Spatial Analysis of Housing Markets Using the Hedonic Approach with Geographic Information Systems

by

Lijian Chen

Submitted to the Department of Urban Studies and Planning in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1994

Signature of Author ....
Department of Urban Studies and Planning
May 1994

Certified by ....
Professor Lyna L. Wiggins
Thesis Supervisor

Accepted by ....
Professor Karen R. Polenske
Chair, Ph.D. Committee

Spatial Analysis of Housing Markets Using the Hedonic Approach with Geographic Information Systems

by

Lijian Chen

Submitted to the Department of Urban Studies and Planning on May 7, 1994, in partial fulfillment of the requirements for the degree of Ph.D.

Abstract

The main objective of this dissertation is to explore the utility of geographic information systems (GIS) to urban and regional spatial modeling in general and housing market analysis in particular. It represents a multi-disciplinary attempt to improve our understanding of the spatial characteristics of urban housing markets by combining the strengths of traditional techniques of spatial analysis and GIS technologies. We focus on constructing improved housing price models and analyzing the impacts on prices of spatial, environmental, and socio-economic characteristics with the aid of GIS.

Most researchers agree that the price of land and property depends a great deal on the spatial characteristics of particular market areas and neighborhood environments. However, these important spatial factors are either largely ignored or oversimplified in most existing empirical pricing studies and in the day-to-day process of land and property assessment. This neglect is not mainly due to theoretical deficiencies but largely to a lack of effective technical tools to deal with these factors. We explore the utility of GIS in the following four aspects. First, we model housing and demographic diversities at different spatial aggregation levels using three commonly used traditional techniques of spatial analysis in combination with GIS. We demonstrate the significance of using block group level data for spatial modeling. Second, we use GIS network analysis capability to improve measurements of accessibility to employment centers. With these improvements in our hedonic price models, GIS-based accessibility measurements are favored over the traditional measurement of straight-line distance. Third, we use GIS to assist the process of segmenting into more homogeneous housing submarkets for the purpose of estimating segmentation price models. Fourth, by graphically displaying model residuals with the aid of GIS, we improve our ability to assess visually model limitations and reformulate the pricing model structure. The research suggests that the finer representation of spatial factors contributes significantly to the improvement of housing price models. Also, it demonstrates that enhanced graphic representations of spatial characteristics of housing markets can improve the model users' understanding of both model inputs and outputs.

The outcomes of this research contribute to our knowledge of critical relationships between housing prices, neighborhood socioeconomic patterns, and environmental characteristics, and illustrate that GIS can be an effective tool for urban and regional spatial analysis as well as the daily practice of assessors, appraisers, and planners.

Thesis Advisor: Lyna L. Wiggins
Title: Associate Professor
Department of Urban Studies and Community Health
Rutgers University

Associate Director
Center for Urban Policy Research
Rutgers University

Former Associate Professor
Department of Urban Studies and Planning
Massachusetts Institute of Technology

To My Mother

Acknowledgments

My thanks, first and foremost, go to the team of my advisors: Professors Lyna L. Wiggins, Joseph Ferreira, Jr., and Jerome Rothenberg. Their continual support and enthusiasm, and their invaluable guidance deserve my special mention.

Lyna provided inestimable help and guidance and threw in a large measure of encouragement from the moment I broached the idea for this undertaking and through every phase of my research. I most appreciated her clear advice and generous accessibility. Despite her many commitments, she took the time to edit every page of this dissertation painstakingly. Words are not adequate to express my gratitude to her.

Joe introduced me to the wonderland of planning support systems in my first week at MIT, and before long I was teaching in his introductory course. His enormous support and wise guidance spanned all these years. I gained many insights in his lectures and office, and I learned about life and career both on the squash court and on his sail boat. For all these, I am deeply grateful.

Jerry taught me the fundamentals of urban economics, and has been a constant source of encouragement and wisdom ever since I visited his spatial economics class. Since then, I have found myself knocking on his office door whenever I get a chance. It has always been a pleasant and enlightening experience to chat with him in his office with its "free" view of beautiful Back Bay and the Charles River. I cannot thank him enough for all the learning opportunities that he has given me.

Special thanks also go to Professor Karen R. Polenske, who talked me into attending MIT, the school that had offered me the least financial support. But I am forever grateful for her persuasion and her invaluable guidance, support, and counsel since then. I thank her for her interest in my academic potential and for taking such good care of me throughout my stay at MIT.

I am indebted to several other mentors who reviewed this research in its various incarnations: Frank Levy, Britton Harris, William Wheaton, Michael Shiffer. I also wish to thank my fellow students, colleagues, and friends, who helped shape many of the ideas in this dissertation and offered frequent and timely advice and technical assistance: John Evans, Qing Shen, Rob Smyser, Phil Thompson, Bijan Azad, Anne Kinsella, Laura Lebow, Kamal Azar, Zhizhong Yang, Paul Cote, Tom Grayson, Alicia Allen, and Greg Rossel. In particular, I would like to thank John Evans, who has been my teacher, office mate, co-instructor for three semesters, and co-author of many problem sets and teaching materials. I have appreciated his continual help in writing and speaking English and learning computers with such honesty, skill, patience, and humor. Most of all, I want to thank him just for being a good friend.

Deep thanks are due to Meg Gross, Mary O'Neil, and Haiyin Yu, who helped me with English editing, photocopying, and data entry on various occasions, and provided support and friendship when needed.

I would like to acknowledge John Avault, Alex Ganz, Greg Perkins, Bob Amatruda, and Mark Johnson of the Boston Redevelopment Authority for their comments on early phases of this research, and for kindly allowing me to use various resources at the Authority.

My greatest debt in writing this dissertation is to my wife Jenny (Zhuangyi). Her most loving encouragement, support, and care made its successful completion possible. I am forever grateful to her for understanding my newfound love for this dissertation, and for graciously tolerating the last few months, as I seemed to live in a time zone that was 12 hours behind hers.

Last but not least, I would like to acknowledge the World Bank for its financial support during the two years when I was taking required courses for my Ph.D. degree.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

There is growing interest in the application of Geographic Information System (GIS) to the field of urban and regional spatial modeling (Geertman and Toppen 1990; Schallaer 1990; Janssen and Rietveld 1990; Anderson 1990; Henry 1991; Landis et. al. 1991). Henry (1991) argues that developments in GIS open renewed opportunities for "authentic" regional science. He contends that using GIS allows better definition and delineation of regions, more accurate measures of spatial relationships, and improved procedures for developing estimators in spatial econometrics. In our view, these three possible contributions of GIS to regional science also can be applied to the study of urban housing markets. Moreover, other capabilities of GIS such as graphic representations of space and manipulation of spatial information provide a new helping hand in the analyses of spatial characteristics and interrelations of urban housing markets.

Housing is a unique commodity. It differs from other consumer goods because it possesses a set of special characteristics: heterogeneity; spatial-fixity; durability; and high search, transaction, and relocation costs. Due to these unique features, the price variations of housing are commonly analyzed by means of the attribute approach, or hedonic price model, and a large urban housing market is often segmented into a set of spatially distinct but related housing submarkets. There exists abundant literature on housing hedonic price models and housing market segmentation. An extensive review of the existing literature has helped us identify a number of deficiencies associated with the measurement of spatially related regressors such as neighborhood socioeconomic characteristics and accessibility to employment centers, and with the traditional approaches to market segmentation. This research demonstrates that these deficiencies can be significantly alleviated by making use of a GIS tool.

The goal of this research is to explore the utility of GIS in urban and regional spatial modeling in general and housing market analysis in particular. The research focuses on constructing improved housing price models and analyzing the impacts of spatial, environmental, socioeconomic characteristics of housing markets on pricing models with the aid of GIS. It represents a multi-disciplinary attempt to improve our understanding of spatial characteristics of urban housing markets by combining the strengths of several commonly used traditional techniques of spatial analysis and GIS technologies. This research represents an experimental step toward the "authentic" spatial analysis of urban housing markets.

1.2 Research Objectives

We are at a juncture where GIS technologies have become sufficiently advanced and sophisticated to help improve our understanding of spatial characteristics of urban housing markets. The measurements of spatial characteristics and relationships are of special importance to housing price modeling and market segmentation. In this research, we set out to explore the potential of GIS in providing a helping hand with spatial analysis of housing market. We investigate and make use of the strength of GIS in the relative ease of manipulation and analysis of geo-referenced information and graphical representation of spatial elements. Specifically, we have the following four research objectives:

The first objective is to improve traditional hedonic price models by refining the measurements of a set of spatially related variables with the aid of GIS. The refined measurement offers an opportunity to reexamine empirically the roles of these variables in affecting the changes of housing price. In this research, we focus on refining two sets of spatially related variables: (1) socioeconomic and demographic diversities measured at different spatial aggregation levels; and (2) accessibility to employment centers. It is generally agreed that these variables are important spatial attributes affecting housing prices. But past empirical research, as discussed in Chapter 2, has shown some controversial results on how these variables behave. Improved measurements of these variables in this research provide a new perspective for examining the estimates of the parameters of the hedonic price functions.

The second objective is to explore the strength of using GIS to integrate and capture spatial data that are important to housing price modeling. Also, we investigate the flexibility of using GIS for spatial data aggregation, market segmentation, and generation of new spatial data. We explore a new style of spatial modeling that allows ease of experimenting with various areal segmentations and data aggregation at different spatial levels. We carry out these explorations by combining GIS technologies with several traditional approaches of spatial analysis: Gini coefficient, location quotient, and the hedonic housing price models.

The third objective is to investigate the spatial patterns related to distributions of housing sales data, housing submarket locations, and the error residuals from housing price models. The GIS spatial representation capability is used to examine whether the results of hedonic models and market segmentation make good spatial sense. By examining visually maps of the spatial distribution of errors (e.g. residuals from the hedonic models) against layers of spatial characteristics over the urban landscape, we may discover new ideas to help us restructure hedonic models and identify important spatial characteristics that are helpful for interpreting model outputs. We compare and contrast the spatial characteristics of housing submarkets that are defined differently using several traditional stratification approaches. This analysis allows us to investigate the implications of the choice of market segmentation in the construction of hedonic price models.

The final objective is to explore and evaluate the promises and difficulties of using GIS technologies to conduct analytical research of urban and regional spatial modeling in general and the utility of GIS for the analysis of urban housing markets in particular. This enables us to reflect on whether using GIS to gain added knowledge of space is a worthwhile endeavor and on whether future research could benefit from our exploration.

1.3 Organization of the Dissertation

This dissertation consists of eight chapters. Chapter 1 introduces the related research fields that this dissertation covers and outlines four research objectives of this dissertation. It also highlights some contributions of this research. Chapter 2 reviews relevant literature on theories and empirical tests of housing hedonic price models, and housing market segmentations. The spatial characteristics of urban housing markets that are the focus of this research include measurements of neighborhood socioeconomic qualities and accessibility to employment centers. Also, a host of traditional spatial and non-spatial approaches toward segmenting housing markets are reviewed. More important, we discuss several weaknesses associated with the conventional ways of incorporating urban spatial features into housing price modeling.

Chapter 3 briefly examines the development of GIS and its capabilities, and reviews a number of pioneer applications of GIS in the field of urban and regional planning. Then, it illustrates the applicability of GIS to housing price modeling. Chapter 4 formulates and discusses research questions and hypotheses and presents the details of the research methodology. This chapter also provides several hypothetical examples which explain conceptually how GIS can play a constructive role in formulating research questions and testing hypotheses that are related to spatial analysis of housing markets.

Chapter 5 describes two study areas: one macro, the greater Boston area, and the other micro, the city of Boston. It then addresses issues related to data requirements, and discusses the sources of data for this research. Also, it illustrates the role of GIS in constructing a spatially integrated database for the housing market analysis in this research.

Chapter 6 consists of two parts. The first part presents and discusses the research findings on how the improved measurements of the spatially related variables impact on modeling neighborhood housing and demographic diversities and variations of median housing values across neighborhoods in the greater Boston area. It shows that GIS can serve as both a helpful complement to several commonly used traditional techniques of spatial analysis and as an important new tool for measuring and analyzing spatial relationships. In the second part of Chapter 6, using a set of housing sales data for the city of Boston, we show how GIS permits use of location data on housing sales, more realistic and localized measurements of accessibility factors, and flexible data aggregation at different spatial levels. We incorporate the added spatial information in hedonic price models for the Boston market. Then, we explore maps of the errors from these models, generated by GIS, to help us through the process of examining model behaviors, interpreting model outputs, and analyzing spatial relationships.

Chapter 7 explores and demonstrates the flexibility of using GIS to assist the process of housing market segmentation and to portray graphically the outcome of the segmentation. It also analyzes results of hedonic price models for three selected schemes of market segmentation, respectively, by accessibility to an employment center, family housing type, and race. The dissertation is concluded with Chapter 8, which offers a summary of our findings and discusses weaknesses of this research. We also discuss what could be done to correct these weaknesses and several possible future research extensions based on our findings.

1.4 Contributions to the Fields

The effort of attempting to integrate GIS technologies with urban and regional spatial modeling is still in its infancy. It has been the hope of many GIS experts and urban and regional researchers to see that the field of spatial modeling can benefit from the advance of GIS technologies, which has been widely considered as a promising tool for handling spatial dimensions with ease, speed, and a high level of realism. To date, however, exploration of this type remains scanty. It is even harder to find applications of GIS technologies to housing hedonic price modeling. In this sense, this research explores uncharted waters. It seems to add valuable contributions to the fields of GIS application and research and housing hedonic price modeling in a number of ways.

First, this dissertation provides a general framework for incorporating GIS technologies into several traditional techniques of spatial analysis such as location quotient and regression analyses. This integration shows how spatial characteristics of the urban and regional environment can be modeled more accurately and intuitively, and illustrates how the outcomes of modeling can be used to effectively assist housing market analysis.

Second, the case studies of both the macro and micro housing markets of Boston and its region with the aid of GIS represents the first effort to empirically test the impact of more accurately measured neighborhood quality variables using GIS technologies on models of housing price variations. Our empirical results indicate that research outcomes and conclusions can vary significantly when different spatial aggregations of census data are employed to measure neighborhood quality. This research also provides new evidence supporting the theory of the urban housing price gradient. Our tests of the four GIS-based alternatives of accessibility measurement reveal that the accessibility to employment center(s), measured in a more realistic manner with the aid of GIS, is a much stronger explanatory variable in housing hedonic price models than measures approximated as a straight-line distance.

Third, the dissertation presents numerous examples both conceptually and empirically of the potential power of GIS technologies in assisting in the analysis of urban and regional spatial environments. We demonstrate how GIS can be used for integrating different sources of data with common spatial components for housing market analysis, analyzing the spatial characteristics of sample data and research output, and effectively displaying various spatial information. In particular, we explore ways of displaying residuals from housing price models and discuss a number of valuable applications of GIS-based residual maps.

Fourth, we contrive and demonstrate a more flexible process of spatial modeling with the aid of GIS. In past research, a specific basic areal unit for data collection and modeling usually needed to be predetermined and maintained through the entire process of spatial modeling. Changing the basic spatial unit after completing data collection was either impossible or too costly and complicated to carry out. By first using GIS to construct a spatially integrated database, data can then be aggregated or disaggregated at different spatial levels and basic areal units further divided or combined in various forms. Thus, it becomes feasible to implement a flexible and non-linear process of spatial modeling.

In short, the key contribution of this dissertation is to demonstrate a number of ways that GIS can be used to assist spatial urban and regional modeling in general and housing hedonic price modeling in particular. The dissertation shows that GIS has a great potential to overcome some inherent problems of conventional spatial modeling techniques. It also shows that GIS can be used to generate value-added information in housing market analysis. In retrospect, we feel that this dissertation leads to more questions than it answers. We hope that this research effort will stimulate more scholars to explore the potentials of integrating GIS technologies with traditional urban and regional spatial modeling.

Chapter 2

Literature Review: Housing Hedonic Price Models and Market Segmentation

The following literature review is divided into two sections. In Section 2.1, we review the literature on housing hedonic price models and related issues. Our emphases are on the measures of neighborhood socioeconomic status and of accessibility to employment centers, and on the regression results for these variables in the existing empirical studies. In Section 2.2, we review the literature on theories of housing market segmentation and various traditional approaches to segmenting a large urban market. In each section, we also discuss several deficiencies associated with the existing literature and indicate areas where GIS might play a constructive role.

2.1 Housing Hedonic Price Models

2.1.1 Theory and uses of housing hedonic price models.

Housing is a multi-dimensional commodity. The purchase of housing is a purchase of a bundle of "attributes" (Lancaster 1966)[1], including a certain number of square feet of living space, different kinds of rooms, a particular structure type, an address, accessibility to employment, a neighborhood environment, a set of neighbors, and a diverse collection of public and quasi-public services such as schools, garbage collection, and police protection. The housing hedonic price models[2] are based on this notion of a housing unit

1. Kelvin Lancaster (1966) developed the concept of viewing differentiated goods as a bundle of characteristics.
2. A good summary of the hedonic technique can be found in Griliches' 1971 book.

as a collection of different attribute commodities. Each of these commodities has an intrinsic value to potential consumers. Therefore, simply stated, the market price of a housing unit is computed as the weighted sum of the values of each of these attributes (Griliches 1971).

The history of the hedonic approach may be dated to as early as 1961 when Griliches pioneered in developing a hedonic index to account for quality changes in automobiles. It was first introduced into economic analysis of housing markets by Ridker and Henning (1967). The main weakness of their research is a lack of a well-defined theoretical model that implies the correct interpretation of the estimates. Then, Rosen (1974) added his contribution to the theoretical underpinnings of the hedonic model. He showed that the hedonic price of a given component is a reduced form measure, an interaction of supply and demand market forces.

A rapid growth of empirical studies using hedonic price models occurred during the last fifteen years. The hedonic approach has been used to analyze a variety of issues related to the operation of housing markets. Some examples are: estimation of price elasticities and households' willingness to pay (Little 1976; Anas and Eum 1986), construction of housing price indices (Goodman 1978; Murray 1978; Palmquist 1980; Greenlees 1982; Mark and Goldberg 1984), analysis of racial market segmentation (Lapham 1971; King and Mieszkowski 1973; Kain and Quigley 1975; Straszheim 1975; Galster 1982; Smith 1982), measurement of benefits from pollution abatement (Harrison and Rubinfeld 1978; Nelson 1978; Freeman 1979; Giannias 1991), and impact of land use and zoning (Grether and Mieszkowski 1980; Mark and Goldberg 1986). In addition, the existing studies also deal with issues of specification of functional form, spatial and temporal stratification of markets, consistency (direction and magnitude) of coefficient estimations, estimating methods, and interpretations of coefficients and error residuals

(Palmquist 1984; Epple 1987; Bartik 1987; Giannias 1991). We make use of many of the findings in this rich body of literature when we build our hedonic models in this research. Generally, the form of the hedonic price equation can be expressed as

$$P = f(C1, \ldots, Cn) \tag{2.1}$$

where P is the sale price of an individual house and C_1 through C_n represent a set of dimensions of housing bundles that are considered to contribute to that price. The hedonic price of the ith dimension of C is defined as $\partial P / \partial C_i$. The hedonic theory provides no basis for a priori determination of functional form. The selection of functional form must be approached from an empirical viewpoint (Halvorsen and Pollakowski 1981). In the existing hedonic models, linear and log-linear forms are most widely used.

A major objective of this research is to derive more accurate measures of the spatial dimensions of housing bundles so as to estimate more reliable coefficients for the hedonic models. Thus, with this historical background information about hedonic price modeling, we now turn to reviewing the literature that has treated those spatially-related variables.

2.1.2 Measurement of neighborhood quality

Most urban housing researchers agree that neighborhood quality of a housing bundle has at least as much effect on housing prices as structural attributes such as number of rooms and lot size. The variables that determine the households' perception of the neighborhood quality may be grouped into three broad categories. First, the quality is influenced by some spatially related elements: accessibility to employment centers, proximity to environmental (dis)amenities (water surface, park, airport etc.), and land use. Second, it is affected by socioeconomic status of the neighborhood. This category includes variables such as racial composition, household incomes, proportion of adults with high education levels, incidence of crime, and education levels of household heads. Third, it is related to municipal services such as school quality, police protection, and garbage collection. In this research we concentrate on reviewing how the variable of accessibility to employment centers and the socioeconomic attributes of neighborhood are measured in the existing hedonic studies. Table 2.1 sums up the main measurement approaches adopted in the hedonic analyses.

(1) Accessibility measurements

Almost every empirical study of housing hedonic price models has tried to account for the effects of households' accessibility to employment centers. Column 3 in Table 2.1 presents several methods of measurement that have been used to generate the accessibility variable. In most studies, the accessibility to the employment center is defined as the straight line distance to the CBD, and is usually obtained from a direct measurement on a paper map. Some researchers (Kain and Quigley 1970; Lapham 1971; Rothenberg et. al. 1991) use the travel time from the house to the CBD as a proxy for the accessibility. Other variations of accessibility measurements include number of block units to the CBD, miles of bus lines or number of stations in a neighborhood, and straight line distance to several employment centers weighted by the size of each center. The examples in the table also include measurements of proximity to parks (Palmquist 1980) and to hazardous waste sites (Michaels and Smith 1990), using straight line distances as proxies. The fact that most analysts have attempted to account for accessibility to employment centers suggests an important role of this variable in households' residential location decisions; however, the reviewed empirical studies indicate no consistent relationship. Roughly a third of the regression models presented in the table end up dropping this accessibility variable after it is tested as statistically insignificant. Another third of the models find it statistically insignificant but still retain the variable in the models, as dictated by the traditional theory of residential location. Still another third of research findings show statistically significant

Table 2.1 Surveyed Articles on Methods of Measuring Accessibility and Neighborhood Socioeconomic Status

Author (by publication date)	Study Area (data year)	Accessibility Measurement	Measurement of Socioeconomic Status
Kain and Quigley (70)	St. Louis (67)	miles to CBD (x) & auto driving time to CBD (x) (not significant for both tenure markets); binary variable (x, -) (1 for entire metropolitan model, and 0 for the city)	census tract average
Lapham (71)	Dallas (60)	bus travel time to downtown (dropped because of correlation with lot size)	not specified
Schnare & Struyk (75)	Boston (70-71)	weighted straight line distance to five employment centers (weighted by number of workers) (dropped, replaced with dummy variables for inner/outer cities)	census tract average
Little (76)	St. Louis (61-71)	access measure (statistically insignificant, dropped) (employment & commercial centers are widely distributed in the city)	census tract average
Goodman (77)	New Haven (67-69)	miles from the CBD (v, -) (measurement method not mentioned)	block group average (1st study on differences between using block and tract level data)
Sumka (77)	North Carolina (66-71)	straight line distance to CBD (2x, 1v, +, -) (+ & - vary by family housing type)	census tract average
Wheaton (77)	San Francisco(65)	a composite index of census tract's proximity to employment center, surveyed time & cost of the respondents (v, x, -, +) (only 5 out of 21 equations have significant estimates for this accessibility variable)	census tract average & community average (the latter is not defined)
Goodman (78)	New Haven (67-69)	not mentioned	block group average (20-25% of the size of tract)

Note: An "v" denotes that the attribute variable is determined statistically significant at either 1% or 5% levels. A "x" denotes otherwise. A "+" or "-" represents the signs for the coefficients of the estimated attributes. When both "+" and "-" are present, it means that more than one equation were constructed and the resulted coefficient signs are inconsistent. The socioeconomic measurements usually include racial compositions, household income, school quality, neighborhood crime, and education levels of household head.

Table 2.1: Surveyed Articles on Methods of Measuring Accessibility and Neighborhood Socioeconomic Status *(Continued)*

Author (by publication date)	Study Area (data year)	Accessibility Measurement	Measurement of Socioeconomic Status
Palmquist (80)	Suburb of Seattle (62-76)	distance to nearest park (v, -)	not specified
Johnson (82)	Santa Clara County (75-78)	straight line distance to employment centers weighted by the size of each center (v, +, -)	census tract average
Brueckner & Colwell (83)	Chicago (79)	number of block unit to the CBD (2x, 1v, all-)	census tract average
Palmquist (83)	7 cities (76-79)	distance to CBD (x, v, +, -) (signs and coefficients fluctuate between cities); number of work place per square miles in each census track of residence (6x, 1v, 6+, 1-)	census tract average
Anas and Eum (84)	Chicago (72-76)	distance to CBD (v, +) (seemingly measured from tract center to CBD)	1/2*1/2 square mile zone & census tract average
Engle, et. al. (85)	San Deigo (73-80)	distance to bus & distance to shopping center (dropped, insignificant) dummy variable for view (x, +)	sample from one neighborhood (disregard neighborhood characteristics)
Anas and Eum (86)	Chicago (72-76)	accessibility to CBD measured by miles of bus lines, number of stations, trips by car and transit (v, +)	1/2*1/2 square mile zone & census tract average
Dubin and Sung (90)	Baltimore (78)	distance to CBD (x, +) (unknown measurement method)	block group average
Micheals and Smith (90)	Boston (77-81)	straight line distance to waste sites (v,x,+,-) weighted distance to employment centers (v, -) (+ & - vary by submarket)	census tract average
Rothenberg, et. al. (91)	Des Moines (60-71)	one way travel time (in minutes) from the centroid of each tract to the CBD on roadway distance (used as a criterion to partition the housing market)	census tract average

Note: An "v" denotes that the attribute variable is determined statistically significant at either 1% or 5% levels. A "x" denotes otherwise. A "+" or "-" represents the signs for the coefficients of the estimated attributes. When both "+" and "-" are present, it means that more than one equation were constructed and the resulted coefficient signs are inconsistent. The socioeconomic measurements usually include racial compositions, household income, school quality, neighborhood crime, and education levels of household head relationships between the dependent and accessibility variables. But they show little agreement in terms of direction and magnitude.

It seems that most of the work that uses the hedonic approach is unsatisfactory because the estimations do not yield consistent coefficients for the employment-accessibility index. This result is at variance with most economic theories of residential location, which would predict the market value of a standardized housing bundle to decline with distance from the city center (Alonso 1960 and 1964; Mills 1967; Harris *et. al.* 1968; Muth 1969). Several possible explanations have been offered. According to Kain and Quigley (1970), these results may be due to inadequate standardization, improper specification of the equations, poor measurement of accessibility, or the inadequacy of existing theories. Brueckner and Colwell (1983) suggest that the locational dummy variables in their model may capture much of the influence of accessibility. Little (1976) explains that the insignificance of accessibility is due to the fact that employment and commercial centers are widely distributed in his study area. Lapham (1971) finds that his representation of accessibility, bus travel time to CBD, is correlated with lot size, resulting in the impossibility of separately identifying their influence on price.

In our view, the inconsistent and unsatisfactory regression results associated with the accessibility variable in most existing hedonic models are perhaps due to the use of a poor proxy, usually the straight line distance to a single center. Even when the accessibility is measured in terms of travel time, the measurement is either a simple linear conversion from straight-line distance to time without taking into account of the conditions of streets and highways, or a mean score representing the entire census tract's accessibility to CBD (from the geometric center of the census tract to CBD). These measurements in many cases are extremely poor proxies for the accessibility of an observed dwelling to the CBD. Moreover, most of these models assume that the employment center is in the CBD. This is a weak assumption given the fact that a significant dispersion of work places has occurred in many cities in the past decades. Given the mixed verdict that emerges from the literature on the effects of the accessibility variable, we need a better measure of accessibility to test its role in affecting the housing prices and the existing theories of residential locations.

(2) Measurements of neighborhood socioeconomic status

Another major concern of this research is the measurement of socioeconomic status of a neighborhood. In Table 2.1, the last column shows that all the hedonic regressions reviewed incorporate socioeconomic variables. Most of these studies use mean scores of socioeconomic characteristics of census tracts to approximate the socioeconomic character of the neighborhood in which the dwelling unit is located. This choice hinges on the fact that the data, aggregated by census tract, are readily obtainable from the Census Bureau. Two papers by Goodman (1977 and 1978) and one by Dubin and Sung (1978) are the only exceptions. Instead of the census tract aggregation, they use block group data as a basic reference unit. Goodman finds that using these block group units, typically 20 to 25% of the size of a census tract, adds some explanatory power over the customary tract aggregation level. Using neither tract nor block group aggregations, Anas and Eum (1984 and 1986) divide their study area into a set of 1/2 by 1/2 square mile zones and then compute socioeconomic variables based on these smaller zones. But judging from their paper, it seems that some of their socioeconomic variables are still based on the census tract average.

A neighborhood is usually defined as having a fairly high degree of homogeneity. Goodman (1977) suggests that a neighborhood should be defined as a small urban area within which the residents receive and perceive a common set of socioeconomic effects and neighborhood services. In particular, a neighborhood should be characterized by an areal aggregation that is large enough to summarize the neighborhood effects that are common to small groups of residents, yet small enough to distinguish significant differences in these effects among neighborhoods across the large urban housing market. However, the final determination of the proper aggregation is limited by availability of suitable data. Goodman (1977) shows that the block group aggregation is a better choice than the census tract. He concludes positively that measuring the neighborhood variables at the block group level contributes some additional explanatory power to the estimated hedonic price models. It is usually difficult to obtain a lower level of aggregation, such as a single block or a parcel group, because of the legal confidentiality requirement for compiling census data. With the aid of GIS, in this research we analyze the differences between using tract and block group levels of aggregation for spatial modeling and explore the flexibility of using GIS for organizing, aggregating, and generating spatial data that are related to housing markets.

Most of these attempts to estimate the implicit market price of accessibility and socioeconomic attributes are deficient because they fail to represent adequately the complexity of accessibility and because they lack adequate measures of socioeconomic status of a neighborhood or in most cases rely exclusively on aggregate census tract data. In this research, we seek to correct these deficiencies by making use of the network analysis capability of GIS to compute a street distance of the shortest path to work (which allows dropping the straight line distance assumption), and by experimenting with different areal aggregations of neighborhood socioeconomic quality. We use both the tract and block group levels of aggregation for our empirical studies.

2.2 Housing Market Segmentation

2.2.1 Importance of housing market segmentation

The significance of housing market segmentation in the construction of the hedonic price model has been widely recognized. In his study of housing market segmentation in the San Francisco Bay Area, Straszheim (1975) finds that there is a large variation in housing characteristics and prices across housing submarkets, which he considers a fundamental characteristic of the urban housing market. In his study of Santa Clara County, Johnson (1982) notes that segmentation should not be viewed as a trivial issue when the intention of the research is to examine the behavior of individual hedonic prices. Therefore, it appears that to treat an entire urban area as a single housing market is often inadequate and misleading. It is necessary to estimate hedonic price models for each identifiable housing submarket. The following literature review shows that many economists concur.

Economists[3] have tested positively the need of estimating hedonic prices within smaller submarkets (Straszheim 1975; Goodman 1978; Johnson 1982; Rothenberg *et. al.* 1991; Michaels and Smith 1990). For example, Goodman (1978) estimates hedonic prices for fifteen geographically segmented submarkets in the New Haven SMSA and rejects the null hypothesis of equality of entire sets of coefficients across space. Straszheim (1975) points out that the distinct neighborhood and environmental attributes in each submarket affect not just the intercept terms (this is the case when dummy variables are used to represent different submarket areas) but also the coefficients of the several qualitative and quantitative indexes describing the individual unit. Also, location rents or other characteristics of prices unique to any given submarket are unlikely to appear only in the intercept term or in the coefficients for such variables as “distance to the CBD”. He

3. Schnare and Struyk’s study (1976) seems to be the only exception. They find little evidence of supporting submarket segmentation for the Boston area, however.

contends that to allow these coefficients to vary across submarket there is no substitute for partitioning an SMSA market into more homogeneous and smaller markets before the estimation of hedonic prices.

Despite the fact that the importance of housing market segmentation is recognized by many economists, according to Michaels and Smith (1990) most hedonic applications have treated the market as coincident with the metropolitan area and estimated a single price function to describe the equilibrium prices. While we favor the view that it is important to estimate separate hedonic models for each submarket, we also endorse Schnare and Struyk's argument that, in some circumstances, a single equation model could be sufficient to provide some serviceable answers about price variations. But it is important to note that the insight that a single equation can offer is limited and should be used with caution. Also, in many situations a single equation is not sufficient. In this research, we underline the importance of market segmentation and investigate several segmenting approaches with the aid of GIS.

2.2.2 Traditional segmentation approaches and their implications

There are a number of theories that have been developed to date by economists to segment a large urban housing market. The choice of theory depends on one's purposes of study. However, a common principle for housing market segmentation is to preserve as much within-submarket homogeneity in terms of a chosen criterion, or a combined set of criteria, as possible. Examining the existing readings on urban housing markets, we find that most researchers employ one or a mix of the following approaches towards segmenting housing markets.

First, urban housing submarkets can be defined based on the mean or median household or family income levels (Straszheim 1975; Wheaton 1979; Rothenberg et al. 1991). Second, submarkets can be categorized according to racial composition, usually measured by the percentage of black population (Crecine, Davis, and Jackson 1967; Muth 1969; King and Mieszkowski 1973; Kain and Quigley 1970 and 1975; Straszheim 1975; Courant 1977). Third, they can be defined by housing types (*i.e.*, rental apartments, multi-family units, single detached, etc.) (Sumka 1977). Fourth, submarkets can be defined in terms of accessibility to employment centers[4] (Straszheim 1975; Goodman 1978; Rothenberg et. al. 1991). Besides these four most commonly used stratification schemes (*i.e.*, by income, race, tenure, and accessibility), a large housing market can also be segmented in terms of building density, population density, or physical conditions (size, age, and quality) of the housing stock (Straszheim 1975). Depending on the research problems in question, a choice of any criterion or criteria might suffice to a certain extent.

The list of stratification methods would not be complete without mentioning two unique and less frequently used approaches. First, Michaels and Smith (1990) partition a metropolitan housing market by eliciting realtors' descriptions of submarkets. Based on a number of credible realtors' knowledge of the housing market, an array of housing submarkets are determined. The second, and probably the most complex approach for market segmentation, is Rothenberg's (1991) market stratification by hedonic price indexes. Each housing unit is grouped into a submarket based on its hedonic value. A set of submarkets for high, medium, and low quality housing units are identified.

Regardless of the approach employed for defining urban housing submarkets, one common shortcoming in most economists' work[5] on urban housing market segmentation is that readers often have little notion about where these submarkets are situated spatially in the urban and suburban landscape and how they relate to each other (Johnson 1982). Thus, houses that are spatially unrelated, or located miles apart in space, are often grouped

4. In most cases, straight line distance to CBD is the stratifier. Thus, segmented submarkets are characterized by a set of concentric rings surrounding the CBD.

into the same submarket. For example, it is possible that two houses that are separately located on opposite edges of a city could be grouped into the same submarket if one (or several) of the above approaches (except by accessibility) is used. Without an effective and efficient tool to portray the results of market segmentation spatially, it is nearly impossible to know whether submarkets are spatially mutually exclusive or overlapping, how submarkets segmented by different stratifiers differ from each other in terms of location, size of market area, population and housing density, *etc*. Thus, the spatial dimensions of the housing submarkets have be analyzed in a highly abstract manner. Consequently, it is difficult to place a high level of confidence on the research results of spatial characteristics of housing submarkets and for these results to be utilized in day-to-day policy analyses and planning activities.

Another important question, which has remained largely unexplored in the existing literature on housing market segmentation, is to what extent the housing submarkets defined using different criteria correspond or contrast to each other. We have been able to locate only one published work (Rothenberg et. al. 1991) that has devoted attention to this question. Rothenberg and his colleagues, in their analysis of Des Moines housing market segmentation, show that the metropolitan housing submarkets can be defined based on a set of quality indexes estimated using the hedonic approach. The types of housing properties within each of these quality-segmented submarkets are then compared with

5. Straszheim's study on housing market compartmentalization in San Francisco Bay Area is an exception. He uses simple schematic maps to show the geography of the housing submarkets and is able to tell a rich story about the spatial patterns such as price, rent and income surfaces. His simplified representations of housing markets show some signs that using a mapping tool for spatial analysis of housing markets can be very fruitful. For example, his examination of the age distribution of the current stock by geographical submarket reveals the approximate timing of new construction over several decades. Also, using maps he is able to find that there is a large difference in prices of comparable housing between submarkets near to each other. Moreover, the maps provide a good visual summary of the overall spatial variation in housing prices in the Bay Area. As many as eight maps are used to analyze the spatial patterns of the housing markets. But due to the lack of an effective tool for graphic representations, most maps are hard to read and spatial patterns not readily identifiable.

those of the housing submarkets defined in terms of racial composition, income levels, and accessibility, respectively. To their surprise, these differently defined housing submarkets do not correspond to each other well. In terms of the quality of housing, each of the spatially distinct submarkets is far from being homogeneous. Usually housing units with a wide spectrum of quality coexist within each of the submarkets. This finding contradicts not only our intuition about what a housing submarket should be but also the homogeneity principle of housing market segmentation.

Again, the paucity of the research on the comparison between the distinctively defined sets of housing submarkets may be primarily attributed to the lack of a good tool to represent the spatially-oriented market areas in an easily comprehensible manner, and to manipulate or compare them with ease. Rothenberg and his colleagues used a cross-tabulation approach. of comparing pair-by-pair the sets of differently defined submarkets. But this did not allow them to examine the spatial relationships among submarkets. Consequently, they can offer only limited explanations to their findings of coexistence of housing with various quality levels within a same submarket. In addition, the cross-tabulation approach allows comparison of only two sets of submarkets at a time. In order to gain more insights into how differently defined housing submarkets are related spatially to each other and how they affect the study of hedonic price model, an effective tool capable of handling the spatial dimensions of urban housing markets is needed. A possible candidate for this tool lies in the increasingly mature GIS technologies.

In next chapter, we review the development of GIS technologies and their growing utility in the field of urban and regional spatial analysis in general. In particular, we discuss how GIS can be used to analyze the spatial characteristics of urban housing markets.

Chapter 3

Potential Contributions of GIS to Housing Market Analysis

In this chapter, we first briefly describe the capabilities of a typical GIS. Then, we review some relevant literature on the application of GIS in the field of urban and regional planning and identify deficiencies in the literature. Next, we discuss the significance and applicability of incorporating GIS into the research of urban housing markets. In the last section of this chapter, we discuss how GIS technologies have made it feasible to implement an exploratory process of spatial modeling that allows greater flexibility in data aggregation and areal segmentation.

3.1 GIS and Its Capabilities

A Geographic Information System (GIS) is a system of computer hardware, software, and procedures that are designed to perform the capture, management, manipulation, analysis, and display of spatially referenced data. In the context of urban space, we can view GIS as a mapped representation of a city or other jurisdiction with a data base consisting of individual and aggregated observations about land, and uses and events located on it. The mapped representation of a city is a particular type of database in itself (Harris 1989). It may consist of many different maps dealing with different human activities such as land ownership, population density, economic activity, and public safety, and with a variety of natural and environmental characteristics, such as slope, elevation, soil type, air quality, and vegetation.

GIS is well known for its ability to store statistical data for the analysis of trends and developments and for the presentation of this information geographically and visually. Besides the ability to maintain and retrieve both spatial and nonspatial data about places, GIS usually offers these two basic capabilities: (1) arithmetic and geometric analysis, allowing calculations of distance, area, and density; (2) overlay facilities, allowing identification and display of areas with common features, and regrouping of geographic data based on specified common features.

In addition, some GIS proprietors claim for their packages two additional capabilities: statistical analysis and impact analysis. The former is a module that can perform a set of statistical manipulations on attribute variables in the spatial databases (or system content variables), such as regression analysis. The latter is usually restricted to a simple arithmetic or geometric operation, such as calculating and displaying the impact of increased congestion on travel times in the peak hours. Compared to the type of information that a traditional modeling approach would provide, most GIS software is rather limited in analytical capability. However, it does provide necessary hooks that link to other computerized analytical tools such as spreadsheets, database management systems, statistical software, and analytic models.

One of the main benefits of GIS is in creating a common spatial environment for data collected from different sources. Data can be linked so that new, previously unknown, or hard-to-obtain information can be generated (through overlay techniques with location information or combining attribute information using statistical or modeling techniques). It is then possible to create new sets of spatially oriented information for sophisticated statistical analysis. By making using of GIS capability of integrating spatial data with common geography, in this research we are able to combine census data, digital data of street network and census spatial units, and housing sales data to form a spatially integrated housing database.

Another advantage of GIS lie in its power of database management. It enables interactive access and manipulation of both spatial and non spatial data for the purpose of exploring spatial relationships. Once various geo-referenced data are integrated, GIS can be used to easily re-aggregate or disaggregate spatial data based on various needs, to further partition or unite basic spatial units to form new spatial entities. These capabilities of GIS have some profound implications in ways that we conduct urban and regional spatial modeling exercises. We elaborate possible implications and discuss a more flexible process of spatial modeling below in Section 3.4.

GIS is still undergoing rapid change and advancement. It still has long way to go before they become truly mature. Software has yet to become standardized and many GIS developers are competing to innovate and excel. This diverse market for GIS technologies presents a good opportunity for those who are interested in exploring their potential. The diversities allow users to couple up GIS technologies with other software and move data between different platforms. For example, in this research we are able to bypass the weakness of current GIS technologies in regression analysis by making using of more developed statistical software.

Many choices of GIS software are available today. Among them, the most widely used is ARC/INFO, developed by Environmental Systems Research Institute, Inc. We have chosen this software for this research because of its many powerful analytical subroutines and our experience in using it in previous research projects.

