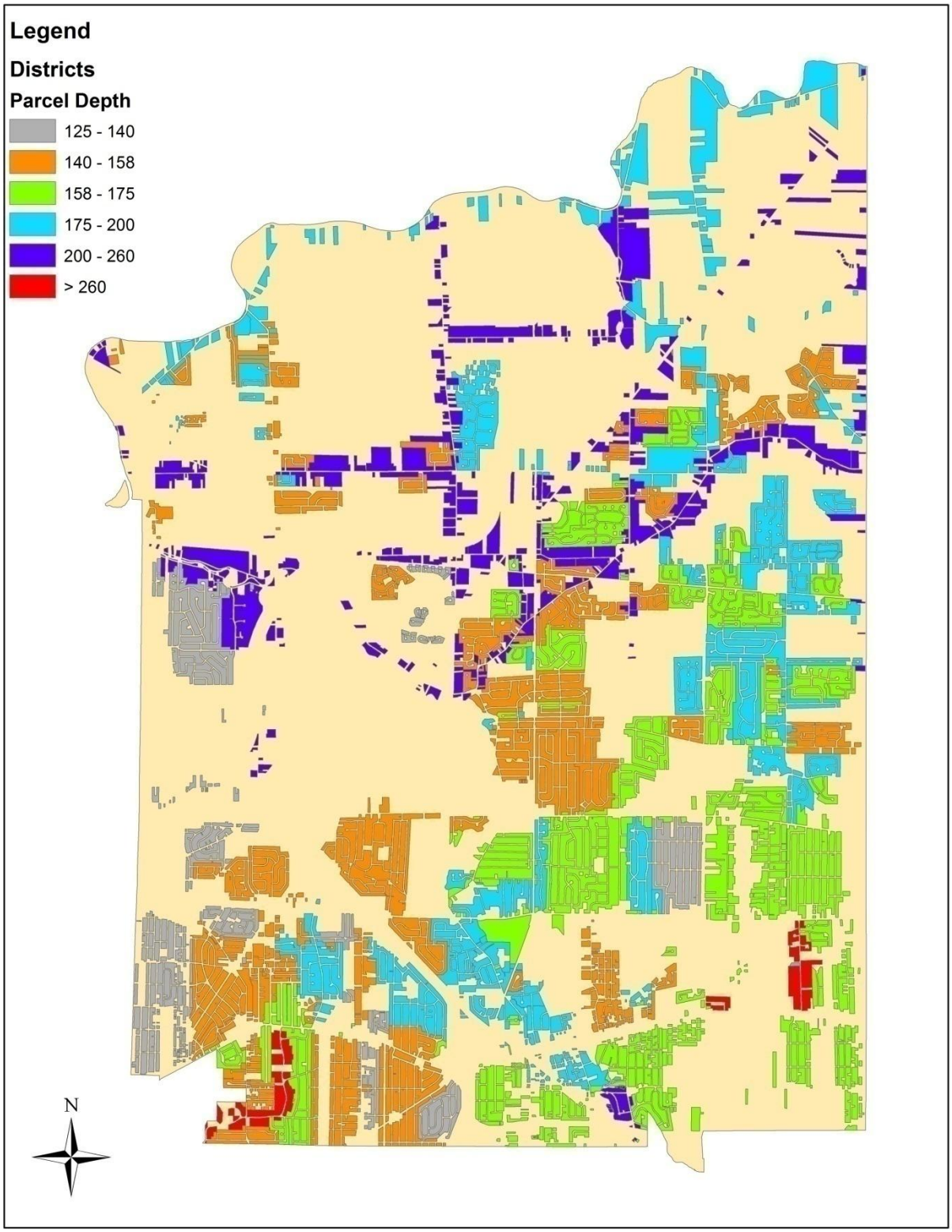
MAP 10.1: Spatial Distribution of Average Assessed Housing Prices ($)



MAP 10.2: Spatial Distribution of House Age



MAP 10.3: Spatial Distribution of Housing Parcel Frontage



MAP 10.4: Spatial Distribution of Housing Parcel Depth

CHAPTER 11: Summary and Conclusions

All municipalities are required by law to re-assess their real estate periodically which they do manually spending large sums. This research has provided alternative methodologies for mass appraisal of residential housing efficiently, cost-effectively, and equitably, without any subjective bias inherent in manual operations. Two statistical models and one AI model were developed and validated using the housing database of a municipality, the Town of Amherst, State of New York consisting of 33,342 single-family houses. The three models were compared and rank ordered for their effectiveness in terms of accuracy. Also, housing price estimates obtained from each of the three models were linked to the GIS layer of the town for visual analysis of the price distributions. This type of linking was never researched before, according to the available literature.

