# Location, Location, Location Choice Models

By
Pablo Posada Mariño

B.Sc. in Civil Engineering
B.Architecture
Universidad de Los Andes
Bogotá, Colombia (2009)

Submitted to the Engineering Systems Division and the
Department of Urban Studies and Planning in partial fulfillment of
the requirements for the degrees of

Master of Science in Engineering Systems
Master in City Planning

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 2015


Author Signature redacted
Engineering Systems Division
Department of Urban Studies and Planning
May 20, 2015

Certified by Signature redacted
Associate Professor P. Christopher Zegras
Department of Urban Studies and Planning
Thesis Supervisor

Accepted by Signature redacted
Munther A. Dahleh
William A. Coolidge Professor of Electrical Engineering and Computer Science
Acting Director, Engineering Systems Division

Accepted by Signature redacted
Professor Dennis Frenchman
Chair, MCP Committee
Department of Urban Studies and Planning

**Location, Location, Location Choice Model**

By
Pablo Posada Mariño

Submitted to the Engineering Systems Division and the Department of Urban Studies and Planning on May 21, 2015 in partial fulfillment of the requirements for the degrees of Master of Science in Engineering Systems and Master in City Planning

## Abstract

Cities are, now more than ever before, the main centers of population and production. The growing demand for limited urban space is increasing urban complexity and magnifying both positive and negative externalities of urban agglomeration: increasing productivity, innovation, and social interaction, but also exacerbating living costs, pollution, inequality, congestion, etc. In order to build sustainable cities and have a net positive balance of urban externalities, we need to better understand the motivations of the different agents competing in the *race for urban space*. Location choice models can help to shine a light on these motivations by providing insights on agents' location preferences. They are also the building blocks of more comprehensive urban models and simulations that can help navigate urban complexity. This thesis explores location choice models for homeowner households and firms in Greater Boston. Specific research questions that these models can help answer include: How do residential location preferences vary with life cycles? What industries value clustering the most? These topics are important given (1) forecasted demographic changes, specifically the aging of the baby-boomers, and (2) the continuing move from a manufacturing-based economy to a service and knowledge-based economy. These changes in population and economy will likely require a change in housing stock in order to better match supply with demand, and changes in the stock of commercial space in order to continue boosting the firms that drive the economy of the region. The thesis also explores the data-related uncertainty of these models (how model estimation changes with different data sources) as well as their temporal transferability (how do preferences change over time). The location choice analysis for households suggests that income has a bigger impact on willingness to pay for location attributes than age of the head of the household or household size. The firm analysis indicates that firms in the professional service and health and education service sector place more value on proximity to jobs in the same industry and density than firms in other sectors. These preferences have strengthened over time. An in-depth analysis, such as the one presented in this thesis, of what city agents look for in a location can, and should, inform planning policies and intervention in order to better match location preferences with opportunities.

Thesis Supervisor: P. Christopher Zegras
Tittle: Associate Professor, Transportation & Urban Planning
Dept. of Urban Studies & Planning, Massachusetts Institute of Technology

Thesis Reader: Victor Rocco
Tittle: Postdoc Dept. of Urban Studies & Planning, Massachusetts Institute of Technology

**Acknowledgments**

I would like to thanks my thesis advisor Professor P. Christopher Zegras, my thesis reader Victor Rocco, Professor Mikel E Murga, and the "Uncertainty" research team (Michael Dowd, Yafei Han, Menghan Li, and Shenghao Wang) for their help, support, generosity, and inspiring work during the process of writing this thesis.

I would also like to extend my deepest gratitude to Professor P. Christopher Zegras for his generous support and the wonderful opportunities he has provided me during my time at MIT.

Finally, I would like to thank my parents, my brother, and Sara for their invaluable love, support, encouragement, and patience.

(This research was supported by MIT and Masdar Institute Cooperative Program)

**Table of Content**

## List of Tables

**List of Figures**

# 1. INTRODUCTION

## 1.1. Motivation

According to the United Nations (United Nations, 2011), in 2010, the population living in cities surpassed the population living in rural areas for the first time in history. This trend is expected to continue. Cities are also the main centers of production. The convergence of multiple and diverse activities at an unprecedented scale often exceeds the capacity of local governments to respond in terms of regulation and provision of services. The competition for space and the increase in complexity magnify both the positive and negative externalities of urban agglomerations, resulting in an increase in productivity, innovation, and in social interaction, but also rising living costs, pollution, inequality, congestion, etc. It is difficult to strike a balance.

Understanding the dynamics within urban systems and the motivations of the different urban agents competing in the *race for urban space* is key to achieving balanced, sustainable cities. A better understanding of the complexity of urban interactions is especially important for achieving effective and efficient urban interventions (e.g. projects and/or policies).

This thesis aims to help untangle urban complexity by analyzing a key component of the urban system: the location preferences of households and firms. Where these two types of urban agents chose to locate, subject to certain constraints (physical, historical, regulatory, market, etc.), underlays the spatial distribution of activities across cities.

Location choice models, based on discrete choice theory, can help decision-makers understand the different trade-offs urban agents make in their process of choosing where to locate. That is, what do urban agents value most when choosing a location? And, equally important given the competitive nature of urban real estate markets, which agent would be willing to pay the most for a certain location characteristic? Cities face variations of these questions all the time: *how do we attract a talented work force? How do we boost the knowledge economy? How do we make successful housing projects? How do we promote diversity and inclusion?* The answers to these questions are not straightforward. Answers based on superficial analysis of observed patterns may lead to wrong decisions.

In addition to providing insights on these issues, location choice models are also the building blocks of more comprehensive urban models that simulate the interactions between agents in a city. These models have the potential to help decision-makers understand and navigate urban complexity and make more informed decisions.

## 1.2. Thesis Objective and Outline

In this thesis I aim (1) to understand the location preferences of households and firms in the Greater Boston area through discrete choice models, and (2) to explore the uncertainty around discrete-choice-based location model estimation.

The research question about residential location preferences are motivated by the claims made in The Greater Boston Housing Report Card 2014-2015[1]:

> *"...Greater Boston is not only experiencing a serious housing shortage, but also an escalating mismatch between the type of available housing and the type of housing most desired by its two fastest growing demographic clusters: aging baby boomers and young millennials. With the metro economy robust and growing, the local housing market is increasingly "out of sync" with demand. As a result, where young millennials are making due by doubling up and tripling up in multi- unit housing in Boston and its nearby communities, working families for which such housing was originally built are being squeezed out. Many aging baby boomers are seeking smaller housing units, but finding it difficult to locate such units at affordable prices in the communities where they have lived for much of their adult lives."*

In this thesis I focus on the questions of:

- How do location preferences vary with the household life cycle?
- Do senior households with no children prefer smaller units compare to younger households?

On the topic of firm's location preferences, the thesis delves into the role of accessibility, clustering, and agglomeration economies. Specific research questions are:

- What industries value clustering the most? (accessibility between firms).
- What industries value being close to customers the most? (accessibility to consumers )
- Have these preferences changed over time? If so, in what way?

---

1 The Grater Boston Housing Report Card 2014-2015 Fixing an Out-of-Sync Housing Market , The Kitty and Michael Dukakis Center for Urban and Regional Policy Northeastern University (2015).

On model estimation uncertainty, I am to answer the following specific research questions:

- Do the model estimation results vary significantly if using different data sources? (data-related uncertainty)
- If location preferences change, how transferable are these models over time?

The thesis is comprised of 5 chapters including this introduction. Chapter 2 reviews the theories of location choice, the methods to tackle the specific research questions, and the empirical context of the thesis research. Chapter 3 presents the analysis of residential location preferences, with a review of historic population dynamics in Greater Boston and the estimation of residential location choice models. Chapter 4 presents the analysis of firm location preferences, with a review of employment evolution in Greater Boston and estimation of firm location choice models. Chapter 6 synthetizes the main research findings, limitations, and possible next steps

# 2. LOCATION CHOICE: THEORY, METHODS, AND CONTEXT

This chapter presents a review of the theory and empirical evidence of the location behavior of households and firms as well as analytical tools and methods to model these behaviors. It finishes by presenting the proposed approach to analyze location behavior in Greater Boston based on the methods presented in earlier in the chapter and the data available.