3.2 Literature Review of GIS Applications

Although GIS is regarded by some (Healey 1991) as having reached an age of maturity, the majority of applications of GIS are currently used primarily to improve the handling and presentation of data. For example, many local authorities have constructed land and property information systems. There are abundant supplies of these types of application examples in the conference proceedings of organizations such as the Urban and Regional Information Systems Association. In the 1990s there has been a rapid increase in journal articles and books contributing to the fundamental knowledge about the theory and practice of modern spatial information systems (Scholten at. al. 1990; Allen et. al. 1990; Huxhold et. al. 1991; Antenucci et. al. 1991; Wiggins et. al. 1990, 1991, and 1992).

There have been a wide variety of GIS applications in the field of urban and regional planning. However, the success of GIS in the field of urban and regional planning thus far has been more in its technological capabilities rather than its contribution to analytical capabilities. Most informed readers would agree that the most impressive and powerful role of GIS has been in visual presentation, mapping, and information integration. It is also widely recognized that most proprietary GIS packages are deficient in terms of their analytical and value adding capabilities (Clarke 1990). Thus, it seems that some form of greater integration between GIS and analytical modeling techniques is needed.

There are very few readings on the use of GIS for housing market analysis. One example is the study of regional housing planning in Holland by Geertman and Toppen (1990). They show how the GIS can be used to help translate general housing policy statements into concrete location decisions. In this application, information on changes in population growth, household size, compositions, and life-style are used to forecast the demand for dwellings. With the aid of GIS, the possible supply is estimated by identifying the opportunities in the developed urban areas and the alternative new sites for housing development. The supply also takes into account the locations of schools and accessibility to railway stations, and the environmental impacts of development.

Another similar GIS-based work is an ongoing U.C. Berkeley research project, the Bay State Population Forecasting Model (Landis et. al. 1991). The model consists of three parts: a demand equation which is an econometric submodel of population growth, a supply component which is a GIS-based submodel of land development, and an allocation submodel which matches the demand with supply. The model is dynamic in the sense that the population forecast results can be adjusted interactively based on the examination of the availability and desirability of land use reality. Thus, the model allows simulations of various growth scenarios. The weakness associated with this GIS-based model seems to be the lack of sound theories for explaining the match of population growth with the available development space.

Most of the existing GIS-related studies are descriptive in nature and are in the environment of "data-rich-theory-poor" (Marble 1991). Perhaps the most obvious deficiency in the existing housing-related studies using GIS is the lack of exploration of how GIS can contribute to the analytical modeling of housing markets. We hope that the results of this research on urban housing markets represent a fruitful effort to correct this deficiency.

3.3 Applicability and Significance of Incorporating GIS

The use of GIS to assist urban and regional spatial analysis in general and housing hedonic price models and market segmentation in particular seems to be still an unexplored territory. Therefore, it is only natural that one would ask how feasible and significant it is to use GIS in the analysis of urban housing markets. We approach this question from the following six perspectives.

(1) GIS is not just a mapping tool that produces remarkable maps.

It is a misconception to view GIS simply as a sophisticated mapping tool. GIS is well known for its ability to produce remarkable map products. But what lies at the heart of GIS is the concept of "The Map as a Database" (Cooke 1991). It is the interlinking of different data for the same location which makes GIS "an analytical and decision making tool fundamentally different from a paper map" (Healey 1991). In essence, GIS-based map data are similar to digital numerical data as both are stored in computers or disks as bits and bytes. GIS is a tool designed to create, update, manage, and manipulate this map database. In this sense, it works in a similar way as computer statistical packages, such as SAS and SPSS, which are capable of handling computer encoded numerical data. This notion of "Map as Database" is important because it means we can apply GIS to solving real problems in addition to producing remarkable maps.

One good example of the use of GIS beyond map making might be the "*Directions*" software in the public domain within MIT's computing network. To use it, a user needs to input two real addresses in the great Boston metropolitan area: a departure point (A) and a destination (B). Within seconds, the user receives a response with detailed instructions of how to drive from A to B. The instruction takes into account street intersections, rotaries, traffic lights, total mileage, one way streets, and on which side of the street one's car is parked before departure. Clearly this application goes beyond producing pretty maps. In fact, the maps used in the software have not been made visible to users. But the application can only be implemented if a complete and accurate map database exists.

This example shows that, with the aid of GIS, we should be able to measure more accurately the variable of accessibility to employment centers, shopping centers, or schools, which are among those important factors affecting housing prices. At the least, we could measure the shortest real distances (or time) of journey-to-work, which is likely a better proxy for the accessibility variables than the straight-line distances to be used in housing hedonic models. We discuss more about the measurement of the accessibility variables in Chapter 4.

(2) GIS permits more realistic spatial analysis.

Recently we have observed a revived interest among economists in the exploration of the importance of space in the areas of regional trade, population mobility, and production locations (Krugman 1991). A consensus among the economists and economic planners is that space matters. This conviction has been substantiated in various research efforts (Alonso 1960, 1964, 1975; Lowry 1967; Mills 1979; Richardson 1979; Wheaton 1979). Most of this work, however, has been based on the assumption that the complex and dynamic regional or urban space consists of lines and points on a featureless plain. This bold simplification has resulted in significant advances in spatial economic theories. Unfortunately, viewed from the perspective of day-to-day economic planning activities, many of these spatial economic theories seem of little practical value. Ultimately, most economic activities are taking place in a very sophisticated and concrete physical space. It is very difficult for economic planners to make use of the much oversimplified and much too abstract spatial models and theories.

In contrast with the economists' traditional highly-simplified way of portraying space, GIS technologies seem to provide an alternative approach to representing space in a more realistic manner. This alternative allows the significance of space to be analyzed more vividly and intuitively, and the complexity of space to become more comprehendible by a much wider range of audience. The GIS allows us to think about variables in spatial terms -- in essentially three dimensions (Henry 1991). It allows us to think in terms of location in a layer system, in terms of distance between entities, and in terms of an "overlay" coincidence of spatial variables. The power of GIS is the ability to work within these multiple dimensions. This strength seems well suited to the analysis of housing, which is usually characterized by multi-dimensional spatially related attributes.

(3) GIS improves existing economic modeling of housing markets and allows us to ask new questions.

GIS can be used to improve the existing urban housing models by achieving better spatial definition of submarkets, more accurate measures of spatial relationship, and an improved approach to developing estimators. By making use of GIS plotting, overlaying, and buffering capabilities, we can compare various methods of market segmentation and search for a more meaningful and rational criteria for defining urban housing submarkets. To segment housing markets, we consider not only those traditionally employed criteria, such as income, race, tenure, and accessibility, but also the information on the uniqueness of spatial features of the studied areas. No cities are identical in space. Therefore, the latter information is of great importance for housing market segmentation. Yet this information was either not available for the previous housing market models or simply ignored. With the aid of GIS, we can add to a great deal of realism to the complex urban housing models.

Moreover, the spatial information is also useful for identifying relevant spatial characteristics to be included and tested in housing hedonic price models. More realistic representation of urban space permits greater precision in the measures of spatial relationship. Thus, the spatially related regressors for housing hedonic models can be estimated more accurately. This improved accuracy provides an opportunity to verify or refute, with a greater degree of confidence, the significance of the spatial features in affecting housing prices and households' residential location decisions, as revealed by the previous hedonic studies and urban structure theories.

GIS enables a new perspective to be added to the examination of the spatial features of urban housing markets. We believe that the changes which GIS has brought about, and will continue to bring about, are profound. Besides being a better tool, GIS will lead to new research questions and problems to be formulated because new solutions are possible. Marble (1990) and Zubrow (1990) contend and illustrate in their papers (using examples of regional economics and archaeology, respectively) how the use of changing tools had a ripple effect and how soon the nature of the questions being asked was altered. We expect similar consequences from the introduction of GIS to the spatial analysis of housing markets. For example, GIS permits more localized and realistic measures of access and easy update of these measures. This makes it possible to test whether a more accurately measured accessibility has impact on housing price.

(4) GIS-based spatial analysis serves as part of verification tool for economic modeling.

Using GIS to portray and display the spatial contents of urban housing markets may serve as an effective check against the possible flaws associated with the traditional pure mathematical analysis of housing markets. It is well known that econometric models generally serve as a better tool for identifying and measuring historical relationship than for predicting the future (Greenberger et. al. 1976). This is due to two possible errors associated with the formulation of an econometric model. First, a statistically significant relationship between variables in question may be obtained by chance. Second, there is always a chance of wrongly establishing a causality between variables that are in reality mutually influencing. For the purposes of planning exercises, we usually are more interested in using models for predicting the future. In order to guard against the possible errors embedded in the econometric models, it seems useful to add to the models a mechanism that allows checking the validity of the forecasting results. In the case of housing market analysis, using GIS to represent the locations of housing submarkets, available land for new development, price-surfaces, and regression residuals, as well as other spatial characteristics, may help us examine our prediction results of housing markets at least visually and intuitively, and perhaps statistically (depending on data availability).

(5) GIS-based housing analysis increases practical values of economic modeling of housing markets.

There are some practical benefits for utilizing GIS to portray various theoretical and empirical research results of economic analysis of housing markets. Virtually all aspects of the efficiency of cities and of their quality of life have spatial dimensions that are defined in the light of the interaction between various parts of the urban system. The information most useful in planning, administration, and policy making processes involves spatial distributions and their interrelationships, as defined in part by the coincidence, contiguity, and propinquity of different activities, events, and conditions. However, this complex spatial information represented in either verbal or numerical forms, or the two combined, can be intimidating to many audiences. This is partly because most of the general public and decision makers, and many urban planners, are visually oriented.

A GIS is capable of representing the complex spatial relationships in a visually appealing graphical form, and adds an important dimension to the communication of information, which is the key to successful public participation and decision making. Therefore, the incorporation of GIS, especially its graphic representation capability, into the economic modeling of urban housing markets is likely to make it easier for housing planners, policy makers, and the public to understand and utilize the results of academic research. In other words, it is hoped that the uses of GIS can make the economists' results more "user-friendly". In this sense, the mapping and display characteristics of a GIS are of real psychological and operational importance. The use of GIS in urban housing studies for interpretations and graphical representation of spatial characteristics will increase the utility of economic research results, and perhaps more important, make scientific research more inviting to the participation of the public and policy makers.

(6) GIS-based housing analysis is becoming cost-effective.

GIS-based housing market analysis has become more cost-effective in recent years. This change can be attributed to two primary factors: the intrinsic nature of housing and rapid progress of computer technologies. First, the special characteristics of housing stock make the housing research an obvious and promising candidate to explore the economy of GIS. Housing is unique because it is generally immobile: it is costly to move a dwelling from one location to another. Put differently, housing is spatially fixed. Also, housing is durable. If a house is maintained properly, it can easily last for many decades. These two unique characteristics of housing give a strong cost-effective reason for municipal planning agencies or private residential real estate companies to invest in the construction of a GIS-based housing information system for their housing markets. The potential gains for building such a housing information system is likely to be much greater than the costs involved, especially in the long run.

Second, the cost function of initial implementation of a GIS project includes three basic elements: the costs of hardware, software, and the creation of a map database, with the last one taking a lion's share. Then, there is a fourth element: that is, the cost of maintaining a sustainable staff to keep the project alive and to ensure the continuation of a long term success. A GIS cost and benefit study is difficult because the initial implementation cost can be overwhelming and because the benefits are difficult to identify since they can be realized in so many different government functions[6] (Huxhold 1991).

6. According to Huxhold (1991), local municipalities that have completed their digital map conversion process have reported that the creation of the digital base map have ranged between 45 to 80 percent of the total cost implementation of the GIS project.

Nevertheless, over the past decade, less expensive computer workstations, improved software for GIS, and the availability of spatial data like the census TIGER/Line files have combined to expose GIS tools to a wider audience of practitioners, researchers, and the public. Consequently, many city planning departments and private firms have completed the initial implementation of GIS projects and are in a process of seeking wider and more sophisticated utilizations of these projects. In addition, GIS technologies have come to the desktop, indicating a new era of increasing use and popularity. These new developments have laid down obviously necessary infrastructure for us to carry out this GIS-based research of the urban housing market. This research has built upon numerous laborious and successful GIS projects completed at the Computer Resource Lab at MIT and benefited from many sources of existing digital map data. More detail on the availability of map databases is presented in Chapter 5.

3.4 An Exploratory Process of Spatial Modeling

Before there was computer statistical software, it took a long time to just estimate one version of a regression model. With today's computation capabilities in statistical packages, the task of model estimation becomes trivial. One can easily obtain many versions of a model in a short time. This allows researchers to concentrate on comparing and interpreting different models, fine-tuning model inputs, and exploring what is a better model. It seems that the most important value of GIS to urban and regional spatial analysis is equivalent to the relationship between computer statistical packages and econometric analysis.

With respect to spatial modeling, GIS is not about making the impossible possible. It is not that various spatial manipulations were not doable before. But they were so complex, tedious, and time-consuming that few ended up trying. Therefore, GIS is about making the possible more feasible and practical, and to enable execution with greater accuracy and higher speed. It is about a greater flexibility of exploring, generating, and manipulating spatial data and spatial analysis. It permits conducting more rounds of analyses, using a greater volume of data, and adding additional spatial dimensions. GIS is about making a new style of spatial modeling feasible and more flexible than the traditional linear style of research.

Traditionally, empirical research on urban and regional economic spatial modeling follows a linear process. First, a basic spatial unit is determined, such as a census tract or a town. After the spatial data are collected, measurements of various variables are aggregated or disaggregated based on the predetermined spatial unit. This spatial level of data aggregation is then maintained throughout the entire process of spatial analysis. It is very difficult to change the level of data aggregation if it is later found to be inappropriate. To correct this inappropriateness, a researcher would have to wait for his or her next research grant to come through or hope that his or her suggestions on future research extensions are taken seriously by others.

Another difficulty often encountered in the traditional process of spatial analysis is to combine or integrate different data sources collected for the same study area. Despite the common geography shared by the data, unless each observation is pre-coded with a common identification, the task of data integration is often tedious and time-consuming. For example, to construct a housing hedonic price model, two sources of data are often collected: housing sales data and census data at the tract level. The sales data often have fields of information for address, zipcode, and town. On the other hand, the census data comes with tract identification numbers. A zipcode area is often bigger than a tract and a tract may falls into two zipcode areas simultaneously. One needs to compare these two spatial units in order to assign each sales observation with a corresponding set of census information. This data matching task surely constitutes one of the most monotonous and uninspiring phases of a research process.

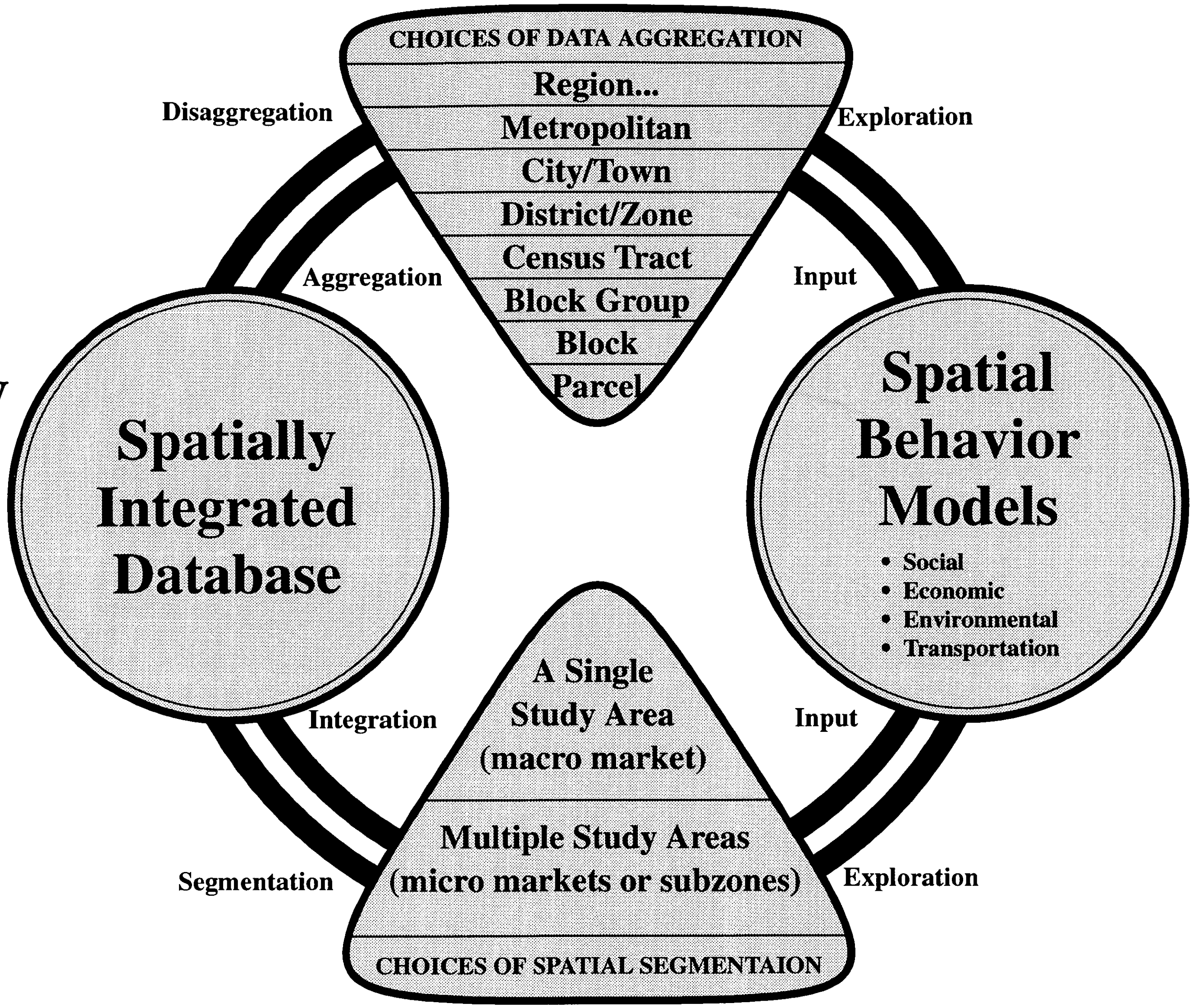
With GIS technologies, a more flexible process of spatial analysis becomes feasible. This new style of spatial modeling process is illustrated in Figure 3.1. First, using GIS, one can easily combine various sources of data, which share a common geography but may be geo-referenced differently, to form a spatially integrated database so that different layers of spatial information are interlinked. This database is represented by the left circle in the figure. We will elaborate this process of data integration in greater detail in Chapter 5 when we describe how a spatially integrated database is constructed for this research.

With the aid of a GIS tool, one is now in a position to take full advantage of this spatially integrated database. Spatial modeling no longer has to be based on a predetermined spatial unit. Spatial data can be aggregated and disaggregated at different spatial levels to serve different purposes, as depicted by the upper triangle in the figure. A pre-selected study area can also be further segmented to form multiple micro markets, or subareas integrated to obtain a macro market, as depicted by the lower triangle. One can then feed data input into spatial behavior models and explore any assortment of model runs. Based on the outcomes of model testing, one can easily modify the choices of data aggregation levels and market segmentation schemes for further exploration. The efficiency and flexibility gained by using GIS allows researchers to explore a much wider range of spatial modeling possibilities. Researchers in the field of urban and regional spatial modeling have yet to probe into the full potential of this exploratory modeling process that has been made feasible by using GIS.

In this research, we set out to explore many of the GIS potentialities discussed in the chapter. In the next chapter, we illustrate in detail how the power of GIS might be utilized to assist in analyzing urban and regional spatial environments, and specifically in testing a

Figure 3.1:

An Exploratory Process of Spatial Modeling



number of hypotheses related to the spatial dimensions of housing hedonic price models and market segmentation.

Chapter 4

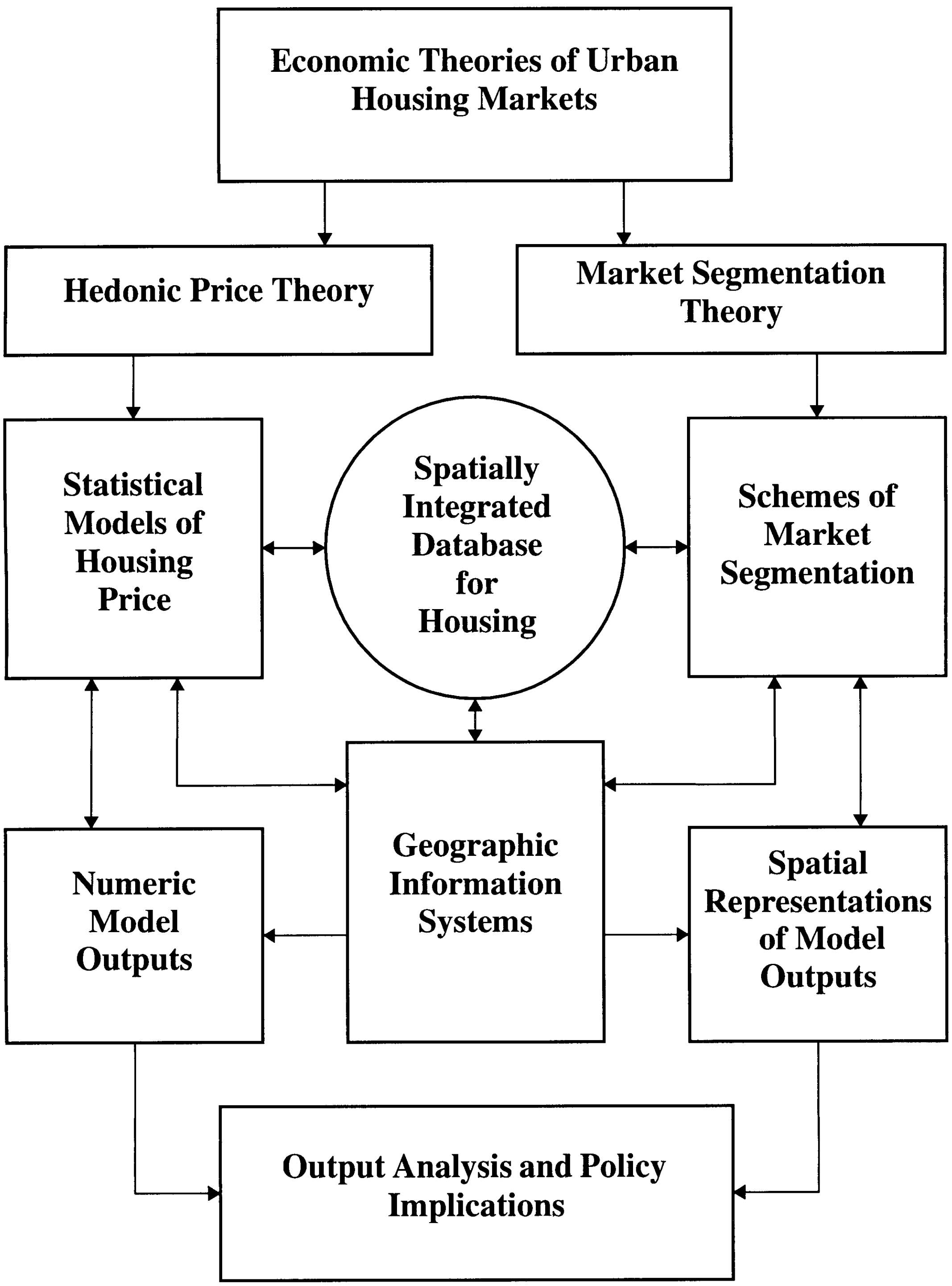
Research Hypotheses and Methodology

We stated four interrelated research objectives in the introductory chapter. The first objective represents a specific application of GIS to preparing more accurate measurements for the accessibility variable and socioeconomic attributes to be included in housing hedonic price models. The second objective relates to the use of GIS in comparing and contrasting the spatial patterns of housing submarkets segmented using several different criteria. The third and fourth objectives are exploratory in nature. The third objective examines how GIS can help us do exploratory spatial analyses of location and interrelationships of housing sales, residuals from price models, and different schemes for market segmentation. The fourth objective focuses on the technical promises and difficulties of using GIS technologies in the field of urban and regional spatial modeling and on lessons that we learn from this research.

In this chapter, in correspondence to each of the four objectives, we formulate a set of research hypotheses and questions. Then we describe in detail the strategies and methodology used to carry out this research. We emphasize how GIS technologies are used to achieve the objectives.

The uniqueness of this research is the incorporation of GIS into the economic modeling of urban housing markets. The interrelationships of housing economic theories, models, data, and the role of GIS as conceptualized in this research are illustrated schematically in Figure 4.1. GIS has been used in these aspects: organizing and generating

Figure 4.1: Interrelationships of Theories, Models, Data, and Uses of GIS



spatial data, computing and analyzing regression variables to be included in the hedonic models, preparing schemes for housing market segmentation, comparing spatial patterns of submarket locations, and representing graphically and examining spatial distributions of housing sales data and residuals from housing hedonic models.

4.1 Improve Traditional Hedonic Price Models

The first objective is to improve traditional housing hedonic price models by refining the measurements of socioeconomic and accessibility variables with the aid of GIS, and to reexamine empirically the roles of these variables in affecting changes in housing prices.

This objective can be translated into a general hypothesis: using GIS technologies to identify and measure how important housing-related spatial attributes will increase the explanatory power and reliability of a housing hedonic price model. Specifically, this can be stated as two separate testable hypotheses:

Hypothesis 1: Socioeconomic attributes measured at the block group level add more explanatory power and more consistent coefficient estimates to housing hedonic price models than those measured at the census tract level.

Hypothesis 2: The accessibility to employment center(s) measured in a more realistic manner with the aid of GIS is a stronger explanatory variable in housing hedonic price models than that approximated as a straight-line distance.

Each of these hypotheses involves a test of equality of coefficients of two regression models. Both models are single-market hedonic price models for the entire urban housing market. For each hypothesis, the two models include an identical set of independent variables. To examine the hypothesis that the coefficients in the two regressions are different, we analyze and compare various statistical tests such as t and F statistics, error residuals, and R-squares. The improvement of a model is defined in terms of the "best-fit" criterion. As in most hedonic studies, "better" is defined in terms of statistical significance of coefficients and explanatory power of the model.

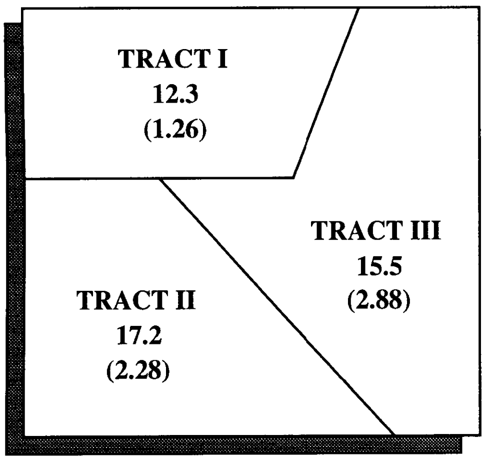
We follow a two-stage procedure of using a GIS tool to prepare and integrate the spatially related socioeconomic and accessibility variables to be included in the housing hedonic models. In the first stage, we use GIS tools to assemble a set of digital maps representing the locations of various human activities, such as employment centers, detailed streets and highways network, census tracts, block groups, and housing location information. These maps are essential for more realistic and accurate measurements of the spatial factors of interest.

In stage two, we use the GIS to measure the socioeconomic and accessibility variables. The traditional measures of accessibility ignore spatial complexities by using the straight-line distance between household residence and employment center(s) (in most cases, one center is assumed). This proxy is based on the rather unrealistic assumption that a human being can fly like a bird. If we imagine this proxy as one extreme of a realism spectrum for the accessibility measurement, closer to the other extreme sits a better approximation of the variable possible with the aid of GIS. Two sets of data are needed for this better approximation: (1) survey data with the addresses of a household head's (or multiple workers') work place and home as well as his or her mode of transportation; (2) the GIS-based map data on street and highway networks and traffic volumes for major roads. With these data, one can use the GIS network analysis module to measure the shortest path (distance or time) of journey-to-work through a city's street and highway network, and also take into account various impedances such as traffic conditions and road quality. By doing this, it is possible to drop the assumption of a single employment center. This GIS-aided measure of accessibility provides a good distance approximation. Additional information on travel times and congestion levels could improve the measurement even more. The GIS network analysis module provides necessary technical capability to carry out the computation. The question is whether the required data are readily available.

Although there are accuracy and completeness problems, the map data of street network are available from the TIGER/Line files. The other two sources of data we use for this research are the 1990 census data and the housing sales data for the city of Boston between 1989 and 1991. The census data are geographically based. The sales data contains detailed information on housing location, which allow matching each sale geographically against the street network map. The network map can then be related to maps of other spatial features. In this research, we are not able to achieve the highest degree of realism of measuring accessibility as described above. Nevertheless, based on the available data, we use GIS to identify and compute the straight-line distance between a housing unit and the primary employment center, the CBD of Boston, as well as the shortest path to work through the street network. The computation of the shortest path for each housing unit may be considered a second best approximation for the accessibility variable. It is a far more realistic measurement than the straight-line distance. More important, as discussed in Chapter 6, it represents a potential prototype which can be easily modified to model accessibility with a greater complexity when more data become available.

Also in stage two is the task of measuring the relevant socioeconomic attributes. In this research, these variables are computed at both the block group and tract levels, and then assigned to related observations of the housing sales data. To illustrate some deficiencies of using the census tract data for hedonic analysis, we use a hypothetical example as presented in Figure 4.2. In this figure, Diagram A shows the mean incomes for three census tracts and Diagram B for 15 corresponding block groups. These two diagrams reveal some large block group variations of mean income within Tracts II and III.

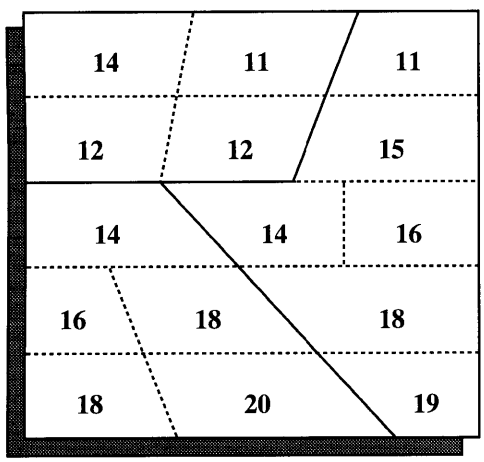
Figure 4.2: A Hypothetical Example of Housing Market Segmentation by Income *(income in thousands of dollars)*



A. Mean Income by Census Tract

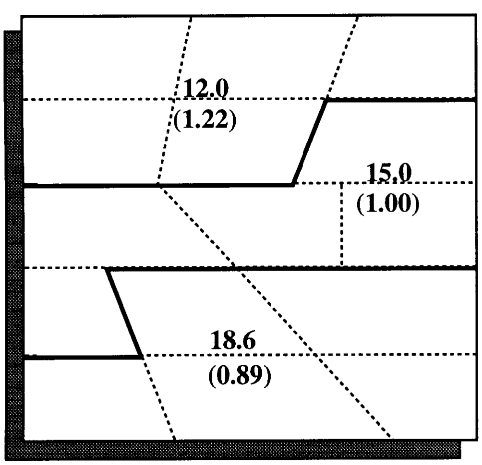
Each tract number is followed by the tract mean household income computed from the Diagram B's block group averages. The numbers in brackets are standard deviations (SD) for income. Traditionally, the three census tracts would be ranked respectively as housing submarkets for high, medium, and low income households.

Sum of SD = 6.42



B. Mean Income by Block Group

Each number represents the mean household income for a block group. This diagram reveals a wide variety of income levels within each census tract.



C. Mean Income by Regrouped Block Group

The fifteen block groups are regrouped into three housing submarkets, based on the examination of Diagram B. The new grouping cuts the previous sum of SD in half. This diagram shows a possible better spatial segmentation of housing markets by income than that shown in Diagram A.

Sum of SD = 3.11

Therefore, in the case of these two tracts, it should improve model accuracy to estimate incomes at the block group level for the purpose of hedonic price studies. Alternatively, given the knowledge of the mean income for each of the block groups, we could re-categorize these groups into three housing submarkets, as depicted in Diagram C, which seem to be characterized by a higher degree of homogeneity in terms of household income. This simple example shows that a great deal of richness of income information can be added to housing value modeling by using block group census information. Also, it shows the potential gain in understanding spatial characteristics and relationships of market areas by using GIS to represent information graphically.

Upon completion of the above two stages of data preparation, reconstruction, and integration, we then explore further the utility of GIS by carrying out a number of tests: First, we model the housing and demographic diversities at different spatial aggregation levels by combining GIS with two simple traditional techniques of analyzing the spatial distribution of socioeconomic factors. These two techniques are: Gini coefficients of spatial distribution and location quotients. Second, we construct housing hedonic models, test the two hypotheses, and analyze regression residuals. Two versions of housing hedonic models are constructed, macro and micro, respectively. The macro models analyze the variations of median housing values in the greater Boston area at two different data aggregation levels, tract versus block group. The micro models, following the traditional housing hedonic price theory, use housing sales prices as a dependent variable. In addition to the accessibility and socioeconomic variables mentioned above, other explanatory variables include individual housing structural attributes such as number of bathrooms and size of living area. The model construction and hypothesis testing follow the standard approaches employed by most housing economic researchers.

To analyze the error residuals from the housing price models, we use GIS to portray the error graphically. This allows examination of the spatial distribution patterns of errors and analysis of how these errors are related to other spatial features. We also discuss several other useful applications of GIS-based maps of residuals.

4.2 Improve Market Segmentation Analysis

The second objective is to explore the use of GIS to assist market segmentation for the purpose of constructing separate hedonic price models for housing submarkets. We examine the stability and consistency in the impacts of housing attributes on prices. Also, we experiment with the exploratory style of spatial modeling that allows ease of testing various areal segmentations and data aggregation at different spatial levels.

The first step of this exploration is to represent the results of market segmentation spatially using GIS map capabilities. We segment the Boston housing market using four commonly used stratifiers: median housing value, median household income, education attainment level of adults, and neighborhood racial composition. We then compare and contrast spatial patterns of segmentation results and other resulting spatial characteristics, including the size of market area for each submarket, and housing and population densities. We also discuss the flexibility of using GIS to explore appropriate segmentation schemes that are coherent with geographic and socioeconomic characteristics of local housing markets.

Next, we investigate how GIS can be used to improve the single-market housing price models by experimenting with three selected approaches of market segmentation: by accessibility, housing type, and race. We construct hedonic models for each of the housing submarkets identified. Then, we compare each submarket model with the single-market model and examine the stability and consistency in hedonic estimates. The comparison should offer some insight into whether there exist distinct housing submarkets in different locations, for different family housing types, and for different ethnic groups.

Specifically, for the study of hedonic estimates and market segmentation by accessibility, based on examining the spatial distribution of residuals from the single market hedonic models, we experiment with two ways of delineating the Boston CBD market area. We investigate how different delineations affect the hedonic estimates and evaluate the flexibility of using GIS to assist in representing spatial features, interpreting model results, and manipulating spatial data.

For both the studies of hedonic estimates and market segmentation by housing type and race, we focus on the use of GIS in helping to understand the spatial patterns of segmentation results and to interpret model results. Our approach is to portray the spatial distributions of housing sales data for each identified submarket using GIS and to estimate separate hedonic price models for all submarkets. Also, we compare each submarket hedonic model with the single-market model.

4.3 Explore the Usefulness of GIS in Analyzing Spatial Patterns

The third objective involves the exploration of using GIS spatial representation capability to examine whether the results of hedonic models and market segmentation make good spatial sense. This exploration is carried out along with the studies on improving hedonic model and segmentations described in the previous two sections. We are interested in the use of GIS-based spatial representation for locations of housing sales cases, residuals from the models, and housing submarket configurations to assist in identifying important spatial factors that might be important to model formulation and interpreting model outcomes.

This objective represents research that is exploratory in nature. We explore the capabilities of GIS in dealing with spatial issues related to housing markets. By using GIS to generate, organize, manipulate, and represent graphically various spatial information related to housing markets, as illustrated skeletally in Figure 4.3, it is possible to examine the information both visually and numerically, using some simple spatial statistics techniques. This GIS-based overlaying and buffering study could potentially shed some light on the following questions. Do landscape features such as parks, water surfaces, airports, and highways affect the results of hedonic price models and housing market segmentation? Are there any spatial regularities or irregularities in terms of housing price variations, population and building densities, and regression error residuals? Are there other research questions about housing markets in particular, and urban spatial economics in general, that might benefit from the use of GIS?

In the research to date, we are not able to empirically investigate the potential of GIS in addressing all these related questions in a comprehensive manner. In the following discussion, we use a number of hypothetical examples to elaborate the applicability of GIS in urban and regional spatial analysis. We demonstrate that such exploration could be potentially fruitful. The GIS-based map information allows us to study how a house is related spatially to environmental disamenities and how the existence of some unique terrains, parks, and human-made structures might alter the impact of the disamenities on the values of housing. For example, suppose we have two houses, A1 and A2, with identical structural features, as shown in Figure 4.4. In terms of straight line distance, A1 is closer to the undesirable waste site No.1 than A2. If this is all the information we know, using the traditional hedonic model we would expect that A1 should be worth less than A2, assuming other things equal. However, since A1 is both visually and spatially separated from the waste site by a hill covered with dense vegetation, and A2 is not, then in reality it is very likely that A1 would be valued higher than A2 (the conventional wisdom has "out of sight out of mind"). This example shows that it is important to know

Figure 4.3: An Illustration of Spatial Analysis of Urban Housing Market Using GIS

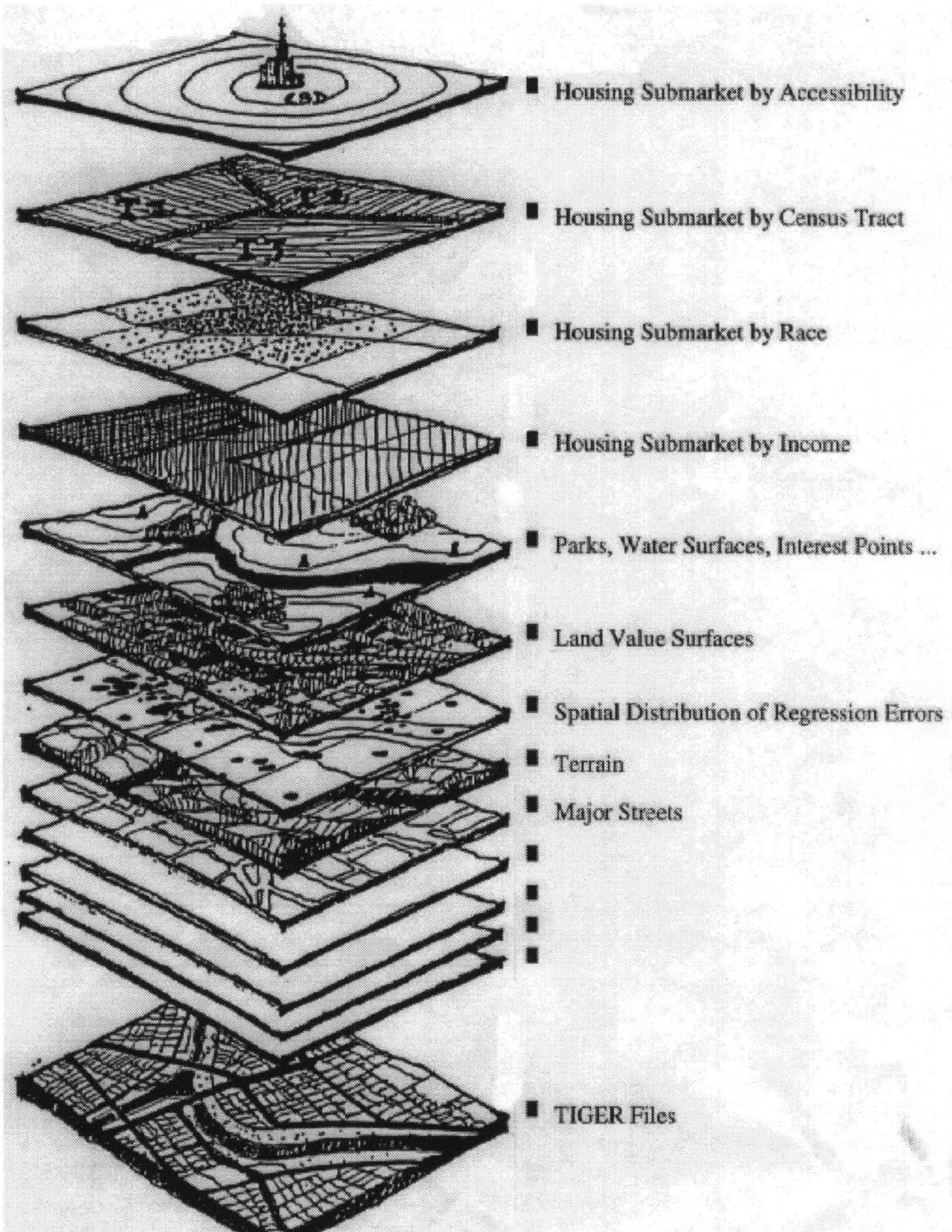
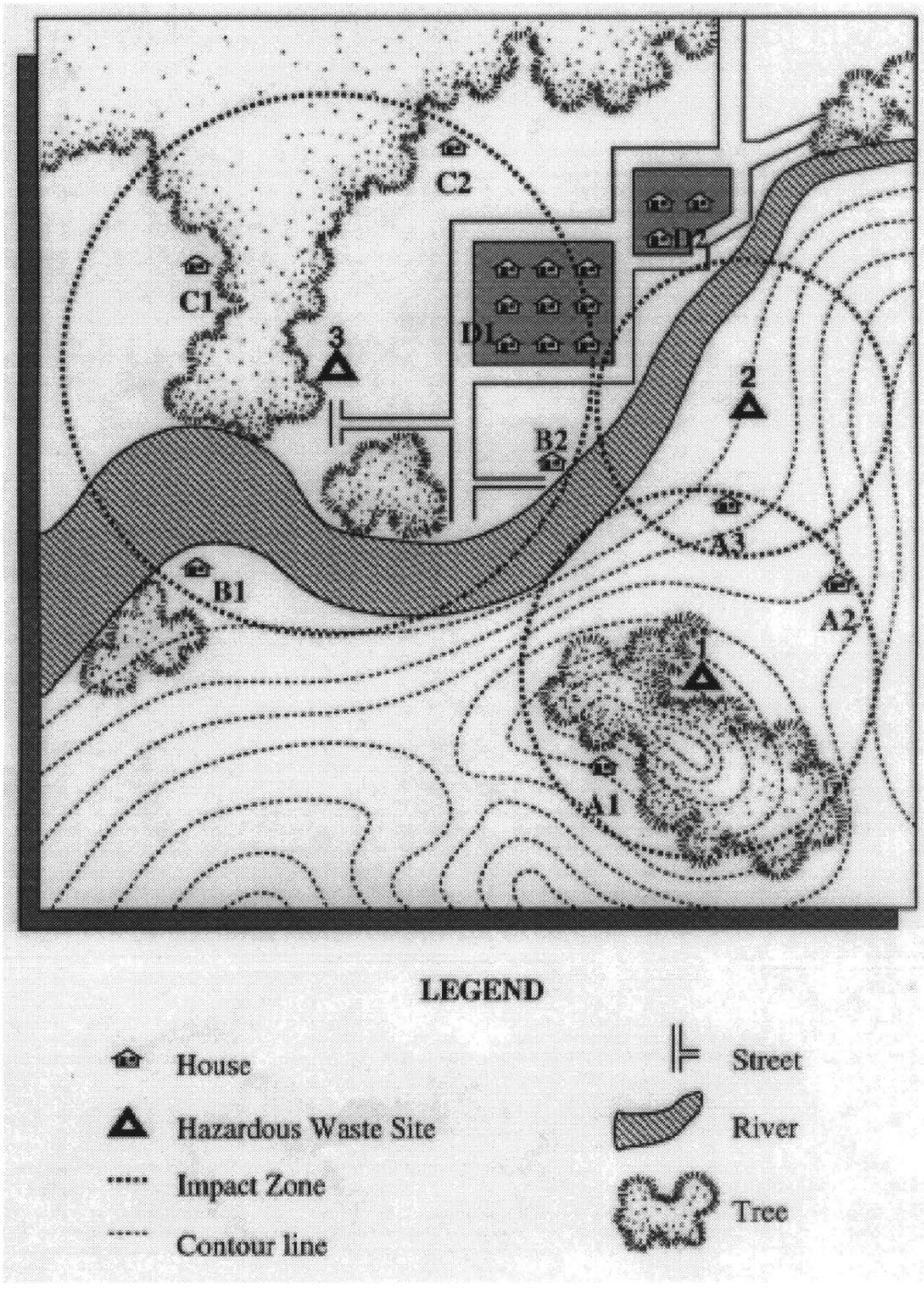


Figure 4.4: An Illustration of the Effects of Spatial Factors on the Values of Housing Properties



where those desirable or undesirable activities and entities are situated in space with respect to a housing property of interest. Given today's GIS technologies, it is possible to construct a 3-D model of housing market to explore the issues such as proximity, visibility, and accessibility in greater realism.