This research also quantified the influences of the major economic indicators on a municipality's housing prices. This research has, also formulated a new methodology for dividing the housing stock of a town into uniform districts, for comparing the data of houses in the same district on an equitable and statistically valid basis. This formulation has not been attempted by any researcher, earlier.

The assessment data of 33,342 single-family residential houses was randomly divided into two parts, 80% of the records (26,674) for the training set and 20% (6,668) for validation. While adjusted R^2 and F-value were used for the model development, their prediction performance was measured using: (i) the root mean squared error (RMSE); (ii) the mean absolute error (MAE); and (iii) the Theil's U statistic.

Nine variables were initially considered for the development of models: *age*, frontage width (*frontage*), depth of parcel (*depth*), square feet of living area (*sfla*), number of bedrooms (*nobedrms*), number of bathrooms (*nobaths*), building style (*bstyle*), number of fire places (*nofirepl*), and the neighborhood (*nghd*). Pearson correlations eliminated: *nofirepl*, *nobedrms*, and *nobaths*. This left six variables to be included in housing price determination models:

A. Quantitative Variables:
- *frontage*
- *depth*
- *age*
- *sfla*

B. Qualitative Variables:
- *bstyle*
- *nghd*

The **stepwise multiple regression model** developed, using the data of 26,674 records, had high accuracy with an R^2 value of 0.9046 and an F-value of 2,838.8, and tested well on validation, and was of the following form:

$$\textbf{\textit{hprice}} = 47{,}049 + 88.60(\textit{frontage}) + 14.18(\textit{depth}) - 383.93(\textit{age}) + 75.97(\textit{sfla}) \pm \textit{bstyle}_i(X_i) \pm \textit{nghd}_j(Y_j)$$

where: $x_1 = -28{,}364$ for bstyle1……, and
$x_{12} = 45{,}915$ for bstyle12, and

$y_1 = -4987.167$ for nghd1,……, and
$y_{66} = -173.93$ for nghd66.

The multiple regression had high prediction performance with the MAE, RMSE, and Theil's U statistic of the validation set being within 3% to 8% of the training set. The difference between the percent of validation set data and the training set data values within 10% of the observed prices was 1%, and also the validation set data within 20% was 1%, both measures showing very high prediction performance of the model.

The **Additive Nonparametric Regression (ANR) Model** was next developed, to analyze the shape of the fitted curve as determined by the data as against the linear relationship in the

stepwise regression. Regression functions for each of the quantitative independent variables: *frontage*, *depth*, and *age* were nonlinear, while for *sfla*, the regression function was almost linear. Like the multiple regression model, the ANR model was tested for its prediction performance using three widely accepted measures. The RMSE of the validation set was +2% of the training set, the MAE +6%, and the Theil's U statistic +5%. Also, the difference between the validation set data and the training set data being within 10% of the observed prices was -2%, and that within 20% was +0.1%. All of the above measures showed a high prediction accuracy of the ANR model.

The **Artificial Neural Network (ANN) Model** was constructed using six (6) neurons, because there was no significant improvement beyond 6 neurons. The correlation coefficient for 6 neurons in the hidden layer was 0.9723, very close to 1.0, which indicated a good fit. The validation set as compared to the training set had percent differences between MAE, RMSE, and Theil's U statistics that were 7%, 26% and 25% higher, respectively, however the three errors were smaller than those in the multiple regression model and the additive nonparametric regression model. Also, percent data within 10% was -2%, and percent data within 20% was -0.5% of the training set, a good prediction. The percentage of data points within 10% and 20% were significantly higher than the two statistical models. Graphs of price per square foot against percent of houses within 10% and within 20% showed that houses costing $54 to $132 per square foot can be predicted with reasonably high accuracy, and those above $88 per square foot can be predicted with high accuracy using the artificial neural network model.