## 2.1. Theory: location behavior of households and firms

Residential Location, Lifecycle, and Lifestyle

According to Rossi (1955), households' decision to relocate is a function of push and pull factors. Push factors are the ones that make a location become inadequate for a household. These factors often refer to changes in the structure of the household associated with lifecycles stages, such as getting married or having a child. The different stages in lifecycles are in turn associated with corresponding consumption patterns for location attributes (Clark et al, 2006). For example, Ström (2010) identifies a positive relationship between homeownership and the number of rooms with the birth of the first child, but no relationship between these factors and the type of dwelling unit. Other push factors include changes in career or workplace, or change in a household's income level or social status. Clark and Deurloo (2006) suggest that housing upgrades are associated primarily with increased wealth and find little evidence that older couples who had purchased large houses move to smaller higher-density housing after their children leave the nest. In the same way, budget constraints can prohibit relocation that would otherwise take place due to lifecycle changes. Pull factors refer to elements that attract households to a specific location, such as the quality of the unit or the built space in the neighborhood. Push and pull can work together; for example a household with children entering school age may push location preferences towards areas with better quality schools. Elements in the broad housing markets, such as interest rates or credit availability, can also constitute push or pull factors.

These push and pull factors, as well as changes in lifecycle stages, are associated with changes in lifestyle. The concept of lifestyle is often used in marketing literature to segment consumption behavior (Cahill, 2006). Different lifestyles are then also associated with different location consumption behaviors. Veal (2001) indicates that the concept of lifestyle can provide a general framework for describing clusters of household choices. The directionality of the relationship or causality between lifestyle and location choice, however, is not obvious. Evidence suggests that households self-select residential location based on their lifestyle preferences, which might

include preferred travel patterns (Mokhtarian and Cao, 2008; Van Wee, 2009). But living at a particular location can also result in the adoption of a particular lifestyle, which then influences decision-making and behavior. Schwanen and Mokhtarian (2005) suggest that households change their lifestyles in response to environmental conditions.

The concept of lifestyle is also vague and difficult to operationalize. Different population groups can have similar lifestyles regarding some specific elements, but may have different behaviors on other aspects. Grouping individual households is not easy and, sometimes may not even be possible in some cases. For example, *knowledge workers* is a concept that has gained growing interest in the last decade. It refers to a segment of the populations associated with the knowledge-based economy.

Recent studies established the relationship between knowledge-workers and economic growth in the Netherlands (Raspe and Van Oort, 2006 and Van Oort et al., 2009), Germany (Wedemeier, 2010) and the U.S. (McGranahan and Wojan (2007). Interest in the location preferences of knowledge workers has grown as cities move (or hope to move) to a knowledge-based economy. Florida (2002) concluded that knowledge workers desire amenities different from traditional ones. Yigitcanlar et al. (2007) later listed location preferences of knowledge-workers, which include elements such as proximity to retail and performance arts. Tomaney and Bradley (2007) examined the preferences of knowledge workers residing in the niche market of top-end housing in gated communities in North-East England and observed that they valued housing size, property value as investment, rural feel, and sense of personal security. Lawton et al. (2013) showed that dwelling size and cost and distance to work were the most relevant factors in knowledge workers' residential choice in Dublin, and that young knowledge-workers did not necessarily prefer the metropolitan core. Frenkel, Bendit, Kapla (2013) modeled housing choices of 833 knowledge-workers in high-technology and financial services and analyze the relative importance of lifestyle and cultural amenities in addition to classic location factors. By their estimates, the most important factors are municipal socioeconomic level, housing affordability, and commuting time, while substantial secondary factors are cultural and educational land-uses and culture-oriented lifestyle of the surrounding area. The difference in findings of the location preferences aligns with the conceptualization made by Kunzmann (2009) of knowledge workers as heterogeneous in nature when it comes to preferences, rather than a homogenous group of workers with prototypical needs.

Firm Location and Agglomeration Economies.

The spatial distribution of employment in a metropolitan area is a main feature of a city's structure. Early models of spatial distribution of activities and of urban equilibrium, such as Alonso's monocentric city model, assumed the urban structure as exogenous. Jobs were concentrated in the city center and this structure determined residential land use distribution and rent and density gradients. But urban structure results from a competition for space between firms (and households) with different preferences and subject to different constraints. Subsequent urban equilibrium models have tried to explain more complex urban structures with different degrees of success.

Like households, firms can be assumed to be utility maximizing agents – more specifically, profit maximizing agents, with the choice of where to locate motivated by a desire to maximize profitability. Location-specific factors that can affect a firm's profitability include: the cost of land, transportation costs for production inputs and outputs, labor costs, costs of utilities, and property and income taxes. Some of these factors are in turn a function the location of other agents. That is, relative proximity to other firms may matter, due to economic spillovers (externalities). Positive economic externalities result in "agglomeration" (i.e., agglomeration economies). Spillovers can also be negative (i.e., "agglomeration diseconomies"), such as where relative proximity creates congestion or other crowding of infrastructures and services.

Marshall (1920) introduced the theoretical underpinnings of agglomeration, arguing that clustering helps reduce three types of transport cost for firms: the cost of moving goods, people, and ideas. From these cost reductions arise the three main benefits from agglomeration: facilitation of goods and services trading, labor market pooling, and knowledge spillovers. Others have argued that clustering encourages competitiveness, increases productivity and specialization which in turn increase wages, trigger economies of scale, reduce some business costs, and otherwise allow firms to achieve better outcomes than they would realize in isolation (Krugman, 1991; Fu & Ross, 2013; Gibbs & Bernat, 1997).

Agglomeration economies are then at the heart of firm location choices. A firm location choice model can help understand the value of agglomeration for different types of firms, as well as the value of agglomeration compared to the value of proximity to other factors. In other words, such models can help identify what types of firms place more value in being close to similar firms, or to their workers, or to their customer base.

Most of the literature on firm location deals with inter-city location. Carlton (1979) used a multinomial logit model to analyze interurban location behavior of industries, and Reif (1981) used a similar approach to analyze industrial location in Venezuela. Hansen E. R. (1985) studied the interurban location behavior of 360 manufacturing firms in the state of Sao Paulo, Brazil using a nested multinomial logit model. He found a strong preference for local agglomeration and no evidence of firm sensitivity to wage level. Carlton D. (2001) modeled location and employment choice of new branch plants across metro areas. He found a large effect of energy costs and no major effects of taxes and state incentive programs.

At the intra-urban scale, Lee (1982) modeled the location of manufacturing firms in Bogotá, Colombia, as the outcome of the competition for urban land between firms. He found that small firms place the highest value on accessibility to local input and output markets while large, export-oriented firms value more plant space and quality of public utility services. Shukla and Waddell (1991) examined the intra-metropolitan location decisions of establishments in major industrial categories in the Dallas-Fort Worth area using a location-choice approach. They evaluated the preference of different industries for two main types of location characteristics: (1) Structural Variables, which refer to general accessibility measures relative to the urban structure such as distance to the CBD or distance to the airport, and (2) Agglomeration Variables, which refer to location characteristics within a given radius such as median income or number of people or jobs. They found that wholesale firms value freeway access considerably and retail is almost exclusively locally oriented (they place high value on agglomeration variables), and therefore, the most decentralized. The finance, Insurance and Real Estate industry present the highest preferences for agglomeration and high-income zones. Recently, Baum-Snow (2013) found evidence that agglomeration economies remain an important incentive for firms to cluster spatially in most industries, even in the face of transportation cost reductions. According to his analysis, finance, insurance & real estate exhibits the strongest preference for density and spatial centrality while wholesale & retail trade exhibits the least.

## 2.2. Methods

Hedonic Approach to Location Choice

Modern location choice models are grounded on basic microeconomic theory first developed for households (i.e. residential location). The household's location decision is driven by utility maximization. A given household ($h$) choses the combinations of consumption goods ($x$) at a

price ($p$) and residential location ($i$) with a set of attributes ($z$) that maximizes its utility given a budget constraint:

$$\max_{x,i} U(x_h, z_i) \tag{2.1}$$

The budget constraint implies that the total cost of the consumption goods ($x*p$) plus the cost of the residential location ($r_i$) have to be less than or equal to the household's income ($I$):

$$I_h \geq px_h + r_i \tag{2.2}$$

If the price of a location is a function of its characteristics, then:

$$I_h \geq px_h + H(z_i), \tag{2.3}$$

where $H$ is the hedonic price function.