More examples are shown in Figure 4.4 to illustrate the importance of understanding spatial characteristics of housing properties. The following comparisons between each pair of residential properties are based on the condition of *ceteris paribus*. A3 is likely to be priced lower than A2 because A3 is located in an area that is adversely affected by both waste sites No.1 and No.2. B1 and B2 are of equal distance to site No.3. Yet, B1 may be of a higher value because a river separates B1 from No.3. The impact of No.3 might not be greater on C1 than C2 because the significance of the proximity factor might be undermined by the presence of the dense cluster of trees. The final example from Figure 4.4 assumes that the groups D1 and D2 housing units happen to be located in an identical census tract. If the housing submarkets are defined according to census tract, we may apply wrongly the effect of No.3 to both D1 and D2. Yet, in fact, the D2 group is not affected by No.3, as shown. These fictional examples demonstrate that information about spatial characteristics of housing properties is important in the study of housing price models and market segmentations.

The knowledge that we gain through this exploration process, either intuitively or visually, should enable us to modify housing price models accordingly and provide helpful insights into other studies of urban and regional spatial structures. Obviously, analysis at this level requires detailed spatial data. It is clear that this type of analysis would not be possible to complete in the time frame of this research.

4.4 Explore Utility of GIS for Urban and Regional Spatial Analysis

The fourth objective is to explore the promises and difficulties of using GIS technologies to conduct analytical research in urban and regional spatial analysis in general and the utility of GIS for the analysis of urban housing markets in particular. This objective is again exploratory and represents broad issues related to the products generated out of this research and the problems and difficulties encountered during the research process. Along with the presentations of our research findings in Chapters 6 and 7, and in the concluding chapter, we comment and evaluate the technical complexity and practical feasibility of applying GIS technologies in the construction of hedonic models, market segmentation, and spatial analysis of housing markets. We also reflect on the outcomes and process of this research and sum up valuable lessons learned. Based on our experience and research outcomes, we point to several possible future research questions that might benefit from the use of GIS.

Chapter 5

Study Areas and A Spatially Integrated Database for Housing

In this chapter, we first define two study areas selected for the research. Then, we describe three sources of data and discuss how we combine these data to construct a spatially integrated database for our housing market analysis.

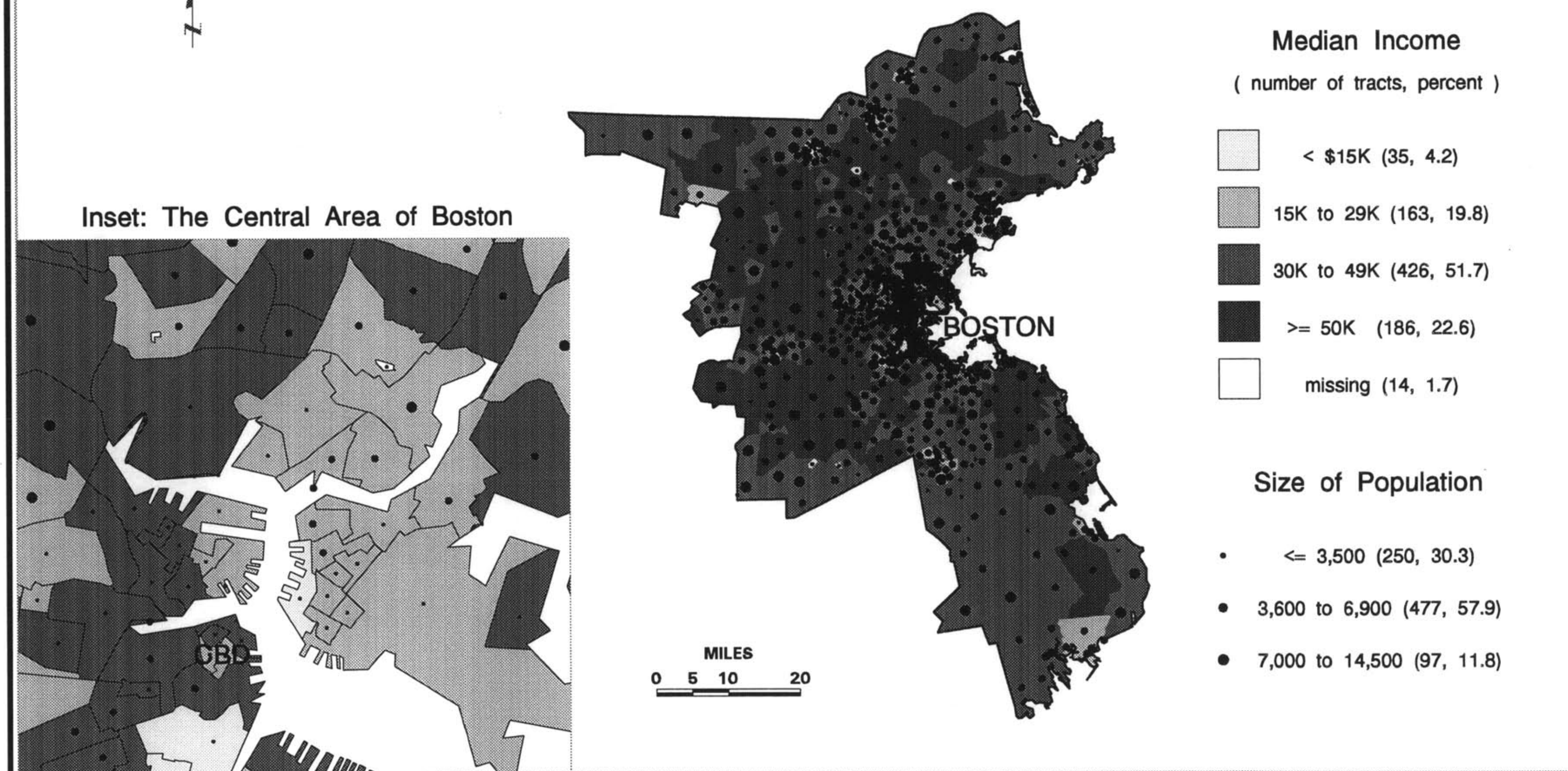
5.1 Study Areas

We select the greater Boston area as a macro housing market and the city of Boston as a micro market for this research. Two factors lead to this choice. First, various GIS-based maps of the areas, including geographical boundaries of block groups, tracts, towns, and counties, and street and public transportation networks, have been prepared through previous research projects. Using these existing digital data saves a significant amount of time and up-front costs. Second, among the abundant literature on housing price models, the Boston residential market is one of the most frequently analyzed and modeled by researchers. Existing research on the Boston markets provides important background information for this research, our new approach offers potential comparisons as well as fresh insights. Moreover, this selection of study areas at two different scales is important as it allows us to keep our exploration of the utility of GIS technologies to a manageable scope.

As defined in this research, the greater Boston area, consisting of the five counties of Suffolk, Essex, Middlesex, Norfolk, and Plymouth, is shown in Figure 5.1. The area is

Figure 5.1 Macro Housing Market: The Greater Boston Area
Population Distribution and Median Income by Census Tract, 1990

(The median income by tract ranges from $5,000 to $103,000. The average population size by tract is 4,571.)



made up of 148 towns, 827 census tracts, and 3429 block groups. In 1990, the population was 3.8 million, with 1.4 million households, and 1.5 million housing units.

Figure 5.1 displays a map of the study area along with some additional income and population information by census tract, serving to acquaint those unfamiliar with the greater Boston area. The shading shows that 23 percent of tracts have $50,000 or higher median household income and 52 percent have between $30,000 and $49,000 income. Most of these wealthy tracts are located in the suburban areas surrounding the city of Boston. The remaining 25 percent of tracts have median income less than $30,000. Most poor tracts can be found in Suffolk County[7] (Boston), as shown in the map inset. A number of other poor tracts with densely clustered population are observable at locations about 20 to 30 miles away from Boston. In this dissertation, our graphic presentations of data and research results are constrained by the size of the page and not being able to use color. However, our exploration has shown that using GIS to display more detail and greater richness of information at the block group level aggregation, in a graphically more attractive manner, is entirely possible.

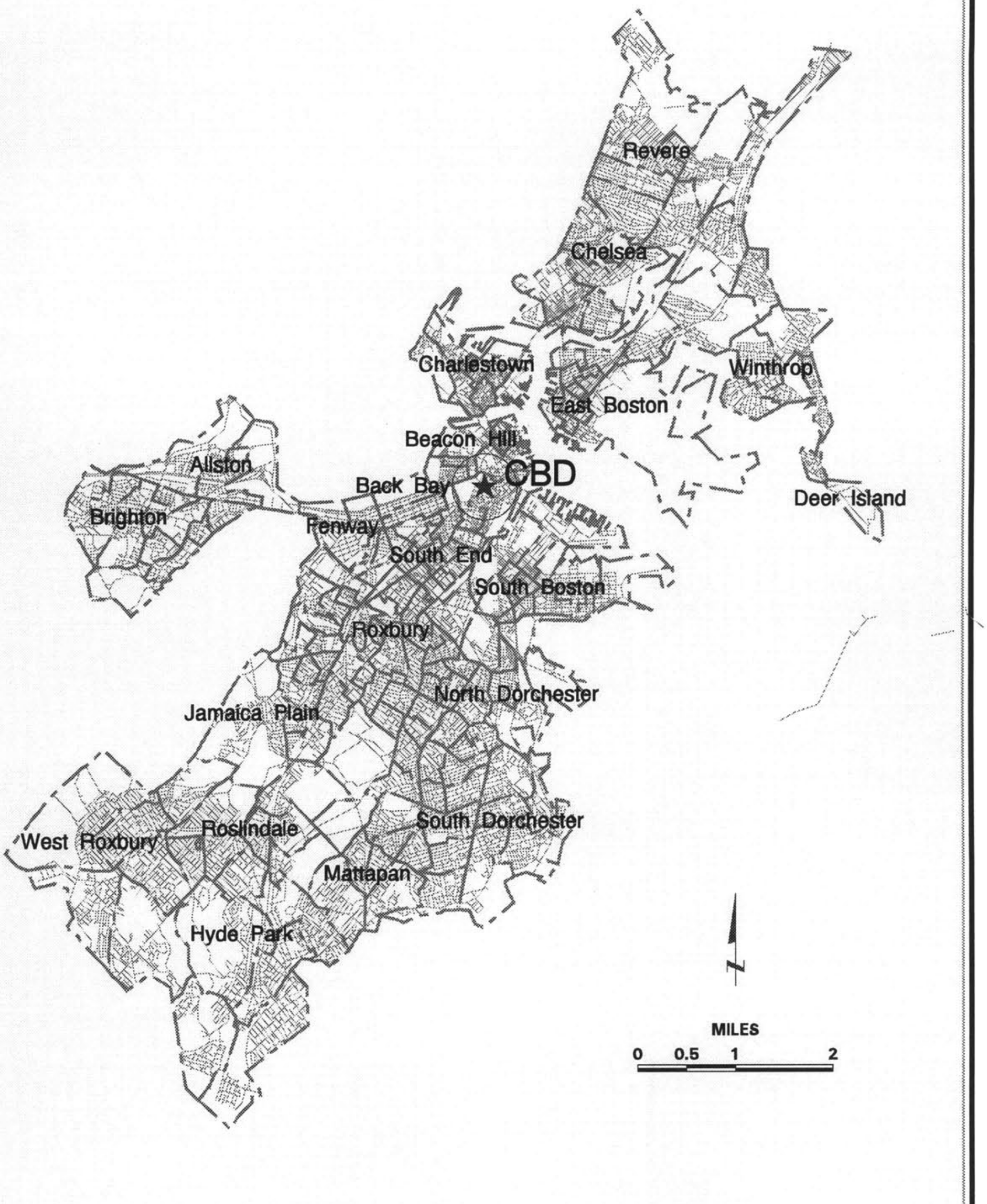
The micro housing market selected for this research is the city of Boston, as shown in Figure 5.2. This area consists of 189 tracts and 689 block groups. According to the 1990 census, the area was inhabited by 664 thousand people, 263 thousand households, and 289 thousand housing units. The median household income by tract ranged from $5,000 to $53,300, and by block group from $5,000 to $150,000. The median housing value by tract fell in the range between $87,500 and $500,000, and by block group between $15,000 and $500,000. In addition to displaying the delineation of tract for the study area, Figure 5.2 shows the complete street network system of the city. To the east of the city lies the

---

7. Suffolk County and the city of Boston are used interchangeably in this dissertation.

Figure 5.2 Micro Housing Market: The City of Boston Neighborhood, Census Tract, and Street Network System

(The city consists of 189 tracts, 689 block groups, and 20,513 street segments.)



Atlantic Ocean. Surrounding Boston in the other three directions are densely populated towns and cities. The total land area for the city is slightly over 390 square miles.

5.2 Sources of Data

A large amount of data are used for this research. We assemble the data from the following sources:

(1) GIS-based map data

Digital maps for counties, towns, tracts, block groups, and street and highway networks of the selected study areas were extracted from the 1992 TIGER/Line files[8]. The street and highway network for the city of Boston, constructed from TIGER using ARC/INFO, consists of 14,657 nodes and 20,513 arcs (lines or street segments).

(2) Census data

Two sets of census data pertaining to neighborhood demographic and socioeconomic characteristics, aggregated at both the tract and block group levels, were extracted from Summary File Tape 3 of US Census 1990 and geo-coded to link to corresponding GIS-based maps.

(3) Housing sales data

We obtained from the Boston Redevelopment Authority a copy of a data set for records of every transaction of land and building properties that occurring between January 1989 and September 1991. During this period, there were 8,073 transactions in the city of Boston, of

---

8. The TIGER/Line files contain information that describes the points, lines, and areas on Census Bureau maps. The files are extracts of selected geographic and cartographic data from the Census Bureau's Census TIGER (Topologically Integrated Geographic Encoding and Referencing) System, which is used to support mapping and other geographic activities of its decennial census and sample survey programs.The TIGER/Line files provide information on streets, rivers, railroads, and other line features, where they intersect and the areas they enclose, in a form that can be processed by a computer. TIGER/Line records contain latitude/longitude coordinates, codes identifying census geographic areas, and address ranges and ZIP Codes.

which 4,260 were residential building and land sales. We extracted from the data set only residential building sales, including four types of housing: single-family, two-family, three-family, and condominium housing. The total number of residential building sales was 1,740, among which 568 were for single-family, 333 for two-family, 383 for three-family, and 456 for condominiums. The detail on data structures, field information, and data cleaning process can be found in Appendix A.1. After cleaning the data, we are left with 1,571 records of sales that are available for this research. This cleaning process should not result in any serious sampling errors because the majority of the deleted sale cases are duplicate sales records for co-buyers of a housing unit. The variables that we use for this research include sale prices, structural characteristics, and addresses (location). The address fields are critical for location matching using the TIGER street maps.

5.3 A Spatially Integrated Database for Housing Market Analysis

Before we are able to move to more interesting and rewarding parts of this research, the first important task is to integrate the data from different sources. A common element among the three sources of data is the geography from which they were collected. Traditionally, when the housing sales data and census data were brought together for the study of housing price models, a researcher first had to match each record of housing sale with its corresponding census tract using paper maps. This matching process can be extremely time-consuming, tedious, and error-prone. If one attempts to match housing sales data with census data aggregated at the block group level, the time needed could easily double or triple. With the aid of GIS, the matching process has become much easier and faster, and error-free provided that correct geographic information is used. Moreover, it makes little difference whether one is matching with census tract or block group data.

More important, GIS technologies can bring out the true value of the information of geographic location contained in each of the three data sets. In the past, the location information had little use except for serving as a common key for data merging. With GIS, that information easily becomes one of the most valuable elements in spatial analysis. GIS creates a common spatial environment for linking data collected from different sources, displaying these data in spatial context, and generating additional spatial information.

In this research, we use GIS to integrate the three sources of data, to portray the geographic data graphically, and to generate new spatial information for constructing housing price models. The process of this integration is depicted in Figure 5.3. The four TV-screen-shaped boxes indicate the initial data sets. Census data are linked to block group and tract geo-coverages using the common location identification codes. Housing sales data are matched with street network coverages based on detailed address information. With the street network coverages and the block group/tract coverages sharing a common geography, we overlay these two sets of map data to create linkages between the housing sales data and the census data. These integrated data are then used to generate new spatial information such as accessibility measurements and market segmentation. The vertical gray bands represents sequential steps, from left to right, of how various data are linked to construct a GIS-based spatially integrated database for housing market analysis.

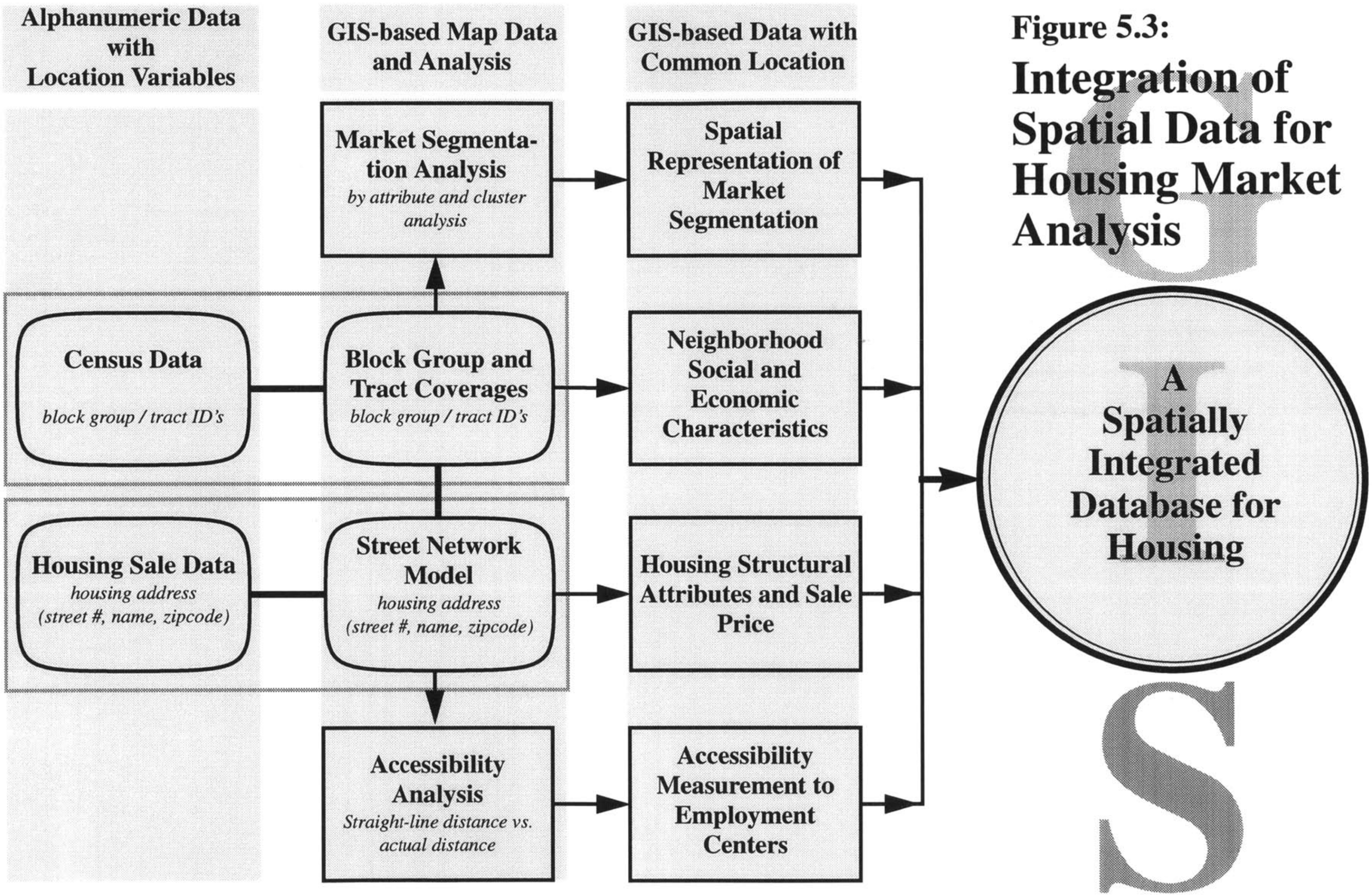


Figure 5.3:

Integration of Spatial Data for Housing Market Analysis

Chapter 6

Spatial Analysis of Housing and Demographic Diversities and Housing Hedonic Price Models

This chapter consists of two major sections. In the first section, we focus on the macro housing market, the five-county greater Boston area. In combination with GIS technologies, we investigate housing and demographic diversities in the region using three commonly used traditional techniques of spatial analysis: Gini index of spatial distribution of poor households, location quotients of low income households and poorly educated adults, and simple regression models of median housing values. In the second section, using the spatially integrated housing database as described in last chapter, we construct housing hedonic price models for the micro housing market, the city of Boston. We emphasize the effects of the improved measurements of socioeconomic and accessibility variables on the estimation of housing hedonic price models and on the uses of GIS-based maps of error residuals for interpreting and understanding the inputs and outputs, and the behavior of hedonic models.

6.1 Modeling Housing and Demographic Diversities

6.1.1 Introduction

The study in this section serves two purposes. First, we investigate housing and demographic diversities in the greater Boston area using 1990 Census data aggregated at two geographic levels: tracts versus block groups. By applying a number of widely used conventional socioeconomic analytical techniques to census data aggregated at the two different levels, we reveal that important diversities are associated with the differently aggregated data and research results can vary significantly. Census data are now available for both tract and block group levels in more easily accessible electronic formats. In the past, most researchers have relied on data aggregated by census tract not because a tract is theorized as an appropriate level to characterize areas differentiated by socioeconomic attributes and demographic compositions, but because more disaggregated data have been difficult to access. Given that 1990 census data for tracts and block groups are equally accessible using today's computing technology, we are also interested in whether census data at the block group level are better choices over tract level data.

The second purpose is to explore the utility of GIS in the analysis of housing and demographic characteristics of a neighborhood, either approximated as a census tract or a block group. We use GIS to portray graphically and spatially the analytical outputs generated by the selected traditional approaches of social and economic research. Also, we employ GIS to generate inputs to the analytical processes.

A census tract has been the most frequently used areal aggregation to define a neighborhood or a housing submarket. The universal use of census tract data aggregation by no means implies that a tract is a good representation of a neighborhood. Rather, its popularity is mostly due to data limitations. It is well known that there exists great diversity of housing and demographic characteristics at the tract level. These diversities are lost because of aggregation. For example, according to the 1990 census, Tract 0705 in Boston is characterized by an average of 27 percent households with annual income less than $15,000 and 51 percent nonwhite population. This tract consists of 4 block groups which actually display substantial variations in income and racial characteristics. For the block groups, the less than $15,000 percentage ranges from 8 to 63, and the nonwhite population from 23 to 86. Such information loss due to aggregation could change one's research results dramatically. As census data aggregated by block group, typically 20 to 30 percent of the size of a tract, have become readily available, it is desirable to look into the potential of using block group data to represent more homogeneous neighborhoods. It is usually difficult to obtain a lower level of aggregation such as a single block or a parcel group because of the legal confidentiality requirement for compiling census data.

6.1.2 Combining Traditional Techniques of Spatial Analysis with GIS

We select three approaches to test the effect of data aggregation levels and portray some of these test results using GIS. These three approaches are Gini coefficients, location quotients, and housing value regression models. All three have been commonly employed by researchers in urban and regional studies. We will describe each of these methods briefly below. But first we need to point out one common weakness that seems to be associated with these methods. Although they all have been frequently applied to analyze spatial and location matters, their ways of representing the distribution of spatial and location attributes are either tabulated alphanumeric information or abstract mathematical formulae. These conventional approaches lack an intuitive and easily comprehendible way of portraying location related data and research results. Thus, we have incorporated GIS into experimenting with possible ways of correcting this weakness.

We explain the three approaches in the context of analyzing housing and demographic variables. For those unfamiliar with these approaches, a fuller account can be found in the references cited.

The first test of aggregation levels is to compute a series of Gini coefficients for the distribution of poor (or rich) households. The Gini coefficient is commonly used to measure the inequality of wealth distribution (Gini 1913; Duncans 1955). The Gini coefficient lies between 0 and 1, with 0 representing complete equality and 1 the most unequal distribution. The calculation and meaning of a Gini coefficient can be easily demonstrated with the Lorenz curve of income distribution which gives the cumulative percentage of the total income accruing to the lowest percentiles of the population (to be graphically illustrated below). Instead of measuring income distribution, we construct Gini coefficients to measure how unevenly tracts or block groups share the proportions of total poor households in the greater Boston area.

The second test is to calculate location quotients of poor households and educational attainment levels by tract and block group respectively, and plot the results using GIS. The location quotient (LQ) is a device frequently used to identify specialization, concentration, or potential of an area in selected employment, industry, or output (Bendavid-Val 1983). Mathematically, the LQ is simply a ratio of ratios, with the top ratio equaling the fractional share of the subject of interest at the local level and the bottom at the regional level. Thus, the LQ can be no less than zero. When the LQ is greater than 1, it means that the local area is more specialized in the subject of interest. The opposite is true when the LQ is less than 1. We use the LQ to identify tracts or block groups that contain higher percentage shares of poor households or less educated population as compared to the greater Boston area as a whole. We then plot and compare the results using GIS mapping capability.

Another test is to construct housing value regression models for the greater Boston area. The regression models allow for testing the effects of several socioeconomic variables simultaneously. Following the spirit of hedonic price theory, we prepare a list of attribute variables based on housing and demographic data from the 1990 Census to model median housing values at the tract and block group levels, respectively. We then compare the tract and block group models statistically. Our emphasis is on demonstrating the benefits of using GIS to assist regression analysis. To analyze how well our models explain the variations in median housing value for a tract or block group, we plot the regression residuals associated with each observation's location. By doing so, we are able to observe the performance of our models by location and detect any spatial patterns. The approach of mapping residuals was first applied by McCarty (1956) and his colleagues to a study of industrial locations at a national level.

6.1.3 Testing Results

(1) Gini coefficients of distribution for poor households

To illustrate how a Gini coefficient of distribution of poor households is calculated, we define poor households as those households with annual income less than $15,000 and calculate the Gini coefficient using tract data. First, we compute the number of the poor households for each tract as a percentage of the total poor households in the entire region and the total households in each tract as a percentage of the total number of households for the region. Second, we rank all tracts in ascending order according to their percentage shares of poor households. Then, as depicted in Figure 6.1, beginning with the tract having the least share of the poor, we plot each tract's cumulative percentage share of poor households along the vertical axis and its cumulative percentage share of the total households along the horizontal axis. By doing so, we obtain a curve, denoted by "Tract Data," that is convex away from the diagonal line. The curve is widely known as a Lorenz curve. The closer to the diagonal that a Lorenz curve is, the more equalitarian is the distribution. The diagonal line represents complete equality. Thus, the "Block Group Data" curve, the Lorenz curve derived from the block group data, represents greater inequality than the tract curve. The Gini coefficient equals twice the area between the curve and the diagonal. The coefficient lies between 0 and 1. In the case of distribution of poor households, a zero Gini coefficient would mean that every subarea (*i.e.*, tract or block group) shares an equal portion of the total poor households in the greater Boston area. A higher Gini coefficient means that poor households are more unevenly distributed across the region. The Gini coefficients for the block group and tract curves are respectively

Figure 6.1: Lorenz Curve for Poor Households in the Greater Boston Area

(Annual Household Income < $15K, Census 1990)

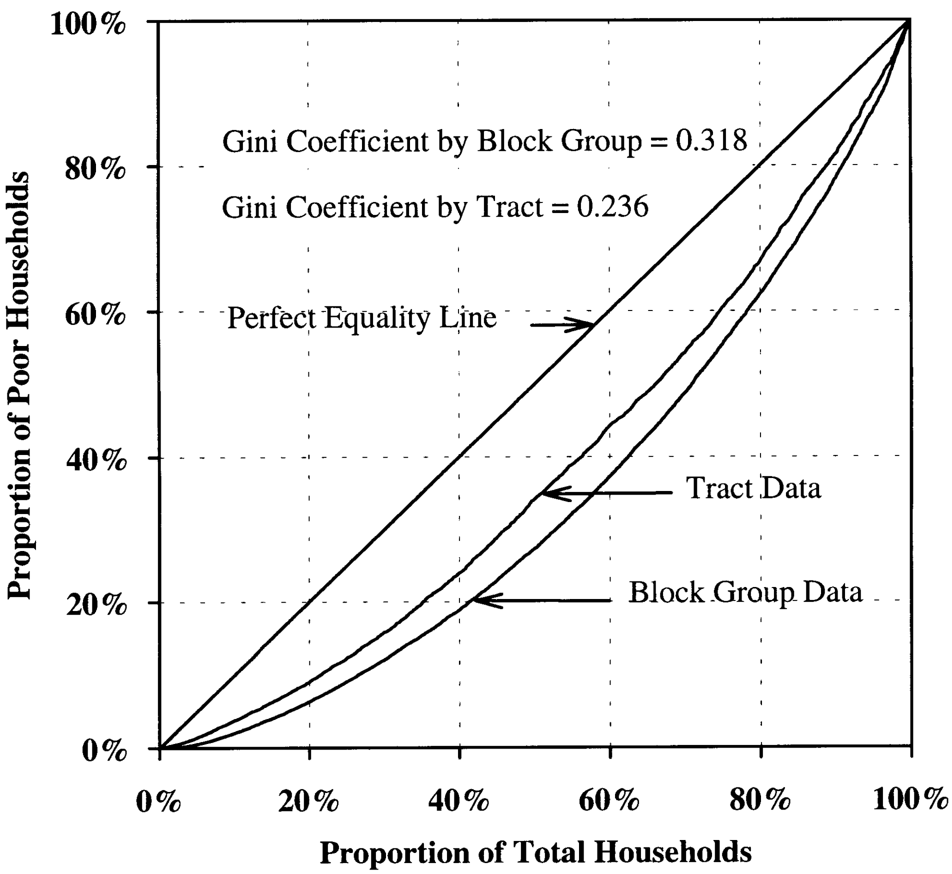


Table 6.1: Comparison of Gini Coefficients of Distribution of Poor Households between the Tract and Block Group Data Aggregations

Income Category	Number of households	As % of total households	Cumulative % of households	Gini coefficient for the poor		
				By Tract	By Block Group	% of Diffe-rence
< $5,000	58,143	4.1	4.1	0.352	0.477	35.5
5,000 to 9,999	118,294	8.4	12.5	0.289	0.390	34.9
10,000 to 12,499	44,696	3.2	15.7	0.260	0.349	34.2
12,500 to 14,999	36,583	2.6	18.3	0.236	0.318	34.7
15,000 to 17,499	43,911	3.1	21.4	0.214	0.289	35.0
17,500 to 19,999	40,398	2.9	24.3	0.198	0.265	33.8
20,000 to 22,499	49,075	3.5	27.7	0.181	0.240	32.6
22,500 to 24,999	39,192	2.8	30.5	0.167	0.223	33.5
25,000 to 27,499	50,112	3.6	34.1	0.152	0.202	32.9
27,500 to 29,999	40,167	2.8	36.9	0.141	0.187	32.6
30,000 to 32,499	54,242	3.8	40.8	0.127	0.169	33.1
32,500 to 34,999	39,440	2.8	43.6	0.118	0.156	32.2
>= $35,000	795,985	56.4	100.0			
Total	1,410,238	100.0				

Source: calculated based on US Census 1990.

0.318 and 0.236.

The 1990 Census reports 25 household income categories, among which 12 are below $35,000, as indicated in Table 6.1 and 13 are greater than or equal to $35,000. We use each of the 12 categories in turn as our bench mark to define poor households. Then we calculate the Gini coefficients of distribution for each of these differently defined poor households by tract and block group, respectively. The results, as presented in Table 6.1, show that the Gini coefficients computed using the block group data are consistently higher than those using the tract data. The differences are in the neighborhood of 35 percent. These results should make one cautious about selecting appropriate data aggregation levels.

(2) Location quotients of poor households and less educated population

In this test of data aggregation levels, we compute the LQs to examine each local area's relative share of poor households and less educated population as compared to those of the whole region. We define poor households as those with annual income less than $15,000 and less educated population as those with less than high school degrees. To illustrate, one formula of computing an LQ for poor households for a tract is:

$$LQ = \frac{\frac{TP_i}{TT_i}}{\frac{RP}{RT}} \tag{6.1}$$

where

TP_i = total number of poor households in Tract i,

TT_i = total number of households in Tract i,

RP = total number of poor households in the greater Boston area,

RT = total number of households in the greater Boston area.

The location quotients calculated using the tract and block group data separately are shown in Table 6.2. For the convenience of comparison, we divide the two distributions of LQ values at <1.0 and >=1.0, with 1.0 being the value where the percentage of the poor or less educated population is the same as that of the region. Two differences between the tract and block group calculations are observable from this table. First, for both LQs, the proportion of census tracts that share a higher percentage of the poor and less educated than the regional counterparts is about 4 percent higher than that of block groups. Second, the poor households are concentrated in fewer tracts or block groups than the less

Table 6.2: Location Quotients of Poor Household and Less Educated Adult by Tract and Block Group

	LQ of poor households				LQ of less educated adults			
Location quotient	Number of Tracts	As % of the total	Number of block group	As % of the total	Number of Tracts	As % of the total	Number of block group	As % of the total
zero	4	.5	141	4.2	11	1.3	145	4.2
< 1.0	453	55.9	1,939	57.4	437	53.0	1,810	52.8
>= 1.0	353	43.6	1,298	38.4	377	45.7	1,471	42.9
Total	810	100.0	3,378	100.0	825	100.0	3,426	100.0
Missing	14		48		2			

Source: calculated based on US Census 1990.

Note: poor households are defined as those with annual income less than $15,000; less educated adults are those 25-year old or above with education attainment less than high school graduates.

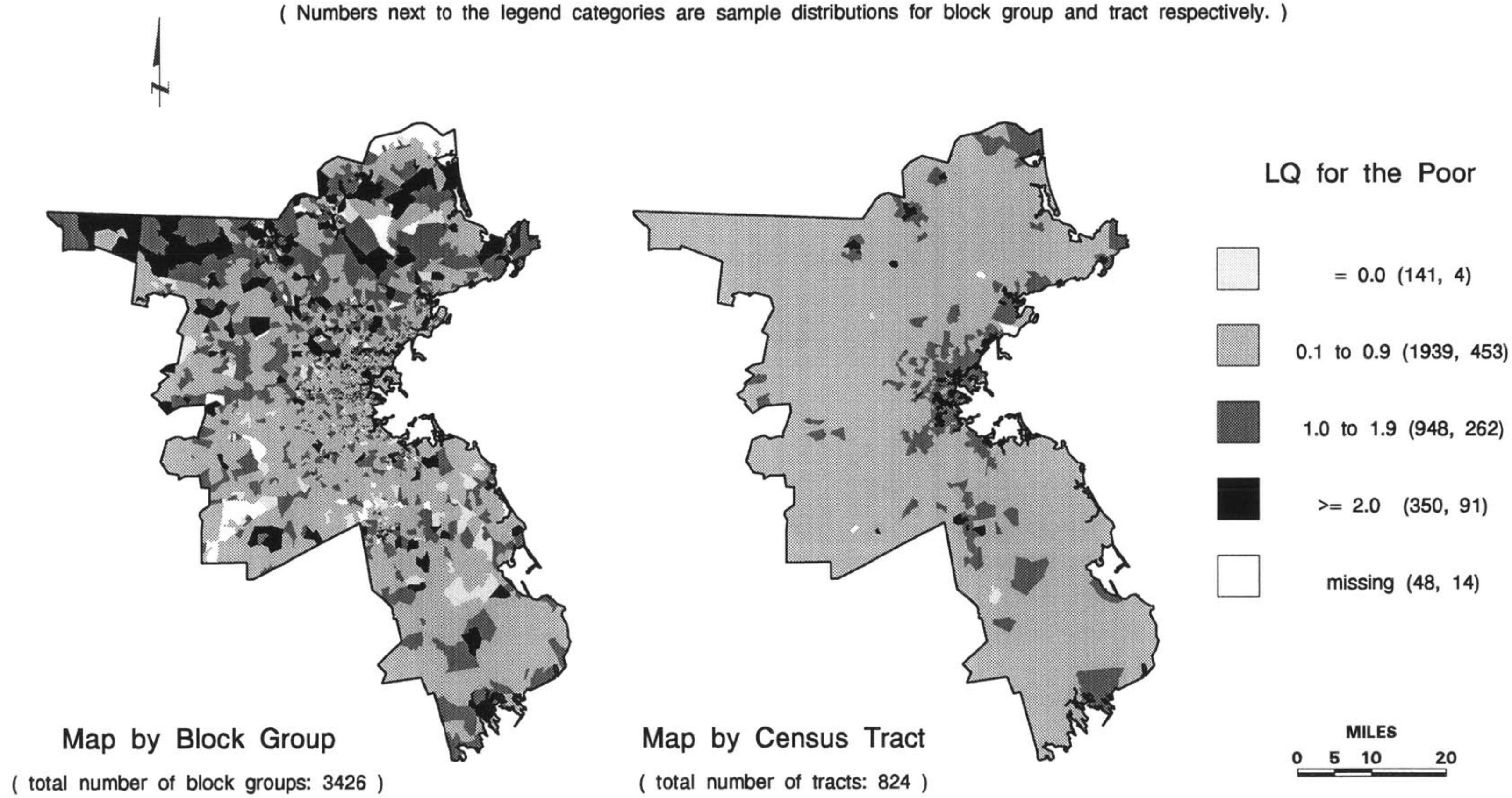
educated adults. These comparisons again indicate that there are important differences between calculations based on different aggregation levels. However, the descriptive statistics in the table tell us little about where these poor and less educated areas are and how they relate to each other spatially. Using a GIS tool, we are able to provide additional important evidences for concerns about the choice of data aggregation levels.

Figure 6.2 shows the spatial distribution of location quotients for the poor households. It is not hard to see that the location quotients plotted by tract can be very misleading. For example, the tract map shows that few areas have LQ equal to zero, indicating those localities with no poor households. In contrast, the block group map shows that there are hundreds of this type of block group scattering across the region. This difference can be attributed to the way that the LQ is calculated. It is much more difficult for a tract to have a zero LQ because each of the block groups within this tract needs to have a zero LQ in the first place. It might not be very difficult to note that in Table 6.2, it is unlikely that all those

Figure 6.2 Location Quotients for Poor Households: Census Track vs. Block Group

(The Greater Boston Area, Census 1990)

(Numbers next to the legend categories are sample distributions for block group and tract respectively.)