Comparison of the Three Mass Appraisal Models

The three models developed in this research: (i) the Multiple Regression (MR) Model, (ii) the Additive Nonparametric Regression (ANR) Model, and (iii) the Artificial Neural Network (ANN) Model, were compared for their accuracy in predicting the housing prices. Again, the three widely accepted measures: RMSE, MAE, and Theil's U statistic were utilized. The percent data within 10% and within 20% of the actual data was also computed.

It was found that the Artificial Neural Network (ANN) Model results in significantly smaller prediction errors than the Multiple Regression (MR) Model or the Additive Nonparametric Regression (ANR) Model due to the many interactive terms between the independent variables underline the Artificial Neural Network (ANN) architecture. The Artificial Neural Network (ANN) Model had about 21% lower MAE, 23% lower RMSE, and 23% lower Theil's U statistic than those of the Multiple Regression (MR) Model. The percent predicted values of the three models within 10% and 20% of the observed prices calculated as below:

Measure of Accuracy	Multiple Regression	Additive Nonparametric Regression	Artificial Neural Network
Percent within 10%	71%	74%	79%
Percent within 20%	92%	93%	95%

The Artificial Neural Network (ANN) Model had significantly higher frequency of houses within 10% and 20% of the observed prices. This exercise ranked the three models in accuracy as below:

Rank No. 1 Artificial Neural Network Model

Rank No. 2 Additive Nonparametric Regression Model

Rank No. 3 Multiple Regression Model

This comparative study of the statistical models and an AI model was the first, never before reported.

The prediction accuracy of the three models was further tested over the per square foot value of houses, and it was found that, for the example municipality:

(i) for houses costing less than $88 per square foot of living area, all of the three models provide almost the same prediction accuracy, while

(ii) for houses costing more than $88 per square foot of living area, the artificial neural network model provides the highest accuracy.

Macroeconomic Indicators Influence

The influence of major macroeconomic indicator has been recognized but not well documented and quantified; this research has attempted to quantify the effect of the following five major macroeconomic indicators on the housing prices, in the context of the Town of Amherst:

(i) Oil Price (*OIL*)
(ii) 30-year Mortgage Interest Rate (*IR*)
(iii) Consumer Price Index (*CPI*)
(iv) Dow Jones Industrial Average (*DJIA*), and
(v) Unemployment Rate (*UR*)

Vector Autoregression was applied on the 1999-2008 monthly housing sales data of the Town of Amherst. Also, unit root tests, Johansen's cointegration test, impulse response function and variance decompositions were employed. The various analyses concluded that the 30-year mortgage interest rate has the highest effect on the housing prices ranging from 4.97 percent in the first month to 8.51 percent in the twelfth month. The Unemployment Rate was next in order, followed by Dow Jones Industrial Average, and Consumer Price Index. The total effect of these

five macroeconomic indicators ranged from 7.3% in the first month to 25.5% in the twelfth month. The conclusions arrived in this paper, along with several related tables and figures will be useful to the housing community and the real estate companies in the region in planning their business in the future years.

Housing Price Diffusion Patterns

This research also studied the ripple effect of changes in housing prices of one size on the prices of other sizes within the same municipality. The various analyses suggested that the two-bedroom housing market is more sensitive to the changes in the three and four-bedroom housing markets. The Granger Causality test concluded that the housing price fluctuations in three-bedroom housing should be considered as the explanatory variable in explaining the four-bedroom housing price dynamics and vice versa. These results will be useful for policy makers in computing the housing demand for the various housing sizes, given some data on any one of the sizes.

Dividing the Housing Stock into Uniform Districts

Housing market is fundamentally spatial in nature, i.e. housing price is a function of the neighborhood in which it is located. Statistical methods have been used in the past to segment housing size markets; this research took a new approach for analyzing the spatial variations in the housing prices using ArcGIS software, along with stepwise regression and cluster analysis. Using the various analyses, this research divided the stock of 33,342 houses into 25 uniform districts as against the currently existing 67 neighborhoods. The methodology used in this research will give a cost effective number of housing districts with similar characteristics. The

municipalities, the consultants, and the planners will be able to work on the minimum set of districts which will give a larger frequency of sales within a uniform district.

The results of this research have proven that:

1) The variables that significantly influence the housing price include: frontage width, parcel depth, age, square foot of living area, building style, and neighborhood.