Rosen (1974) proposed a 2-stage method to model a consumer's choice process. In the first stage, the consumer maximizes her utility function subject to a budget or income constraint with $x_h$ and $z_i$ held fixed. The objective function then becomes:

$$V_h\left(p, z_i, I_h - H(z_i)\right), \tag{2.4}$$

where $V$ is an indirect utility function conditional on a given location; that is, the utility that a consumer can achieve at a price $p$ if she is residing in a location with characteristics $z_i$, and has a budget constraint $I_h$.

In the second stage, the consumer chooses a location with the characteristics that maximize her indirect utility function:

$$\max_{i} V_h(p, z_i, I_h - H(z_i)). \tag{2.5}$$

The first order conditions for a maximum are the derivative of the indirect utility function with respect to location characteristics:

$$\frac{\partial V_h}{\partial z_{ij}} = \frac{\partial V_h}{\partial I_i} \frac{\partial H}{\partial z_{ij}}, \tag{2.6}$$

where j are the different characteristics or attributes of a location. This formulation allows the introduction of the consumer's bid-choice function, $B$, which determines the price a given consumer is willing to pay for a given location with characteristics $z_i$ at a constant utility level $\bar{u}$:

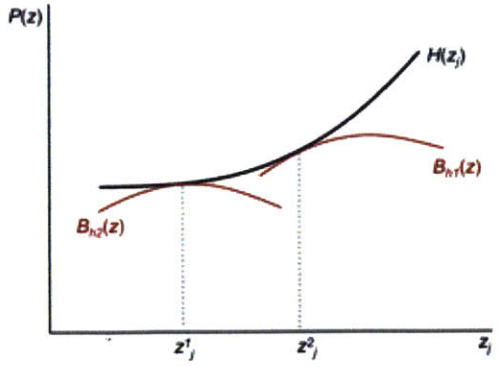
$$B_h(p, z_i, I_h, \bar{u}). \tag{2.7}$$

The change in the bid-choice function with respect to changes in the location characteristics are:

$$\frac{\partial B_h}{\partial z_{ij}} = \frac{\frac{\partial V_h}{\partial z_{ij}}}{\frac{\partial V_h}{\partial I_h}}. \tag{2.8}$$

This provides a direct link between hedonic pricing theory and consumer's Bid-choice. Consumers select the $z_i$ for which its marginal willingness to pay for more of each characteristic $j$ ($B$ curve) is equal to the marginal cost of obtaining that characteristic in the market (Hedonic price function $H$). Figure 1 illustrates this relationship of tangency condition.

**Figure 1 Relationship between the hedonic price function H(z) and bid curves B(z)**




The curve $H$ represents the change in price for a change in characteristic $z_j$, holding other location characteristics constant. Curves $B_{h1}$ and $B_{h2}$ represent the Bid-choice function of households 1 and

2 as a function of changes in characteristic $z_j$. The utility maximizing amount of characteristic $z_j$ for households 1 and 2 are $z^1_j$ and $z^2_j$ respectively.

The Rosen 2-stage method to determine consumers' bid-choices implies specifying consumers' utility functions and a market's hedonic price function and using this information to derive consumer's demand for characteristics. This approach has been criticized for its inability to treat multiple housing characteristics simultaneously (Ellickson, 1979). Additional critiques suggest that the method suffers from simultaneity biases due to the joint determination of the supply (hedonic function) and demand (bid-choice function) and to the nonlinearity of the price function.

An alternative to the two-stage method is to estimate the parameters of the bid function directly using the discrete choice models framework. This alternative, in turn, can be divided into two different approaches: the Price-Choice approach and the Bid-Rent approach.

Price-Choice Approach to Location

Under the price approach (McFadden 1978, Anas 1982) the indirect utility function for a household type h with attributes k (e.g. income, household size, race, etc.) for living in a location type $z$ (which incorporates the characteristics of the housing structure as well as the characteristics of the area in which the structure is located) can be written as:

$$V_h\left(z, H(z)\right). \tag{2.9}$$

The price of goods $p$ and the income $I_h$ have been suppressed given that they are assumed to be the same for all households of type $h$ within a given metropolitan area. The indirect utility can be expressed as a combination of the deterministic indirect utility and a stochastic or random component representing idiosyncratic differences within locations of a same type:

$$V_h\left(z, H(z)\right) + \varepsilon_z = V_{hz} + \varepsilon_z. \tag{2.10}$$

The probability that a household type h will choose a location $z_i$ is the probability that the indirect utility of a household type h at location $z_i$ will be greater than the indirect utilities of the other household types that could choose that location:

$$P(z|h) = Prob\{V_{hz} + \varepsilon_z > V_{hz'} + \varepsilon_{z'}, z' \neq z; z, z' \epsilon K\}, \quad (2.11)$$

where $K$ is the set of all the locations that household type h could choose. If the random variables for the different location types are independently and identically distributed Weibull, this probability takes the following form (McFadden, 1978):

$$P_{z|h} = \frac{e^{(V_{hz})}}{\sum_{z' \epsilon K} e^{(V_{hz'})}}. \quad (2.12)$$

If the indirect utility functions are linear, then the probability becomes:

$$P_{z|h} = \frac{e^{(\beta_h z + \gamma_h H(z))}}{\sum_{z' \epsilon K} e^{(\beta_h z' + \gamma_h H(z'))}}. \quad (2.13)$$

The parameters $\beta_h$ and $\gamma_h$ can be estimated via maximum likelihood.

Bid-Rent Approach

The Bid-Rent approach, proposed originally by Ellickson in 1981, estimates the Bid function directly, circumventing the utility function (and therefore the hedonic function *H(z)*) entirely. It does so by determining the probability of a given household being located in a given location rather than a given location being chosen by a given household to locate. Location (or landlords) choosing tenants rather than tenants choosing location. The change in the direction of the conditional probability is grounded on the notion of the real estate market working as an auction process that goes back to Alonso (1964). The location will choose the tenant with the highest bid (the one who is willing to pay the most for the location).

As with the indirect utility function, the bid function can be expressed as a combination of a deterministic bid component and a stochastic or random component representing idiosyncratic differences within households of the same type:

$$B_h(z) + \varepsilon_h = B_{hz} + \varepsilon_h. \quad (2.14)$$

The probability of a given household type $h$ being located in a given location type $z$ is the probability that the bid of household type $h$ for location $z$ is higher than the bids of the other household types that could occupy the location:

$$P(h|z) = Prob\{B_{hz} + \varepsilon_h > B_{h'z} + \varepsilon_{h'}, h' \neq h; h, h \in G\}, \tag{2.15}$$

where $G$ is the set all the household types participating in the bid for $z$.

If the random term follows an Extreme Value Distribution, the best bid probability can be expressed as a logit model (McFadden 1978):

$$P_{h|z} = \frac{e^{(\mu B_{hz})}}{\sum_{h' \epsilon G} e^{(\mu B_{h'z})}}. \tag{2.16}$$

Under the auction assumption, the rent or price $r_z$ of a location $z$ will be the highest bid. The extreme value distribution assumption allows the expected maximum bid to be expressed as the logsum of the bids (Ben-Akiva and Lerman 1985):

$$r_z = \frac{1}{\mu} \ln\left(\sum_{h' \epsilon G} e^{(\mu B_{h'z})}\right) + C, \tag{2.17}$$

where $\mu$ is a scale parameter and $C$ is an unknown constant indicating that, given that the logit model is under-identified, the maximum value of the bids cannot be measured, only the relative highest bid (or the difference between bids).

The original formulation of Ellickson considered a linear bid function where parameters are estimated through maximum likelihood:

$$\mathcal{L} = \prod_{z \epsilon S}\left(\prod_{h \epsilon G}\left(P_{h|z}\right)^{y_{hz}}\right), \tag{2.18}$$

where $y_{hz}$ is a binary indicator equal to one if household $h$ is observed in location $z$ and zero otherwise, G is the set all the household types participating in the bid for $z$, and $S$ the total set of units available in the market.