141 block groups with zero LQ of poor households would all fall within the 3 tracts of the same characteristics. But it would be very difficult to gain a good understanding how these block groups differ from those tracts in the real space without representing these locations graphically using GIS. The comparison of the two maps indicates that the differences of LQ for poor households between displaying the tract and block group data can be very dramatic. The same conclusion can be drawn for the LQs of less educated adults, as shown in Table 6.2.

The plotting of LQs also allows us to understand where those tracts and block groups with higher proportions of poor households and less educated population than the regional average are located, how these localities relate to each other (clustering around a number of centers or spreading all over), and whether some of these localities cluster to form a critical geographic mass which spans across administrative boundaries, such as those of counties and towns. Our correct answers to these questions would be of important value for further research design and many policy and program initiatives.

(3) Housing value regression models

We wish to estimate linear regression models to explain the variations of median housing values by census tract and block group respectively for the greater Boston area. Our goal is two-fold. First, we hope to explore how census data aggregated at different levels affect the results of regression analysis of housing value. Second, and more important, we attempt to illustrate that GIS can be a very useful tool in the process of formulating research questions, constructing models, and interpreting results. Hence, it is important to make it clear that the regression models presented below do not necessarily represent the best possible regression models in explaining areal variation of median housing values.

To construct housing value models, we first created a long list of housing, socioeconomic, and demographic variables based on the 1990 census data. A selected list

Table 6.3: List of Variables Included in the Estimated Housing Value Models

Variable	Definition
HOUSING_VALUE	Median value of owner-occupied housing units (dependent variable)
LOW_INCOME%	Percentage of households with annual income less than $25,000
HIGH_EDUCATION%	Percentage of adults (>=25 years) with Bachelor's degree or higher
DISTANCE	Distance (in miles) from the CBD, Boston (the CBD is defined as the center point of the block group where Boston Common is located)

of these derived variables along with a correlation coefficient matrix are presented in Appendix A.2. Various socioeconomic indicators from the census can be used as possible independent variables to explain the areal variations of median housing value. However, many of these possible variables are highly correlated with each other. For example, variables measuring educational attainment are highly correlated with income; the percentage of non-white population with the proportion of low income households; and the distance away from the CBD with the population density related variables. To avoid the potential problem of multicollinearity among the independent variables and to keep the regression models simple, after numerous model runs, we developed a fairly good fitting and simple model including only four variables, as described in Table 6.3. Among these variables, distance is the only variable that is not directly derived from the census data. We compute distance from the geographic center of each block group or tract to the CBD of Boston based on the geographic coverages of block group and tract converted from the TIGER/LINE files. The computation of distance was completed by using GIS built-in functionalities to first derive the coordinates of the centroid for each block group and tract respectively, and then calculating the distance of each centroid to the assumed CBD coordinates.

Table 6.4: Housing Value Models for the Greater Boston Area: Tract versus Block Group Data, Census 1990

(dependent variable = HOUSING_VALUE)

Independent Variable	Parameter Estimates			Pearson Correlation Coefficients		
	Coefficients	Standardized Coefficients	t-statistics	Coefficients	Standardized Coefficients	t-statistics
Using census tract data (N =784):						
Constant	94,790		11.829			
LOW_INCOME%	78	0.015	0.589	-0.410		
HIGH_EDUCATION%	3,377	0.811	31.066	0.818	-0.544	
DISTANCE	-586	-0.077	-3.451	-0.232	-0.205	-0.187
F-statistics:	540.16					
Adjusted R-square	0.674					
Using block group data (N = 3,172):						
Constant	110,256		30.042			
LOW_INCOME%	-101	-0.021	-1.586	-0.385		
HIGH_EDUCATION%	3,076	0.752	56.670	0.775	-0.498	
DISTANCE	-621	-0.080	-6.893	-0.190	-0.128	-0.150
F-statistics:	1,623.92					
Adjusted R-square	0.606					

The regression results, including the regular and the standardized coefficients, are shown in Table 6.4. The coefficients in the second column tell us the estimated effect of a unit change in each of the independent variables on the dependent variable. The standardized coefficients in the second column compare the importance of the independent variables in determining the median housing values. The most important determinant of housing values is the percentage of adult population with higher education degrees. Also presented in the table are the correlation coefficients between those variables included in the models.

All estimated coefficients are significant at the one percent level except for those associated with the variables of percentage of low income households. In the block group model, as expected, the sign on this income variable is negative and nearly statistically significant at the 10 percent level. However, the same variable has a wrong sign and is statistically insignificant in the tract model. The significance levels for all independent variables but the percentage of low income households are considerably higher for the block group data than for the tract. Therefore, our regression results seem to suggest that using the block group data contribute to higher significance levels and greater consistency of individual coefficients. In terms of overall explanatory power, the tract model outperforms the block group slightly. The adjusted R-square declines from 0.675 for the tract model to 0.606 for the block group model. This decline is not surprising because it is generally true that when data are aggregated at higher levels, the variance for each regression variable is reduced. Therefore, models estimated for higher aggregation levels usually achieve higher R-square values, all other things equal.

The magnitude of impact of the independent variables on the median housing value differs between the tract and the block group models. According to the tract model, if two tracts differ by one percentage point in proportion of adult population with Bachelor's degree or higher education attainment, the median housing value of the tract would differ by \$3,377. For the block group model, the difference would be as high as \$3,076 for each percentage point difference in the variable for adults with higher education between two block groups. In contrast, the distance variables show a slightly stronger negative impact on the housing values in the block group model than in the tract. For each mile away from the CBD (Boston), median housing value for a block group would drop by an estimated

$621, noticeably higher than the $586 coefficient for a tract. The ways in which the distance variables behave in these models very much reflect the spirit of residential location theory formulated by Alonso (1964) and Muth (1969). This theory states that the market value of a standardized housing bundle declines with distance from the city center. We test this theory empirically using the housing sales data for the micro market area in next section.

Using GIS, we are able to make further comparisons between tract and block group aggregations by plotting the standardized error residuals for the two models, as shown in Figure 6.3. Tracts and block groups are shaded on the basis of the size of their corresponding standardized residuals estimated by the models. The following differences between the two maps of residuals are worth noting.

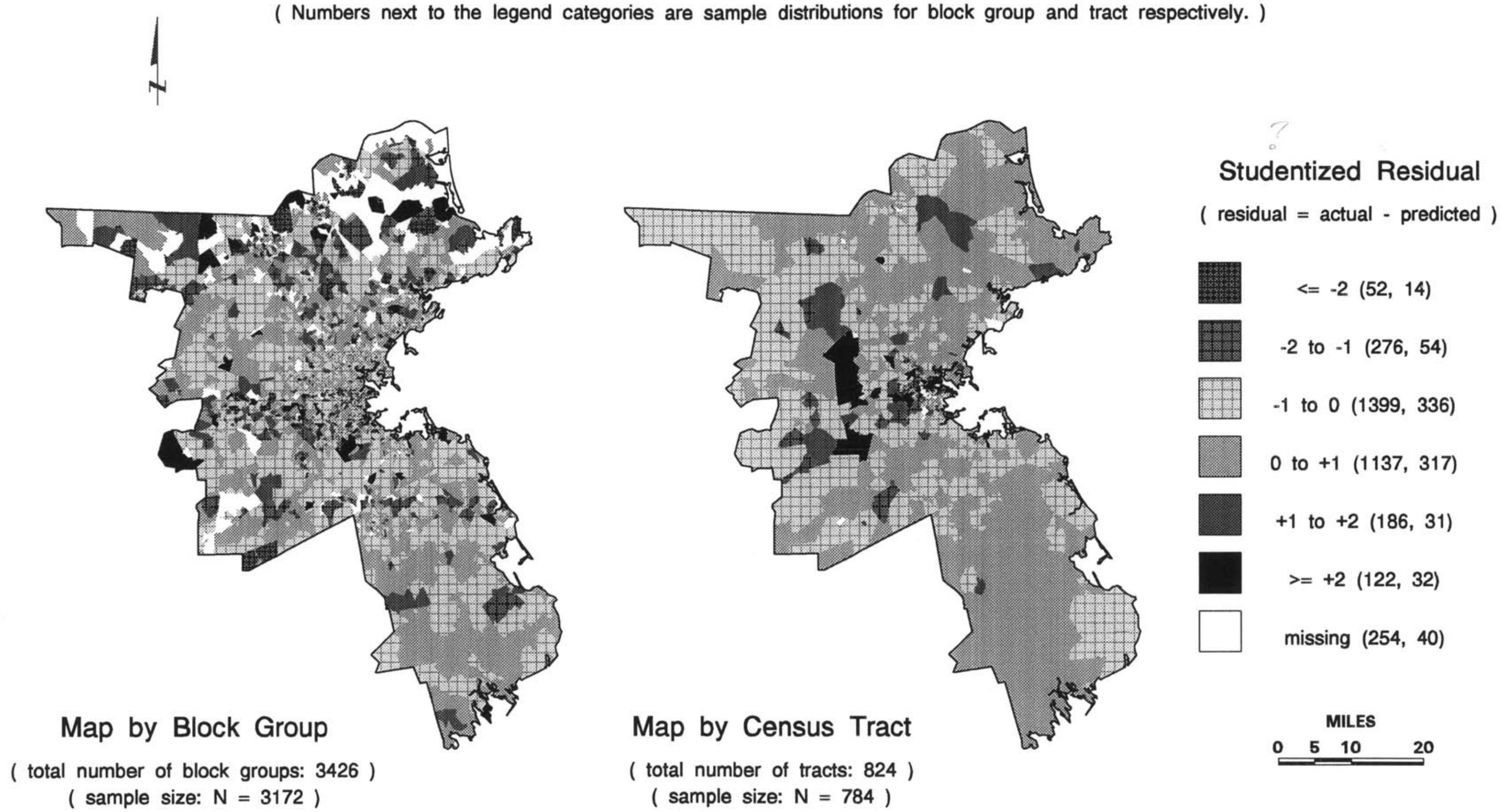
First, the block groups where median housing values are overestimated or underestimated by more than one unit of standardized error residual equal slightly over 20 percent of the total number of block groups. For the tract model, overestimations and underestimations occur in less than 17 percent of the total tracts. One might not conclude that there exist significant differences in the numerical distributions of error residual between the tract and block group models. However, it is readily observable that the patterns of spatial distributions of overestimations and underestimations are very different, as seen in Figure 6.3. On the block group map, many overestimations and underestimations occur in the peripheral areas of the region. In contrast, on the tract map, the areas of large estimation errors are mostly situated in the smaller tracts close to the center of the region. It is clear from these comparisons that the display of the residuals is useful in understanding the behaviors of the models and interpreting the model results.

Second, in a dozen tracts situated to the west of the CBD (Boston), and forming a roughly vertical band, median housing values are consistently underestimated by more

Figure 6.3 Error Residuals of Housing Value Models: Census Track vs. Block Group

(The Greater Boston Area, Census 1990)

(Numbers next to the legend categories are sample distributions for block group and tract respectively.)



than one unit of standardized errors by the tract model. The use of GIS to portray the error residuals by location lends us an easy and intuitive way of identifying areas where the estimated models seem to fall apart. It enables us to quickly focus our attention on the problematic areas, initiate extra efforts to investigate other possible causes for the problems, and to reformulate research hypotheses. We soon found out that these tracts form the six or seven most wealthy towns in the greater Boston area. Our tract model predicts poorly for these towns because many factors that are unique to these locations have not been considered in the models, including zoning, minimum lot size, and public service choices and quality. It is also readily apparent from the maps that the block group model is more successful than the tract model in explaining the change of housing values for those wealthy towns.

Third, an area where housing values are overestimated using the tract data may change to one with housing being undervalued when using the block group data, and *vice versa*. For example, the areas that are situated on the far left side of the region show overestimated median housing values on the tract map and underestimated value on the block group map. Thus, it seems that models estimated using data aggregated at different levels generate different predictions.

6.1.4 Summary

We have used two sets of census data aggregated at the tract and block group levels, respectively, to examine the diversity of housing and population characteristics. Three conventional analytical techniques, Gini coefficient, location quotient, and regression analysis, have been used to test the differences between using the tract and block group data. We find that different data aggregations can lead to significantly different research results and conclusions. The Gini coefficient analysis reveals a nearly 35 percent greater inequality of distribution of poor households in the greater Boston area at the block group level compared to the tract. The location quotients calculated using the tract and block group data also show alarming differences. By plotting the location quotients by area using GIS, we find that much richer and more complicated spatial pattern of distribution of poor households and less educated population are revealed at the block group level. The analysis of housing value regression models sheds additional light on the important dissimilarities in different levels of data aggregation. The model constructed using the block group data reduces the overall explanatory power slightly, but contributes to higher significance levels and greater consistency of individual coefficients. From these three sets of analysis, it can be concluded that research results are very sensitive to the definition of the areal entities. Also, it can be concluded that the block group data seem to be most valuable when one's research goal is to reveal and take into consideration as much of the diversity within a locality as possible. The plotting of model error residuals appears to be a very instructive way of examining the performance of regression models by location. It provides a useful basis for identifying ways to further improve regression models.

The above analysis also shows that GIS technologies can make useful value-added contributions to the conventional analysis of income distribution and housing market by examining them spatially and by improving the construction of regression models that use areas as their unit of analysis.

6.2 Improving Housing Hedonic Price Models with the Aid of GIS

6.2.1 Introduction

In this section, we construct several housing hedonic price models with the aid of GIS. We use GIS to combine different sources of relevant data based on their common location attributes and to measure accessibility variables to be included in the models. We demonstrate that GIS offers a potential of modeling the impact of accessibility in a more realistic manner. Also, we employ GIS to display the error residuals spatially and design maps that illustrate the behavior of the models graphically and intuitively. We then discuss several possible uses of GIS-based maps of error residuals.

The section is organized as follows. First, we discuss the processes of data aggregation and correction, and address matching of the housing sales data. Then, we describe four different approaches to measuring accessibility using GIS. We test statistically the differences between each pair of measurements. Also, we discuss the potential of further improvement towards a more realistic measurement of accessibility. Next, we present and analyze our empirical results of housing hedonic price models. The results show that improved measurements of accessibility contribute significantly to the improvement of the hedonic models. Finally, we display the error residuals from two selected hedonic price models spatially. The magnitude and direction of the residual for each housing sale case are linked to the exact location of the housing. The graphical representation of errors provides an intuitive and informative way of conveying the outcomes of the models and a useful analytical aid to model application.

6.2.2 Data aggregation, address matching, and spatial distributions of housing sales

Our hedonic housing price models include an array of variables that are created from these three different sources: the 1990 Census data aggregated at both tract and block group levels, the 1989-91 housing sales data for the city of Boston, and GIS-generated distance measurements. In the first section of this chapter, we have already revealed some significant differences between using the two aggregation levels of census data. In this section we test further the impact of the data aggregations on housing price in combination with micro housing sales data. In this subsection, we focus on explaining how the location information in the micro housing sales data are processed to merge with the Census data and to provide necessary spatial information for constructing accessibility variables. The keyword for this process is address matching[9].

In addition to various housing structural variables such as size of living area, number of bathrooms and bedrooms, and lot size, the housing sales data contain detailed information on each property's location (*i.e.*, address). A typical address contains a street number, street name, city or town name, state, and ZIP code. With this address information, we can match most housing unit to its location on a geo-coded street map of the city. This digital map, prepared from the 1992 TIGER/Line file, contains every street segment of the city and its associated location information, such as street name, range of street numbers, and ZIP code. The result of address matching creates a spatial linkage between the individual housing unit and its neighborhood environment.

We began with the housing sales data that contains 1,571 residential sales. Our first attempt at address matching using the ARC/INFO software resulted in 1,433 successful matches, a rate of 91.2 percent. This is a very high matching rate in view of the fact that there are several common problems associated with TIGER/Line files: positional accuracy, data accuracy with regard to street names and address ranges, duplicate records, and missing ZIP codes. We carefully diagnosed the 138 sales that failed to match. We corrected various errors such as misspelled street names, improperly coded numbers and

9. Address matching is a process that compares two addresses to determine whether they are the same. If they are, a data relationship between the two addresses is established. This relationship permits geographic coordinates and attributes to be transferred from one address to the other.

street suffix based on our *priori* knowledge of the study area. In a few cases, we had to assign an approximate street number of a housing unit because the incorrect address number ranges in the TIGER file. With these corrections, our second attempt led to an additional 136 successful matches. Thus, we ended up with a geo-coded housing sales data set of 1,569 records of housing sales. The successfully matched sales are portrayed in Figures 6.4 and 6.5, respectively by housing type and by price. It should be noted that each successful match represents only an approximation of where a housing unit falls on a particular street segment.

A number of observations can be made based on the spatial distribution of the housing sale. First, it is clear from Figure 6.4 that the majority of condominium units are concentrated in four neighborhoods in the proximity of Downtown Boston and the waterfront. These neighborhoods are Beacon Hill, Back Bay, Charlestown Navy Yard, and North End. The concentrations are in reality much more intense than those depicted in the figure because in many cases, one dot on the map represents multiple condo units in the same building. The high concentration of these condo sales and their proximity to the employment center suggests that there exists very small variations in their accessibility to the center measured in terms of distance. Thus, it is difficult to use the condo sales to test the hypothesis regarding the effect of accessibility on housing price. More important, condominiums as a type of housing differ in many ways from other housing types. The pooling of condominium sales data with the data of other housing sales makes little theoretical sense. Therefore, we decided to omit the condo case from our study. Nevertheless, this observation shows the value of being able to use address information and the importance of understanding the spatial distribution of data samples.

Second, Figure 6.4 also shows that most single-family housing sales are located further away from Downtown than most of the three-family housing sales, with the sales

Figure 6.4: Spatial Distribution of The Housing Sales Cases By Type

(Housing Sales In The City of Boston, 1989.1 - 1991.9)

(the total number of housing sales matched and displayed: N = 1569)

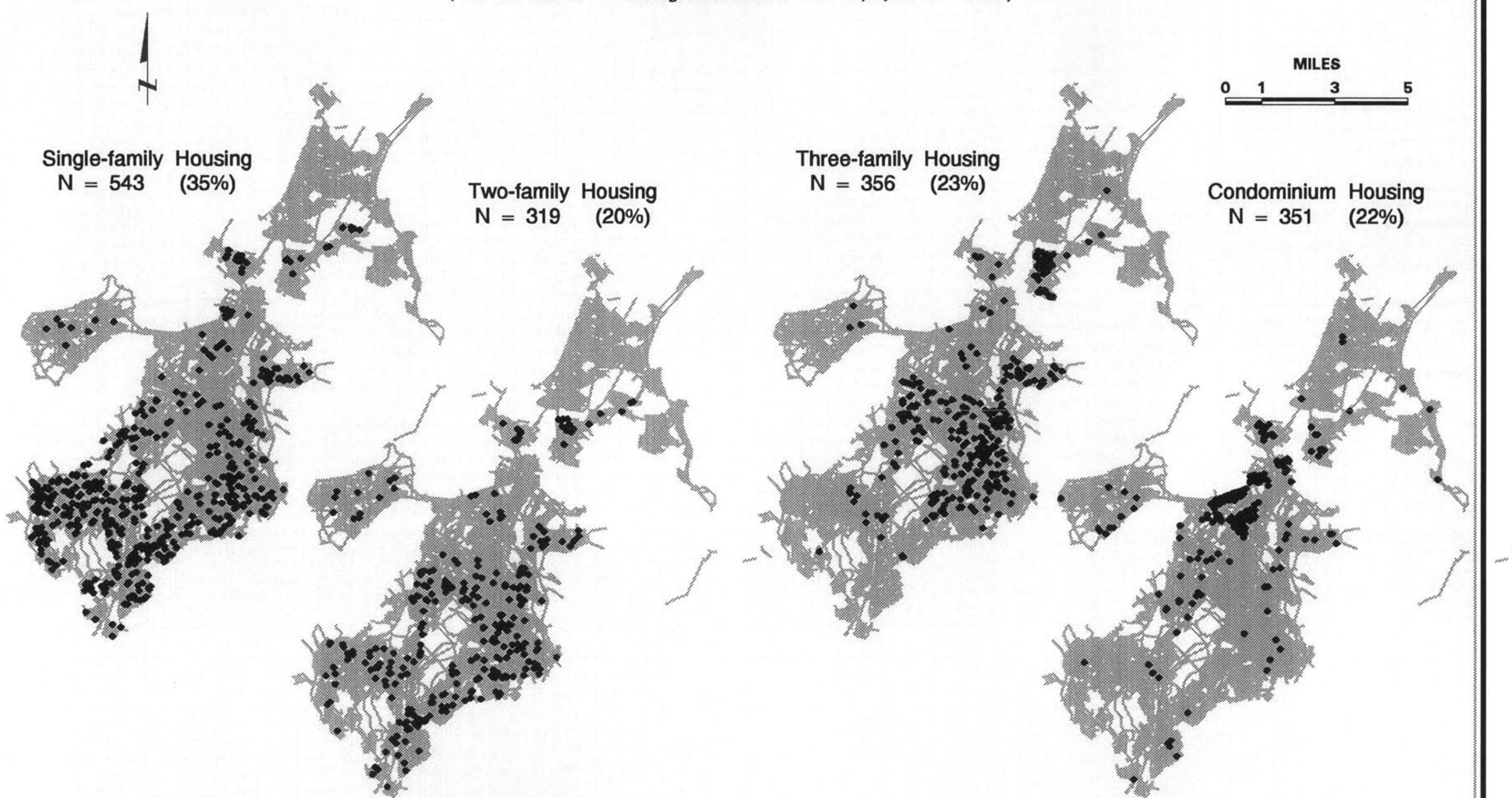
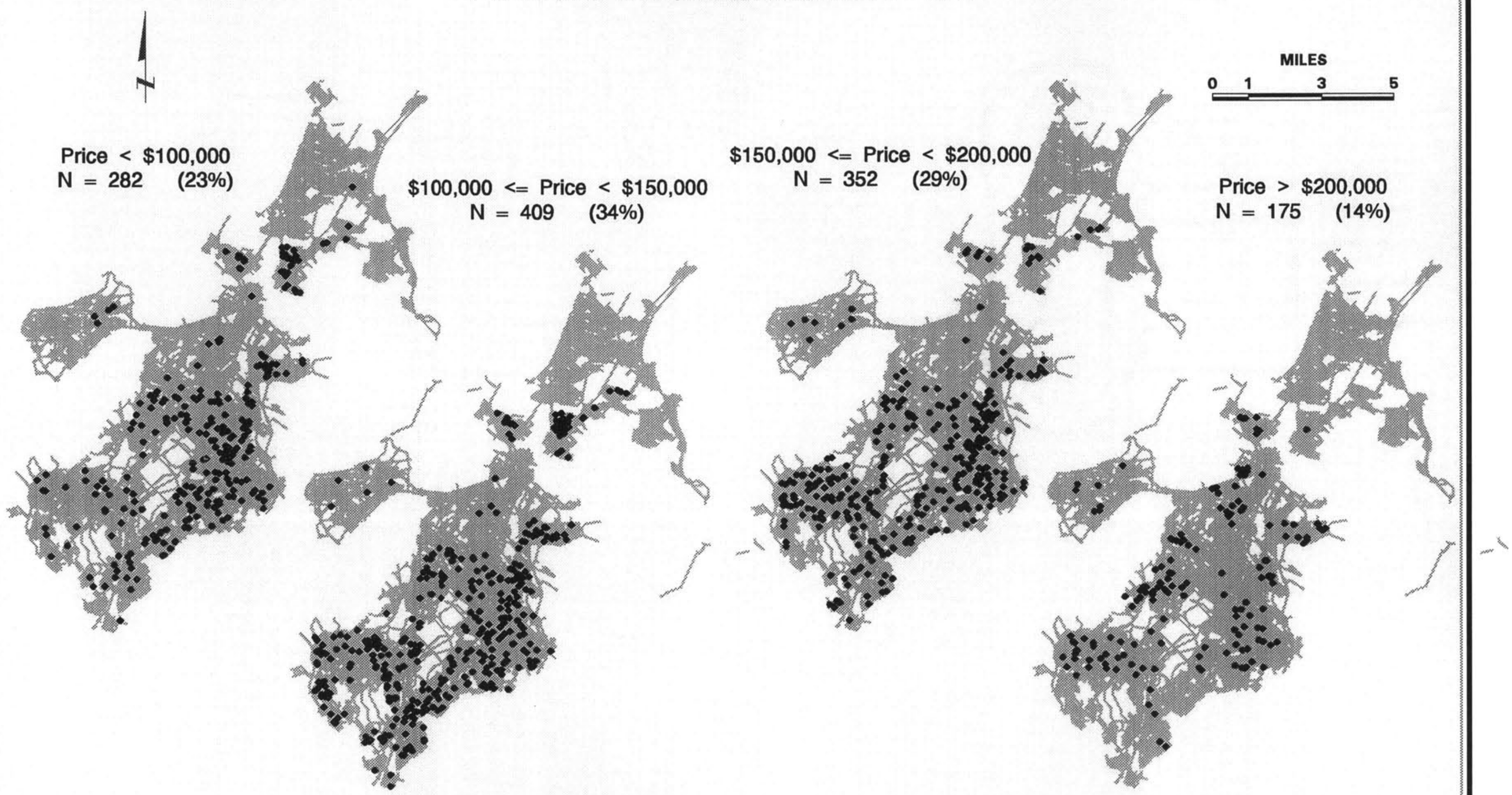


Figure 6.5: Spatial Distribution of The Housing Sales Cases By Price

(Housing Sales In The City of Boston, 1989.1 - 1991.9)

(the total number of family housing sales displayed: N = 1218)



of two-family housing mingling between the two. This simply confirms our common sense that building density increases in the direction of proximity to the employment center. This spatial display of housing sales by type offers an intuitive way of examining whether the results of address matching, which occurred inside a "colorful computer screen," or a "black box" when a screen saving procedure is activated, make spatial common sense. In fact, in the process of this research, a simple display of the spatial distribution of the housing sales by ZIP code has led to a discovery of many data coding errors of location attributes in the 1990 TIGER/Line files. After examining a number of similar exercises of displaying matched sales data, we have concluded that the 1992 TIGER/Line files are much more accurate than its predecessors.

Our third observation emerges from eyeballing the four maps of spatial distribution of the family housing sales by price, as shown in Figure 6.5. Each map depicts one price-category of housing sales. The maps provide a quick-and-dirty approach to learning whether there exist any salient patterns of spatial clustering in housing price. These maps seem to suggest no strong reason to suspect the existence of spatial autocorrelation. Obviously, visual inspection of spatial patterns cannot substitute for more rigorous statistical tests of spatial autocorrelation, most of which require complex and tedious mathematical computation. Nonetheless, even this simple and "primitive" way of checking would have been unlikely without GIS technologies. No empirical testing of spatial autocorrelation of housing price in any form has been found in the reviewed literature on housing hedonic price models. The question of spatial autocorrelation is beyond the scope of this research. However, it is worth noting that GIS technologies seem to provide an effective tool to pursue various research questions related to spatial statistics.

Once the sales data are matched with a digital map, there are many possibilities for displaying the spatial data quickly and easily using GIS. For example, given the available variables in the data set, we could also display the spatial distribution of the housing sales by size, date of transaction, and mortgage lender. All these graphic representations of data by location provide an intuitive and luminous way of understanding various important characteristics of a large volume of spatial data.

After the address matching, it is a simple task to use GIS to identify each housing sale's corresponding tract and block group. This identification process establishes a spatial linkage between each house and its neighborhood. Thus, the census data on neighborhood socioeconomic characteristics, aggregated at both the tract and block group levels, can be merged easily with the housing sales data. In this newly combined data set, each record of housing sale contains a set of structural attribute variables as well as numerous socioeconomic variables aggregated at two different geographic levels. In this case, the use of GIS technologies largely liberates researchers from the laborious tradition of merging the two sources of data by manually encoding each observation with its corresponding tract or block group identification.

In the process of constructing housing hedonic price models, we have made the following necessary adjustments to the housing sales data. First, all condo sales are excluded from the data set. Thus, we are left with 1,220 records for the family-type housing sales. Then, three observations are removed because of missing values for the socioeconomic variables. All sales with a recorded price less than or equal to $50,000 are excluded. A close examination shows that the majority of these low-priced sales are non-arm's length transactions. These are often non-competitive sales and their prices do not reflect the demand and supply forces of the housing market at work. Thus, we are left with 1,098 observations in the final data set. Then, two more adjustments are made. First, for 49 observations with missing values for the variable of median housing value, we replace the missing observations by the sample mean of the available observations. Second, we adjust the housing sale prices to reflect the impact of inflation, resulting in all costs measured as 1990 dollars[10].

In addition to combining the housing sales data with the census data, we use GIS to improve the measurement of accessibility. Address matching of the housing sales data with the street network coverage derived from the TIGER/Line file also provides a necessary building block for the possibility of improving the measurement of accessibility variables. We now turn to this subject.

6.2.3 Accessibility measurements: assumptions and alternatives

(1) Traditional approach and assumptions

As indicated in the literature review of housing hedonic studies in Chapter 2, traditionally, it is a common practice to use the straight line distance, or *Euclidean* distance, between the centroid of a tract to a designated employment center as a proxy to represent the accessibility of a residence to work. The measurement was usually carried out on a paper map. Then, all sale samples located in a same tract would be assigned an identical value of the distance measured from the tract center. This approach is tedious and laborious. Worse yet, depending on the size and shape of a tract, the proxy could be a badly distorted representation of the real distance. We suspect that the use of a poor measurement of accessibility is a primary cause for the insignificant and inconsistent performance of the accessibility variables in many of the existing housing hedonic price models.

In the real world, the concept of accessibility of a residence is very complex. It could take many possible notions: for example, accessibility to employment centers, schools,

10. The price adjustment is made based on the numbers in the column of "Total Housing" in Table B-56, Consumer Price Indexes for Major Expenditure Classes, 1950-92, in Economic Report of the President. January 1993. Washington: United States Government Printing Office.

major public transit terminals, supermarkets, and shopping malls. These notions are further complicated by the fact that there is usually more than one member in a family who travel to various locations by different modes of transportation and with different frequencies. Even more intricately, each trip could occur at different hours of a day, at or off the peak traffic, and follow streets and highways with different speed limits and road quality. On top of all these complexities, we might add the calculation of how one values his/her time spent travelling and how one perceives and values the degree of comfortableness in travelling. In short, some forms of abstraction and simplification are necessary if one wants to model the impact of perceived accessibility by a household on housing price.

The traditional approach toward measuring accessibility of a residence, as reviewed above, has made the following assumptions, explicitly or implicitly. First, journey to work is the most important and frequent trip made by a household. This assumption has become weaker over time as society moves towards shorter working hours and contemporary communication technology allows many workers to work at home. Second, there is an assumption of a single major employment center in the study area and that every working member of a household travels to this center to work. This is also a weak assumption as there is abundant evidence supporting the trend of population suburbanization and the growth of suburban employment centers in the past two decades. Nevertheless, these two assumptions are necessary to make a complex concept manageable. The third common assumption, and the most unrealistic of all, is that workers can travel to the center on a featureless and uniform terrain and follow a straight-line path. This idealistic assumption is primarily due to the inability of research tools at hand to measure the distance travelled by workers in a more realistic manner.

Like most of the existing empirical studies of housing hedonic price models, this research continues the tradition of making the first two assumptions to keep our exploration manageable. However, we illustrate below that with the aid of GIS technologies, it is possible to relax or drop both the second and third assumptions. We then drop the assumption of featureless terrain and use GIS to measure the actual street distance travelled by workers. We test the differences between the three more disaggregate and localized measurements and the traditional proxy and their impacts on hedonic housing prices.

(2) GIS-based measurements of accessibility

Most of today's sophisticated GIS software contains the capability of network analysis. We use the ARC/INFO network analysis module[11] to illustrate how the measurement of accessibility can be improved. This module provides tools to find the shortest or least-cost path through a street network. The cost may be determined by any attribute of the network components that is expressed in numeric terms. For example, distance or travel time may be used to compute the shortest path in a transportation network, or a combination of factors may be used to calculate a monetary cost value. The module also computes accessibility as an aggregate measure of how reachable a location is from other locations given the attractiveness of other locations.

The pathfinding tool allows the calculation of actual distance travelled in a street network or terrain model of a city. This tool makes it possible to drop the assumption of the straight line distance. We test below how much difference it makes to substitute the traditional approach to measuring accessibility with the GIS-based approaches. It is important to point out that the impact of distance on a particular activity may be linear or

11. The material on the network analysis module relies primarily on ARC/INFO User's Guide: Network Analysis, Environmental Systems Research Institute, Inc. April 1992.

non-linear. Using the network analysis module, this phenomenon of viewing distance as a non-linear deterrent to movement can be modeled with a distance-decay function. Moreover, other possible impedances to travel, such as quality of a street, speed limits, and traffic volume, can also be easily modeled if these data are available.

The module's capability of computing accessibility in an aggregate fashion allows the relaxation of the assumption of a monocenter of employment. Mathematically, the computation of a weighted measure for accessibility can be specified by the following equation:

$$A_i = \sum_{j=1}^{n} \frac{W_j}{d_{ij}^{\beta}} \tag{6.2}$$

Where

A_i is the accessibility index at location i,

W_j is the attractiveness of center j,

d_{ij} is the distance between location i and center j,

β is the exponent for distance decay,

n is the number of centers in the market area.

In the case of modeling accessibility to employment centers, we could use the total number of employees to represent the attractiveness of a center. The distance would be computed in terms of the minimum impedance from location i to center j. The accessibility index is called *the potential for interaction* in some geographic or urban and regional economic textbooks. The computed index can be further multiplied by some known factor of the origin locations to yield a weighted accessibility measure. This is often called the gravity model. The contribution added by GIS technologies to this conventional gravity model is that real distance or time travelled can be modeled in a much more realistic manner once a GIS based street network model is constructed. In our case study of the micro housing market, the city of Boston, we maintain the traditional assumption of a single employment center since it closely represents reality. We use a GIS-based street network model to compute accessibility of each housing unit in the data set to a designated employment center.

(3) Four alternatives of measuring accessibility

In this research, we assume Downtown Crossing, or the intersection of Washington Street and Summer Street, to be the employment center for the purpose of distance computation. Given the available data, we select four alternatives of measuring accessibility in terms of distance, compare their differences, and test their impacts on modeling housing price variation. The four alternatives are defined as follows:

(i) TRACT_DISTANCE: the *Euclidean* distance from the centroid of a tract to the designated employment center;

(ii) BLOCKGRP_DISTANCE: the *Euclidean* distance from the centroid of a block group to the designated employment center;

(iii)HOUSING_DISTANCE: the *Euclidean* distance from the location of a house to the designated employment center;

(iv)STREET_DISTANCE: the distance of the shortest path measured using the GIS-based street network model from the location of a house to the designated employment center.

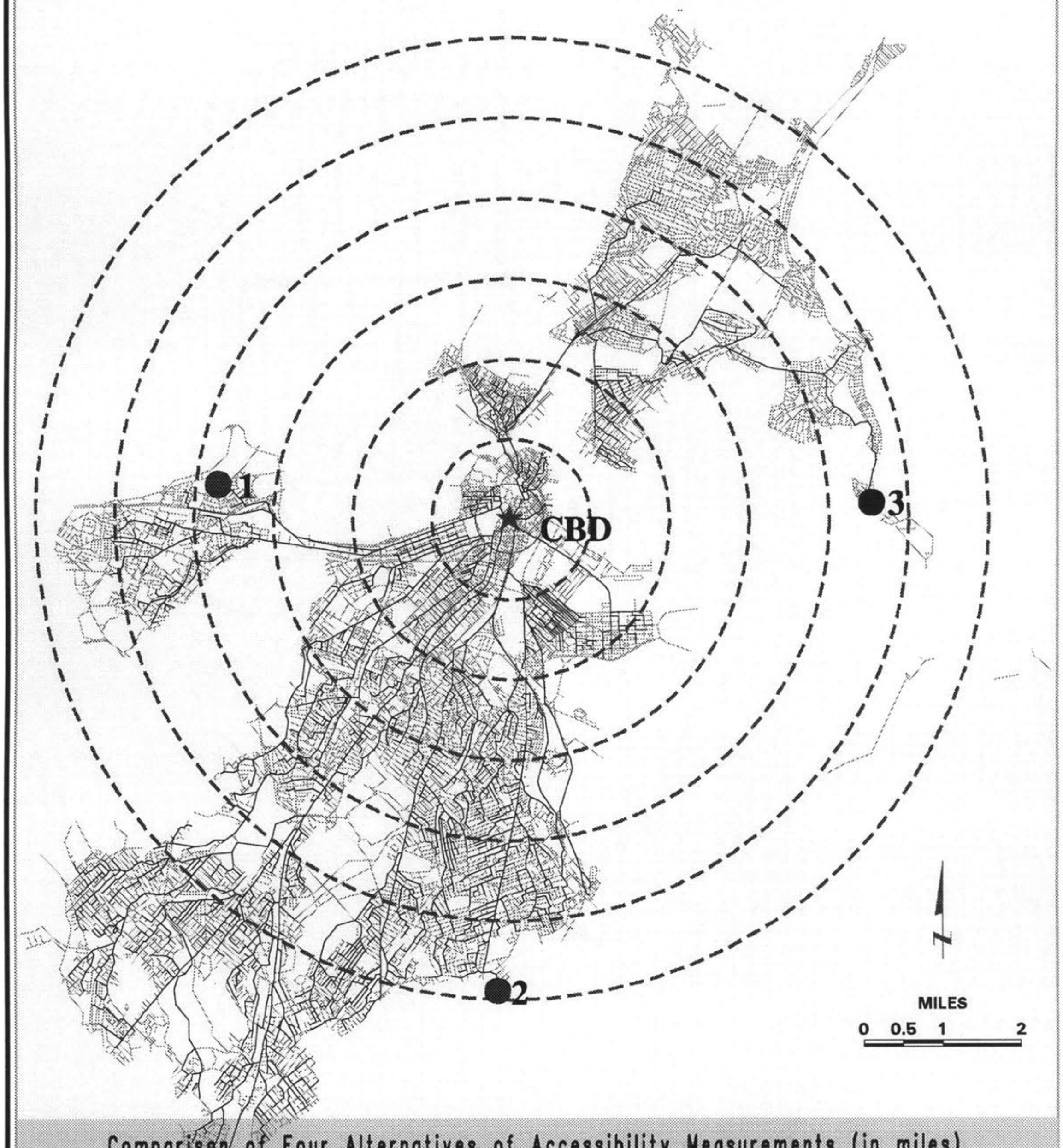
All four measurements are carried out using the ARC/INFO software (see Appendix B for details). After we have completed the address matching of the housing sales data to the street network of Boston, we obtain a geographic coverage of points, each of which represents a housing sale. The first three measurements involve the computation of

Euclidean distance between point(s) of one coverage to point(s) of another coverage. Essentially, distance between each pair of coordinates is computed. The first and second alternatives use the centroids of tract and block group polygons, respectively, to represent the location of each housing unit. The first is identical to the most commonly used traditional measurement of accessibility except that ARC/INFO can complete the calculation of the distances of 189 tracts in the city to the center in a few seconds. The fourth alternative, which is a more realistic measurement of street distance travelled, is a bit more complicated in computation. We first convert each point in the housing location coverage into a node in the street network coverage. Next, using the pathfinding tool in the ARC/INFO network analysis module, we compute the distance of the shortest path through the street network between each housing location and the center. This approach requires extensive data manipulation and significantly more computing time. For each pair of nodes, it takes nearly 30 seconds to find the shortest path in the street network model of Boston, using an IBM PowerStation 6000. Thus, it takes approximately 9 hours to complete the calculation for the 1098 housing sales cases. Figure 6.6 illustrates the results of the shortest pathfinding for each of the housing sales cases. It also shows a few extreme examples of differences between the Euclidean distances and the shortest street distance.

From the first alternative to the last and in this order, the approximation of accessibility moves closer towards the actual distance that a worker would travel from his/her house to the center. Table 6.5 summarizes the results of these calculations. It is clear that the first three means are close. The mean for the shortest street distance is a half mile greater than those of the others. Table 6.6 shows the statistical tests of differences between the STREET_DISTANCE and each of the three other measurements. For example, the difference variable of "STREET_DISTANCE-TRACT_DISTANCE" is computed as follows. First, we transform the tract distance and the street distance for each housing unit

Figure 6.6: GIS-based Accessibility Measurements
An Illustration of The Shortest Street Distance

(The radius increases at 1 mile per concentric circle.)



Comparison of Four Alternatives of Accessibility Measurements (in miles)

Location ID	Location of the examples	Shortest street distance	Euclidean distance from the house	Euclidean distance from block group centroid	Euclidean distance from tract centroid
1	Allston	4.35	3.79	3.59	3.87
2	South Dorchester	6.26	5.81	5.46	5.29
3	Deer Island	8.58	4.58	4.72	4.83

Table 6.5: Analysis of the Distance Variables (in miles)

Variable	N	Mean	Std. Dev.	Minimum	Maximum
TRACT_DISTANCE	1098	4.49	2.00	0.40	8.41
BLOCKGRP_DISTANCE	1098	4.48	1.99	0.26	9.08
HOUSING_DISTANCE	1098	4.54	2.03	0.18	9.18
STREET_DISTANCE	1098	5.09	2.09	0.18	9.83

Table 6.6: t-test for Paired Comparisons of Distance Variables in Deviations Form

Variable	Mean	Std Error	t	Prob>|t|
STREET_DISTANCE - TRACT_DISTANCE	0.31	0.010	32.62	0.0001
STREET_DISTANCE - BLOCKGRP_DISTANCE	0.29	0.010	28.85	0.0001
STREET_DISTANCE - HOUSING_DISTANCE	0.27	0.011	24.58	0.0001

to deviation forms by expressing computed distances for each observation in terms of deviations from their respective means. Then, we obtain the difference variable by calculating the differences between the two distance variables in deviations forms. The null hypothesis for this pair of comparison is that there is no difference between the *Euclidean* distance measured from the centroid of a tract to the center and the distance measured as the shortest street path through the network. All t-statistics indicate that significant differences exist and we reject all three null hypotheses.