2) The statistical models and the artificial neural network model using computers can be used for mass appraisal, with accurate results.

3) Any of the three models can be used with similar prediction accuracy for lower and medium priced houses. For the municipality that provided the data of their 33,342 houses, this limit was the houses costing less than $88 per square foot of living area.

4) For higher priced houses, the artificial neural network (ANN) model gives higher accuracy than the two statistical models, and

5) Linking the predicted data to the GIS layers of the town will be of great benefit to the assessors and the planners in the following ways:

    a) Because of similarity in housing prices within a neighborhood, any deviations in assessed values, within that neighborhood, can be identified, reviewed and corrected and

    b) The planning of future neighborhoods can be organized using more optimal approaches so that the region can grow according to the society's goals.

These results are relevant to the example municipality whose assessment data was used in this research; however similar formulations can easily be done for any municipality.

APPENDIX A: Description of Fields Provided in the Data Set Used for Training and Validating the Models

Field	Description
ParcelID	consecutive number (internal use only)
swis	village or town
sbl	sbl (padded)
printkey	sbl (not padded)
eastgrid	parcel centroid provided by state (east)
nogrid	parcel centroid provided by state (north)
stnum	parcel address street number (text)
stnumval	parcel address street number (numeric)
stname	parcel address street name
Unit	parcel unit indicator
StCode	parcel address street code (internal use only)
frontage	parcel frontage (feet)
depth	parcel depth (feet)
acres	parcel acreage (partial information only)
newnpcls	property classification code (overall)
OwnCode	owner code
oname1	property owner's name
mailadd1	property owner's mailing address (field 1)
mailadd2	property owner's mailing address (field 2)
streetadd	property owner's street address
City	property owner's city
State	property owner's state
Zip	property owner's zipcode
Country	property owner's county
schdist	parcel school district
status	parcel status
cutotav	Erie County total assessed value
culandav	Erie County land only assessed value
tntaxble	Town of Amherst taxable value
schtaxble	school district taxable value
viltaxble	Village of Williamsville taxable value
sitnpcls	property classification code (site)
siteuse	site use (incomplete - not used)

APPENDIX A: Continued

Field	Description
ovayrblt	year built (incomplete - not used)
nghdcode	neighborhood code
bldggrade	building grade (not used)
bldstyle	building style
bldyrblt	building year built
sfla	square feet of living area
heatype	heating type
basmtype	basement type
nofirepl	number of fireplaces
nobedrms	number of bedrooms
zoning	zoning (not used)
nobaths	number of bathrooms
SaleDate	last sale date (after 1987 only)
SaleType	last sale type (after 1987only)
SalePrice	last sale price (after 1987 only)
ArmLength	arm's length sale flag

APPENDIX B: Building Style Description

1. **Ranch:**

The ranch is a uniquely American single story structure. The ranch house is usually a long, close-to-the-ground profile, a low angle gable, a hip style roof, or a flat roof. The style is noted for its minimal use of exterior and interior decoration.

2. **Raised ranch:**

The raised ranch home is the variation of the ranch style. Usually, the raised ranch has two stories. The basement walls are usually elevated four feet or more above ground level. The lower story is used as living area. There is usually a full flight of stairs leads to the upper story.

3. **Split level:**

A split level home is a type of house in which the floor levels on one side of the house is about one-half way between a floor and its ceiling of the other part of the house. These levels are accessed by a half flight of stairs.

4. **Cape Cod:**

The Cape Cod is a simple frame house with a shingled, steeply-pitched roof, central chimney, and one or one-and-a-half floors. The style originated in colonial Cape Cod, Massachusetts, and has evolved over hundreds of years into several popular variations.

5. **Colonial:**

It is a style which has many variations and has changed to meet the needs of the marketplace. The only similar appearance between the various styles is a symmetrical façade, a center entry-hall floor plan, and usually a gable roof style.

6. **Contemporary:**

The category is applied to describe a catch-all style that can take on many different shapes. It is usually the customized buildings. Normally, it is a structure given to large, open spaces and odd, irregular shape.

7. **Mansion:**

A mansion is an extremely large and imposing dwelling house. The house is designed and built with the cost is out of consideration. Normally, the dwelling is unique and the number of rooms will exceed 10, with approximately 1 bathroom for every bedroom.