Lerman and Kern (1983) complemented Ellickson's approach with three contributions. First, they included the observed rent price paid for a given unit into the estimation, more precisely into the probability density function of a given household being located in a given unit:

$$P(h|z) = Prob\{B_{hz} + \varepsilon_h = P^* and\ B_{hz} + \varepsilon_h > B_{h'z} + \varepsilon_{h'}, h' \neq h; h, h \in G\}, \tag{2.19}$$

where $P^*$ is the observed price or rent. If the random term is independent and identically Gumbel distributed (IIGD), then the probability becomes:

$$P(h|z) = f_\varepsilon(P^* - B_{hz}) \prod_{\substack{h' \in G \\ h' \neq h}} F_\varepsilon(P^* - B_{h'z}), \tag{2.20}$$

where $f_\varepsilon$ is the probability density function and $F_\varepsilon$ the cumulative density function given by:

$$\begin{aligned} f_\varepsilon &= \mu e^{(-\mu\varepsilon)} e^{(-e^{(-\mu\varepsilon)})} \\ F_\varepsilon &= e^{(-e^{(-\mu\varepsilon)})}. \end{aligned} \tag{2.21}$$

The parameters of the bid function can then be estimated through the following likelihood function:

$$\mathcal{L} = \prod_{z \in S} \left(-\mu e^{(-\mu(P^* - B_{hz}))}\right) \prod_{h' \in G} \left(e^{(-\mu(P^* - B_{h'z}))}\right)^{y_{hz}}. \tag{2.22}$$

The inclusion of the observed price or rent allows the identification of the scale parameter, $\mu$, thereby solving the under-identified nature of the logit model. It also allows for the interpretation of the bid function as a direct monetary willingness-to-pay of a given household for a change in a given location attribute.

Second, Lerman and Kern (1983) pointed out that if the random term of the bid function is IIGD, the mean of the random terms for a household type $h$ depends on the size of the group. Therefore, the logit models have to be normalized by the size of the groups:

$$P_{h|z} = \frac{e^{(\mu B_{hz}) + ln(N_t)}}{\sum_{h' \in G} e^{(\mu B_{h'z}) + ln(N_{t'})}}, \tag{2.23}$$

where $N_t$ is the number of household type $h$ can bid for the unit. This normalization reflects the fact that, all else equal, larger groups are more likely to win a bid than smaller groups.

Finally, Lerman and Kern (1983) commented on the loss of accuracy in the choice set that comes with an estimation based on household groups. They propose an alternative approach based on McFadden's work (1978) in which the parameters of bid functions are estimated for individual households using a randomly drawn sample of households in order to reduce the set of bidders to a manageable size.

Hurtubia and Bielaire (2013) proposed treating the expected maximum bid as a latent variable, which can be adjusted through a measurement relationship using observed prices in the area. They applied and validated this approach for a case study in Brussels and compared the results to Lerman and Kern's and Ellickson's specification. They found their approach predicted more accurately the spatial distribution of agents than Lerman and Kern's approach, while also adjusting expected bids to reflect realistic values.

While the presentation above uses the residential sector as an example (households and dwelling units), an analogous approach applies for the non-residential sector, with profit, not utility, maximization the implied objective. For example, analogous to Ellickson's work on residential location choice, Lee (1982) formulated the probability that a certain type of firm is located at a given site with a multinomial logit specification. The utility functions of the multinomial logit are the bids that firms make for a given location based on its characteristics such as the lot size or the distance to the CBD, which are in turn inputs for the firms' profit functions.

The bid-rent approach has also been the basis for more comprehensive land use models that seek to simulate the spatial distribution of urban agents as the result of a real estate market interaction and clearing process. Examples of such models include RURBAN (Miyamoto and Kitazume, 1989), MUSSA (Martínez 1996), IRPUD (Wegner 2008), ILUTE (Salvini and Miller 2005), and CUBE Land (CITILABS, Martinez 2010). Other models such as UrbanSim (Waddell et al. 2003) use a price-choice approach, with the price of the units calculated using a hedonic price function. The real estate market is represented as the interaction of different type of agents, usually households and firms, which compete for locations. Since these models are often used to forecast future development scenarios in aggregate models (not microsimulation), they mostly use a group-based formulation, which requires less information on the future number of agents. Table 1 presents a review of some relevant location choice models.

**Table 1 Review of location choice models**

| Source | Study Area | Resolution | Sample Size | Specification | Explanatory variables | Agent categorization criteria | Categories |
|---|---|---|---|---|---|---|---|
| **Households** | | | | | | | |
| 1 | San Francisco Bay Area | Census block | 28000 | Ellickson | - log(lot size in SF)<br>- log(num. of rooms)<br>- log(age of unit)<br>- log(travel time to SF in min.)<br>- log(median tract income in 1960), proxy for neighborhood quality<br>- log(elementary median income) as proxy for school quality<br>- % of black students in elementary school<br>- % of black students in junior high<br>- % of black HOUSEHOLDS in census tract<br>- Hedonic residual (Price vs. all variables). Proxy for other traits | - Race (black, white)<br>- Tenure (owner, renter)<br>- Family Type (children, no children)<br>- Income (<$7000, $7000-$9999, $10000<) | 24 |
| 2 | Bogota | Neighbor-hood | | Lerman & Kern | - log(Num. of rooms)<br>- log(Total Living Area)<br>- Floor quality index (1=Earth, 2=Cement, 3=Tile)<br>- Roof quality index (1=Scrap or veg, 2=Clay, Zinc, or tile, 3=Concrete)<br>- log(Mean neighborhood income in $) (proxy, neighborhood quality)<br>- log(Accessibility to 13 employment centers)<br>- Toilet index (2= none or shared, 3=Exclusive latrine, 4=flush toilet)<br>- Monthly Rent including utilities | - Income (rich, poor)<br>- Size (large, small) | 4 |
| 3 | Chicago | Census Tract | 3044 | Lerman & Kern<br><br>2-stage hedonic regression (Ross) | - Num. Of Rooms<br>- Age of unit<br>- Area in SF<br>- AC system (no AC=0, window or wall AC=1, central AC=2)<br>- Garage (no=0, on-site parking=1, not built-in gar=2, carport=3, Built-in gar=4)<br>- Property TAX<br>- % of whites in census tract<br>- Median income of census tract<br>- Accessibility to employment (distance downtown)<br>- Cook county dummy<br>- Proximity to Airport (5 mile radius dummy)<br>- Particulate matter (PM-10) reading<br>- (Sale price) | Evaluates different cross-categories:<br>- Income (<$40K>$40K) vs. Presence of Children (with, without)<br>- Income (<$40K >$40K) vs. Race (white, non-white)<br>- Income (< $30K;$30K-$50K;$50K<)<br>- Income (< $30K;$30K-$60K;$60K<) | 4,3 |

**Table 1 (continued)**

| Source | Study Area | Resolution | Sample Size | Specification | Explanatory variables | Agent categorization criteria | Categories |
|---|---|---|---|---|---|---|---|
| **Households** | | | | | | | |
| 4 | Brussels | 4945 Zones | 1367 | Ellickson<br><br>Lerman & Kern<br><br>Own (latent variable) | - Ave. Area by unit type (m2)*log(households size)<br>- Unit Type (dummy)*households size 2 or more (dummy)<br>- % of higher-education people in zone*high-education people in HHs<br>- % of higher-income households in zone*mid/high-income households<br>- % of low-income households in zone*high-income households<br>- Public transportation access. (facilities/km2)*no car in households<br>- Public transportation access. (facilities/km2)*2+ cars in households<br>- Car access. (gen travel cost to all zones) * 1+ cars in households<br>- Industry jobs (jobs/m2) * high-income households<br>- office jobs (jobs/m2) * workers in households | - Unit type: fully-detached, semi-detached, attached, apartment<br>- Income (low, mid, high) | |
| 5 | Boston | 2727 TAZs | | Lerman & Kern (undefined) | - Households size<br>- age of head of households<br>- income dummy for low income households<br>- car access to retail jobs (index based on general cost gravity function)<br>- car access to retail jobs (index based on general cost gravity function)<br>- Car access to base jobs (index based on general cost gravity function)<br>- Transit access to all jobs (index based on general cost gravity function)<br>- Num. of rooms in unit<br>- Dummy for units multifamily | - Age of head of households and households Type (15-34 non-family, 15-44 family, 35-64 non-family, 45-64 family, 65+) | 5 |
| 6 | Seoul | 74 TAZ | | Lerman & Kern (undefined) | - Ave. age of unit<br>- Ave. num. of rooms<br>- Multifamily dummy<br>- log(accessibility to jobs) (gravity function)<br>- log(average zonal income)<br>- log(households density)<br>- log(employment density)<br>- Seoul dummy | - Income (<$1000, $1000-$3000,$5000-$5000,$5000<) | 4 |