6.2.4 Empirical results of hedonic price models

The purpose of hedonic analysis is to explain the equilibrium variation in housing prices as a function of the attributes of dwellings, locations, and neighborhood and environmental amenities. In this research, using the micro data of housing sales for the city of Boston combined with census data and the GIS-based accessibility measurements, we attempt to construct improved housing hedonic price models and test the following two hypotheses:

Hypothesis 1: Socioeconomic attributes measured at the block group level add more explanatory power and more consistent coefficient estimates to housing hedonic price models than those measured at the census tract level.

Hypothesis 2: The accessibility to employment center(s) measured in a more realistic manner with the aid of GIS is a stronger explanatory variable in housing hedonic price models than that approximated as a straight-line distance.

Six hedonic price models, shown in Table 6.7, are constructed to test these two hypotheses. The dependent variable is PRICE (in 1990 dollars). In each model, the price of each house is regressed against two variables for characteristics of the dwelling structure, five aggregate variables of neighborhood quality based on the 1990 census data, and one variable of accessibility. The variables of dwelling structural attributes (*i.e.*, number of bathrooms (BATHROOM#) and size of living area (SIZE_LIVING) in square feet), are common in all six models. The five neighborhood quality variables are the median value of housing stock (MEDIAN_VALUE), the percentage of adult (25 years or older) with college or higher education (HIGH_EDUCATION%), the percentage of non-white population (NONWHITE%), the percentage of households with annual income less than $15,000 (LOW_INCOME%), and the median age (in years) of housing stock (MEDIAN_AGE). In the first three models shown in the upper half of Table 6.7, these neighborhood quality variables are aggregated at the tract level and the last three models at the block group level. Models I, II, and III differ in terms of the accessibility variable used, respectively, in each of which the distance from a tract to the center (TRACT_DISTANCE), the distance from housing to center (HOUSING_DISTANCE), and the street distance from housing to center (STREET_DISTANCE) are included. Models IV, V, and VI differ in the same fashion except that Model IV uses the distance

Table 6.7: Comparison of Hedonic Estimates with Census Tract and Block Group Aggregations and GIS-based Accessibility Measurements

(Dependent Variable = Housing Price; N = 1098)

	Models with Data Aggregated at Tract								
Attributes	I (Tract_Distance)			II (Housing_Distance)			III (Street_Distance)		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	15,146.00		0.51	44,113.00		1.52	50,386.00		1.70
BATHROOM#	14,479.00	0.11	3.45	12,901.00	0.10	3.09	13,623.00	0.10	3.31
SIZE_LIVING	23.25	0.22	7.23	23.52	0.23	7.35	22.97	0.22	7.20
MEDIAN_VALUE	0.93	0.48	13.32	0.89	0.46	12.90	0.89	0.46	12.86
HIGH_EDUCATION%	646.11	0.09	2.16	674.96	0.09	2.26	622.31	0.08	2.08
NONWHITE%	-216.99	-0.07	-2.63	-196.23	-0.06	-2.38	-224.21	-0.07	-2.73
LOW_INCOME%	-974.96	-0.08	-2.76	-1,188.70	-0.10	-3.41	-1,154.95	-0.10	-3.38
MEDIAN_AGE	-1,328.13	-0.08	-3.43	-1,421.04	-0.08	-3.69	-1,393.78	-0.08	-3.64
Ln(STREET_DISTANCE)	-	-	-	-	-	-	-20,237.00	-0.10	-3.46
Ln(HOUSING_DISTANCE)	-	-	-	-17,981.00	-0.09	-3.30	-	-	-
Ln(TRACT_DISTANCE)	-9,544.22	-0.05	-1.69	-	-	-	-	-	-
Adjusted R-sq	0.4998			0.5035			0.5040		
F-statistics	138.04			140.05			140.32		

	Models with Data Aggregated at Block Group								
Attributes	IV (BlockGrp_Distance)			V (Housing_Distance)			VI (Street_Distance)		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	149,899.00		5.76	147,669.00		5.76	161,322.00		6.10
BATHROOM#	12,610.00	0.09	2.82	12,410.00	0.09	2.77	13,835.00	0.10	3.15
SIZE_LIVING	25.79	0.25	7.47	25.84	0.25	7.49	24.74	0.24	7.21
MEDIAN_VALUE	0.46	0.27	8.57	0.47	0.28	8.84	0.46	0.27	8.60
HIGH_EDUCATION%	1,099.75	0.17	4.74	1,134.03	0.17	4.89	1,060.49	0.16	4.57
NONWHITE%	-244.77	-0.08	-2.98	-230.22	-0.08	-2.80	-273.82	-0.09	-3.34
LOW_INCOME%	-1,253.76	-0.14	-4.89	-1,250.91	-0.14	-4.89	-1,217.35	-0.13	-4.80
MEDIAN_AGE	-2,010.17	-0.13	-5.54	-2,008.91	-0.13	-5.54	-1,966.82	-0.13	-5.46
Ln(STREET_DISTANCE)	-	-	-	-	-	-	-38,882.00	-0.18	-6.78
Ln(HOUSING_DISTANCE)	-	-	-	-34,198.00	-0.18	-6.48	-	-	-
Ln(BLOCKGP_DISTANCE)	-34,135.00	-0.18	-6.38	-	-	-	-	-	-
Adjusted R-sq	0.4299			0.4305			0.4325		
F-statistics	104.39			104.68			105.51		

Note:
STREET_DISTANCE = street distance of the shortest path from the location of a house to the CBD;
HOUSING_DISTANCE = Euclidean distance from the location of a house to the CBD;
TRACT_DISTANCE = Euclidean distance from the centroid of a tract to the CBD;
BLOCKGP_DISTANCE = Euclidean distance from the centroid of a block group to the CBD.
Sources: a spatially integreted data set based on 1990 Census, 1992 TIGER/Line File, and 1989-91 Housing Sale Data from
County Data Corporation, VT.

from a centroid of block group to the center. All the distance variables are in logarithmic form.

The models in Table 6.7 have been estimated with ordinary least square assumptions, since the price data are not observations on the same housing unit over time and thus, the estimated equations do not exhibit serial correlation. There are other possible candidates for housing structural attribute variables, such as number of bedrooms and lot size, that can be included in the models. However, analysis of correlation coefficients among the attributes shows that these candidates are highly correlated with the two attributes selected. The same thing can be said about a myriad of socioeconomic variables derived from the census data. For example, the variable of median income is highly correlated with the percentage of low income households and the median rent with the median housing value. We have chosen to keep the models simple and not to include variables that are highly correlated with each other. As a result, the six models constructed do not show multicollinearity to any significant degree. Correlation analyses for the housing structural attribute variables and the selected hedonic variables are presented in Appendices A.3 and A.4.

Each of the coefficients has correct signs. All the hedonic attributes except TRACT_DISTANCE are statistically significant at the 1 percent level. The log distance variable makes good spatial sense in this market area given the spatial distribution patterns of housing sales. As shown in Figure 6.4, most of the housing sales cases are located to the south of the employment center and spread fairly evenly over an area as far as five or six miles away from the center. The significance of the coefficients for the log distance variables suggests that the slope of price increase is much steeper as the location approaches the center of the city. In the following, we focus our discussion of model results on what evidence has been shown to support or reject the two hypotheses.

First, we examine the differences between the three tract models and the three block group models. Magnitudes of influence from the majority of the hedonic attributes on price are comparable between the tract and block group models. The only striking difference between the two sets of models is the impact of the variables of distance and median housing value. The impact of MEDIAN_VALUE on price in the block group models is only half of its impact in the tract models. To illustrate, we compare Models III and VI. Since the means of housing MEDIAN_VALUE for both tracts and block groups are around $166,000, according to the coefficients of MEDIAN_VALUE in Model III, a house located in a mean value tract would add an additional $150,000 to the price, assuming other things equal; while Model VI shows only half of this amount of impact resulting from the MEDIAN_VALUE variable. The nearly exact opposite is true for the impact of the distance variable. For example, in Model III, *ceteris paribus*, for every mile away from the center, the housing price would decline by $20,000. In contrast, Model VI shows a nearly $40,000 per mile impact of the distance variable. The large differences in these estimated coefficients seem to reflect the fact that in a large city, such as Boston, many tracts are characterized by very diversified block groups in terms racial composition, educational attainment level, family income, and median housing values. Like the results shown in the study of the macro housing market, the socioeconomic indicators aggregated at the block group level capture a more diversified image of urban reality. The models estimated using census data aggregated at the tract level would have a higher tendency to either overestimate or underestimate a housing price.

The regression results show that all three tract models explain around 50 percent of the variation in housing price. The block group models explain around 43 percent of price variations, lower than that of the tract. Judging from the adjusted R-squares, it would seem that we should reject the first hypothesis on the impact of data aggregation. This is not uncommon because it is generally true that variables aggregated at a higher aggregation level have smaller variation than those of their counterparts at a lower level. Therefore, a small decrease in R-square does not necessarily imply a poorer model. It is important to examine also other differences in the two groups of models. A more salient difference between the two groups of models is the t-statistics associated with the five socioeconomic regressors. Except for MEDIAN_VALUE, t-statistics are higher for all the other four regressors in the block group models than in the corresponding tract models. Therefore, the disproportionally high impact of MEDIAN_VALUE occurring in the tract models is reduced by half in the block group models. The other difference between the tract and block group models is that the constant terms are statistically insignificant in all the tract models.

Next, we turn to the second hypothesis. The regression results seem to suggest that the hypothesis of the choice of accessibility variables can not be rejected. In fact, according to our models, there is sufficient evidence to support the statement that the traditional accessibility variable measured as the distance from a tract centroid to the center is the worst regressor among the four tested alternatives of accessibility measurement. Model I shows that the t-statistic for TRACT_DISTANCE is insignificant. For the other tract models, both coefficients for STREET_DISTANCE and HOUSING_DISTANCE are significant and the former is slightly superior to the later in terms of t-statistics. This is also true for all the block group models. Another noticeable finding is that BLOCKGRP_DISTANCE, the distance measured from the centroid of a block group to the center, performs almost as well as STREET_DISTANCE and HOUSING_DISTANCE. This suggests that BLOCKGRP_DISTANCE is a much better proxy than TRACT_DISTANCE for estimating the effect of accessibility on housing price. As we have pointed out previously, in an urban area, the size of a block group is generally in the neighborhood of 25 percent of a track. Thus, BLOCKGRP_DISTANCE naturally should be a better proxy than TRACT_DISTANCE to represent each individual housing unit's accessibility. Our micro market models use the data of housing sales that occurred only in Boston, which as a highly populated metropolitan hub is representative of many tracts and block groups that are much smaller than those in the cities and towns located further away from the center. Thus, the distortion of the housing accessibility represented by either TRACT_DISTANCE or BLOCKGRP_DISTANCE would seem to increase as a house is located further away from the center. Following this line of reasoning, it seems predictable that BLOCKGRP_DISTANCE would be an even better regressor than TRACT_DISTANCE if the market area selected for a hedonic study of housing price were to be extended to include suburban areas.

Therefore, given GIS technologies and provided that the necessary data are available, we should make efforts to measure the accessibility as the shortest street distance. If one prefers to save a few hours of computer time and does not want to execute the computer pathfinding procedure, the substitution of STREET_DISTANCE with HOUSING_DISSTANCE seems to suffice. However, unlike the other two approximations, both of the housing location-oriented measurements need to deal with the preparation of location data for address matching and the construction of a GIS-based street network model. This should not pose many technical difficulties and budgetary concerns as the GIS-based map data becomes more standardized and more widely available, and GIS software becomes more user-friendly and economical. The two more accurate estimates do, however, require a higher level of program know-how and greater computing power to maintain and manipulate a large amount of digital data. For the other two measurements, BLOCKGRP_DISTANCE and TRACT_DISTANCE, a simple, low budget, and friendly desk-top computer and low-cost mapping software are sufficient to accomplish the work. It appears that one should at least try to estimate the accessibility at the block group aggregation level. Using the distance aggregated at the tract level to represent the accessibility of each housing sample appears to be both erroneous and misleading.

As seen from the above discussion, with the aid of GIS the improved measurements of accessibility lead to much more significant estimates of its effects on housing price. In addition, the less aggregate data on neighborhood socioeconomic factors seem to provide new insights into the effects of the socioeconomic variables. In both of these exercises of hypothesis testing, GIS technologies have served as an effective tool for data manipulation and construction of new input of spatial variables.

The comparison between the tract and block group models does not seem to suggest a conclusion that the block group aggregation is better than tract or *vice versa*. The higher R-squares for the tract models and the better t-statistics for most coefficients in the block group models indicate that a better model might be one that incorporates variables aggregated at multiple spatial aggregation levels. There are many socioeconomic variables that have been theorized to be important factors in affecting housing price. However, there is no reason to believe that their impacts on housing price should be perceived at an identical spatial aggregation level. In fact, there are more reasons to consider that different socioeconomic variables are likely to have different spatial effects. For example, for analyzing the effect of school quality, the aggregation at a town level is probably most appropriate and adequate. For summarizing the effect of appearance and age of overall housing stock, it is probably better to go down to a more localized neighborhood level. People are likely to be more concerned about the housing quality in a few nearby blocks of residence than those that are a dozen blocks away. For analyzing the effect of crime incidence on housing value, a town level's summary of crime rate per thousand households is probably less meaningful to potential buyers than the information on the exact locations and frequency of crime. In addition, it is likely that the effects of various socioeconomic factors change from town to town.

Therefore, it appears that the choice of data aggregation level for hedonic price modeling should be determined empirically. In the past research on housing hedonic studies, it took so much time to assign to each observation a set of socioeconomic factors measured at one predetermined spatial aggregation level, few researchers ended up testing the impacts of different spatial aggregations on housing price. The difficulty of changing spatial aggregation levels can be resolved to a great extent by making use of GIS technologies. We have already shown that socioeconomic variables aggregated at both the tract and block group levels can be easily derived with the aid of GIS. We have also shown that the accessibility can be measured at a more localized level in three different ways. With these preparations of data at multiple spatial aggregation units, we explore and find a model that seems more superior than all the six models in Table 6.7, each of which adopts a single census data aggregation level. This model is presented in Table 6.8 and labelled as "Model VII." It is also labelled as "the Single-market Model" for convenience of comparison later in Chapter 7.

A higher R-square is achieved for Model VII, which explains nearly 52 percent price variation in the housing sales cases. All coefficients are statistically significant at 1 percent level and have signs as expected. Three of the five socioeconomic variables aggregated at the tract level are MEDIAN_VALUE, HIGH_EDUCATION%, and NONWHITE%. They are tested to be more power explanatory variables than their counterparts that are aggregated at the block group level. The opposite is true for the two other socioeconomic variables, LOW_INCOME% and MEDIAN_AFE. The logarithm variable of street distance is also an important factor in explaining the price variations. In the next chapter,

Table 6.8: Hedonic Estimates with Variables Measured at Different Spatial Aggregation Levels

(Dependent Variable = Housing Price; N = 1098)

Attributes	The Single-market Model (Model VII)		
	Estimates	Std. Estimates	t-stat
Constant	68,411.00		2.71
BATHROOM#	12,513.00	0.09	3.08
SIZE_LIVING	25.15	0.24	7.91
MEDIAN_VALUE (t)	0.88	0.46	12.99
HIGH_EDUCATION% (t)	698.70	0.09	2.56
NONWHITE% (t)	-241.36	-0.08	-2.98
LOW_INCOME% (b)	-1,036.02	-0.11	-4.58
MEDIAN_AGE (b)	-1,954.73	-0.13	-5.88
Ln(STREET_DISTANCE)	-20,165.00	-0.10	-3.65
Adjusted R-sq	0.5159		
F-statistics	147.16		

Note: (b) = data aggregated at the block group level; (t) = data aggregated at the tract level.
Sources: a spatially integreted data set based on 1990 Census, 1992 TIGER/Line File, and 1989-91 Housing Sale Data from County Data Corporation, VT.

we continue to explore the possibilities of using GIS to help improve this housing hedonic price model.

Next, we turn to analysis of error residuals from the estimated hedonic price models. We demonstrate below that GIS can also serve as a very useful tool for error analysis, verification of predictions, and an effective aid to model application.

6.2.5 Uses of error residual maps

In the analysis of the macro housing market in the previous section, we use GIS to display the error residuals for the regional regression models of median housing values (Figure

6.3). We illustrate that maps of residuals are useful for interpreting the outputs of the models, understanding how models behave spatially, and identifying and comparing spatial patterns of residuals. In this subsection, we further explore the usefulness of error residual maps. We first use several maps of residuals to analyze two representative hedonic price models: Models I and VI in Table 6.7, respectively the "traditional" model and the "best" model at the block group level. Then, we discuss and illustrate a number of possible extensions for the use of residual maps.

On the residual maps of the macro models depicted in Figure 6.3, each basic spatial unit *(i.e.*, each tract and block group), is shaded based on the size of their corresponding error residuals. One weakness associated with this type of mapping scheme is that a large error that is linked with a spatial unit of large size could be visually overexaggerated. One possible correction is to use a circular dot to represent the size of an error and plot the dot on the centroid of each basic spatial unit. In the following analysis, we explore two ways of displaying the error residuals for the hedonic price models. One way is to display the residuals by grouping them according to their magnitude (Figures 6.7 and 6.8). The other is to show each of the residuals using a circular dot whose size is proportional to the magnitude of error (Figure 6.9).

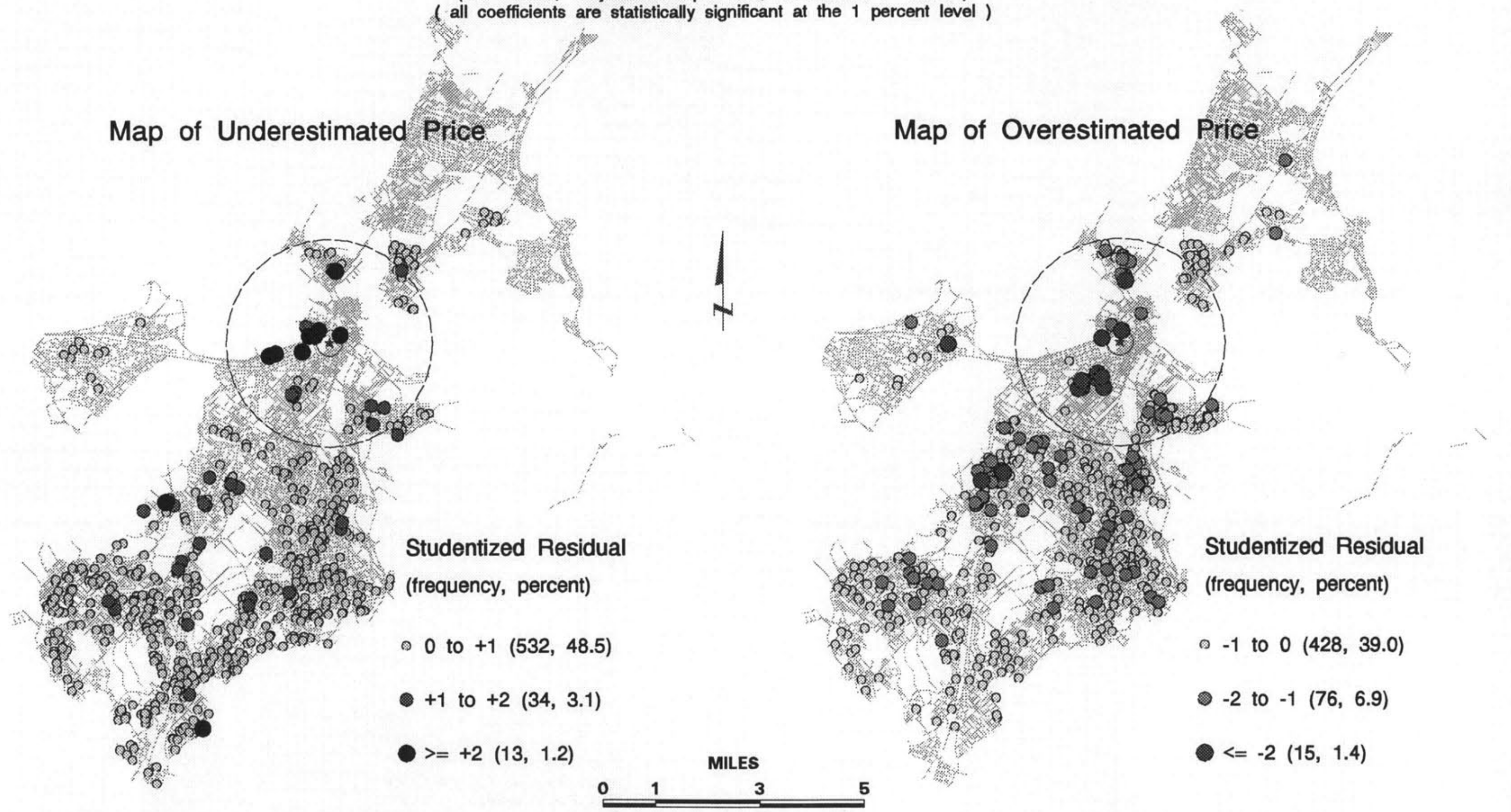
First, Figure 6.7 shows the error residuals for Model VI, using two maps: positive studentized residuals (*actual - predicted*) on the left and negative on the right. The positive and negative are displayed separately because otherwise too many dots would be overlapping each other due to the large number of housing sales used. At least two distribution patterns are noticeable from the maps. First, the left map shows that outside the CBD ring, large errors spread over the market area in a fairly random fashion. In contrast, the right map shows at least one significant cluster of large errors in the neighborhood of Jamaica Plain, to the southwest of the center. This provides a sufficient

Figure 6.7: Studentized Residuals from the Housing Hedonic Price Model (VI)

(Data Aggregated at the Block Group Level, The City of Boston)

MODEL VI: PRICE = 161300 + 13800 BATHROOM# + 25 SIZE_LIVING - 2000 MEDIAN_AGE + 0.5 MEDIAN_VALUE
- 1200 LOW_INCOME% + 1100 HIGH_EDUCATION% - 300 NONWHITE% - 38900 Ln(SHT_DISTANCE)

(N = 1098; Adjusted R_sq = .433; F-statistics = 105.51)
(all coefficients are statistically significant at the 1 percent level)



warning for caution and calls for extra effort to examine carefully the sale cases represented in this cluster. There might be some common factors affecting the prices of these houses that have not been captured by the model and are worthy of further investigation. Second, only a relatively small number of the cases are located in the areas that are within 2 to 3 miles of Downtown, as indicated by the density of dots that falls into the ring with a 2-mile radius on the maps. Yet, there are a disproportionally large number of big dots of error associated with these areas. This reminds users that the model is likely to be unsuitable for accurate price prediction of a house in the neighborhoods proximate to Downtown. This makes good sense because the sales data include only owner-occupied housing sales and most of them tend to be located in the areas that are more than 3 miles away from the center.

Figure 6.8 displays the error residuals for Model I. In the case of positive residuals, the spatial distribution patterns of Models I and VI are very similar. However, Model I has more cases with large negative errors (<=-2). As a result, it seems that a greater cluster of large errors is identifiable in the CBD ring. We choose to show only maps of residuals for Models I and VI because in terms of frequency and spatial distributions of error residuals, Models II and III are nearly identical to Model I, and Models IV and V are similar to Model VI. Thus, the analysis of error residuals shows that the tract models have a greater number of large errors than the block group models and the tract models seem to exhibit more serious problems of error concentration.

Figure 6.9 illustrates the second way of displaying error residuals. The maps show the residuals for Model VII, or the single-market model in Table 6.8. The center of each dot on the maps represents the location of a housing unit. The size of the dot is proportional to the size of error residuals. Again the positive and negative residuals are displayed separately. This approach of displaying errors enables both model builders and users to

Figure 6.8: Studentized Residuals from the Housing Hedonic Price Model (I)

(Data Aggregated at the Tract Level, The City of Boston)

MODEL I: PRICE = 15150 + 14480 BATHROOM# + 23 SIZE_LIVING - 1330 MEDIAN_AGE + 1 MEDIAN_VALUE
- 980 LOW_INCOME% + 650 HIGH_EDUCATION% - 220 NONWHITE% - 9550 Ln(SHT_DISTANCE)

(N = 1098; Adjusted R_sq = .499; F-statistics = 138.04)
(all coefficients are statistically significant at the 1 percent level except the intercept and the distance variable)

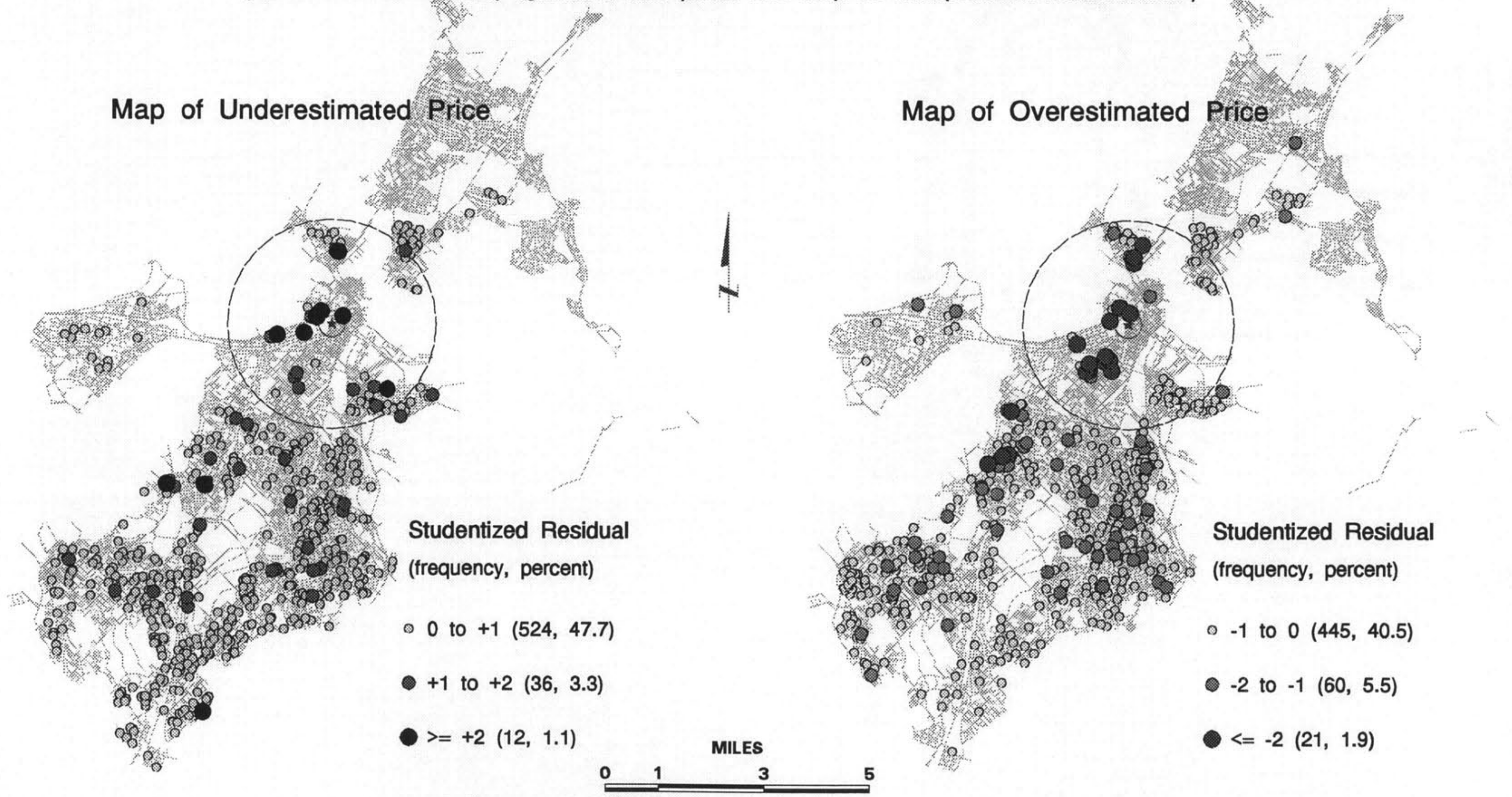
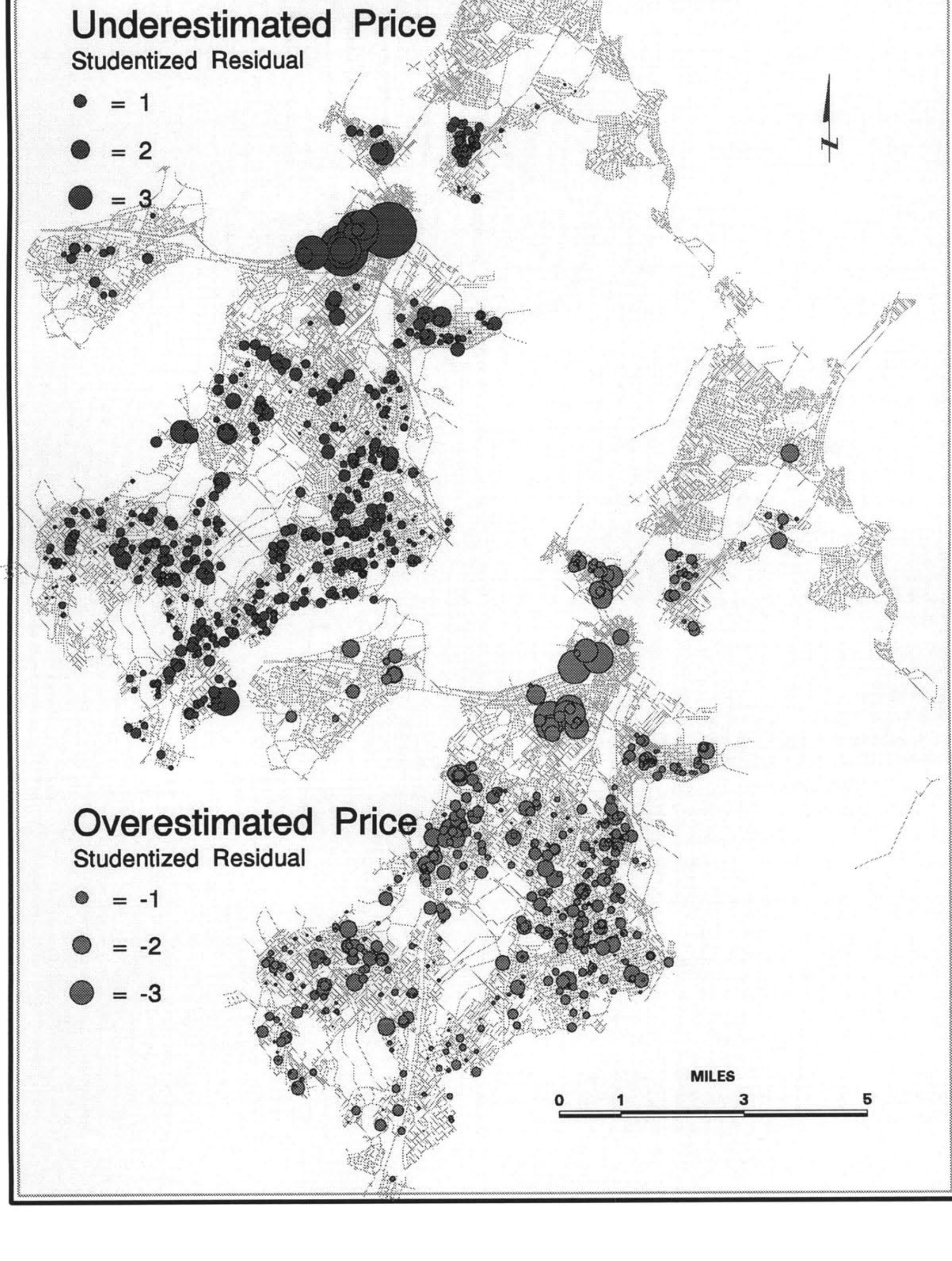


Figure 6.9: Spatial Analysis of Error Residuals

MODEL VII: PRICE = 68400 + 12500 BATHROOM# + 25 SIZE_LIVING - 2000 MEDIAN_AGE(b)
+ 0.9 MEDIAN_VALUE(t) - 1040 LOW_INCOME%(b) + 700 HIGH_EDUCATION%(t)
- 240 NONWHITE%(t) - 20170 Ln(STREET_DISTANCE)

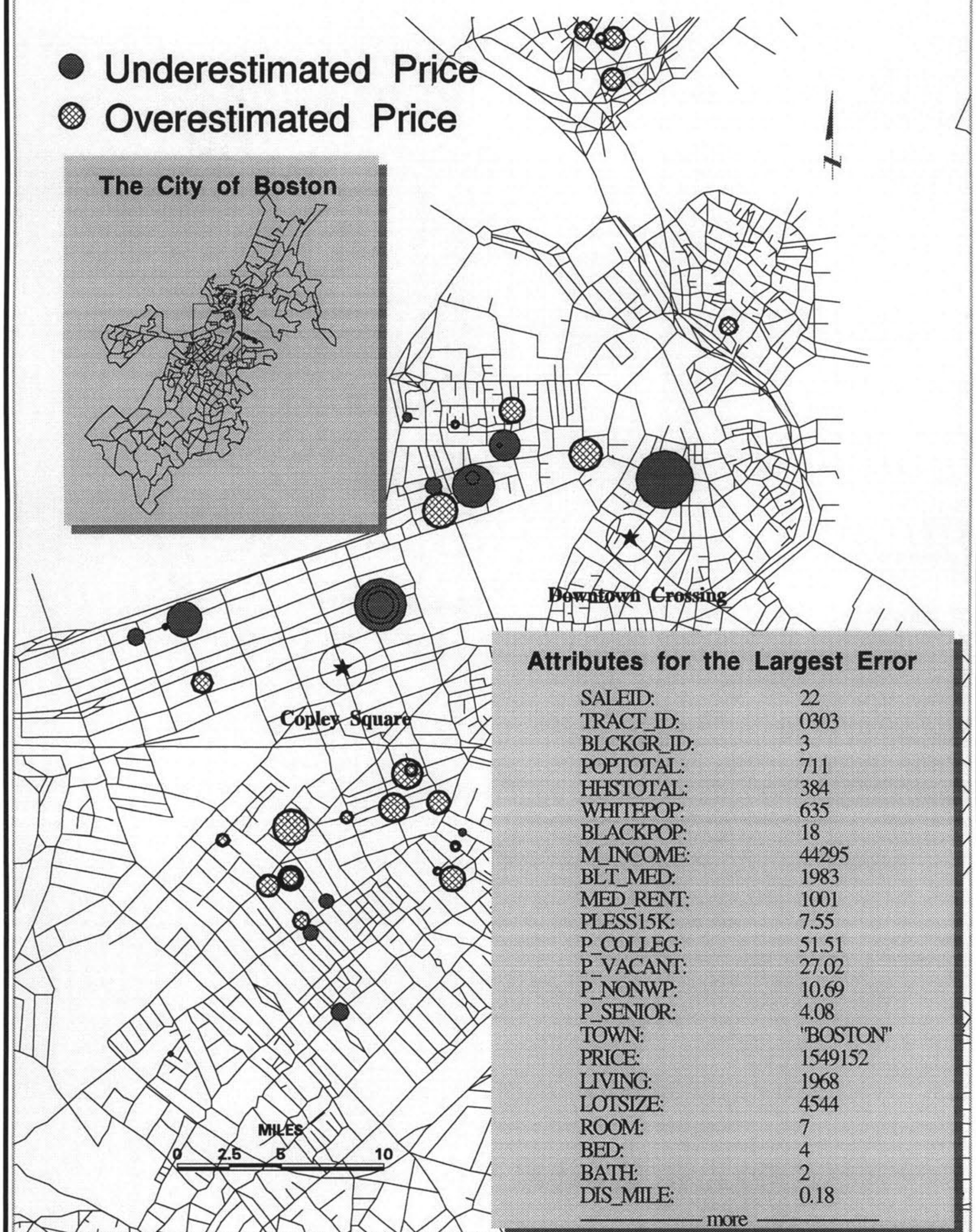
(Note: (b)= data aggregated at block group; (t)= data aggregated at tract.)
(N = 1098; Adjusted R_sq = .516; F-statistics = 147.16;)
(all coefficients are statistically significant at 1 percent level)



gain an intuitive understanding of the spatial distributions of samples and error residuals as well as the precise and relative magnitude of each residual. These maps readily reveal those alarmingly large errors and effectively expose various locations where the model fails to perform well.

Another important value of a GIS-based display of hedonic model residuals lies in its capability to enable model builders to access instantly, selectively, and interactively the spatially integrated database of the housing market. What is displayed in Figure 6.9 may seem not much different from what a traditional paper map could offer. In reality, the GIS behind the figure provides more info than a paper map. What is offered by a GIS-based map is an on-line information system. Using GIS, we could easily query any spatial and nonspatial information associated with each dot on the maps. By simply pointing and clicking on a dot (or selecting multiple dots) with a computer mouse, we could reveal instantly its attribute information. For example, selecting the largest dot of error on the maps, we disclose a "hidden mine" that causes the model to tumble. It happens to be a single-family housing unit sold at one of the highest prices in the sales data. A partial list of attribute information for this selected sale are shown in Figure 6.10. Although its structural attributes and associated neighborhood quality variables would lead to a prediction of high price, the estimate based on the model is still far lower than the actual price. Of course, we do also do this in traditional non-graphic statistical analysis as well, by always looking at our cases of highest error and examining the values of the independent variables. However, a GIS-based residual map allows the examination of not only the extreme cases but also their surrounding cases and other layers of spatial information. On a computer screen, we zoom in to take a closer look at nearby cases and also the surrounding physical environment, as illustrated in Figure 6.10. We discover that the house is located within the financial district of the city. This location information

Figure 6.10: Spatial Distribution of Residuals for the Boston CBD Area



points to several possibilities for explaining the large residual associated with this sale. First, there are very few residential units in this district. Therefore, the census information for this area is not likely to be accurate due to limited census procedures. Second, a family housing unit in the prime location of Downtown and surrounded by a forest of skyscrapers is totally out of place. Perhaps the unit is a historical building. Its land value would probably account for most of its price. Third, it might be simply a sample error. This example illustrates that a GIS-based map offers a convenient and efficient way of querying spatial and nonspatial information to assist spatial analysis and decision making. Also this example shows that residual maps can be used for identifying patterns of "spatial outliers" in the sample.

Outliers are usually defined as the data points that lie more than some arbitrary distance from a typical value. We did not exclude the so-called "hidden mine" from the final sample for hedonic analysis because initial statistical analysis did not seem to indicate that this sale differed much from other housing units sold at high prices. Also, experts in econometrics (Pindyck and Rubinfeld 1991) have pointed out that outliers may represent important information about the relationship between several variables and should not be thrown out without further analysis. For the spatial analysis of housing hedonic price models, a GIS-based map of residuals seems to bring about a new perspective of understanding what an outlier could mean. An observation may possess features that are typical in terms of all the variables included in a model. But it may still be considered as an outlier because of its unique location and surrounding environmental amenities. A GIS-based map lends a convenient way of examining each observation right where it is located. For example, a sale observation may be "normal" in every sense captured by the variables included in the model and the model may yield an estimated price that is close to the actual. Nonetheless, this observation could still be an outlier if it is a lonely observation lying far away from any other observations in the sample. Knowing its location and its spatial relationships with other cases should comprise an important element in analyzing its effect on housing price.

This use of GIS-based maps to understand intuitively the spatial relationships between each observation and its error residual is also very helpful to professionals, such as assessors and real estate agents, who make use of housing hedonic price models in their practice to predict a potential value or price of a housing unit. In this process of price prediction, besides the mathematical components of the model, the location information about a house, the GIS-based maps displaying the spatial distribution of other sales cases, and the spatial distribution and magnitude of error residuals from the model could provide users with valuable information in determining their estimates of potential prices for the house. To illustrate, suppose that an assessor needs to assess a potential value of a house that has the following attributes: 2.5 bathrooms, 3,500 square feet of living space, and an address of 3 Devon Street, Dorchester, MA 02122. In order to use the GIS-based hedonic models, she first uses the address to identify its actual location in the market area and its associated information on neighborhood quality. Suppose that the first step determines that the house is located in a block group that is characterized by a median age of housing property of 25 years, a median housing value of $350,000, 10% low income families, 60% college-educated adult, and 20% nonwhite population. Then, she measures its accessibility to the employment center using the GIS-based street network model. Suppose that the shortest street distance to the center is found to be 3 miles. Feeding all the above numbers into Model VII, the single-market model in Table 6.8, she would derive that a potential price for this property is around $300,000.

Is this a good estimate based on the available information on how the model behaves spatially? The GIS-based residual map for the single-market model offers a possibility for gaining a spatial sense of how reliable this estimate might be. If the property is located in an area that is clustered with many of the model's estimates with high error residuals, or if the surrounding area has very few sales cases that have been used to construct the model, she would intuitively feel suspicious about the reliability of the estimate. On the other hand, if the area shows many cases with small error residuals, she could place a higher level of confidence on the estimate.