8. **Old style:**

The house of this classification is a complex, older structure, built prior to 1950 and often exceeds 100 years in housing age. A typical old style shows sign of physical and functional obsolescence throughout and is of average construction quality.

9. **Cottage:**

The cottage residence type is used for weekend or summer getaways by city dwellers. Usually, they are used as a place to spend holidays with friends and family. Cottages are often located next to lakes, rivers, and the ocean.

10. **Log home:**

A log home is typically made from logs that have not been milled into conventional lumber. There are two kinds of log homes. One is the handcrafted home; another is the milled home.

11. **Duplex:**

This style of residence is a house with two separate dwelling units. An apartment with rooms on two floors connected by a private staircase.

12. Town house:

Town houses are multiple single family dwellings which are typically 2 stores in height and share common walls. Sometimes, this style residence might have common stairwells.

13. Other:

If a specific dwelling is not close to any style descriptions, this classification should be used.

References

Anglin, P. M. and Gencay, R. (1996), "Semiparametric estimation of a hedonic price function," *Journal of Applied Econometrics*, 11(6), pp. 633-648.

Anselin, L. (1988), *Spatial Econometrics: Method and Models*, Kluwer Academic Publishers, Norwell, MA.

Abelson, P., Joyeux, R., Milunovich, G. and Chung, D. (2005), "Explaining house price in Australia: 1970-2003", *The Economic Record,* Vol. 81, No. 255, pp. s96-s103.

Abraham, J. M., and Hendershott, P. H. (1996), "Bubbles in Metropolitan Housing Markets." *Journal of Housing Research*, 7(2), 191–207.

Alexander, C. and Barrow, M. (1994), "Seasonality and cointegration of regional house prices in the UK", *Urban Studies* **31**(10), 1667–89.

Ashworth, J. and Parker, S. C. (1997), "Modelling regional house prices in the UK", *Scottish Journal of Political Economy*, 44(3), pp. 225–246.

Baffoe-Bonnie J. (1998), "The dynamic impact of macroeconomic aggregates on housing prices and stock of houses: A national and regional analysis", *Journal of Real Estate Finance and Economics*, **17**, 179-197.

Blettner, Robert A. (1969), "Mass Appraisals Via Multiple Regression Analysis", *Appraisal Journal*, Oct. 1969, Vol. 37 Issue 4, pp. 513-521.

Basu, A., and T.G. Thibodeau. (1998), "Analysis of Spatial Autocorrelation in House Prices", *Journal of Real Estate Finance and Economics*, 17:1, pp. 61–85.

Bourassa, S. C., F. Hamelink, M. Hoesli, and B. D. Macgregir. (1999), "Defining housing submarkets", *Journal of Housing Economics* 8: 160-83.

California Property Taxes, boe.ca.gov/proptaxes/faqs/changeinownership.

Chen, Z. et al. (2009), "Forecasting Housing Prices under Different Market Segmentation Assumptions", *Urban Studies*, January 2009, 46(1):167-187.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing of Scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

Cleveland, W. S., Devlin, S. J., and Grosse, E. (1988), "Regression by Local Fitting," *Journal of Econometrics*, 37, 87-114.

Coulson, N. Edward, (1992), "Semiparametric Estimates of the Marginal Price of Floorspace," *Journal of Real Estate Finance and Economics*, Springer, vol. 5(1), pages 73-83, March.

Cho, M. (1996), "House Price Dynamics: A Survey of Theoretical and Empirical Issues." *Journal of Housing Research*, 7(2), 145–172.

Clapp, John M., and Dogan Tirtiroglu. (1994), "Positive Feedback Trading and Diffusion of Asset Price Changes: Evidence from Housing Transactions". *Journal of Economic Behavior and Organization* 24:337-55.

Cook, S. (2005), "Detecting long-run relationships in regional house prices in the UK", *International Review of Applied Economics*, 19(1), pp. 107–118.

Dickey, D. and W. Fuller (1979), "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American Statistical Association 74* (366), 427–431.

Dubin, R.A. (1998), "Spatial Autocorrelation: a Primer", *Journal of Housing Economics*, Vol. 7, pp. 304-327.

Dale-Johnson, D. (1983), "An alternative approach to housing market segmentation using hedonic price data", *Journal of Urban Economics*, 11: 311-32.