**Table 1 (continued)**

| Source | Study Area | Resolution | Sample Size | Specification | Explanatory variables | Agent categorization criteria | Categories |
|---|---|---|---|---|---|---|---|
| **Firms** | | | | | | | |
| 7 | Bogota | 27 Comunas | 126 | Ellickson | - Percentage of products sold in Bogota<br>- Percentage of inputs bought in Bogota<br>- Airline distance in km form the CBD (centroid to centroid)<br>- Percentage of production workers living in the south<br>- Percentage of administrative workers living in north<br>- Frequency of electricity interruptions<br>- Population density of comuna<br>- Location quotient (share of employment in industry sector j relative to total employment in manufacturing in comuna)<br>- Year of initial operation at the present location<br>- Ownership dummy | - 2 Industry sector (textile and fabricated metal)<br>- firms size (large, small) | 4 |
| 8 | Dallas Fort Worth | 141 zips in the Dallas Fort Worth area | | Ellickson (sampling 5 alternatives for each location) | - Distance to CBD in miles<br>- Distance to CBD squared<br>- Distance to the airport in miles<br>- Dummy if zip contains or borders major roads<br>- Zip's share of region's total usable area<br>- % of zip area that is developed<br>- Total population within decay radius<br>- Median households income within a decay radius<br>- % of black population within decay radius<br>- Empl. in construction, manufacturing, & wholesale within decay radius<br>- Empl. in mining, transportation, utilities, and F.I.R.E within decay radius<br>- Employment in retail trade and services within decay radius | - Industry sectors | 5 |
| 5 | Boston | 2727 TAZs | | Lerman & Kern | - Ave. pay per worker by industry<br>- Ave. establishment size (num. of workers)<br>- Lots between 0-2,499 SF (Dummy)<br>- Lots between 2,500-9,999 SF (Dummy)<br>- Lots between 10,000-39,000 SF (Dummy)<br>- FAR<br>- Parcels low density retail, entertainment, service, medical, office, or hosp.<br>- Parcels that are warehouses, industrial, or utilities (dummy)<br>- Car accessibility to low-income households (gravity-based function)<br>- Car accessibility to medium-high income households (gravit-based)<br>- Car accessibility to high-income households (gravit-based)<br>- Accessibility to selected highways (index based on skim distance) | - Industry sector | 11 |

Notes: (1) Ellickson, 1981; (2) Gross, 1986; (3) Chattopadhyay, 1997; (4) Hurtubia & Bierlaire, 2013; (5) MAPC et al., 2013; (6) Myung-Jin Jun, 2013; (7) Lee, 1982; (8) Shukla & Waddell, 1991;

Empirical model development

In order to evaluate a location choice model a combination of formal (statistical) and informal tests are required. Informal tests examine how well the model results align with *a priori* knowledge of the phenomena the model is trying to represent. The most common informal test is to examine the value of the coefficients to see how they compare to *a priori* assumptions and expectations. In location choice models, the coefficients estimated represent the value that a specific agent (household or firm) places on a given location attribute relative to the other agents. The informal test then examines if the relative valuation of location characteristics by the different agents are in line with *a priori* expectations. For example, households with no cars are expected to value proximity to transit higher than households with cars.

The formal, or statistical, tests are used to examine individual coefficients or the models as a whole. For example, the asymptotic *t* test is used for hypothesis testing; i.e., to test a null hypothesis that a particular parameter differs from a value (usually zero). The significance level of the *t* statistic determines the level of confidence with which one can reject the null hypothesis. The Goodness-of-Fit measures are used to compare model estimations to determine which one bests fits the data. The most common measure of goodness-of-fit is the likelihood ratio index or rho-square, which is defined as:

$$\rho^2 = 1 - \frac{\mathcal{L}(\hat{\beta})}{\mathcal{L}(0)}, \tag{2.24}$$

where $\mathcal{L}(\hat{\beta})$ is the final log likelihood with the estimated coefficients and $\mathcal{L}(0)$is the final log likelihood fixing all coefficients to zero. When comparing two models, all else equal, a model with a higher likelihood ratio index is preferable. Similar to the regression statistic $R^2$, the likelihood ratio index either stays the same or increases when new explanatory variables are added to the model. Akin the adjusted-$R^2$, the adjusted likelihood ratio index adjusts for this by accounting for the parameters estimated:

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\hat{\beta}) - K}{\mathcal{L}(0)}, \tag{2.25}$$

where $K$ is the number of unknown parameters in the model.

To compare different specifications, the most common test is the likelihood ratio test (LRT). The null hypothesis of this test is that the restricted model (e.g. all coefficients are equal to zero) is the true model. Rejecting this hypothesis supports the unrestricted model. The LRT statistic is defined as:

$$-2\left(\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U)\right), \tag{2.26}$$

where $\mathcal{L}(\hat{\beta}_R)$ is the final likelihood of the restricted model and $\mathcal{L}(\hat{\beta}_U)$ is the final likelihood of the unrestricted model. The statistic is distributed chi-square with $(K_U - K_R)$ degrees of freedom, where $K_U$ and $K_R$ are the number of estimated coefficients of the unrestricted and restricted model, respectively. This test is also used to determine the difference in model results with different datasets. In this case, the restricted model assumes that, for the same specification, the coefficients are the same for the two datasets. To implement this test, one pools the two datasets together to estimate the restricted model. The unrestricted model can be individual estimations of the two datasets (in this case, $\mathcal{L}(\hat{\beta}_U)$ would be the sum of the individual final log likelihoods) or a pooled estimation in which some of the coefficients are unrestricted (allowed to be different for the two datasets).

In order to determine which coefficients might vary between datasets (stability of preferences), the assumption that the coefficients are the same can be lifted for some of the variables in the pooled model. For a direct comparison of the coefficients of these variables, the scale parameter $\mu$ must be adjusted based on the relationship of variances between the different datasets and fixing one scale parameter to 1, as shown in equations 2.27 to 2.29.

$$Var(\varepsilon) = \frac{\pi^2}{6\mu^2} \tag{2.27}$$

$$\frac{Var(\varepsilon^{model\ 2})}{Var(\varepsilon^{model\ 1})} = \frac{(\mu^{model\ 1})^2}{(\mu^{model\ 2})^2} = \frac{1}{\sqrt{\mu^{model\ 2}}}\ (fix\ \mu^{model\ 1} = 1) \tag{2.28}$$

$$\beta^{model\ 1} = \beta^{model\ 2} * \mu^{model\ 2} \tag{2.29}$$

To evaluate if the possible differences in coefficients (after adjusting for scale) are significant, the following $t^*$ statistic is calculated (Galbraith and Hensher, 1982)

$$t^* = \frac{\beta_1 - \beta_2}{\sqrt{\left(\frac{\beta_1}{t_1}\right)^2 + \left(\frac{\beta_2}{t_2}\right)^2}} \tag{2.30}$$

If this statistic is significant, we can reject the null hypothesis of the corresponding coefficient being the same for both datasets.