No assessors or real estate agents would be comfortable in making their assessment decision based solely on the results obtained from a hedonic price model. Obviously, in practice, they would also look into other assessment approaches and information. Information such as local and regional housing demand and supply conditions, financial market situation, and incidences of comparable housing sales would also need to be carefully examined in order to place an appropriate price tag on a property. Nonetheless, a GIS-based hedonic price model should provide a sound and important first step to a good assessment of housing value (or price). It is well known in the real estate industry that the most important information on a property is: location, location, and location. Thus, it seems that the most valuable contribution of a GIS-based approach towards housing price modeling is the addition of an accessible spatial dimension and the incorporation of enhanced location information.

6.2.6 Summary

The information on the location of a housing unit is crucial in estimating its potential market price. In the previous empirical hedonic research, the commonly available information of housing address has not been used to its full strength to assist in the estimation and prediction of housing price. This important spatial information has been cold-shouldered because of the lack of effective tools to explore its potential. Using GIS technologies, this research demonstrates that the spatial information can play an important role in constructing, interpreting, and using housing hedonic price models.

This section begins by describing how GIS can be used to combine different sources of data with common spatial components, display spatial distributions of cases by various categories, and measure accessibility variables to be incorporated in hedonic models. It then illustrates the possibility of modeling accessibility in a more realistic manner using GIS technologies and prepared four alternative measurements for the accessibility variable. Respectively, the four alternatives are GIS-based measurements of *Euclidean* distances by the centroids of a tract and block group, and by housing location, and of the shortest street distance. Each of these alternatives has been tested for its impact on housing prices. According to the empirical results of the first six housing hedonic price models formulated in this section, important differences exist between the models estimated using census data aggregated at the tract level and those at the block group level. All three tract models have slightly higher adjusted R-squares than the three block group models. However, most of the coefficients for individual explaining variables have much higher t-statistics in the block group models than in the tract models. The empirical results also show that the accessibility of a housing unit to an employment center measured as the shortest street distance is a much stronger explanatory variable in both the tract and the block group models than the accessibility measured using the traditional *Euclidean* distance. Based on the comparison between the tract the block group models, we argue and demonstrate and a better hedonic price model is achievable when multiple spatial aggregation levels are used to measure socioeconomic variables and accessibility variables. Finally, we explore two different ways of mapping model residuals and illustrate various uses of the maps in constructing and applying the housing hedonic price models.

Chapter 7

GIS-based Market Segmentations and Hedonic Price Models

Our primary goal in this chapter is to explore the use of GIS for segmenting the Boston housing market and estimating separate housing hedonic price models for each of the segmented submarkets. The exploration shows the advantage of using GIS to display and analyze the spatial patterns of market segmentation by several commonly used stratifiers, and the flexibility of using GIS for data aggregation and space delineation. It also tests the stability and consistency in the impacts of those explanatory variables selected for constructing the single-market hedonic price models in Chapter 6, especially the impacts of the accessibility variable measured as actual street distance.

We first describe a GIS-based spatial representation of market segmentation, exemplified by four commonly used stratifiers in the literature of housing hedonic analyses, and discuss the advantages of the GIS-based visualization of spatial patterns. Then, we construct housing hedonic price models for three sets of submarkets, segmented respectively by accessibility to the CBD, type of family housing, and race. These exercises demonstrate the value of GIS in assisting the hedonic analysis and its flexibility in spatial data integration and market segmentation for model testing.

7.1 Spatial Representation of Market Segmentations

The literature review in Chapter 2 shows that researchers of housing markets have employed many different approaches towards market segmentation. Among them, the most commonly used approaches include stratifying market areas by neighborhood quality indicators such as median housing value, median household income, race, and education levels. Traditionally, most of these indicators are measured at the tract level of aggregation. Another commonly used approach is grouping subareas, usually tracts, into an array of housing submarkets based on the accessibility of each area to an employment center. Still another commonly employed approach is to categorize a large market area by housing types[12], such as family housing versus condominiums. One missing important element in nearly all the existing empirical studies on market segmentation analysis for the purpose of estimating separate hedonic models is to represent spatially the results of segmentations. Frequently, in the traditional studies of segmentation, researchers were concerned more with the size of the housing sales sample that fell into each of the segmented submarkets than with the locations of housing sales and submarkets as well as their spatial interrelationships. Many researchers have indicated the importance and value of understanding spatial characteristics of housing markets (Straszheim 1975) for housing price modeling. But little has been done to make use of location information of housing data and to capture the spatial characteristics of the market areas due to the deficiency of tools for spatial representation.

Therefore, the first step of this study of market segmentation is using GIS to portray spatially the results of segmenting the Boston market into four submarkets using four stratifiers, respectively. The selected stratifiers include median housing value, median household income, percentage of college (or higher) educated adults, and percentage of nonwhite population. We choose block group as a basic spatial unit for market segmentation analysis and measure all these four variables at this spatial aggregation

12. Using housing type as a stratifier could also include housing tenure (owner-occupied versus rental), housing styles (modern versus classic), housing structure (brick versus concrete), to list only a few. In this study, housing type is used to refer to different types of family housing such as single-family, two-family, and three-family housing units.

level. This choice of a more disaggregated spatial unit than census tract should ensure the segmentation of more homogeneous housing submarkets as characterized by the stratifiers used. One of many possible results of segmentation by each of the four factors is represented spatially using GIS in Figure 7.1.

The figure shows that the Boston market area consists of 689 census block groups[13], 532 of which have reported census information for the four selected variables. The size of block groups ranges from 0.08 to 15 square miles. Four submarkets are identified in each of the four maps in the figure. The breakpoints for segmentation are determined in such a way that a roughly equal number of block groups fall into each of the four submarkets, in GIS jargon -- classification "by equal numbers." For example, the left map represents four housing value submarkets using median housing value aggregated at the block group level as a stratifier. The block groups with greater than and equal to $188,000 median housing value are shaded with the darkest color and others with lower values are shaded lighter. On this map, the spatial pattern for each submarket seems to place some contiguous block groups with similar median housing values together to form several clusters at varied distances away from the center of the city. Each map may or may not show readily identifiable spatial patterns of submarkets. However, the graphic representation of market areas allows researchers or users to see and compare visually and intuitively the results of market segmentation.

Some numerical results of the four alternatives of segmentation are summarized in Table 7.1. The table shows a number of spatial characteristics for each submarket, such as the physical size of each submarket, and population and housing densities. The Boston market depicted in the maps is characterized by a total land area of approximately 311

13. Most block groups with missing census data were in 1990 vacant land, parks, and primarily non-residential land use areas where very few numbers of households resided.

Figure 7.1: Four Alternatives of Spatial Segmentations for the Boston Housing Market

(Data aggregated at the block group level; Total number of block groups = 708; Nummber of block groups with data = 532)
(Four submarkets are represented in each map. Each submarket contains 25% of the total number of block groups with data.

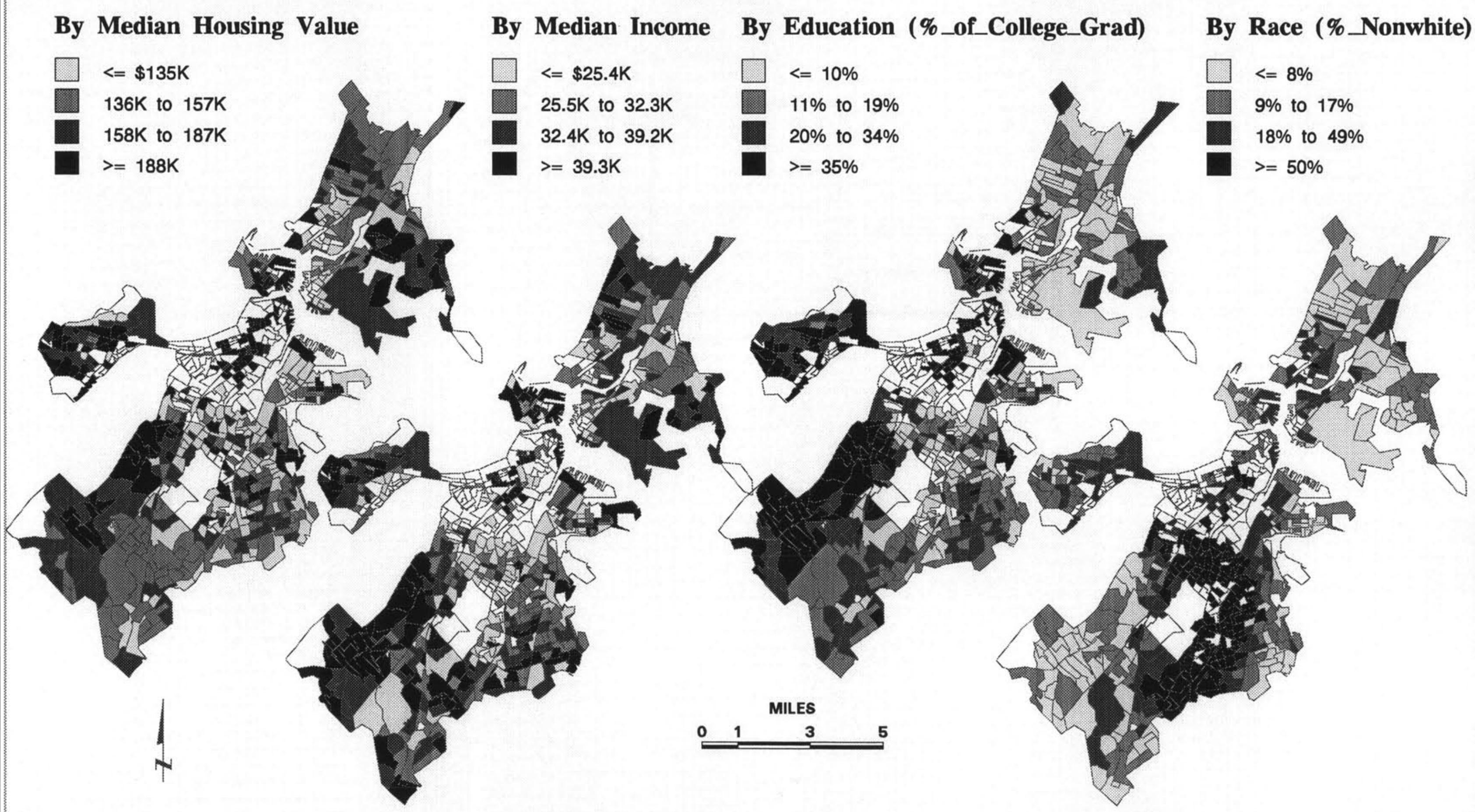


Table 7.1: Summary Statistics for the Four Alternatives of Market Segmentation

Stratifiers for Housing Market Segmentation	Breakpoints	Number of Block Groups	Land Area (sq.mile)	% of Land Area by Submarket	Population	Number of Housing Units	Population Density (#/per sqm)	Housing Density (#/per sqm)
Total		532	311		525,235	225,837	1,687	725
Submarkets by Median Housing Value	<= $13.5K	134	61	20	123,840	50,702	2,030	831
	13.6K to 15.7K	133	94	30	134,356	55,080	1,431	587
	15.8K to 18.7K	133	66	21	129,315	64,235	1,962	975
	>= 18.8K	132	91	29	137,724	55,820	1,522	617
Submarkets by Median Income	<= $25.4K	133	65	21	140,130	62,792	2,153	965
	25.5K to 32.3K	134	96	31	134,147	55,399	1,403	580
	32.4K to 39.2K	132	88	28	120,147	52,532	1,368	598
	>= 39.3K	133	63	20	130,811	55,114	2,083	878
Submarkets by % of Adult with College Education	<= 10%	133	85	27	130,490	51,829	1,533	609
	11% to 19%	133	74	24	130,617	54,652	1,756	735
	20% to 34%	133	75	24	131,158	65,858	1,739	873
	>= 35%	133	76	25	132,970	53,498	1,741	700
Submarket by % of Nonwhite Population	<= 8%	133	108	35	110,275	46,591	1,023	432
	9% to 17%	133	66	21	140,489	64,234	2,125	972
	18% to 49%	133	50	16	140,041	51,210	2,809	1,027
	>= 50%	133	88	28	134,430	63,802	1,535	729

Source: Calculated based on 1990 Census and 1992 TIGER/Line files.

square miles, 525 thousands of population, and 225 thousands of housing units. The four submarkets defined by education and by its four breakpoints are by and large comparable in terms of the spatial attributes measured in the table. In contrast, spatial characteristics for some of the submarkets defined by the other three alternatives and their respective breakpoints differ distinctively from each other. For example, the submarket with the smallest percentage of nonwhite population has the largest land area and the lowest population and housing density. It is not difficult to imagine that by manipulating choices of the breakpoints, very different spatial results of market segmentation can be obtained.

Therefore, the spatial characteristics of location patterns, size of market area, and densities for each of the submarkets provide valuable information for understanding and interpreting hedonic estimates and model behaviors. We will illustrate the important use of this spatial information in the following sections when we estimate separate hedonic price models for a set of differently defined housing submarkets.

The above presentation shows two important advantages of using a GIS-based segmentation approach. First, it provides an effective tool for spatial representation and visualization of market segmentation. Second, it can be used to capture and generate many areally related information such as size of market areas and density attributes. The capture and generation of simple spatial information can be performed easily by GIS with many segmentation schemes.

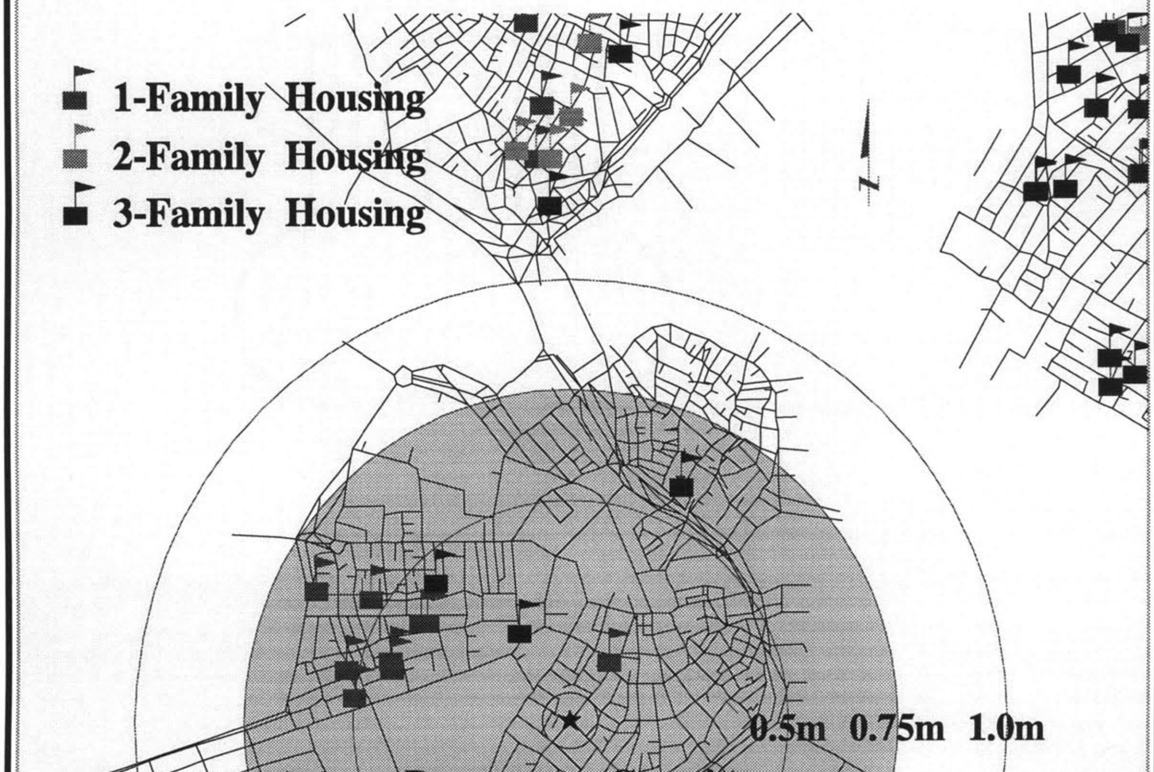
A third advantage of using GIS in the analysis of market segmentation goes beyond the paper maps shown in Figure 7.1. The figure shows only one of the nearly unlimited possibilities of segmentation. What the figure really represents is a housing information system that permits an instant display on a computer screen of segmentation results based on any stratifiers selected from the integrated housing spatial database, on any number of submarkets, and any breakpoints used for classification. This feature of GIS-based market segmentation offers a great deal of flexibility of exploration in selecting appropriate segmentation schemes that are coherent to geographic characteristics of local housing markets. The users can see and analyze various possibilities of segmentation in spatial context and explore final schemes of segmentation that are relevant to policy purposes, or based on either a ground of reasonableness or possibly some optimization criteria. Although this research has not developed such a GIS-based market representation system with a high degree of flexibility, a study on visualization of spatial patterns of land use and demographic features in the Washington D.C. area using the GIS mapping capability (Ferreira and Wiggins 1990) has already demonstrated this valuable flexibility of GIS-based spatial representation. In next section, we further illustrate the flexibility of using GIS to assist housing market segmentation.

7.2 Segmentation by Accessibility

The analysis of spatial distribution of error residuals in Section 6.2.5 of the previous chapter reveals that most of the large residuals are associated with the sales of housing units located in the neighborhoods proximate to the CBD. This is an good indication of the necessity to model prices for the housing units in the downtown area differently.

One way of modeling houses in the central location differently is to use a dummy variable to capture location impact on their prices. To create location dummy variables, traditionally a large market is first segmented into a series of concentric rings based on their accessibility to a designated center. Then, each ring is assigned a dummy variable. Following this traditional approach, we create several concentric rings of housing submarkets for the Boston CBD area based on their accessibility to downtown. The CBD submarkets are depicted in Figure 7.2. After testing a number of hedonic price models that are different from each other in terms of the number of rings and each ring's distance to

Figure 7.2: Spatial Distribution of Housing Sales and Delineation of the CBD Submarkets I



the center, we select a simple hedonic price model with a dummy variable representing one CBD submarket for further discussion. In this case, the CBD submarket is defined as the circular area within 0.75 mile distance[14] from Downtown Crossing, a designated representation of the employment center. The introduction of this CBD submarket leads to an improved hedonic price model for explaining the housing price variations in the sampled sales data. The model is presented in Table 7.2, labelled as "Model for a One-center CBD Submarket." For the convenience of comparison, all hedonic price models estimated in the chapter for various housing submarkets include an identical set of structural and socioeconomic variables as used in the single-market hedonic model presented in Table 6.8. Also, this single-market model is reproduced in Table 7.2.

This hedonic model with a one-center CBD submarket shows a significant improvement over the single-market model. The new model shows a considerably higher adjusted R-square of 0.577, which represents an increase of slightly over 6 percentage points from that of the one-market model. The t-statistics for the coefficients of all structural and socioeconomic variables remain significant at 1 percent level and signs of these coefficients are as expected. The sign for the dummy variable of "Submarket 1" is positive and statistically significant. It is the most important variable in the model in explaining the price variations. A house located in the CBD submarket gains a premium of nearly $308,000. However, the accessibility variable of street distance in logarithm is no longer a significant factor. According to this model, there exists a plateau of housing price in the CBD submarket that is sharply higher than those outside the submarket. The model, however, does not support the theory of a housing price gradient as a function of distance away from the center.

14. We have tested many models using various combinations of radii for segmenting the Boston market into a series of concentric rings. Using a radius of 0.75 mile for defining the central submarket turns out to produce a model with the highest R-square and significant coefficients for nearly all independent variables.

Table 7.2: Comparison of Hedonic Estimates with Different Delineations of the CBD Housing Submarkets

(Dependent Variable = Housing Price; N = 1098)

Attributes	Model for One-center CBD Submarket			Model for Two-center CBD Submarkets			The Single-market Model		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	65,990.00		2.80	54,288.00		2.33	68,411.00		2.71
BATHROOM#	14,316.00	0.11	3.76	14,792.00	0.11	3.94	12,513.00	0.09	3.08
SIZE_LIVING	24.61	0.24	8.28	23.09	0.22	7.85	25.15	0.24	7.91
MEDIAN_VALUE (t)	0.56	0.29	8.11	0.78	0.41	9.99	0.88	0.46	12.99
HIGH_EDUCATION% (t)	774.75	0.10	3.04	663.34	0.09	2.63	698.70	0.09	2.56
NONWHITE% (t)	-225.86	-0.07	-2.99	-190.66	-0.06	-2.55	-241.36	-0.08	-2.98
LOW_INCOME% (b)	-980.02	-0.11	-4.63	-1,058.51	-0.12	-5.06	-1,036.02	-0.11	-4.58
MEDIAN_AGE (b)	-1,549.13	-0.10	-4.96	-1,617.72	-0.11	-5.25	-1,954.73	-0.13	-5.88
Ln(STREET_DISTANCE)	-2,417.38	-0.01	-0.45	-11,369.00	-0.05	-2.07	-20,165.00	-0.10	-3.65
Submarket 1	307,721.00	0.32	12.59	226,727.00	0.24	8.12			
Submarket 2				-129,143.00	-0.15	-5.73			
Adjusted R-sq	0.5771			0.5891			0.5159		
F-statistics	167.35			158.30			147.16		

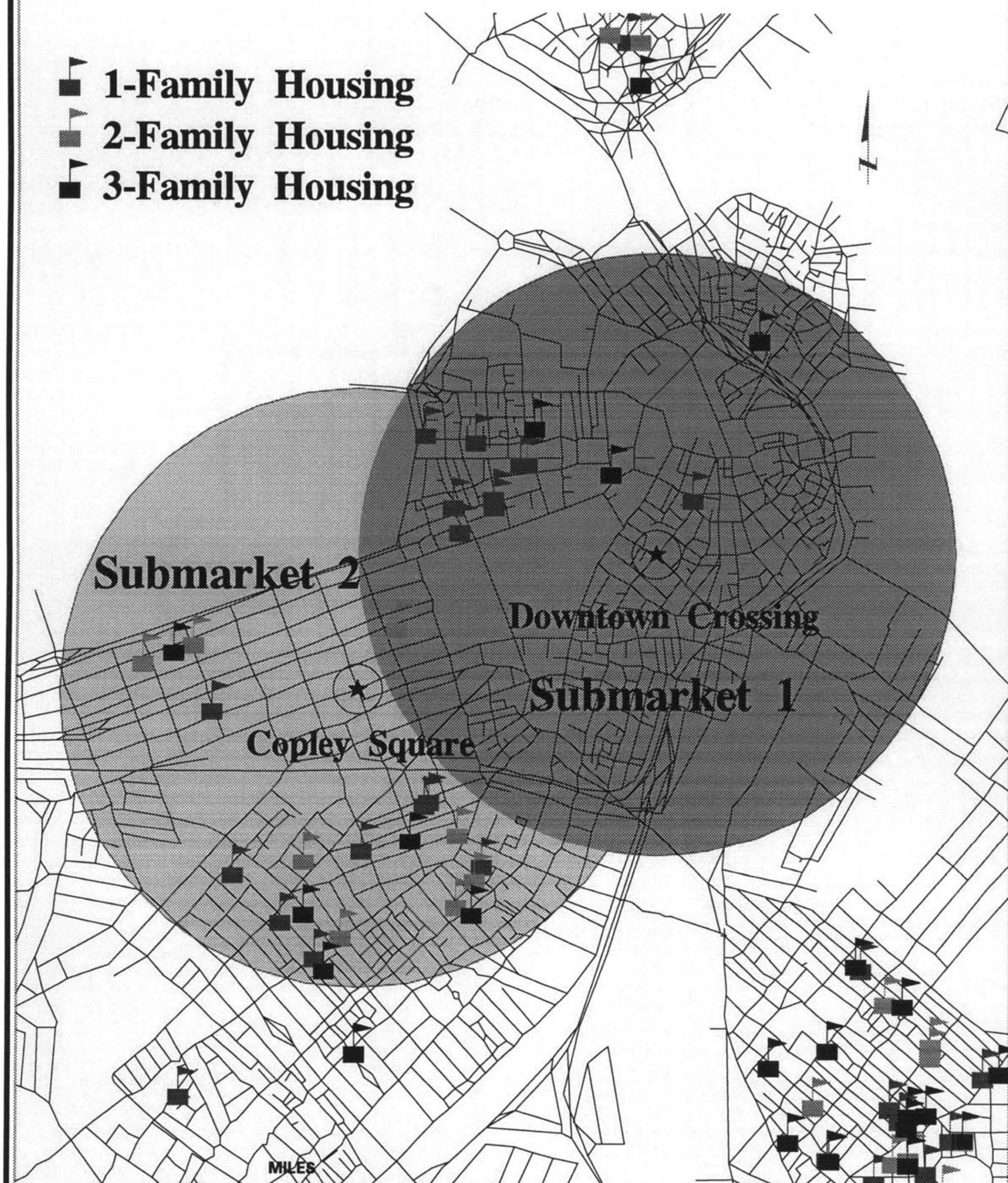
Note: (b) = data aggregated at the block group level; (t) = data aggregated at the tract level.

Sources: a spatially integreted data set based on 1990 Census, 1992 TIGER/Line File, and 1989-91 Housing Sale Data from County Data Corporation, VT.

This model is unsatisfactory because there are still some large negative residuals associated with many of the sales cases to the southwest of the center and immediately outside the one-center submarket, as depicted in Figures 6.9 and 7.2. We have tried using an additional concentric submarket to model these sales. However, this new ring-shaped submarket would also include those sales occurring in the areas to the south of the center. These areas are a group of distinctively different neighborhoods, where most of the sales can be fairly accurately estimated by the single-market model. This shows that the traditional approach of using concentric rings to model accessibility is rather arbitrary and rigid. The choice of this approach is largely due to the convenience of measuring accessibility as a straight line distance and of regrouping data according to the submarkets segmented as a series of concentric rings. To further improve this hedonic price model, we construct a hedonic model for a two-center CBD housing submarket. The selection of two centers may seem as arbitrary and idiosyncratic as the traditional one-center approach. Nonetheless, this exercise helps to illustrate the flexibility of using GIS for model exploration.

In fact, our choice of two-center submarkets are not entirely arbitrary. Arguably, the Back Bay area of Boston may be considered a secondary employment center, next to the primary center of financial institutions and commercial complexes near Downtown Crossing. At the least, the skyline of the Back Bay manifests an employment concentration. Thus, we assume Copley Square as the central point for this secondary employment center for our price modeling purposes. The exact delineation of the two-center CBD housing submarket is portrayed in Figure 7.3. Two dummy variables are used to represent the two submarkets, respectively "Submarket 1" and "Submarket 2." The hedonic estimates for a model of the two-center CBD submarket are presented in Table 7.2.

Figure 7.3: Spatial Distribution of Housing Sales and Delineation of the CBD Submarkets II



This model has a slightly higher adjusted R-square than that of the one-center model, a 1.2 percentage point improvement. More noticeable is the fact that t-statistics are significant for all explanatory variables included. All signs are as anticipated. The coefficient for the "Submarket 2" dummy variable is negative, indicating a deduction of nearly $130,000 in price if a house is located in this submarket, with all other things equal. This is unlikely to be true for those neighborhoods along the Charles River, which are among the most amiable residential areas in the center of the city. A possible reason for this negative coefficient is that most of the sales cases happen to be located in the south of "Submarket 2" and in the neighborhoods that are bordering Roxbury, an inner city neighborhood with a high concentration of nonwhite population and crime incidences. This proximity might have affected the housing sale prices adversely and yet are not captured in the model. In terms of the impact on price, the dummy variable "Submarket 1" becomes next in importance to the variable "MEDIAN_VALUE" of housing. All other coefficients in the two-center model are comparable to those in the one-center model except the variable of accessibility. In the two-center model, the impact of accessibility is significant statistically and has a negative sign, which again provides evidence supporting Alonso's theory of residential location.

The combined effects of the dummy variable of "Submarket 1" and the accessibility variable indicate that the price gradients of housing are characterized by a sharp jump for the housing units that fall in the submarket and by a slowly declining slope in the areas outside most of this submarket and away from the primary center. For those areas that fall into "Submarket 2," the pattern of price gradient is slightly more complex. It is further characterized by a price plunge between the price plateau in "Submarket 1" and the price slope outside the two submarkets.

Comparing the model for the two-center CBD submarket with the single-market model, we find that all hedonic estimates and statistics are fairly consistent in both despite the introduction of the two dummy variables. This consistency seems to provide further evidence supporting the importance of measuring the accessibility variable in a more realistic and localized manner.

This exploration of housing submarket segmentation for hedonic modeling illustrates the possibility of using GIS to model the geography of a local market and capture spatial realities in more realistic manner. In fact, the submarket does not have to be circular in shape. They can be in any irregular form determined to be coherent to local geographic actualities. The buffering and overlapping capabilities of GIS allow reaggregating of spatial data based on selected segmentation schemes for model estimation.

It is important to point out that the process of using GIS to carry out the tasks of segmenting space and re-grouping or reaggregating spatial data is far from easy and mature. It still requires complex command sequences for specifying analysis and display options. Also, it requires importing and exporting spatial data between GIS and statistical software for reestimating hedonic price models. Ideally, a mature GIS tool would allow using a computer mouse to mark out a submarket area and immediately a built-in statistical module would reestimate a model on the fly. The realization of such a mature GIS tool would indeed make the exploration of market segmentation based on spatial and geographic realities a flexible, practical, friendly, and pleasant process. It would also make the testing of the robustness of a model much more convenient and speedy.

In the next section, we continue to examine whether the CBD submarkets defined by accessibility still make sense when we further stratify the Boston market into three different family housing submarkets.

7.3 Segmentation by Family Housing Type

This analysis of market segmentation by family housing type and hedonic price models is organized as follows. First, we again use GIS to display the results of segmentation and set up a spatial context for modeling price by housing type *(i.e.*, single-family, two-family, and three-family housing units). Second, we examine whether there exist distinctive housing submarkets for the three different types of housing by comparing hedonic estimates for each of the submartkets. Third, we continue to evaluate the impact on housing price of the accessibility variable measured as actual street distance. We show that the understanding of spatial distribution of the sales cases is crucial for interpreting the behaviors of the accessibility variables.

In order to represent spatially the housing submarkets by type, we need to match each of the housing sales to the GIS-based street network of the city using the location information of each housing unit. We have completed this process and displayed the results of the address matching by type in Figure 6.4. Using a full page of space for displaying each submarket, we again represent the spatial distribution of the housing sales by type in Figures 7.4, 7.5, and 7.6, respectively for single-family, two-family, and three-family submarkets. Unlike Figure 6.4, these three displays include only sales that have been used to estimate the hedonic price models. Figure 7.4 shows clearly that a majority of the cases for single-family housing units concentrate in the areas more than 5 miles away from the CBD. The single-family housing sales in most of the areas close to the CBD are rather sparse. For the two-family housing sales, as shown in Figure 7.5, the distribution is fairly even across the landscape of the city. In contrast to the spatial distribution of single-family housing sales, almost all sales of three-family housing units, depicted in Figure 7.6, are within 5 miles from the CBD, with two neighborhoods, East Boston in the northeast and North Dorchester in the south, taking up the lion's share. In each of these three figures,

Figure 7.4: Spatial Distribution of Single-Family Housing Sales Cases

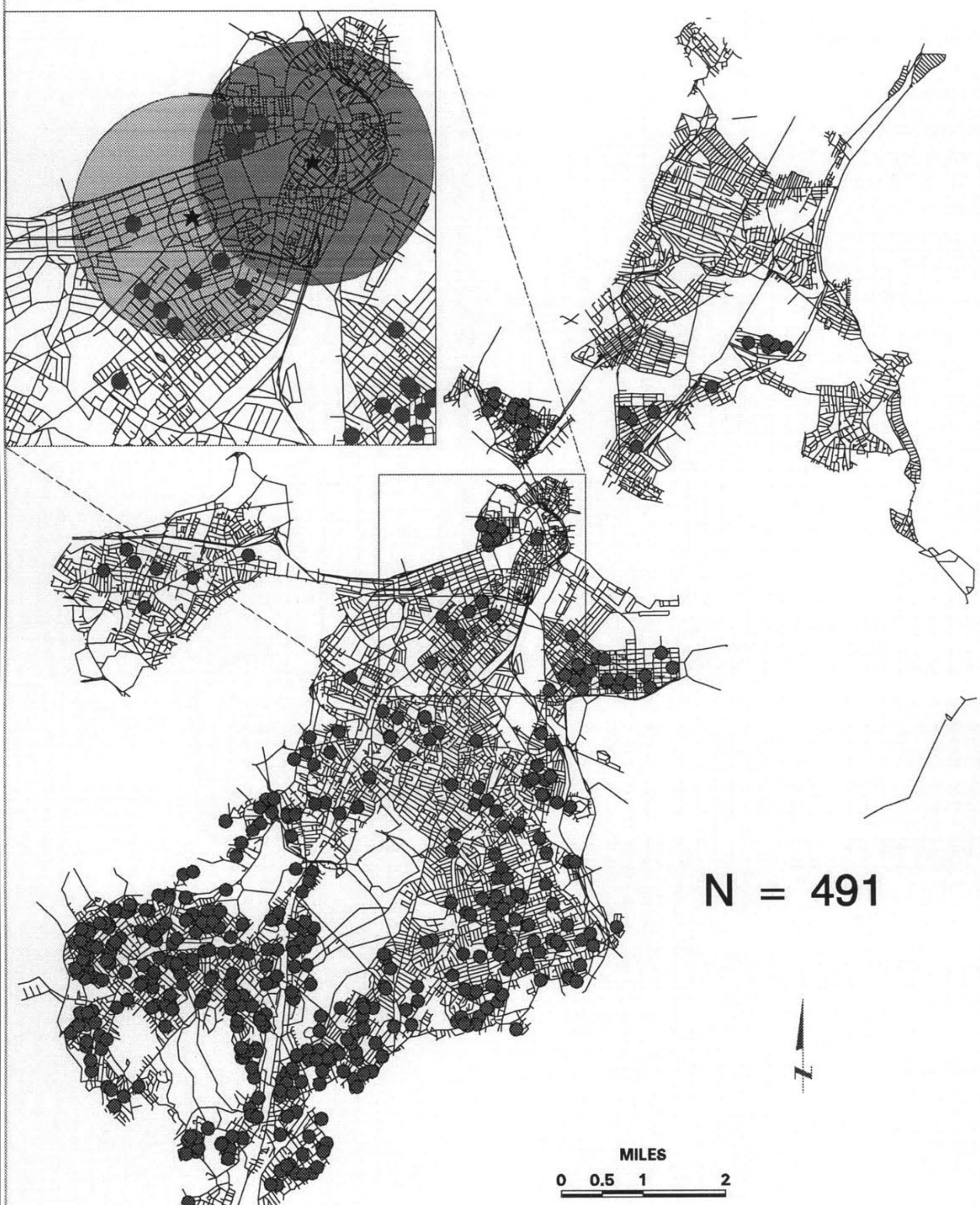


Figure 7.5: Spatial Distribution of Two-Family Housing Sales Cases

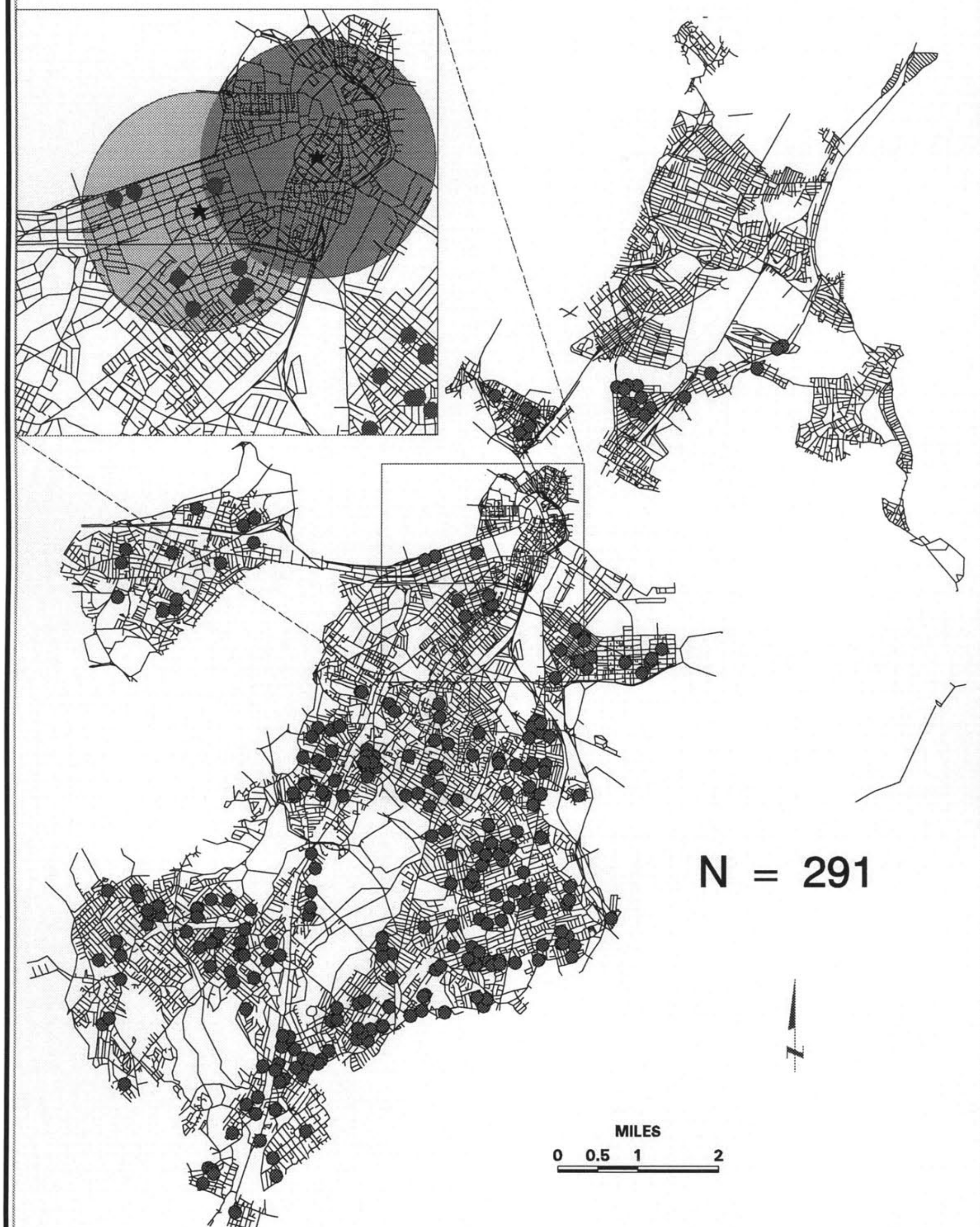
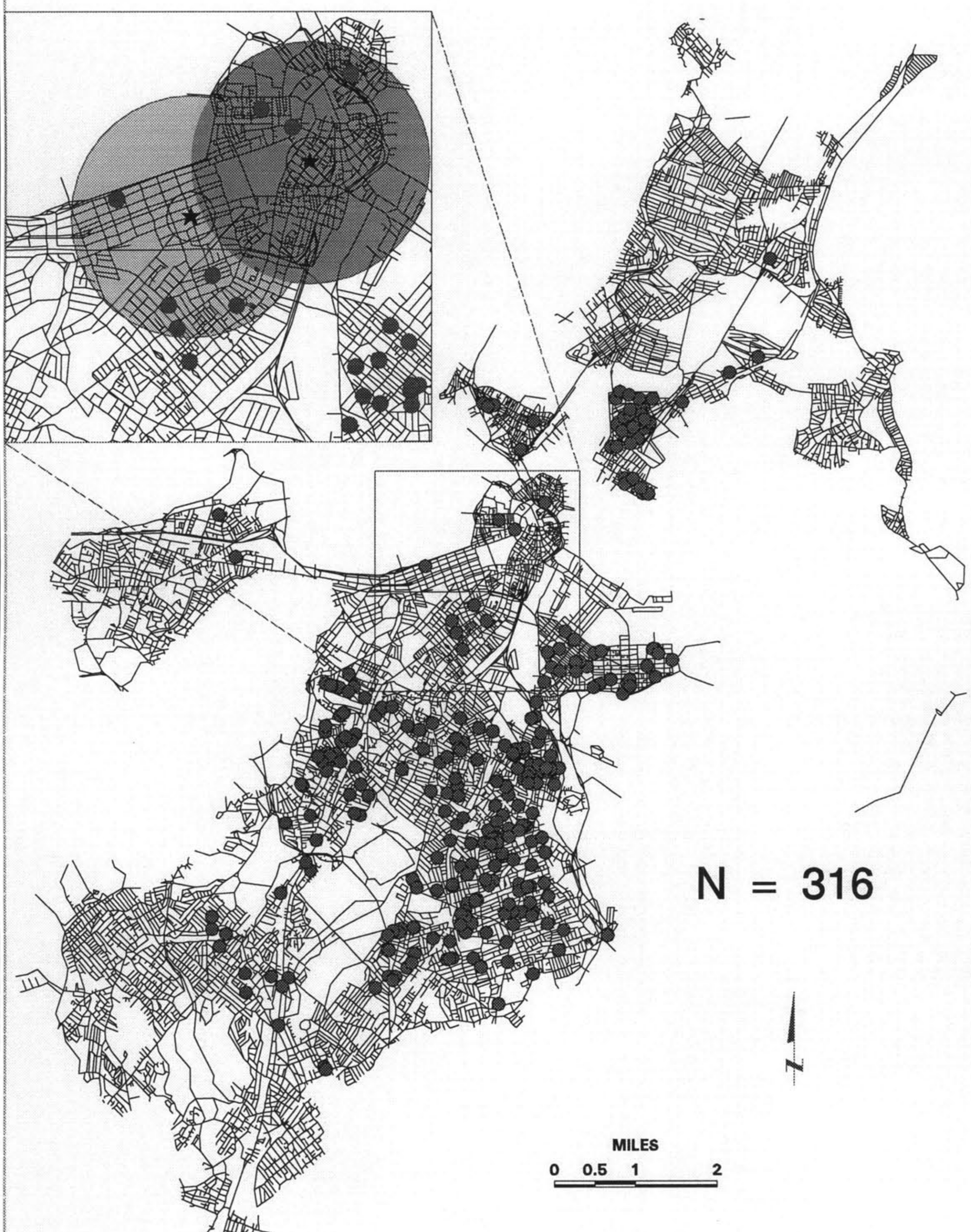


Figure 7.6: Spatial Distribution of Three-Family Housing Sales Cases



a magnified view showing the delineation of CBD submarkets is placed in a map inset. One caution for reading the two-family housing sales in the CBD “Submarket 1” is in order. The single dot in this submarket actually represents three housing sales that share an identical street address but different unit numbers.