Din, A., M. Hoesli & A. Bender, (2001), "Environmental Variables and Real Estate Prices," *Urban Studies*. 38(11): 1989-2000.

Do, Q. and G. Grudnitski, (1993), "A Neural Network Analysis of the Effect of Age on Housing Values," *Journal of Real Estate Research*, 253–64.

Friedman. J. and W. Stuetzel. (1981), "Projection Pursuit Regression." *Journal of the American Statistical Association*, 76:817-23.

Grasa, A. A. (1989), *Econometric Model Section: A New Approach,* Kluwer Academic, Boston.

Goodman, A. C., T. G.Thibodeau. (1998), "Housing market segmentation", *Journal of Housing Economics*, 7: 121-43.

Goodman, A. C., T.G. Thibodeau. (2003), "Housing market segmentation and hedonic prediction accuracy", *Journal of Housing Economics* 12:181-201.

Gencay, R. and Yang, X. (1996), "A forecast comparison of residential housing prices by parametric versus semiparametric conditional mean estimators," *Economic Letters*, 52(2), pp. 129-135.

Granger, C.W.J. (1969), "Investigating Causal Relationships by Econometric Models and Cross-Spectral Methods." *Econometrica*, 37: 424–438.

Granger, C.W.J. (1981), "Some Properties of Time Series Data and their Use in Econometric Model Specification." *Journal of Econometrics*, 16: 121–130.

Hartzell, D., Eichholtz, P. and Selender, A. (1993), "Economic diversification in European real estate portfolios", *Journal of Property Research*, **10**, 5-25.

Hastie, T. and R. Tibshirani. (1990), Generalized Additive Models. Chapman and Hall-New York.

Iwata, S., H. Murao and Q. Wang. (2000), "Nonparametric Assessment of the Effects of Neighborhood Land Uses on the Residential House Values" *Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics*. Vol. 14. JAI Press.

Jack F. Eisenlauer. (1968), "Mass versus Individual Appraisals." *Appraisal Journal*, October 1968, 532-540.

Johnes, G., and Hyclak, T. (1999), "House Prices and Regional Labor Markets." *Annals of Regional Science*, 33(1), 33–49.

Johansen, S. (1991), "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models." *Econometrica*, 59, No. 6: 1551 – 1580.

Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* Oxford University Press: Oxford.

Johansen, S., and K. Juselius. (1990), "Maximum Likelihood Estimation and Inference on Cointegration: With Applications to the Demand for Money." *Oxford Bulletin of Economics and Statistics*, 52 No. 2: 169 -210.

Kang H. B. and Alan K. Reichert, (1987), "An Evaluation of Alternative Estimation Techniques and Functional Forms in Developing Statistical Appraisal Models", *Journal of Real Estate Research* 21 (Fall 1987): 1-27.

Kearl, J. R. (1979), "Inflation, mortgages, and housing", *Journal of Political Economy*, Vol. 87, No. 5, Part 1 (Oct., 1979), pp. 1115-1138.

K. Hornik, (1991), "Approximation Capabilities of Multi-layer Feed-forward Networks," *Neural Networks*. 4: 251 – 257.

Kwok, Tin-Yau and Dit-Yan Yeung, (1997), "Constructive Algorithms for Structure Learning in Feedforward Neural Networks for Regression Problems" IEEE Transactions on Neural Networks: 8(5), 1131 – 1148.

Lenk, M., Worzala E. & Silva, A. (1997), "High-Tech Valuation: Should Artificial Neural Networks Bypass the Human Valuer?", *Journal of Property Valuation & Investment*, Vol.15 pp8-26.

Masters, T. (1993), Practical Neural Network Recipes in C++. Boston: Academic Press.

Ma, Zheng-Gui. (2006), "Residential Property Assessment." Master's Thesis, Department of Civil, Structural and Environmental Engineering, State Univ. of New York at Buffalo.

Meese R. & Wallace N.,(1991), "Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices." *Journal of the American Real Estate and Urban Economics Association*, 19(3):309-331.

Mays, J.E. (1995), "Model robust regression: combining parametric, nonparametric, and semiparametric methods." Unpublished Ph.D. dissertation. Virginia Polytechnic Institute and State University. Blacksburg, VA. 200p.