The tests mentioned previously examine the specification of utility (or bid) functions and take the model structure as a given. But the model structure itself should also be tested, to see if the assumptions of the logit model hold. One of the main assumptions of the logit model is the condition of independence from irrelevant alternatives (IIA), which means that the relative choice probability between two alternatives is independent of the other available alternatives in the choice set. A given model specification can violate the IIA condition when (1) alternatives share unobserved attributes, or (2) the error terms of the alternatives are not identically distributed (have different variances). The test to examine possible violation of the IIA condition involves comparing models estimated with subsets of alternatives available in the choice set. One such test is McFadden's omitted variables test, which examines if cross-alternative variables enter the model. If this is the case, the IIA condition is violated. This is implemented by adding an auxiliary variable to the utility function of the subset of the choices that may be correlated. If the logit assumptions hold, the coefficients of the auxiliary variable are zero (coefficient will not be significant). Given a subset $\tilde{C}_n$, the auxiliary variables are defined as:

$$z_{in}^{\widetilde{C_n}} = \begin{cases} \hat{V}_{jn} - \dfrac{\sum_{j \in \widetilde{C_n}} \hat{P}(j|C_n)\, \hat{V}_{jn}}{\sum_{j \in \widetilde{C_n}} \hat{P}(j|C_n)} & \text{if } i \in \widetilde{C_n} \\ 0 & \text{if } i \; not \in \widetilde{C_n} \end{cases} \tag{2.31}$$

where $\hat{V}_{jn}$ is the systematic utility from the base model (without the auxiliary variables) and $\hat{P}(j|C_n)$ is the corresponding location calculated from the estimated base model.

## 2.3. Context

Grater Boston is a dynamic region that has experienced significant demographic and economic changes in the last decades. From 1880 to 1920, the population in the City of Boston more than doubled, from 363,000 resident to 745,000. This growth was a consequence of City annexations

and migration from Europe. The population was accommodated mostly in small multi-unit housing (triple-deckers). After World War II, the returning veterans who were looking to form a family spurred a wave of single-family suburban housing developments. The exodus to the suburbs led to deterioration the urban core. Between 1950 and 1980, the city of Boston's population decreased from 801,000 to 563,000 and real estate value plummeted. By 1980, Boston was a declining city in a middle-income metropolitan area (Glaeser, 2004). However, the region's skilled workers and historical linkage to educational institutions allowed Boston to take advantage of the booming information and knowledge economy. The economy turned from manufacturing to high tech, finance, health, and education services (Glaeser, 2004). According to The Greater Boston Housing Report Card 2014-2015, Greater Boston has the third highest metro-area-wide rents in the country. The metro area attracts young millennial and aging suburban baby boomers. Increasing demand for both residential and commercial space continues pushing real estate value up and boosting new real estate developments. The demographic composition of the area is expected to continue changing. According to population projections made by the Metropolitan Area Planning Council (MAPC), by 2030 the Boston region will have 282,000 new households headed by someone age 65 or older. Given the linkage between life stages and location preferences, the development of location choice models for Greater Boston can inform housing policies and projects aimed at adapting the region to the foreseeable demographic changes

### 2.3.1. Analysis strategy

Location choice modes are data intensive, requiring a large amount of information to properly characterize all the possible locations available for all the possible types of agents. Features such as the area of a residential unit, the cost of utilities, the lighting in the unit, the material of the floors, the way it looks from the outside, the number of bathrooms, or proximity to the potential tenants' social circle are part of what households evaluate when choosing a location. Similarly, when looking for commercial space, firms evaluate the status of the building, the quality of the space, whether it is on the first floor or the top floor, and its proximity to similar firms. Data on residential and commercial real estate stock at this level of detail are hard to come by. Even when available, data on the locations are not enough. Since the estimation of these models is based on revealed preferences, information on both the locations and their corresponding occupants is necessary. This is even harder to come by.

The estimation of a location choice model requires data on three levels:

- The locating agents (e.g., households, firms).
- The commercial space or residential unit.
- The location of the commercial space or residential unit.

For this thesis, I obtained information on the agents from two different sources. One is the 2012 Massachusetts Travel Survey. Since the survey does not include information on the residential unit, I geolocated the records and matched them to specific parcels in the MassGIS Level 3 Assessors' Parcel Mapping data. This database includes information on individual parcels and buildings, which I used to construct and approximate attributes of the residential units.

The other source is a database on customers and businesses compiled by Infogroup, a private business data provider. Infogroup provided data for businesses for the years 2010 and 2000 and households for the year 2010. Unlike the Travel Survey, the Infogroup data include information on the commercial space and residential unit.

With both sources of information geolocated, I constructed zonal variables to characterize the zones. Chapters 3 and 4 present more details on the different datasets. Table 2 summarizes the different sources of data used in this thesis.

**Table 2 Summary of data sources**

| Level | Agent Type | Data Set | Source | Year |
|---|---|---|---|---|
| Agent | Households | Consumer Data | Infogroup | 2010 |
| | | 2012 Massachusetts Travel Survey | MassDOT and MPOs | 2010-2011 |
| | Firms | Bussines data | Infogroup | 2010, 2000 |
| Unit | Households | Consumer Data (Owners) | Infogroup | 2010 |
| | | Lavel 3 Parcel Data | MassGIS, MAPC | 2009-2013 |
| | Firms | Bussines data | Infogroup | 2010, 2000 |
| | | Lavel 3 Parcel Data | MassGIS, MAPC | 2009-2013 |
| Zone | Households, Firms | Demographic and Housing data | American Fact Fincer, Geolytics, US Census Bureau | 2010, 2000 |
| | | Employment | CTPP | 2010, 2000 |
| | | Lavel 3 Parcel Data | MassGIS, MAPC | 2009-2013 |
| | | Crimes | FBI | 2010 |
| | | Travel time, CUBE Voyager Trasportation Model | MIT | 2010, 2000 |
| | | SAT scores | MassESE | 2010 |

In order to answer the research questions stated in Chapter 1, I use the following location choice model estimation strategy:

- Estimate residential location choice models by categories of households for 2010 with both the Travel Survey data and the Infogroup data. Apply the same model specification and formulation to both data sources in order to analyze data-related uncertainty.
- Estimate a residential location choice model by individual households to compare this type of specification with a category-based specification.
- Estimate firm location choice models for firms for 2010 and 2000 to analyze if/how location preferences have changed.

# 3. RESIDENTIAL LOCATION CHOICE – HOUSEHOLDS

This chapter consists of two main sections: a general analysis of housing and population dynamics, and the estimation of residential location choice models. The section on model estimation, in turn, contains three subsections. The first subsection describes the data used. The second subsection presents the model specification and results for the estimation by categories of agents. The third subsection presents the model estimation for individual households.

## 3.1. Population and housing dynamics in Greater Boston

Population

To understand better the broad urban dynamics at play, I divided the study area in three concentric sub regions: Boston, Metro Boston, and Greater Boston (**Figure 2**).

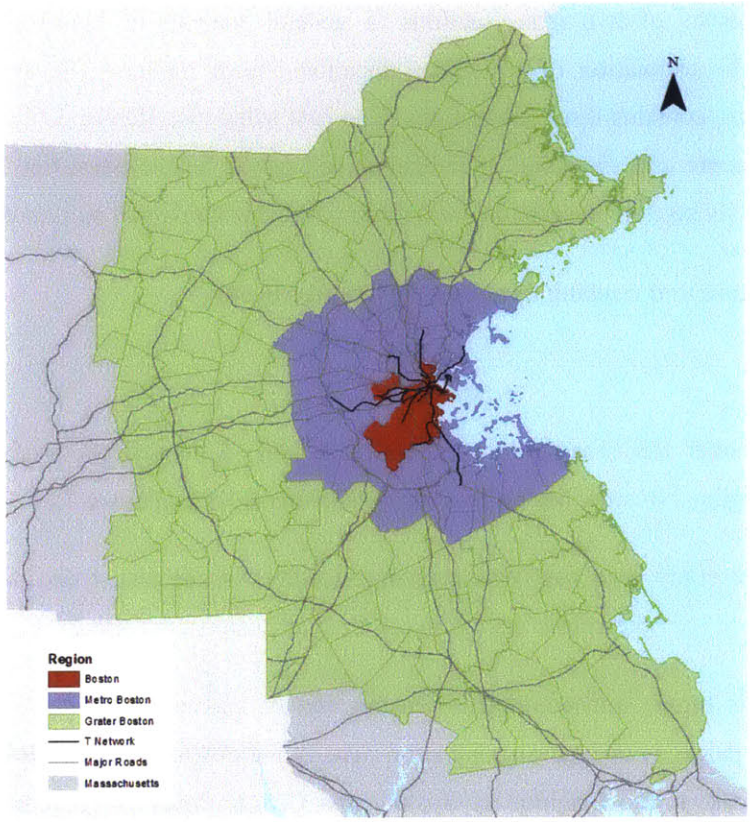
Population in the study area has been growing steadily for the last 4 decades.[2] The largest population increase has taken place mainly in the Greater Boston sub region.