Two hedonic price models are constructed for each of the three-family housing submarkets: one for using a dummy variable to represent the one-center CBD submarket and the other two dummy variables two-center CBD submarkets. The hedonic estimates for all six models are presented in Table 7.3. All four single-family and two-family housing models show substantial increase in adjusted R-square, 65 percent for the single and over 76 percent for the two-family, as compared to the 59 percent R-square for the model of all family housing with two-center CBD submarkets in Table 7.2. In contrast, a large portion of price variations in the three-family housing units can not be explained by either of the two submarket models, both having an adjusted R-square in the neighborhood of 36 percent. Based on both the R-squares and F-statistics, it seems that no significant difference exists between each pair of models for the one-center and two-center CBD submarkets in each of the three different family housing markets. Therefore, in the following comparison of hedonic estimates across the three-family housing submarkets, we focus on the three models for the one-center CBD submarket (*i.e.*, the models presented in the upper half of Table 7.3).

All coefficients for the variables in the single-family housing model have correct signs and are statistically significant at the 1 percent level, except for the variable of percentage of educated adult, which is significant only at the 10 percent level. In this model, the two location variables combined, the street distance and the dummy variable for the CBD, explain as much of the price variation as all the five neighborhood socioeconomic variables combined together. The housing structural variables are important, but not as

Table 7.3: Comparison of Hedonic Models by Family Housing Type with Different Delineation for the CBD Housing Submarkets

(Dependent Variable = Housing Price)

Attributes	Models by Type of Housing for One-center CBD Submarket								
	One-Family (N=491)			Two-Family (N=291)			Three-Family (N=316)		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	161344.00		5.07	-69441.00		-1.37	-22299.00		-0.49
BATHROOM#	15537.00	0.08	2.85	21970.00	0.10	2.37	28512.00	0.15	2.75
SIZE_LIVING	31.49	0.17	5.53	12.02	0.08	2.07	28.02	0.37	6.93
MEDIAN_VALUE (t)	0.34	0.18	2.94	0.85	0.35	6.57	0.20	0.17	2.48
HIGH_EDUCATION% (t)	643.91	0.09	1.72	1042.71	0.11	2.15	732.25	0.14	2.00
NONWHITE% (t)	-395.25	-0.11	-3.46	16.68	0.00	0.12	-248.32	-0.13	-2.41
LOW_INCOME% (b)	-777.24	-0.08	-2.55	-897.62	-0.07	-2.09	-397.43	-0.07	-1.43
MEDIAN_AGE (b)	-2215.37	-0.16	-5.77	-118.52	-0.01	-0.19	15.85	0.00	0.03
Ln(STREET_DISTANCE)	-22389.00	-0.11	-3.10	12140.00	0.04	1.17	-12646.00	-0.08	-1.30
Submarket 1	353955.00	0.43	9.89	603764.00	0.43	10.52	-65349.00	-0.09	-1.70
Adjusted R-sq	0.6518			0.7675			0.3548		
F-statistics	102.9100			107.3390			20.2440		

Attributes	Models by Type of Housing for Two-center CBD Submarkets								
	One-Family (N=491)			Two-Family (N=291)			Three-Family (N=316)		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	142354.00		4.25	-42033.00		-0.86	-50820.00		-1.09
BATHROOM#	15275.00	0.08	2.81	14655.00	0.06	1.63	37447.00	0.19	3.42
SIZE_LIVING	31.58	0.17	5.56	9.02	0.06	1.61	26.57	0.35	6.55
MEDIAN_VALUE (t)	0.54	0.28	3.34	1.14	0.47	8.32	0.29	0.24	3.24
HIGH_EDUCATION% (t)	459.94	0.06	1.18	727.47	0.08	1.55	811.34	0.15	2.23
NONWHITE% (t)	-377.28	-0.11	-3.29	72.92	0.02	0.56	-218.74	-0.11	-2.12
LOW_INCOME% (b)	-765.74	-0.08	-2.52	-1285.10	-0.10	-3.07	-416.09	-0.08	-1.51
MEDIAN_AGE (b)	-2271.27	-0.17	-5.91	-351.46	-0.02	-0.59	-10.98	0.00	-0.02
Ln(STREET_DISTANCE)	-26006.00	-0.13	-3.47	-2835.99	-0.01	-0.27	-17762.00	-0.11	-1.80
Submarket 1	289633.00	0.35	5.69	532671.00	0.38	9.37	-92811.00	-0.13	-2.33
Submarket 2	-71165.00	-0.08	-1.77	-188762.00	-0.17	-5.01	-77003.00	-0.16	-2.38
Adjusted R-sq	0.6533			0.7858			0.3645		
F-statistics	93.3440			107.3920			19.0670		

Note: (b) = data aggregated at the block group level; (t) = data aggregated at the tract level.

Sources: a spatially integreted data set based on 1990 Census, 1992 TIGER/Line File, and 1989-91 Housing Sale Data from County Data Corporation, VT.

important as the neighborhood quality and location variables. In the two-family housing model, three socioeconomic factors, median housing value, education, and income, account for nearly all of the impact of neighborhood quality on housing price. The racial composition and age of neighborhood housing stock are less important. The distance variable is insignificant. The dummy variable for the CBD submarket is an important factor. This is entirely due to the three high-priced sales that share an identical address and fall into this CBD submarket, as we have noted above. When excluding these three sales and the dummy variable from the model, the remaining variables explain only 45 percent of the price variation.

For the three-family model, the dominant factors in explaining price variation are the two structural variables. The variables of income and age of housing stock are insignificant. The other three socioeconomic variables contribute approximately equally to the change of housing price. The location variables are not statistically significant. The insignificance of the distance variable seems explainable by the map of spatial distribution of the three-family housing sales (see Figure 7.6), which shows that most sales are in the neighborhoods close to the employment center and thus do not differ much in their accessibility. This observation is supported by the fact that the variance of the distance variable for the three-family housing sales is only half as much as that for the single-family housing sales. The above analysis seems to indicate clearly that there exists three distinctively different family housing submarkets in the Boston area although different types of family housing are intertwined with each other across the urban landscape.

Based on the hedonic estimates by family housing type, we find that the accessibility variable of street distance is a very significant factor in affecting the price of single-family housing. The statistical tests for the distance variable in the single-family housing models strongly support the traditional theory of residential location. The same variable is barely significant at the 10 percent level in the three-family housing model with two-center CBD submarkets and insignificant in the model with a one-center CBD submarket. According to both of the two-family housing models, the distance variable is not a significant factor.

7.4 Segmentation by Race

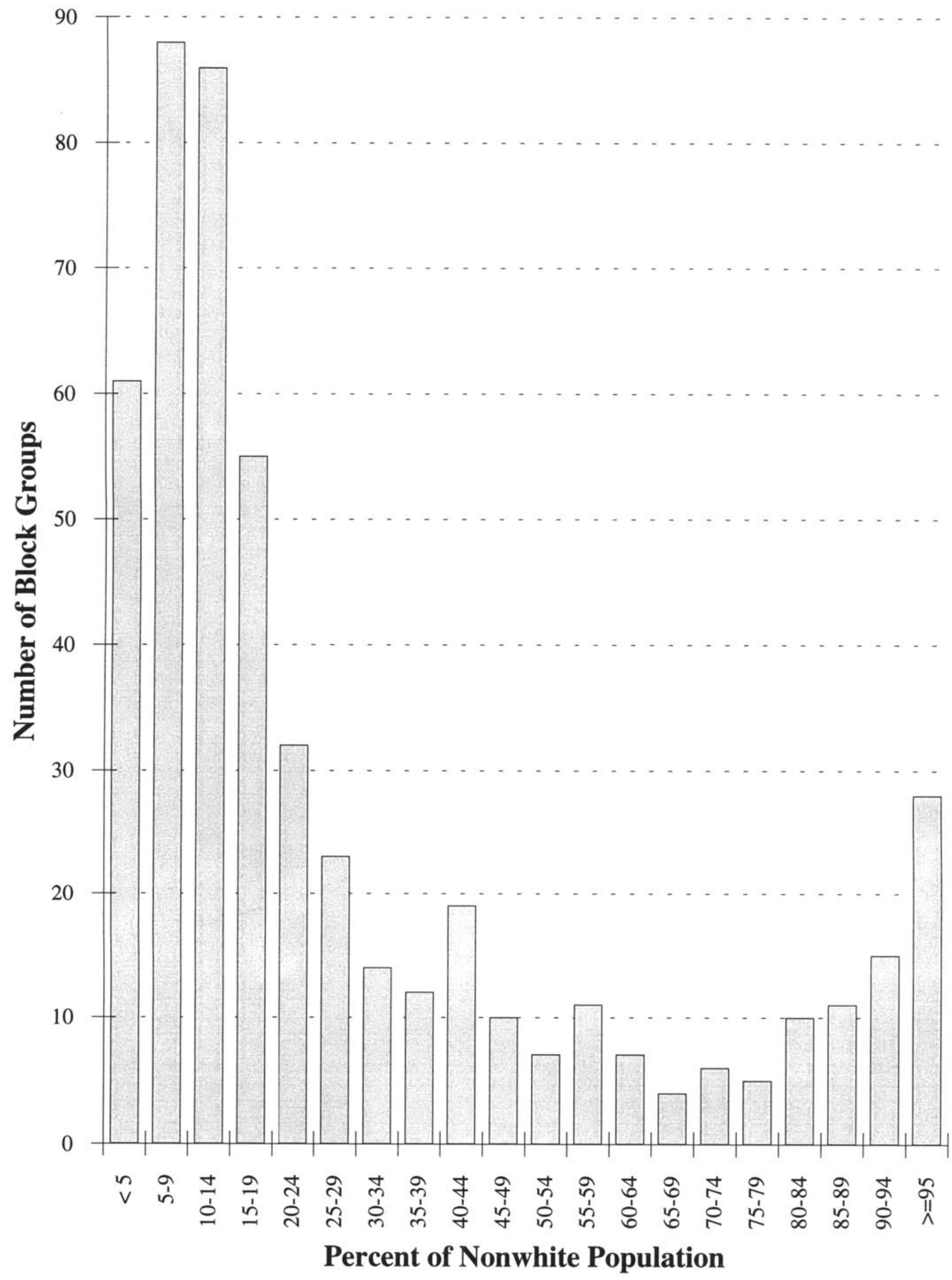
We begin by discussing how a natural breakpoint for the racial variable, percentage of nonwhite population, can be identified to segment the Boston market into two submarkets. Based on this selection, we use GIS to display the spatial results of market segmentation. Then, we estimate separate hedonic models for the two submarkets. The results of hedonic estimates show that the two submarkets differ distinctively from each other. Moreover, the distance variable seems to affect the housing prices in the opposite directions in the two submarkets. Possible explanations may be found by examining the spatial patterns of the housing sales.

The map of housing market segmentation by race, shown in Figure 7.1, portrays a clear spatial pattern for the location of block groups with greater than 50 percent nonwhite population. Most of these block groups are spatially contiguous and clustered within four neighborhoods: Roxbury, North Dorchester, South Dorchester, and Mattapan. This location pattern seems to suggest two spatially distinct housing submarkets, one with a majority nonwhite population and the other with a majority white. This perception is also supported by a histogram showing the frequency distribution of block groups by percentage of nonwhite population (see Figure 7.7). The dumbbell-shaped distribution seems to indicate a clear split. Therefore, we choose 70 percent as a natural breakpoint[15]

---

15. We have also used 50 percent as a breakpoint to segment the Boston market and constructed hedonic models based on this selection. The results do not differ significantly from the following analysis that uses 70 percent as a breakpoint.

Figure 7.7: Distribution of Census Block Groups by Percentage of Nonwhite Population, Boston 1990



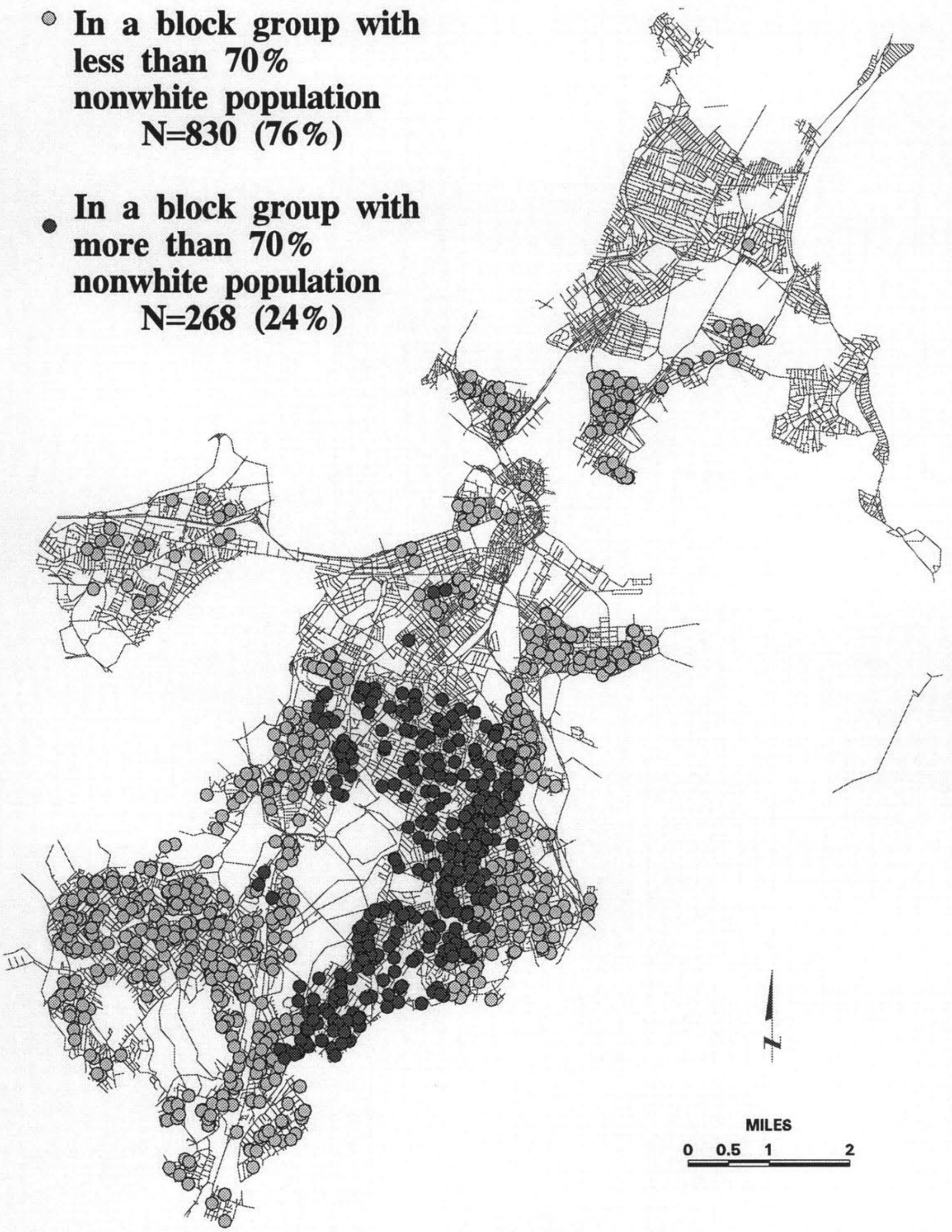
to segment the Boston housing market into two submarkets for estimating separate hedonic price models.

As a result of grouping the housing sales cases based on the racial characteristics of their associated block groups, 830 sales are located in the block groups with less than 70 percent of nonwhite population and the remaining 268 in the block groups with more than 70 percent of nonwhite population. This result is depicted spatially in Figure 7.8. Unmistakably, two spatially exclusive housing submarkets can be identified. The sales in the block groups with more than 70 percent of nonwhite population form a nearly vertical strip of neighborhoods. By mapping the sales data spatially, we are able to understand not only the location and size of housing submarkets but also the spatial configuration of the subarea. These spatial features serve as important background information for constructing and interpreting housing hedonic price models.

Given such a large volume of housing sales data, without mapping the sales spatially, this distinctive spatial pattern is unlikely, if not impossible, to be within any commoner's imagination. In the real world, these are cases with spatial patterns for some socioeconomic phenomena that are so complex and obscure that even the most advanced and sophisticated spatial statistical techniques seem inadequate and hopeless for pattern identification. However, more often than not, many socioeconomic phenomena do show some readily identifiable spatial patterns. This is because the growth patterns of a human settlement are shaped by history and constrained by geography, usually with rationale rather than casting a die. Therefore, surprisingly, a simple and thoughtful spatial portray of some socioeconomic phenomena is often sufficient and essential to reveal patterns and to help explain possible spatial relationships. It seems important not to underestimate what can be revealed by a simple map. In this study of housing market segmentation by race, the map of sales is at least worth a thousand words.

Figure 7.8: Spatial Distribution of Housing Sales
Two Housing Submarkets by Race

- In a block group with less than 70% nonwhite population N=830 (76%)
- In a block group with more than 70% nonwhite population N=268 (24%)



Two submarket models for the housing submarkets by race are shown in Table 7.4. The adjusted R-square for the model of the majority white submarket is 0.528, more than twice as high as that for the majority nonwhite submarket. All the coefficients for the majority white model are statistically significant and with expected signs. The magnitude of all coefficients for the seven independent variables is approximately on a par with that of the corresponding estimates in the single-market model, which is also presented in the table for the convenience of comparison. Reducing the size of the housing sales cases by nearly a quarter and excluding the variable of race do not seem to show much change in the coefficients for the remaining variables in the majority white model. The similarity between the two models indicates that the coefficients estimated for each of the variables are fairly consistent, stable and somewhat robust.

The hedonic model for the submarket with majority nonwhite population does poorly in explaining the price variations in the 268 sales. Only the variables of size of living area and percentage of low income households are significant at the 1 percent level. The variable of number of bathrooms is significant at the 5 percent level and the distance variable at the 10 percent level. The rest of the variables are statistically insignificant. The two structural variables make the most contribution to the impact on housing price. The significant estimate for the income variable suggests that in the majority nonwhite submarket, a house would be sold at a higher price in a high income neighborhood than in a low income one, all other things equal. The most surprising outcome in this model for the majority nonwhite submarket is the positive sign for the coefficient of the distance variable. In this submarket, the further away from the center the more expensive the housing, *ceteris paribus*. This impact of the accessibility to the employment center on housing price works in contradiction to the traditional theory of residential location. Examining the spatial distribution pattern for the housing sales cases for this submarket,

Table 7.4: Comparison of Hedonic Models for the Housing Submarkets with Different Racial Compositions

(Dependent Variable = Housing Price; N = 1098)

Attributes	Model for Submarket with < 70% of Nonwhite Population			Model for Submarket with >= 70% of Nonwhite Population			The Single-market Model		
	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat	Estimates	Std. Estimates	t-stat
Constant	60,225.00		1.81	66,815.00		2.76	68,411.00		2.71
BATHROOM#	12,351.00	0.09	2.46	9,389.42	0.16	1.97	12,513.00	0.09	3.08
SIZE_LIVING	27.62	0.23	6.93	15.24	0.38	4.61	25.15	0.24	7.91
MEDIAN_VALUE (t)	0.90	0.46	10.75	0.01	0.00	0.05	0.88	0.46	12.99
HIGH_EDUCATION% (t)	803.73	0.10	2.50	205.47	0.03	0.49	698.70	0.09	2.56
LOW_INCOME% (b)	-1,022.74	-0.09	-3.42	-490.43	-0.14	-2.14	-1,036.02	-0.11	-4.58
MEDIAN_AGE (b)	-2,145.10	-0.12	-4.81	-246.87	-0.05	-0.76	-1,954.73	-0.13	-5.88
NONWHITE% (t)							-241.36	-0.08	-2.98
Ln(STREET_DISTANCE)	-19,195.00	-0.09	-2.86	17,795.00	0.12	1.83	-20,165.00	-0.10	-3.65
Adjusted R-sq	0.5242			0.2016			0.5159		
F-statistics	131.46			10.63			147.16		

Note: (b) = data aggregated at the block group level; (t) = data aggregated at the tract level.

Sources: a spatially integreted data set based on 1990 Census, 1992 TIGER/Line File, and 1989-91 Housing Sale Data from County Data Corporation, VT.

we know that the distance variable for the sample should contain sufficient variance for testing its impacts on housing price. The significant effect with a positive sign may have to do with other factors, such as crime incidences, public services, school quality, and public transit access, which are not captured in the model. Nonetheless, the spatial configuration of the majority nonwhite submarket and the distribution patterns of the sales cases seem to add some degrees of confidence in the estimated coefficient for the distance variable.

It seems safe to conclude that the neighborhoods with majority nonwhite population seem to form a housing submarket that is distinctively different from the neighborhoods with majority white population. Unlike in the majority nonwhite submarket, in the majority white submarket, the neighborhood quality factors matter a great deal. Moreover, what matters most in affecting housing price in the majority nonwhite submarket are the housing structural variables. The study of hedonic estimates for both submarkets suggests that the more disaggregated measurement of accessibility matters.

7.5 Summary

The focus of this chapter is to explore the potential of using GIS to assist segmenting urban housing markets spatially and modeling housing prices. We show that GIS is a useful tool for representing the results of market segmentation spatially, generating additional spatial information, and regrouping and aggregating spatial data. A GIS-based representation of market segmentation and other spatial information is often a crucial step in the process of hedonic price model construction and interpretation. It is also an essential step for spatial pattern identification and providing important background spatial information for modeling when a large volume of spatial data are involved.

We use two different approaches to segmenting the CBD market area to illustrate the flexibility of spatial exploration with the aid of GIS. We also segment the Boston market by family housing type and neighborhood racial composition. The analysis of hedonic estimates shows that distinct housing submarkets, as defined by housing type and race respectively, exist in the Boston area. There is a need to estimate separate housing hedonic price models for different housing submarkets in order to capture correctly how housing attributes affect price differently in each submarket. Moreover, the accessibility variable measured as actual street distance from a housing unit to the employment center appears to be a significant factor in affecting housing price in most of the submarket hedonic models. It seems that modeling accessibility in a more localized and realistic manner provides strong evidence supporting the economic theory of residential location.

Chapter 8

Conclusions and Future Research

8.1 Summary of Research Findings

The primary purpose of this dissertation research is to explore the utility of geographic information systems in the fields of urban and regional spatial analysis in general and the analysis of urban housing market in particular. The research shows clearly that GIS has potential to support a new style of spatial modeling process that is characterized by flexibility in data aggregation at various spatial levels, delineation of a study area suitable for different neighborhood definitions and market segmentations, and graphic representation of inputs and outputs of spatial models. The dissertation demonstrates the value of GIS in helping develop a better understanding of housing markets, especially their spatial characteristics and interrelationships. It illustrates how GIS can be used to combine various sources of data of common geography to construct a spatially integrated database for housing analysis. It provides strong confirmation of the important impacts of the accessibility factor, measured at a more localized and realistic manner, on the value of residential properties, and tests positively the necessity of segmentation price modeling in a large urban housing market. The exploration of segmentation modeling seems to support the notion that there exist distinct housing submarkets by race, family housing type, and accessibility in the Boston area. The research also demonstrates the effectiveness of using GIS for dynamic graphic representations of inputs and outputs of spatial modeling. In particular, it explores the usefulness of mapping housing hedonic model residuals in conceptualizing and modifying models and interpreting model outcomes.

8.2 Promises and Difficulties of GIS Applications

This research demonstrates that GIS technologies have become sufficiently advanced and sophisticated to make important contributions to the spatial analysis of urban housing markets. At the conclusion of this dissertation research, we are more optimistic than ever before about the potential contributions promised by GIS technologies to the various fields concerned with the spatial elements of urban and regional activities. However, the road that we have travelled through our exploration of using GIS to assist this housing market analysis has been bumpy and sometimes frustrating. The difficulties reflect the immaturity of GIS technologies and highlight opportunities for shaping the technologies in future courses of development.

GIS has experienced rapid advancement in recent years and is believed by some to have nearly reached its maturity. But the fact is that it is still undergoing rapid growth and fast changes. The technology involves many software developers, numerous suppliers of geo-referenced data, and a huge and diverse pool of users including government agencies, private sector firms, research and educational institutions, and community and neighborhood development groups. GIS is far from reaching a stage of standardization. Especially as a tool to assist spatial modeling, GIS software is still a long way from being truly friendly and flexible.

We have experienced several difficulties in using ARC/INFO, widely believed to be the most sophisticated GIS software, to assist in our spatial modeling exercises. First, since the software does not have a built-in capability of carrying out regression analysis, it is rather cumbersome to move data in and out of the software in order to make use of computer statistics packages. Second, the software requires a fairly high degree of programming and design skills for an effective application to spatial modeling. Most of the spatial analyses involving GIS in this research require writing complicated sequences of commands. Third, there are still some limitations in terms of the size of data that can be processed by the software, which makes the application of GIS to the analysis of large urban and regional networks less straight forward. For instance, in our computation of the shortest street distance for each of the housing sales cases, we had to divide our data set into two subsets in order to circumvent the processing limit set by the ARC/INFO network analysis module. This division unnecessarily complicated our process of exploration. This could pose an extra layer of difficulty when one attempts to use GIS for carrying out a larger scale of urban and regional spatial modeling. Finally, there are still serious bugs associated with the software, which are not uncommon for any fast growing and changing computer software.

Also, we have found that there are some significant errors and incompatibilities associated with various sources of geo-referenced data. For example, there are still many erroneous codings of location information in the 1990 TIGER/Line files. The 1992 files have shown significant improvement, but are still far from being perfect. Suppliers of housing sales data usually have their own established ways of coding the location information, which are often incompatible with the formats required by GIS software. Thus, tedious checking becomes necessary to insure data accuracy. These problems have hampered our research to a varying degree. Like other quantitative research, extra caution should be exercised in checking and double-checking the accuracy of data. Ideally, various mechanisms should be developed to test the results of spatial data integration.

In short, GIS technologies have become very sophisticated and powerful. But the learning curve is rather flat. In other words, a substantial cumulation of experience in using GIS is essential before one can move to seek more meaningful applications. We hope to see greater improvement in statistical analysis functionality in some leading GIS software in the near future. For example, we hope to see a GIS tool that would make it possible to "cut out" a submarket area, shaped either regularly or irregularly, and then give spatial data that are reaggregated instantly and with related models reconstructed accordingly. This type of capability should make a new style of exploratory modeling truly flexible.

For now, since GIS technologies still represent a frontier that is full of promises, we have to be creative in mixing and matching various technologies to solve complex problems of spatial modeling, and confront various challenges that are commonly associated with any changing and immature technology.

8.3 Recommendations for Future Research

Our exploration of using GIS to assist spatial modeling of urban housing markets seems to open up a number of possible avenues for future research. Some thoughts on future research extensions are outlined below.

(1) Expanding the study area

Although the data and the empirical results represent a significant improvement over previous studies, the analysis is deficient in several important respects. Perhaps the most serious shortcoming is the limited size of the housing market area selected for this study of hedonic price modeling. The study area does not cover other neighboring central cities and towns, and the suburban towns. For most of the households, the relevant housing market is the entire region. This is because much of the variation in the bundles of residential services is in the suburban areas of the regional market. Budget constraints on the data collection in this research precluded using the entire region as a study area for exploring how GIS can contribute to the improvement of housing hedonic price models and other spatial analyses of the housing markets. The other weakness in not including the entire region as a study area is the incompleteness of the street network and its associated data. As a result, some of the measurements of the shortest street distance are inaccurate, although fortunately it is unlikely that they are significantly distorted due to the configuration of the selected micro market area. In view of these weaknesses, a natural and promisingly fruitful extension of this research is to collect much larger and more extensive bodies of housing sales data and include the entire region in a study area. This extension is crucial for developing truly useful housing hedonic price models.

(2) Modeling accessibility with greater complexities

The four GIS-based alternatives for accessibility measures tested in this research still involve a high degree of simplification of reality. In order to provide stronger and more convincing evidence to support the significant impact of more accurately measured accessibility variables on residential properties, we need to construct a GIS-based regional street and highway network and also include the public transit networks. We need to drop the monocentric assumption and compute the interaction of a housing unit's accessibility to multiple centers. We also need to take into account as much as possible various traffic impedances associated with major streets and street segments. These extensions require a huge effort of data collection and integration. What can be offered by GIS is an effective helping hand to process and analyze a large volume of spatial data. The improvement of accessibility measurements will not only advance our understanding of the housing markets, but could also benefit many other fields of research and application, such as transportation planning, growth management, and land use planning.

(3) Exploring additional data aggregation levels and spatial factors

This research explores the impacts on housing price of several socioeconomic and environmental variables at only three levels of data aggregations (*i.e.*, tract, block group, and a more localized level for accessibility measurement). We have shown the flexibility of using GIS to aggregate data at different spatial levels, and organize and regroup data based on different delineations of a study area. Also, we have argued and demonstrated that different socioeconomic and environmental factors are likely to have different spatial scope of impact on housing price. Therefore, to take advantage of the strength of GIS, in future research, we could explore and test the effects of a number of other socioeconomic and environmental factors at different spatial aggregation levels. For example, one extension is to test the impact of school quality. This variable is likely to vary from the central parts of a city to suburban areas. It is probably more appropriate to measure it at a neighborhood level in the former areas and at a town level in the latter. Another extension is to look into the spatial effect of neighborhood crime. In the past, a commonly used proxy to measure the impact of crime on housing value is a rate that measures crime incidences per thousand household or the like. This is obviously a rather poor proxy because what really matters is the exact location of crime and frequency of crime associated with a location. With GIS, we could identify the exact location of each crime incidence and measure its proximity and intensity with respect to the location of a housing unit of interest. In this case, the most localized measurement of neighborhood crime should yield more accurate estimates for the effect of crime on residential properties. Another similar factor that is worth testing is neighborhood arson, which has important impact on residential value and should also be studied at more localized levels in relation to locations of the arson incidence.

Other important spatial factors that could be examined in more localized fashions with the aid of GIS include proximity to environmental amenities and dis-amenities, such as, parks, water surfaces, views, and hazardous waste sites. GIS permits potentially much better measurements of these variables.

(4) Modeling temporal changes of housing disequilibrium price surfaces

Using the 3-D modeling capacity of GIS, we could extend the results of this research to model temporal changes of disequilibrium price surfaces for housing. Based on the results of estimated hedonic prices, we could compute price per hedon for each observation, which equals the actual housing price divided by the estimated price. Then, a mean price per hedon can be computed for each block group. This mean can then be used as a vertical value to map out a disequilibrium price surface for the entire market area. If sufficient temporal housing sales data can be gathered, it is possible to visualize the changes and movement of the disequilibrium price surfaces over space and time. This extension of housing market research using GIS might lead to some valuable insights into the process of how the forces of demand and supply adjust in space and time. Another value of this sort of 3-D price surface model is to help planners and assessors visualize the performance of housing price models in their spatial context and to understand the locations and spatial interrelationships of those neighborhoods whose residential properties are frequently or consistently overestimated or underestimated by a price model.

(5) Defining more homogeneous housing submarkets

We have shown in this dissertation research that three different market segmentation approaches lead to very different segmentation price models. This is a useful and important finding in helping develop a good understanding of the characteristics of the housing market. However, a more interesting research question worth pursuing seems to be: which segmentation approach would create housing submarkets that are most homogeneous, and which allows construction of the best hedonic price models? GIS seems to offer an effective tool to help answer these questions. The more accurate identification of homogeneous housing submarkets would permit the examination of how the forces of demand and supply operate at a more appropriate spatial scope.

(6) Conceptualizing neighborhood

City planners, economists, sociologists, geographers, politicians, and many others have been interested in the study of socioeconomic status of neighborhoods. Yet, just what constitutes a neighborhood spatially? Or to put it differently, how should a neighborhood be defined and where are its physical boundaries? It seems that no agreement on the conceptualization of an urban neighborhood exists among different disciplines and even among researchers in the same field. Most existing studies have adopted the census tract as a spatial proxy or working definition for measuring neighborhood quality due to the convenience of using census data. A tract in proximity to the center of a city is usually characterized by diverse demographics, housing conditions, income, and racial composition. We have shown that a census tract is usually too large an areal unit to summarize neighborhood effects and using block groups as a proxy can lead to very different research conclusions. Both tract and block group boundaries are highly arbitrary. At least two important groups of factors should be taken into account when we try to conceptualize what constitutes a neighborhood spatially. First, it is likely for a neighborhood to be bounded by parks, highways, cemeteries, ponds, rivers, and many other natural and human-made landmarks. These environmental and geographic entities are important elements in defining a neighborhood. Second, a neighborhood should be homogeneous in terms of its demographic and housing characteristics. In view of these two considerations, it is conceivable that with the aid of GIS we could derive a better definition for an urban neighborhood. Using GIS, it is possible to employ a bottom-up approach by adding homogeneous and spatially contiguous blocks to form a neighborhood while at the same time taking into consideration the local environmental and geographic features.

(7) Designing an interface for easy application of housing hedonic models

With a modest amount of effort, one could design a friendly user interface for applying housing hedonic models. For instance, the user might enter the structural attributes and the address of a housing unit. Based on the address information, the housing unit would appear on a GIS-based map of socioeconomic factors, with an estimate of its accessibility to a particular employment center. Then, the user could choose one of several housing hedonic price models to predict the property's potential market price. The user could easily assess the reliability of this result by displaying it on a map of error residuals from the model, and against various layers of related spatial information. Placing the prediction in a spatial context would allow for holistic comparisons with housing properties in surrounding neighborhoods. This kind of user interface (a rudimentary "planning support system") would be valuable to planners, assessors, appraisers, and real estate agents who frequently use housing price models in their practice.

8.4 Concluding Remarks

We consider that the outcome of this research represents a meaningful step toward filling a gap in our understanding of the spatial characteristics of the housing market and provides some valuable indications on how useful the incorporation of GIS into the traditional urban and regional spatial analysis could be. It is our hope that this experimental step will stimulate further research and analysis of housing markets using GIS.

In view of the rapid advancement and growing application of GIS technologies, it seems that there exists some urgency in exploring a promising integration of traditional approaches of urban and regional spatial analysis with GIS technologies. This type of basic research is much needed and promises great potential. Marble (1990), a well-known geographer and a pioneer of applying GIS in his field, goes as far as to state that GIS as a research tool is beginning to revolutionize those portions of the social sciences which are concerned with the spatial aspects of human society to a point where many of the things that we do must be looked at from a completely different viewpoint. Most contemporary theories of spatial behavior, including those in the field of spatial economics, rest upon a foundation which has been simplified to the extent of unreality. A major reason for this oversimplification is that we have lacked the tools which would permit organizing and comprehending the spatial data that define the real and extremely complex environment in which human behavior takes place. This dissertation research has demonstrated that using GIS, it is possible to model space and spatial activities in a more sophisticated manner. We believe that the utilization of GIS technologies will make theoretically sophisticated and spatially complex data models not only possible, but a common practice of the next decade.

Appendix A

Documentation of Data Cleaning and Analysis

This appendix consists of four parts. The first part, A.1, documents the housing sales data employed for this dissertation research and the process of cleaning this data set. Each of the next three parts contains a correlation analysis for a particular set of variables. Specifically, A.2 contains an correlation coefficient matrix for the socioeconomic variables used for modeling median housing value in the greater Boston area. A.3 presents a correlation analysis of the housing structural attributes that are available in the housing sales data. A.4 is a correlation analysis of the variables included in the housing hedonic price models constructed for the Boston housing market.

A.1 The Housing Sales Data

The housing sales data obtained from the Boston Redevelopment Authority contain real estate transactions occurring between January 1989 and September 1991 in the city of Boston. The total number of transactions during this period was 8,073, of which 4,260 were coded as residential building and land sales and the others were commercial real estate sales or unclassified. The following table summarizes the distribution of sales by type for the data set.

Table A.1.1: Summary of the Real Estate Transactions for the city of Boston: January 1989 - September 1991

Transaction Type		Code	Count
Residential Building (subtotal):			**1,740**
	Single-family Housing	101	568
	Two-family Housing	104	456
	Three-family Housing	105	333
	Condo	102	383
Residential Land Use (subtotal):			**2,520**
	Residential 1	R1	729
	Residential 2	R2	485
	Residential 3	R3	483
	Residential 4	R4	104
	Condo Dwelling	CD	719
Total Residential Sales:			**4,260**
Total Non-residential Sales:			**3,813**
Total:			**8,073**

For this research, we extracted from this data set only residential building sales, including four types of housing: single-family, two-family, three-family, and condominium housing. The following is a partial list of fields that are contained in this data set.

Table A.1.2: A Partial List of Fields in the Housing Sales Data

Field Name	Field Definition	Data Type
COUNTY_C	county code	C
TOWN	town name	C
STNUM	street number	C
STREET	street name	C
UNIT	unit number	C
SELLER	seller's name	C
BUYER	buyer's name	C
PRICE	sale price	N
TDATE	date of transaction	C
LOTSIZE	lot size	N
CLASS	type of residential use	C
ASSESSED	assessed property value	N
LIVING	size of living area	N
FLOOR	size of floor area	N
ROOM	number of rooms	N
BED	number of bedrooms	N
BATH	number of bathrooms	N
MORTGAGE	amount of mortgage	N
LENDER	mortgage lender's name	C
ZIP	ZIP codes	C

Note: C = character; N = numeric.

To prepare a clean sales data set for this research, we processed the 1,740 residential building sales as follows. For the 1,284 cases of the family-type housing, we excluded 6 with missing data, 3 with mis-coded fields, and 55 duplicate records for co-buyers. We ended up with 1,220 sales for the family-type housing. For the 456 condo sales, we kept

351 sales after deleting 83 with missing data, 1 with mis-coded fields, and 21 duplicate records. Thus, after this data cleaning, we constructed a housing sales data set of 1,571 records. This is the data set that was used for address matching.

A.2 Correlation Coefficient Matrix of Socioeconomic Variables for Modeling Median Housing Values

Table A.2.1: Correlation Analysis for Socioeconomic Variables Used in the Median Housing Value

By Tract: N=784	By Block Group: N=3172 MEDVALUE	M_INCOME	BLT_MED	PLESS15K	PLESS25K	P_COLLEG	P_R2O	P_VACANT	P_NONWP	P_SENIOR	P_I2V	P_H50OLD	DENSITYP	DENSTYHH	DIS_MILE
MEDVALUE		0.63 0.00	0.01 0.41	-0.32 0.00	-0.39 0.00	0.77 0.00	-0.19 0.00	-0.11 0.00	-0.17 0.00	0.03 0.13	-0.31 0.00	0.00 0.97	-0.04 0.03	0.05 0.00	-0.19 0.00
M_INCOME	0.57 0.00		0.20 0.00	-0.71 0.00	-0.81 0.00	0.64 0.00	-0.66 0.00	-0.25 0.00	-0.35 0.00	-0.16 0.00	0.41 0.00	-0.24 0.00	-0.42 0.00	-0.38 0.00	0.07 0.00
BLT_MED	0.01 0.86	0.32 0.00		-0.15 0.00	-0.22 0.00	0.03 0.13	-0.34 0.00	-0.04 0.03	-0.15 0.00	-0.09 0.00	0.18 0.00	-0.90 0.00	-0.42 0.00	-0.36 0.00	0.43 0.00
PLESS15K	-0.35 0.00	-0.82 0.00	-0.26 0.00		0.90 0.00	-0.44 0.00	0.62 0.00	0.24 0.00	0.40 0.00	0.26 0.00	-0.50 0.00	0.19 0.00	0.39 0.00	0.35 0.00	-0.11 0.00
PLESS25K	-0.41 0.00	-0.89 0.00	-0.33 0.00	0.96 0.00		-0.50 0.00	0.70 0.00	0.27 0.00	0.45 0.00	0.24 0.00	-0.53 0.00	0.25 0.00	0.44 0.00	0.40 0.00	-0.13 0.00
P_COLLEG	0.82 0.00	0.65 0.00	0.05 0.21	-0.49 0.00	-0.54 0.00		-0.22 0.00	-0.12 0.00	-0.23 0.00	-0.07 0.00	-0.07 0.00	0.00 0.88	-0.06 0.00	0.03 0.12	-0.15 0.00
P_R2O	-0.16 0.00	-0.78 0.00	-0.45 0.00	0.75 0.00	0.81 0.00	-0.22 0.00		0.16 0.00	0.44 0.00	0.07 0.00	-0.51 0.00	0.42 0.00	0.62 0.00	0.59 0.00	-0.38 0.00
P_VACANT	-0.18 0.00	-0.41 0.00	-0.11 0.00	0.38 0.00	0.43 0.00	-0.23 0.00	0.27 0.00		0.19 0.00	-0.06 0.00	-0.14 0.00	0.08 0.00	0.13 0.00	0.12 0.00	0.24 0.00
P_NONWP	-0.18 0.00	-0.47 0.00	-0.22 0.00	0.52 0.00	0.56 0.00	-0.25 0.00	0.52 0.00	0.30 0.00		-0.21 0.00	-0.21 0.00	0.16 0.00	0.43 0.00	0.31 0.00	-0.25 0.00
P_SENIOR	0.01 0.79	-0.11 0.00	-0.12 0.00	0.22 0.00	0.17 0.00	-0.09 0.01	0.09 0.01	-0.10 0.01	-0.30 0.00		-0.25 0.00	0.07 0.00	-0.08 0.00	-0.02 0.19	-0.11 0.00
P_I2V	-0.32 0.00	0.52 0.00	0.37 0.00	-0.66 0.00	-0.66 0.00	-0.10 0.00	-0.76 0.00	-0.26 0.00	-0.41 0.00	-0.21 0.00		-0.23 0.00	-0.34 0.00	-0.37 0.00	0.25 0.00
P_H50OLD	0.01 0.81	-0.37 0.00	-0.92 0.00	0.30 0.00	0.37 0.00	-0.04 0.28	0.53 0.00	0.17 0.00	0.22 0.00	0.10 0.00	-0.43 0.00		0.49 0.00	0.43 0.00	-0.44 0.00
DENSITYP	0.04 0.21	-0.51 0.00	-0.45 0.00	0.49 0.00	0.54 0.00	-0.02 0.58	0.70 0.00	0.22 0.00	0.46 0.00	-0.13 0.00	-0.59 0.00	0.54 0.00		0.95 0.00	-0.50 0.00
DENSTYHH	0.15 0.00	-0.44 0.00	-0.40 0.00	0.41 0.00	0.46 0.00	0.07 0.05	0.63 0.00	0.19 0.00	0.33 0.00	-0.07 0.04	-0.58 0.00	0.48 0.00	0.96 0.00		-0.47 0.00
DIS_MILE	-0.23 0.00	0.16 0.00	0.47 0.00	-0.18 0.00	-0.20 0.00	-0.19 0.00	-0.47 0.00	0.14 0.00	-0.30 0.00	-0.11 0.00	0.45 0.00	-0.49 0.00	-0.53 0.00	-0.50 0.00	

Note: Included in the bottom-half triangle are correlation coefficients estimated by census tract, and the top-half by block group.
For each pair of numbers in the table, the top is correlation coefficient and the bottom p_value.