McGreal, S., A. Adair, D. McBurney & D. Patterson, (1998), "Neural Networks: the Prediction of Residential Values," *Journal of Property Valuation & Investment*. 16(1):57-70.

McCluskey, W. J. & R. A. Borst, (1997), "An Evaluation of MRA, Comparable Sales Analysis and ANNs for the Mass Appraisal of Residential Property in Northern Ireland," *Assessment Journal*. 4(1):47-55.

Munro, M. and Tu, Y. (1996), "The dynamics of UK national and regional house prices", *Review of Urban and Regional Development Studies* **8**, 186–201.

MacDonald, R. and Taylor, M.P. (1993), "Regional house prices in Britain: long-run relationships and short-run Dynamics", *Scottish Journal of Political Economy* **40**(1), 43–55.

Meen, G. (1999), "Regional house prices and the ripple effect: A new interpretation", *Housing Studies*, 14(6), pp. 733–753.

New York State, Real Property Tax Laws (RPTL) 301 and 305.

Nguyen, N. & A. Cripps, (2001), "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Network," *Journal of Real Estate Research*. 22(3):313-336.

Pace, R. K. (1993), "Nonparametric methods with applications to hedonic models," *Journalof Real Estate Finance and Economics*, 7(3), pp. 185-204.

Pace, R. K. (1995), "Parametric, semiparametric, and nonparametric estimation of characteristic values within mass assessment and hedonic pricing models," *Journal of Real Estate Finance and Economics*, 11(3), pp. 195-217.

Pollakowski, H.O. and Ray, T.S. (1997), "Housing price diffusion patterns at different aggregation levels: An examination of housing market efficiency", *Journal of Housing Research*, 8(1), pp. 107–124.

Robert A. Blettner.(1969), "Mass Appraisals Via Multiple Regression Analysis," *Appraisal Journal*, October 1969, 513-521.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1986), "Laerning representations by back-propagating errors," *Nature*, vol. 323, pp. 533-536.

Rapach, D. E., and Strauss, J. K. (2007), "Forecasting Real Housing Price Growth in the Eighth District States. Federal Reserve Bank of St. Louis." *Regional Economic Development*, 3(2), 33–42.

Rapach, D. E., and Strauss, J. K. (2009), "Differences in Housing Price Forecast ability Across U.S. States." *International Journal of Forecasting*, 25(2), 351-372.

Straszheim, M. (1974), "Hedonic estimation of housing market prices: a further comment", *Review of Economic and Statistics*, Vol. 56 No. 3, pp. 404-6.

Sims, Christopher A. (1980), "Macroeconomics and Reality," *Econometrica*. 48, pp. 1-48.

Smith, B. A., and Tesarek, W. P. (1991), "House prices and regional real estate cycles: Market adjustment in Houston", *Journal of the American Real Estate and Urban Economics Association*, **19**, 396-416.

Sims, C.A. and Zha, T. (1999), "Error bands for impulse responses." *Econometrica*. 67: 1113-1156.

Sternlieb, G., and Hughes, J. W. (1977), "Regional market variations: The Northeast versus the South", *Journal of the American Real Estate and Urban Economics Association*, 44-68.

Sutton, G. D. (2002), "Explaining changes in house prices", BIS Quarterly Review, September, 46–55.

Stevenson, S. (2004), "House price diffusion and inter- regional and cross-border house price dynamics", *Journal of Property Research*, 21(4), pp. 301–320.

Town of Amherst Property Assessment Database.

Tirtiroglu, Dogan. (1992), "Efficiency in Housing Markets: Temporal and Spatial Dimensions." *Journal of Housing Economics* 2(3):276-92.

Tu, Y. (2000), "Segmentation of Australia housing market: 1989-98", *Journal of Property Research*, 17(4), pp. 311–327.

USACE, U.S. Army Corps of Engineers (2005). *Town of Amherst Soils and Residential Foundation Study*.

Visit L., G. Christopher & M. Lee, (2004), "House Price Predication: Hedonic Price Model vs. Artificial Neural Network," *American Journal of Applied Science*. 3:193-201.

Worzala, E., Lenk, M. & Silva, A. (1995), "An Exploration of Neural Networks and its Application to Real Estate Valuation", *Journal of Real Estate Research*, Vol. 10, No. 2, pp. 185-202.