The population is highly concentrated in Boston and its surrounding towns. Within that the central area, the places with highest population density are downtown, Back Bay, East Boston, North End, Chinatown and Brighton in Boston, and Central Square in Cambridge (Figure 4). These places also show the highest increases in population density over the last 4 decades, along with Chelsea and some areas of South Boston. Outside the central area, the increases in population density have been concentrated in Marlborough, Brockton, Newburyport and Amesbury. The areas where population density has decreased concentrate mainly in southern neighborhoods in Boston (Dorchester, Roxbury, and Mission Hill) and in Somerville (Figure 5). The average household size in the region has decrease from 2.84 persons per household in 1990 to 2.59 in 2010. This causes the number of households to increase at a higher rate than population growing 9.5% from 1980 to 1990, 9.1% from 1990 to 2000, and 6.1% from 2000 to 2010 (corresponding to 4.0% 6.2%, and 5.0% population growth).

---

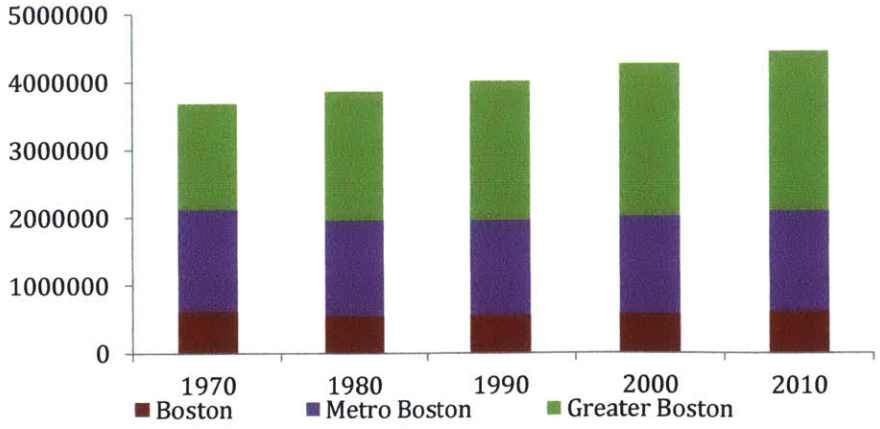
[2] *US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