MEDVALUE: Median value of owner-occupied housing units
M_INCOME: Median household income
BLT_MED: Median year structure built
PLESS15K: Percent of household with annual income less than $15K
PLESS25K: Percent of household with annual income less than $25K
P_COLLEG: Percent of adults (>=25 years) with Bechelor's degree or higher
P_R2O: Percent of renter over owner occupied housing units
P_VACANT: Percent of vacant housing units
P_NONWP: Percent of non white population
P_SENIOR: Percent of senior population (>=65 years)
P_I2V: median income as a percentage of median housing value
P_H50OLD: Percent of housing stock older than 50 year
DENSITYP: Population density (number of people per acre)
DENSTYHH: Household density (number of households per acre)
DIS_MILE: Distance (in miles) from the CBD

A.3 Correlation Analysis of Housing Structural Attribute Variables for Modeling Housing Price Variation

Table A.3.1: Correlation Analysis for Housing Structural Attribute Variables

Simple Statistics (N = 1220):

Variable	Mean	Std. Dev.	Sum	Minimum	Maximum
PRICE	149,137	111,688	181,947,127	5,000	1,559,550
ASSESSED	72,730	40,211	88,731,000	0	605,200
BATH	2.05	0.83	2,506.00	0.50	6.00
BED	4.61	1.92	5,629.00	1.00	12.00
LIVING	2,298.57	1,066.47	2,804,255.00	425.00	9,900.00
LOTSIZE	4,466.39	2,701.11	5,448,992.00	575.00	28,450.00
ROOM	10.34	3.95	12,616.00	4.00	27.00

Pearson Correlation Coefficients (N = 1220):

	PRICE	ASSESSED	BATH	BED	LIVING	LOTSIZE	ROOM
PRICE	1.000						
	0.000						
ASSESSED	0.420	1.000					
	0.000	0.000					
BATH	0.283	0.061	1.000				
	0.000	0.034	0.000				
BED	0.100	-0.065	0.681	1.000			
	0.001	0.023	0.000	0.000			
LIVING	0.292	0.069	0.752	0.749	1.000		
	0.000	0.017	0.000	0.000	0.000		
LOTSIZE	0.026	0.024	-0.174	-0.100	-0.013	1.000	
	0.365	0.408	0.000	0.000	0.649	0.000	
ROOM	0.140	-0.051	0.802	0.868	0.830	-0.105	1.000
	0.000	0.074	0.000	0.000	0.000	0.000	0.000

Note: Refer to Table A.1.2 for variable definitions. For each pair of numbers in the bottom half of the table, the top number is correlation coefficient and the bottom p_value.

A.4 Correlation Coefficient Matrix for the Variables Included in the Housing Hedonic Price Models

Table A.4.1: Correlation Analysis for the Variables Included in the Housing Hedonic Price Models

Simple Statistics (N = 1098):

	Data Aggregated at the Census Tract Level					Data Aggregated at the Census Block Group Level					
Variable	Mean	Std. Dev.	Sum	Minimum	Maximum	Mean	Std. Dev.	Sum	Minimum	Maximum	Variable Definition:
PRICE	162,001	109,983	177,877,277	50,945	1,559,550	162,001	109,983	177,877,277	50,945	1,559,550	Housing sale price in 1990 dollars
BATHROOM#	2.16	0.83	2,367.00	1.00	6.00	2.16	0.83	2,367.00	1.00	6.00	Number of bathrooms
SIZE_LIVING	2,304.34	1,064.35	2,530,164.00	425.00	9,900.00	2,304.34	1,064.35	2,530,164.00	425.00	9,900.00	Size of living area in square feet
MEDIAN_AGE	47.93	6.37	52,629.00	17.00	51.00	47.37	7.27	52,013.00	7.00	51.00	Median Age of housing stock in a tract or block group
MEDIAN_VALUE	165,834	57,173	182,085,483	87,500	500,001	166,249	66,028	182,541,873	14,999	500,001	Median housing value in a tract or block group
LOW_INCOME%	23.99	9.38	26,342.00	8.60	69.07	22.17	11.98	24,345.00	0.00	69.23	Percentage of households with annual income less than or equal to $15,000
HIGH_EDUCATION%	24.09	14.65	26,451.00	1.80	84.51	25.14	16.75	27,600.00	0.00	86.83	Percentage of adult (>=25-year-old) with Bechelor's degree or higher education
NONWHITE%	34.80	34.28	38,205.00	0.00	99.17	34.21	36.00	37,563.00	0.00	100.00	Percentage of nonwhite population in a tract or block group
Ln(STREET_DISTANCE)	1.52	0.52	1,664.83	-1.71	2.29	1.52	0.52	1,664.83	-1.71	2.29	The shortest street distance from a housing unit to the employment center in log
Ln(HOUSING_DISTANCE)	1.38	0.57	1,516.05	-1.71	2.22	1.38	0.57	1,516.05	-1.71	2.22	The Euclidean distance from a housing unit to the employment center in log
Ln(TRACT_DISTANCE)	1.37	0.57	1,505.14	-0.92	2.13						The Euclidean distance from a tract centroid to the center in log
Ln(BLOCKGRP_DISTANCE)						1.37	0.57	1,500.98	-1.35	2.21	The Euclidean distance from a block group centroid to the center in log

Pearson Correlation Coefficients (N = 1098):

By Tract: \ By Block Group:	PRICE	BATHROOM#	SIZE_LIVING	MEDIAN_AGE	MEDIAN_ VALUE	LOW_ INCOME%	HIGH_ EDUCATION%	NON_ WHITE%	Ln(STREET_ DISTANCE)	Ln(HOUSING_ DISTANCE)	Ln(BLOCKGRP _DISTANCE)
PRICE		0.324	0.322	0.009	0.529	-0.192	0.460	-0.158	-0.374	-0.338	-0.358
		0.000	0.000	0.755	0.000	0.000	0.000	0.000	0.000	0.000	0.000
BATHROOM#	0.324		0.699	0.126	0.148	0.202	-0.002	0.145	-0.374	-0.397	-0.398
	0.000		0.000	0.000	0.000	0.000	0.935	0.000	0.000	0.000	0.000
SIZE_LIVING	0.322	0.699		0.173	0.122	0.199	-0.015	0.257	-0.279	-0.255	-0.262
	0.000	0.000		0.000	0.000	0.000	0.615	0.000	0.000	0.000	0.000
MEDIAN_AGE	0.013	0.151	0.141		0.075	-0.086	0.068	-0.108	-0.175	-0.195	-0.196
	0.677	0.000	0.000		0.013	0.004	0.023	0.000	0.000	0.000	0.000
MEDIAN_VALUE	0.646	0.170	0.124	0.028		-0.236	0.642	-0.190	-0.373	-0.322	-0.367
	0.000	0.000	0.000	0.357		0.000	0.000	0.000	0.000	0.000	0.000
LOW_INCOME%	-0.178	0.226	0.209	-0.070	-0.239		-0.476	0.311	-0.163	-0.203	-0.190
	0.000	0.000	0.000	0.021	0.000		0.000	0.000	0.000	0.000	0.000
HIGH_EDUCATION%	0.505	-0.024	-0.045	0.060	0.730	-0.547		-0.437	-0.180	-0.127	-0.165
	0.000	0.426	0.132	0.048	0.000	0.000		0.000	0.000	0.000	0.000
NONWHITE%	-0.147	0.165	0.284	-0.135	-0.213	0.373	-0.448		-0.027	0.020	0.012
	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.371	0.515	0.693
Ln(STREET_DISTANCE)	-0.374	-0.374	-0.279	-0.128	-0.442	-0.285	-0.176	-0.041		0.964	0.967
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.179		0.000	0.000
Ln(HOUSING_DISTANCE)	-0.338	-0.397	-0.255	-0.155	-0.396	-0.336	-0.119	0.011	0.964		0.995
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.716	0.000		0.000
Ln(TRACT_DISTANCE)	-0.331	-0.398	-0.257	-0.156	-0.430	-0.331	-0.151	-0.013	0.964	0.986	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.670	0.000	0.000	

Note:
Included in the bottom-half triangle are correlation coefficients estimated by census tract, and the top-half by block group. For each pair of numbers in the table, the top is correlation coefficient and the bottom p_value.

Appendix B

ARC/INFO AML for Measuring Accessibility

This appendix contains two ARC/INFO AML files. The AML, standing for ARC Macro Language, is the ARC/INFO software's programming language that provides a set of tools to tailor user interfaces and allows automation of frequently performed actions. The two AML files are written to carry out computations of Euclidean distance and the shortest street distance between locations, respectively. These two AML files allow us to easily prepare the four alternatives of accessibility measurements tested in this dissertation research.

B.1 Computation of Euclidean Distance Using ARC/INFO

The following AML file computes Euclidean distance between two points. One of these two points is always the point that represents a designated employment center and forms a single-point geographic coverage. The other point should be in another point coverage, which may contain many points, representing locations of housing units, census tract centroids, or block group centroids. In this research, we used this AML to compute Euclidean distance between tract and block group centroids and the CBD, Boston, and between each housing location and the center.

```
/**********************************************************************/
/*
/* File:          /mit/crldata/lijian/cal_eud.aml
/*
/* Author:        Lijian Chen, Thu Mar 31 21:58:50 1994
/*
/* Purpose:       calculate Euclidean distances of points to CBD, BOSTON
/*
/* To execute:    [Arc]
/*                &r cal_eud <in_point_coverage> <out_distance_infotable>
/*
/* For example:   &r cal_eud SALEALLN SN_EUD.DAT
/*                &r cal_eud TRACTCEN TR_EUD.DAT
/*
/* Note:          "bos_cbd" is a single point coverage of CBD
/*
/**********************************************************************/


/* take 2 arguments from terminal
&args inputcov infotab


/* prompt a warning message
&if [null %infotab%] &then &do
 &type USAGE:
 &type &r cal_eud <in_point_coverage> <out_distance_infotable>
 &return
&end


/* calculate a straight line distance from the location of
/* a housing unit to CBD, Boston
pointdistance %inputcov% bos_cbd %infotab%
```

```
/* convert distance measurement unit from feet to miles
additem %infotab% %infotab% MILE 8 8 N 2

&DATA ARC INFO
ARC

SELECT %infotab%

CALCULATE MILE = DISTANCE / 5280


Q STOP
&END

&return


/*------------------ End of the Main AML: cal_eud.aml -----------------
```

B.2 Computation of the Shortest Street Distance Based on the Street Network for the City of Boston Using ARC/INFO

The following AML file computes the shortest street distance between two points. One of these two points is always the point that represents a designated employment center. The other point represents the location of a housing unit. In this research, we used this AML to compute the shortest street distance between the location of a housing unit and the CBD, Boston.

```
/************************************************************************/
/*
/* File:          /mit/crldata/lijian/path.aml
/*
/* Author:        Lijian Chen, Wed Mar 16 02:50:16 EST 1994
/*
/* Purpose:       This aml finds the shortest path between each of
/*                the given set of housing units (nodes) and the
/*                center of Boston, defined as Downtown Crossing.
/*
/* Files called:  stopfile.aml, route_id.aml
/*                |
/*                +-->> stopfile.tab
/*
/* To execute:   [Arcplot] &r path <number_of_nodes> <NODE_COVERAGE>
/*
/************************************************************************/

/* take 2 arguments from terminal
&args number .nodecov


/* prompt a warning message
&if [null %.nodecov%] &then &do
 &type USAGE:
 &type &r path <#_of_housing_unit> <NODE_COVERAGE>
 &return
&end


/* turn on performance timer
&pt &on


/* set up an info stop file using the data file: stopfile.tab
 /* call Subroutine 1
 /*************************
 &r stopfile.aml
 /*************************
```

```
/* create a relate between %.nodecov%.STP and %.nodecov%.NAT
relate add
stop_rel
%.nodecov%.STP
info
%.nodecov%-ID
%.nodecov%-ID
linear
rw
[unquote ' ']


/* start a map composition
/* comment out the "killmap" if this is 1st time running this AML
killmap path1.map
map path1.map


/* set up map environment
pagesize 8.5 11
mape %.nodecov%
maplimits 1.25 1 7.25 10
mapposition cen cen
linecolor orange


/* draw the street network coverage
arcs %.nodecov%


/* begin pathfinding
&sv route_id = 1

&do &until %route_id% eq ( %number% + 1 )
mbegin

 /* specify network coverage and stop file
 netcover %.nodecov% s%route_id%
 stops %.nodecov%.stp in_order route_id stop_imp ~
                transfer # cum_imp cum_trans

 /* select the pair of nodes for pathfinding
 reselect %.nodecov%.stp info ROUTE_ID = %route_id%

 /* it takes 30 seconds to find one route; so be patient!
 &type ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 &type
 &type 'I''m working on pathfinding. Please don''t log me out!'
 &type
 &type ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

 /* find the path and draw on screen
 path stops
```

```
 aselect %.nodecov%.stp info
 routelines %.nodecov% s%route_id% 4

 /* report the execution status
 &type ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 &type
 &type ROUTE [value route_id] is done!
 &type [calc ( %number% - %route_id% )] more to go. PLEASE, WAIT!
 &type
 &type ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

 /* update the stop info table
 &sv route_id = %route_id% + 1


 /* call Subroutine 2
 /************************
 &r route_id.aml %route_id%
 /************************

mend
&end
/* end pathfinding


/* draw the center of the city
mbegin
 markerset oilgas.mrk
 markersymbol 228
 markercolor red
 markersize .4
 reselect %.nodecov% nodes %.nodecov%# = 4772
 nodes %.nodecov%
 aselect %.nodecov% nodes
mend

map end

/* be nice to other users
&type
&type ********************************************
&type PATH FINDING AML COMPLETED!!!
&type ********************************************
&type
&type You may log me out if you need this machine.
&type --------------------------------------------

&return

/* turn off the timer
&pt &off


/*----------------- End of the Main AML: path.aml ------------------
```

```
                /***************/
                /* Subroutine 1 */
                /***************/

/*----- Save the following in a separate file named: stopfile.aml ------

/**********************************************************************/
/*
/* File:          /mit/crldata/lijian/stopfile.aml
/*
/* Purpose:       This aml defines a stop INFO table and adds data to it.
/*                The table is then used by "path.aml" for pathfinding.
/*
/* Files called: stopfile.tab /* an example of this file is shown below
/*
/* Called by: path.aml
/*
/* To execute: [Arcplot] &r stopfile <NODE_COVERAGE>
/*
/**********************************************************************/


&DATA ARC INFO
ARC

/* comment out the next 3 lines if this AML is executed for the 1st time
SELECT %.nodecov%.STP
ERASE %.nodecov%.STP
Y

DEFINE %.nodecov%.STP
%.nodecov%-ID,7,7,I
IN_ORDER,7,7,I
ROUTE_ID,7,7,I
STOP_IMP,7,7,I
TRANSFER,7,7,I
DISTANCE,8,16,F,2

ADD FROM /mit/crldata/lijian/stopfile.tab


/**********************************************************************
/* note on how to create a data file "stopfile.tab:"
/*
/* %.nodecov%-ID has to be the node id that exists in the %.nodecov%.NAT;
/* ID comes from the results of address matching to the street network
/* coverage; each successful matched point of housing units is then
/* converted to a node in the street network coverage; the conversion
/* process assigns each housing record a node ID; then, the IDs are
/* exported to create the tab file;
/* set IN_ORDER equal 1 for the first record and 2 for all other records;
/* set ROUTE_ID equal 1 for the first record and from 1 to the total
```

```
/* number of housing units for all other records;
/* set STOP_IMP, TRANSFER, and DISTANCE equal 0 for all.
/*
/*
/* an example of "stopfile.tab:" with 10 nodes
/* (locations of housing units)
/* the first record represents the center of the city
/*
/*
/* 4807,2,1,0,0,0
/* 2954,2,1,0,0,0
/* 3313,2,2,0,0,0
/* 4352,2,3,0,0,0
/* 4614,2,4,0,0,0
/* 5427,2,5,0,0,0
/* 6219,2,6,0,0,0
/* 6364,2,7,0,0,0
/* 9955,2,8,0,0,0
/* 10173,2,9,0,0,0
/* 10648,2,10,0,0,0
/*
/**********************************************************************


/* list a few to check whether the data file has be added successfully
LIST 1,4 BOS_STN1-ID, IN_ORDER, ROUTE_ID, DISTANCE


Q STOP
&END
&return


/*----------------- End of Subroutine 1: stopfile.aml ------------------



                /****************/
                /* Subroutine 2 */
                /****************/

/*------ Save the following in a separate file named: route_id.aml ------

/*********************************************************************/
/* File: /mit/crldata/lijian/route_id.aml
/*
/* Purpose: This aml updates the value of ROUTE_ID
/* in the info table: %.nodecov%.STP
/* and save the computed DISTANCE.
/*
/* Called by: path.aml
/*
/*********************************************************************/
```

```
/* take one argument from terminal
&args route_id

&DATA ARC INFO
ARC

/* change ROUTE_ID in %.nodecov%.stp table
SELECT %.nodecov%.STP
UPDATE
1
ROUTE_ID=%route_id%


/* copy the computed impedance to the DISTANCE variable
&sv copydis = %route_id% - 1
RESELECT FOR ROUTE_ID EQ %copydis%
CALCULATE DISTANCE = CUM_IMP / 5280


/* display result on screen
&type Path = [value DISTANCE] miles.
ASELECT


/* kill the created route system to save storage space
SELECT %.nodecov%.RATS%copydis%
ERASE %.nodecov%.RATS%copydis%
Y

SELECT %.nodecov%.SECS%copydis%
ERASE %.nodecov%.SECS%copydis%
Y

Q STOP
&END

&return


/*----------------- End of Subroutine 2: route_id.aml ------------------
```

References

[1] Allen, Kathleen M. S., Stanton W. Green, and Ezra B. W. Zubrow. 1990. Interpreting Space: GIS and Archaeology. London: Taylor & Francis.

[2] Alonso, W. 1960. "A Theory of the Urban Land Market," in Papers and Proceedings, Regional Science Association, 6, p. 149-177.

[3] Alonso, W. 1964. Location and Land Use, Harvard University Press, Cambridge, Massachusetts.

[4] Alonso, William. 1975. "Location Theory," Regional Policy, edited by J. Friedmann and W. Alonso. Cambridge, MA: The MIT Press, p. 35-63.

[5] Anas, Alex and Sung Jick Eum. 1984. "Hedonic Analysis of A Housing Market in Disequilibrium," Journal of Urban Economics. Vol. 15: 87-106.

[6] Anas, Alex and Sung Jick Eum. 1986. "Disequilibrium Models of Single-Family Housing Prices and Transactions: The Case of Chicago, 1972-1976," Journal of Urban Economics. Vol. 20: 75-96.

[7] Anderson, Dean R. 1990. "The Application of GIS in the Spatial Analysis of Crime" in Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p.143-152.

[8] Antenucci, John C. et. al. 1991. GIS: A Guide to the Technology. New York: VNR.

[9] Arnott, R. and F. Lewis. 1979. "The Transition of Land to Urban Use," Journal of Political Economy, Feb.

[10] Bailey, Martin J., Richard F. Muth and Hugh O. Nourse. 1963. "A Regression Method for Real Estate Price Index Construction." Journal of the American Statistical Association. Vol. 58: 933-942.

[11] Ball, Michael J., 1973. "Recent Empirical Work on the Determinants of Relative House Prices." Urban Studies 10 (June): pp. 213-231.

[12] Bartik, T. J. 1987. "The Estimation of Demand Parameters in Hedonic Price Models," Journal of Political Economy. Vol. 95: 81-95.

[13] Batty, Michael and P. Longley. 1985. "The Practical Simulation of Urban Structure," Papers in Planning Research, No. 92. Department of Town Planning, University of Wales Institute of Science and Technology, Wales, UK.

[14] Batty, Michael. 1980. "A Perspective on Urban System Analysis," Papers in Planning Research, No. 4. Department of Town Planning, University of Wales Institute of Science and Technology, Wales, UK.

[15] Batty, Michael. 1980. "Computers and Design: The New Challenge to Urban Planning," Papers in Planning Research, No. 2. Department of Town Planning, University of Wales Institute of Science and Technology, Wales, UK.

[16] Batty, Michael. 1983. "Technical Issues in Urban Model Development," Papers in Planning Research, No. 66. Department of Town Planning, University of Wales Institute of Science and Technology, Wales, UK.

[17] Batty, Michael. 1985. "The Spatial Impact of Computer Culture," Papers in Planning Research, No. 91. Department of Town Planning, University of Wales Institute of Science and Technology, Wales, UK.

[18] Bendavid-Val, Avrom, 1983. Regional and Local Economic Analysis for Practitioners. New and Expanded Edition. New York, NY: Praeger Publishers, pp. 67-78.

[19] Blomquist, G. and L. Worley. 1981. "Hedonic Prices, Demands for Urban Housing Amenities, and Benefits," Journal of Urban Economics, Vol. 9: 212-221.

[20] Box, F. E. P. and D. R. Cox. 1964. "An Analysis of Transformations," Journal of the Royal Statistical Society. Series B, Vol. 26: 211-252.

[21] Bradbury, Katherine, Karl Case and Constance Dunham. 1989. "Geographic Patterns of Mortgage Lending in Boston, 1982 to 1987," The New England Economic Review. September/October. p. 3-30.

[22] Brueckner, Jan K. and Peter F. Colwell. 1983. "A Spatial Model of Housing Attributes: Theory and Evidence," Land Economics. Vol. 59: 58-69.

[23] Burrough, R. A. 1986. Principles of Geographical Information System for Land Resources Assessment. New York: Clarendon Press.

[24] Case, Karl E. 1986. "The Market for Single-Family Homes in the Boston Area." The New England Economic Review. (May/June), p. 38-48.

[25] Case, Karl E. and Robert J. Shiller. 1987. "Prices of Single Family Since 1970: New Indexes for Four Cities," The New England Economic Review. September/ October. p. 45-56.

[26] Chen, Lijian. 1993. "Modeling Housing and Demographic Diversity at Censustract Versus Block Group Levels of Aggregation." Proceedings of Urban and Regional Information System Association (URISA), Vol. III, p. 55-72, Atlanta, Georgia, July 26-29.

[27] Chinloy, Peter T. 1977. "Hedonic Price and Depreciation Indexes for Residential Housing," Journal of Urban Economics. Vol. 4: 469-482.

[28] Chow, Gregory C. 1960. "Test of Equality between Sets of Coefficients in Two Linear Regressions," Econometrica. July Vol. 28: 591-605.

[29] Clark, W.A.V. and W.F. J. Van Larrup. 1986. "Residential Mobility and Household Location Modeling," Handbook of Regional and Urban Economics, Vol. 1. Regional Economics, Chapter 3, p. 97-132.

[30] Clarke, Martin. 1990. "Geographical Information Systems and Model Based Analysis: Towards Effective Decision Support Systems," Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p.165-176.

[31] Cooke, Donald F. 1991. "The Invisible Map: Computers and The Death of Cartography." The Second Annual Meeting of the Northeastern Map Organization. June 8, New London, New Hampshire.

[32] Courant, Paul N. 1977. "On Models of Racial Prejudice and Urban Residential Structure," Journal of Urban Economics, Vol. 4: 272-291.

[33] Crecine, John, Otto Davis, and John Jackson. 1967. "Urban Property Markets: Some Empirical Results and Their Implications for Municipal Zoning." Journal of Law and Economics, Vol. 10: 79-100.

[34] Dubin, Robib A. and Chein-Hsing Sung. 1990. "Specification of Hedonic Regressions: Non-nested Tests on Measures of Neighborhood Quality," Journal of Urban Economics, Vol. 27: 97-110.

[35] Duncan, Otis D. and Beverly Duncan, 1955. "A Methodological Analysis of Segregation Indexed," American Sociological Review, April, Vol. 20: 210-217.

[36] Edmonds, R. G. 1984. "A Theoretical Basis for Hedonic Regression: A Research Primer," AREUEA Journal, Vol. 12, No. 1: 72-85.

[37] Edmonds, R. G.?. "Travel Time Valuation Through Hedonic Regression," Southern Economic Journal, Vol. 50: 83-98.

[38] Engle, Robert F., D. M. Lilien and M. Watson. 1985. "A Dynamic Model of Housing Price Determination." Journal of Econometrics. Vol. 28: 307-326.

[39] Environmental Systems Institute, Inc. 1990. Understanding GIS: The ARC/INFO Method. Redlands, CA.

[40] Epple, Dennis. 1987. "Hedonic Price and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," Journal of Political Economy. Vol. 95: 59-80.

[41] Ferreria, Joseph, Jr. and L. L. Wiggins. 1990. "The Density Dial: A Visualization Tool for Thematic Mapping," Geo Info System, November/December 1990, p. 69-71.

[42] Fisch, Oscar. 1977. "Dynamics of the Housing Market," Journal of Urban Economics, Vol. 4: 428-447.

[43] Fisher, Franklin M. 1970. "Tests of Equality between Sets of Coefficients in Two Linear Regressions: An Expository Note," Econometrica. Mar. Vol. 38: 361-366.

[44] Freeman, A. M. III. 1979. "Hedonic Prices, Property Values and Measuring Environmental Benefits: A Survey of the Issues." Scandinavian Journal of Economics.

[45] French, S. P. and L. L. Wiggins. 1990. "Planning Agency Experiences with Automated Mapping and Geographic Information Systems," Environment and Planning B, Vol. 17: 441-450.

[46] Geertman Stan C.M. and Fred J. Toppen. 1990. "Regional Planning For New Housing In Randstad" in Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p. 95-106.

[47] Giannias, Dimitrios A. 1991. "Housing Quality Differentials in Urban Areas," Journal of Urban Economics, Vol. 29: 166-181.

[48] Gini, Corrado, 1913. "Sulla misure della concentration e della variabilita dei carattere" ATTI DEL REALA INSTITUTO VENETO DI SCIENCE, LETTERE ED ARTI, Tome 73, pt2 (1913-1914), 1207ff.

[49] Glaster, G. C. 1982. "Black and White Preferences for Neighborhood Racial Composition," AREUEA Journal. Vol. 10:39-66.

[50] Goodchild, M. and Sucharita Gopal. 1989. The Accuracy of Spatial Databases. New York: Taylor & Francis.

[51] Goodman, A. C. 1977. "A Comparison of Block Group and Census Tract Data in a Hedonic Housing Price Model," Land Economics, Vol. 53, No. 4: 483-487.

[52] Goodman, Allen C. 1978. "Hedonic Prices, Price Indices and Housing Markets," Journal of Urban Economics, Vol. 5: 471-484.

[53] Greenberger, M, et. al. 1976. Models in the Policy Process: Public Decision Making in the Computer Era. New York: Russell Sage Foundation.

[54] Greenlees, J. S. 1982. "An Empirical Evaluation of the CPI Home Purchase Index, 1973-1978," AREUEA Journal, Vol. 10: 1-24.

[55] Grether D. M. and P. Mieszkowski. 1980. "The Effects of Nonresidential Land Uses on the Prices of Adjacent Housing: Some Estimates of Proximity Effects." Journal of Urban Economics. January, Vol. 8: 1-15.

[56] Griliches, Z. 1961. "Hedonic Price Indices for Automobiles: An Econometric Analysis of Quality Change," in Models of Income Determination, Studies in Income and Wealth. Vol. 28. Columbia University Press for NBER, New York.

[57] Griliches, Z. 1967. "Hedonic Price Indexes Revisited: Some Notes on the State of the Art," 1967 Business and Economics Statistics section, Proceedings of the American Statistical Association.

[58] Griliches, Z. 1971. Price Indexes and Quality Change: Studies in New Methods of Measurement. Harvard University Press, Cambridge, Massachusetts.

[59] Harris, Briton. 1989. "Beyond Geographical Information Systems: Computers and the Planning Professional," APA Journal. Winter, p. 85-90.

[60] Harris, R. N. S., G. S. Tolley, and C. Harrell. 1968. "The Residence Site Choice," Review of Economics and Statistics, Vol. 50: 241-247.

[61] Harrison, D. and D. L. Rubinfeld. 1978. "Hedonic Housing Prices and the Demand for Clean Air," Journal of Environmental Economics and Management. Vol. 5: 81-102.

[62] Healey, M. J. 1991. "Improving the Local and Regional Economy Information Base" in the Policy Review Section, Regional Studies, Vol. 25:4, p. 363-369.

[63] Henk, J. S. and C.H. Stillwell. 1989. "Geographical Information Systems: the Emerging Requirements," Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p.3-14.

[64] Herzog, Henry W., and Richard Olsen. 1977. "Shift-Share Analysis Revisited: The Allocation Effect and Stability of Regional Structure," Journal of Regional Science. Vol. 17, No. 3, p. 441-454.

[65] Henry, Mark S. 1991. "Modeling Regional Economic Change with the Aid of GIS: Authentic Regional Science?" The Review of Regional Studies. Vol. 21, No. 2, Summer.

[66] Hoverson R. and H. O. Pollakowski. 1981. "Choice of Functional Forms for Hedonic Price Equations" Journal of Urban Economics. Vol. 10: 37-49.

[67] Huxhold, William E. 1991. An Introduction to Urban Geographical Information Systems. Oxford University Press.

[68] Janssen, Ron and Piet Rietveld. 1990. "Multicriteria Analysis and GIS: An Application to Agricultural Land Use in the Netherlands" in Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p.129-139.

[69] Johnson, David Dale. 1982. "An Alternative Approach to Housing Market Segmentation Using Hedonic Price Data," Journal of Urban Economics, Vol. 11: 311-332.

[70] Jud, G. 1980. "The Effects of Zoning on Single-family Residential Property Values: Charlotte, North Carolina," Land Economics, May, Vol. 56: 142-154.

[71] Kain, John F. and John M. Quigley. 1975. Housing Markets and Racial Discrimination: A Microeconomic Analysis. New York: National Bureau of Economic Research, 1975.

[72] Kain, John F. and John M. Quigley. 1972. "Note on the Owner's Estimate of House Value." Journal of the American Statistical Association. Vol. 67: 803-806.

[73] Kain, John F. and John M. Quigley. 1970. "Measuring the Value of Housing Quality." Journal of the American Statistical Association. Vol. 65: 532-548.

[74] Kain, J. F. and J. M. Quigley. 1969. "Evaluating the Quality of the Residential Environment," Environment and Planning, Vol. 2.

[75] Kennedy, Peter. 1985. A Guide to Econometrics. 2nd ed. Cambridge, Mass.: MIT Press.

[76] King, Thomas K. and P. Mieszkowski. 1973. "Racial Discrimination, Segregation, and the Relative Price of Housing," Journal of Political Economy. Vol. 81: 590-606.

[77] Krugman, P. 1991. Geography and Trade. MIT Press, Cambridge, Massachusetts.

[78] Lancaster, Kelvin. 1966. "A New Approach to Consumer Theory," Journal of Political Economy. April. Vol. 74: 132-157.

[79] Lapham, Victoria. 1971. "Do Blacks Pay More for Housing," Journal of Political Economy. Vol. 79: 1244-1257.

[80] Landis, J. et. al. 1991. "BASS II: Design and Development Notes" Institute of Urban and Regional Development, University of California at Berkeley, (Draft), Paper presented at the Second International Conference on Computers in Urban Planning and Urban Management, Oxford, U.K.

[81] Linneman, P. 1982. "Hedonic Prices and Residential Location," in The Economics of Urban Amenities, edited by D. B. Diamond and G. S. Tolley. p. 69-88. Academic Press.

[82] Little, James T. 1976. "Residential Preferences, Neighborhood Filtering and Neighborhood Change," Journal of Political Economy. Vol. 3: 68-81.

[83] Lowry, I. 1964. A Model of Metropolis. Santa Monica, Calif.: Rand Corporation.

[84] Lowry, I. 1967. Seven Models of Urban Development: A Structural Comparison, RAND.

[85] Marble, Duane F. 1990. "The Potential Methodological Impact of Geographical Information Systems on the Social Sciences," in Interpreting Space: GIS and Archaeology, edited by Kathleen M. S. Allen, Stanton W. Green and Ezra B. W. Zubrow. London: Taylor & Francis. p. 9-21.

[86] Marble, F. D. 1991. "Book Review" in Geographical Analysis. Oct. Vol.23, No. 4.

[87] Mark, J. H. and M. A. Goldberg. 1984. "Alternative Housing Price Indices: An Evaluation," AREUEA Journal, Vol. 12, No. 1: 30-49.

[88] Mark, J. H. and M. A. Goldberg. 1986. "A Study of the Impact of Zoning on Housing Values Over Time," Journal of Urban Economics. Vol. 20: 257-273.

[89] McCarty, H. H., et al. 1956. The Measurement of Association in Industrial Geography. Connecticut: Greenwood Press, Publishers.

[90] McFarlan, F. W. 1984. "IT Changes the Way You Compete," Harvard Business Review, No. 3: 98-103.

[91] Michaels, R. Gregory and V. Kerry Smith. 1990. "Market Segmentation and Valuing Amenities with Hedonic Models: The Case of Hazardous Waste Sites," Journal of Urban Economics, Vol. 28: 223-242.

[92] Mills, E. S. 1979. "Economic Analysis of Urban Land Use Controls" in Current Issues in Urban Economics, edited by P. Mieszkowski and M. Straszheim. John Hopkins University Press, Baltimore. p. 511-541.

[93] Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area," American Economic Review, May. p. 197-211.

[94] Murray, M. P. 1978. "Hedonic Prices and Composite Commodities," Journal of Urban Economics. Vol. 5: 188-197.

[95] Muth, Richard F. 1969. City and Housing: The Spatial Pattern of Urban Residential Land Use. Chicago and London: The University of Chicago Press.

[96] Nelson, J. P. 1977. "Accessibility and the Value of Time in Commuting," Southern Economic Journal, p. 1321-1329.

[97] Nelson, J. P. 1978. "Residential Choice, Hedonic Prices and the Demand for Urban Air Quality," Journal of Urban Economics. Vol. 5: 357-369.

[98] Nijkamp, P., R. Rietveld and F. Snickars. 1986. "Regional and Multiregional Economic Models: A Survey," Handbook of Regional and Urban Economics, Vol. 1. Regional Economics, Chapter 7, p. 257-294.

[99] Olsen, Edgar O. 1969. "A Competitive Theory of the Housing Market," American Economic Review, Vol. 59. p. 612-622.

[100] Olsen, Edgar O. 1987. "The Demand and Supply of Housing Services: A Critical Survey of the Empirical Literature," in E. Mills (ed.) Handbook of Regional and Urban Economics, Vol. 2, Urban Economics, Chapter 25, Rotterdam, Elxevier.

[101] Palmquist, Raymond. B. 1979. "Hedonic Price and Depreciation Indexes for Residential Housing: A Comment," Journal of Urban Economics. Vol. 6: 267-271.

[102] Palmquist, Raymond B. 1980. "Alternative Techniques for Developing Real Estate Price Indexes," Review of Economics and Statistics. Vol. 62: 442-480.

[103] Palmquist, Raymond. B. 1982. "Measuring Environmental Effects on Property Value Without Hedonic Regressions," Journal of Urban Economics. Vol. 11: 333-347.

[104] Palmquist, Raymond B. 1984. "Estimating The Demand for the Characteristics of Housing," Review of Economics and Statistics. Vol. 66: 394-404.

[105] Parsons, George R. 1990. "Hedonic Prices and Public Goods: An Argument for Weighting Locational Attributes in Hedonic Regressions by Lot Size," Journal of Urban Economics. Vol. 27: 308-321.

[106] Pindyck, Robert S. and Daniel L. Rubinfeld. 1991. Econometric Models and Economic Forecasts. 3rd edition, New York: McGraw-Hill, Inc.

[107] Porter, M. F. and V. E. Millar. 1985. "How Information is Transforming the Nature of Competition," Harvard Business Review, NO. 5, p. 149-160.

[108] Proceedings. 1991. The Second International Conference on Computers in Urban Planning and Urban Management, Oxford, U.K.

[109] Proceedings of Urban and Regional Information System Association (URISA) Conference. 1989, 1990, 1991, and 1992.

[110] Rao, Potluri and Roger Miller. 1971. Applied Econometrics. Belmont, California: Wadsworth Publishing Co. Inc.

[111] Richardson, Harry, W. 1979. Regional Economics. Urbana, IL: University of Illinois Press.

[112] Ridker, Ronald and John A. Henning. 1967. "The Determinants of Residential Property Values with Special Preference to Air Pollution," The Review of Economics and Statistics. Vol. 49: 246-257.

[113] Rosen, Sherwin. 1974. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," Journal of Political Economy. Vol. 82: 34-55.

[114] Rothenberg, Jerome et al. 1991. The Maze of Urban Housing Markets: Theory, Evidence, and Policy. Chicago: University of Chicago Press.

[115] Schaller Jorg. 1990. "GIS Applications in Environmental Impact Assessment" in Geographical Information Systems for Urban and Regional Planning, edited by Henk J. Scholten and John C. H. Stillwell, p. 107-117.

[116] Schnare, Ann B. and Raymond J. Struyk. 1976. "Segmentation in Urban Housing Markets," Journal of Urban Economics. Vol. 3: 146-166.

[117] Scholten, Henk J. and J. C. H. Stillwell. 1990. Geographical Information Systems for Urban and Regional Planning. Boston: Kluwer Academic Publishers.

[118] Smith, B. A. 1982. "Racial Composition as a Neighborhood Amenity," in The Economics of Urban Amenities edited by D. B. Diamond and G. S. Tolley. p. 165-181. Academic Press.

[119] Straszheim, Mahlon R. 1975. A Econometric Analysis of the Urban Housing Market. New York: National Bureau of Economic Research.

[120] Sumka, Howard J. 1977. "Measuring the Quality of Housing: An Econometric Analysis of Tax Appraisal Records," Land Economics, No. 3: 298-309.

[121] Weicher, J. and R. Zerbst. 1973. "The Externalities of Neighborhood Parks: An Empirical Examination," Land Economics, Feb. Vol. 49: 99-105.

[122] Wheaton, W. 1979. "Monocentric Models of Urban Land Use: Contributions and Criticism," Journal of Urban Economics, July.

[123] Wheaton, William C. 1977. "A Bid Rent Approach to Housing Demand," Journal of Urban Economics, Vol. 4: 200-217.

[124] Wiggins, Lyna L. et. al. 1990. Expert Systems: Applications to Urban Planning, New York: Springer-Verlag.

[125] Wiggins, Lyna L. and S. P. French. 1991. "Geographic Information Systems: Assessing Your Needs and Choosing a System," Planning Advisory Service Report, American Planning Association.

[126] Wiggins, Lyna L. and J. Ferreira. 1992. "MIT's Computer Resource Lab: A Research and Educational Facility for GIS in Urban Planning," International Journal of Geographic Information Systems.

[127] Wright, Jeff R., L. L. Wiggins, R. Jain, and T. J. Kim. 1992. Expert Systems in Environmental Planning. New York: Springer-Verlag.

[128] Wrigley, N. and F. Brouwer. 1986. "Qualitative Statistical Models for Regional Economic Analysis," Handbook of Regional and Urban Economics, Vol. 1. Regional Economics, Chapter 11, p. 443-492.

[129] Zubrow, Ezra B. 1990. "Contemplating Space: A Commentary on Theory" in Interpreting Space: GIS and Archaeology, edited by Kathleen M. S. Allen, Stanton W. Green and Ezra B. W. Zubrow. London: Taylor & Francis. p. 68-72.