**Figure 2 Subdivision of Greater Boston in 3 Sub regions**



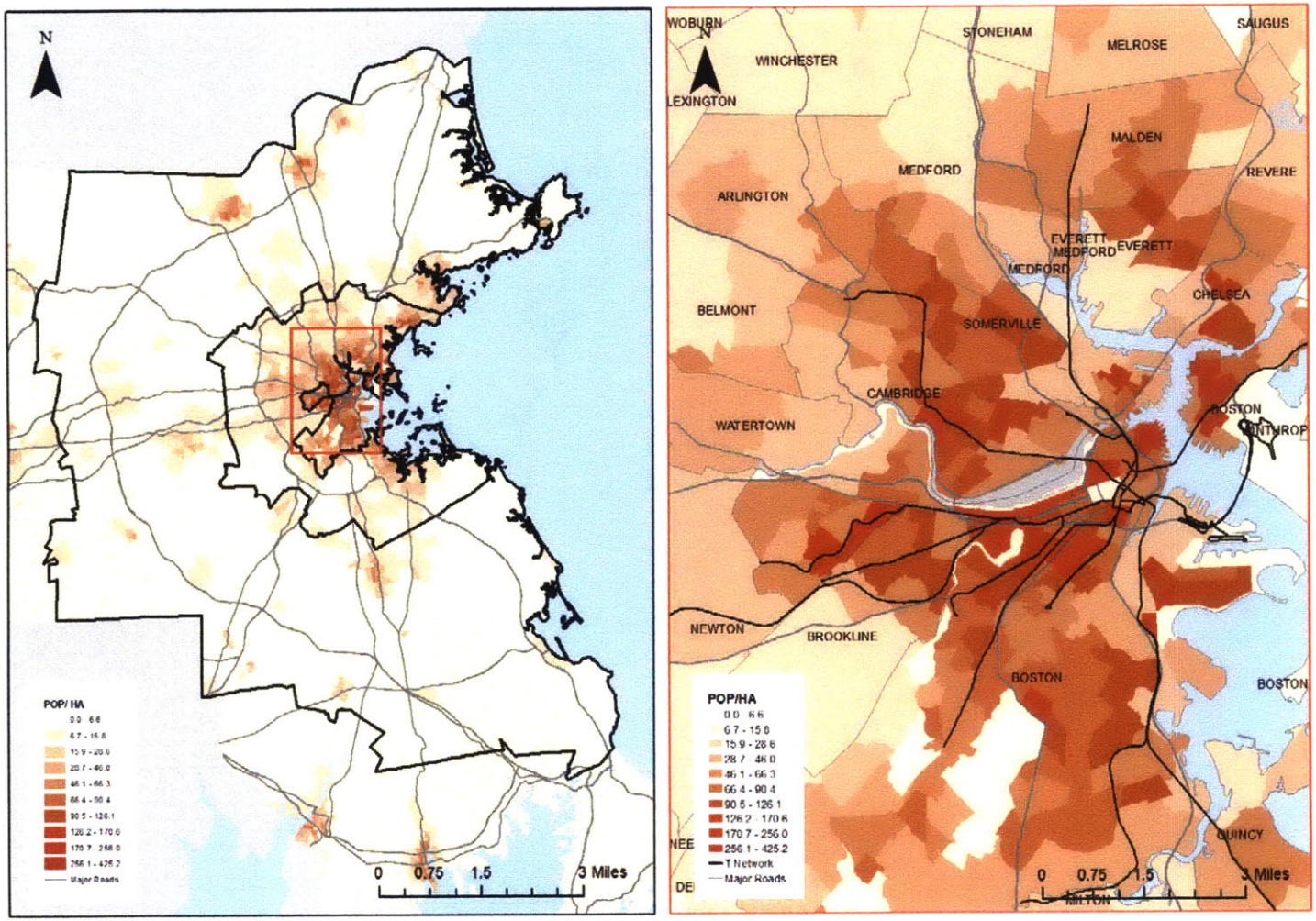

*Source: Own with MassGIS shapefiles*

**Figure 3 Population evolution by sub region**



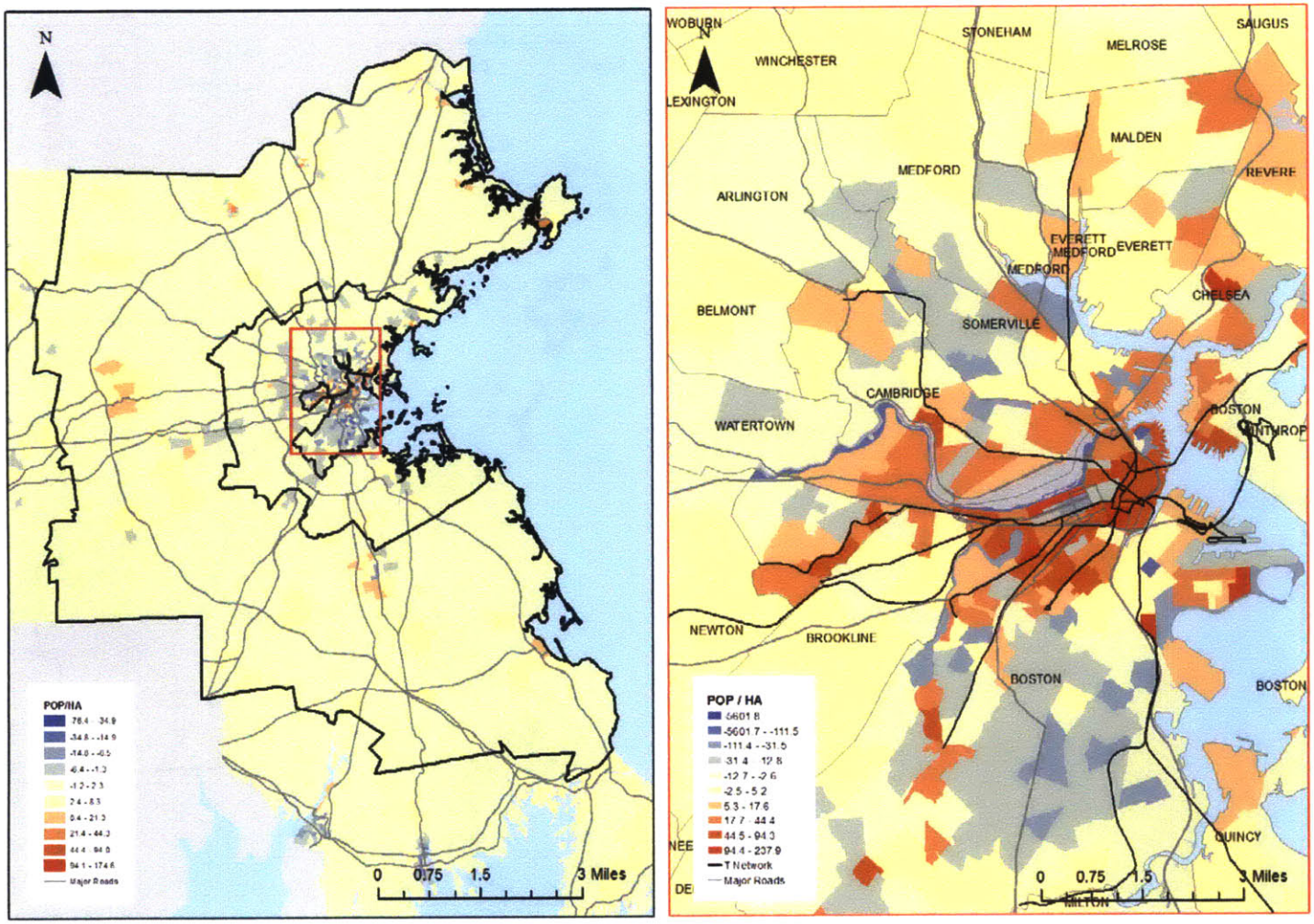

*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

**Figure 4 Population density in 2010**




*Source: American Community Survey for 2010*

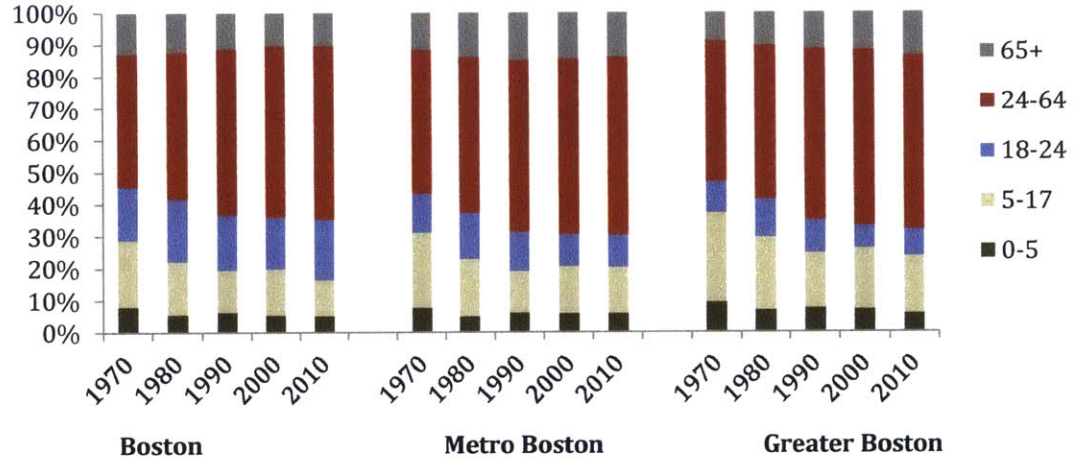
**Figure 5 Change in population density from 1970 to 2010**




*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

Median age in the study area increased from 33.4 in 1990 to 38.6 in 2013,[3] with the 24 to 64 year-old age group increasing the most. More specifically, since 1990 the 45 to 64 year-old age group is the only one that has increased, growing 3.4% between 1990 and 2000 and 22.8% between 2000 and 2010. The three sub-regions have experienced a similar demographic change, resulting in an overall increase in the share of the population over 24 years old. Boston, followed by the Metro Area has the largest share of young adults (between 18 and 24 years old), while the Greater Boston sub-region has the highest share of children (under 17 years old) (Figure 6)

**Figure 6 Share of total population by age group**




*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

Racial diversity has increased in the last decades. The share of population that is white decreased from 88.1% in 1990 to 77.0% in 2013, while the African-American share grew from 6.2% to 8.3% and the Asian and Hispanic populations more than doubled to 7.2% and 10.1% respectively during same period.

In spite of a total increase in the employment rate, real median household income has remained relatively unchanged, from $67,002 in 1990 to $69,206 in 2010 (in 2010 dollars). However, income inequality has widened. Households with median annual income below $35,000 have increased by 21% between 1990 and 2010. In the same period, households with median annual income above $100,000 increased 19%.

---

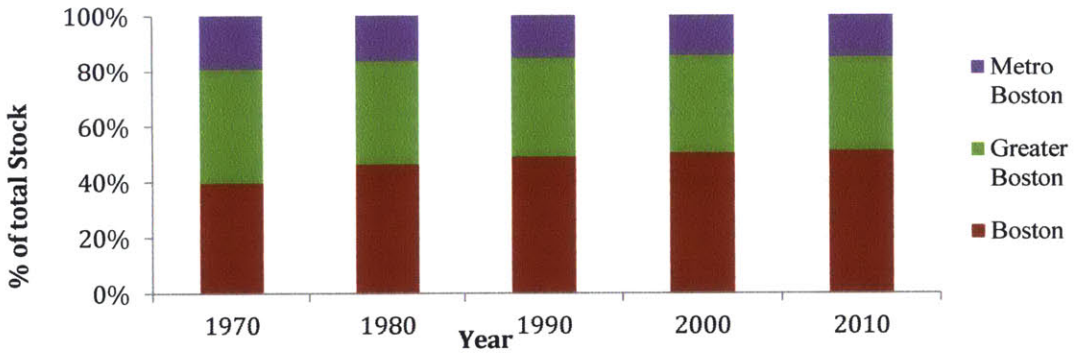
[3] *The Greater Boston Housing Report Card 2014-2015*

Housing Market and Housing Stock

Boston's share of the total housing units in the study area has grown from 40% in 1970 to over 50% in 2010, while the Metro Boston sub region's share decreased from 19% to 15%. The outer towns (Greater Boston) decreased their share from 41% to 34% (Figure 7).

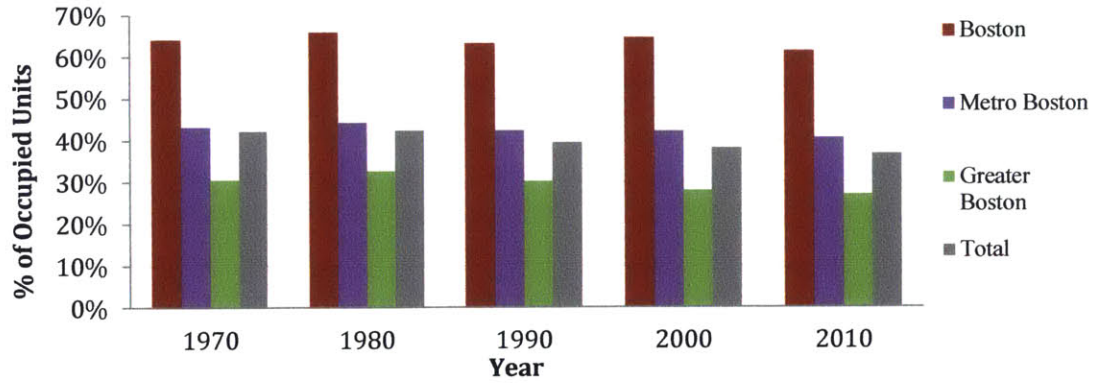
Housing tenure status varies widely within the region. The majority of people living in Boston are renters; this number has modestly decreased from 64% in 1970 to 61% in 2010. Outside of Boston, the majority of residents own the units they occupy (Figure 8).

**Figure 7 Percentage of total housing stock**




*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

**Figure 8 Renter Occupied Units**



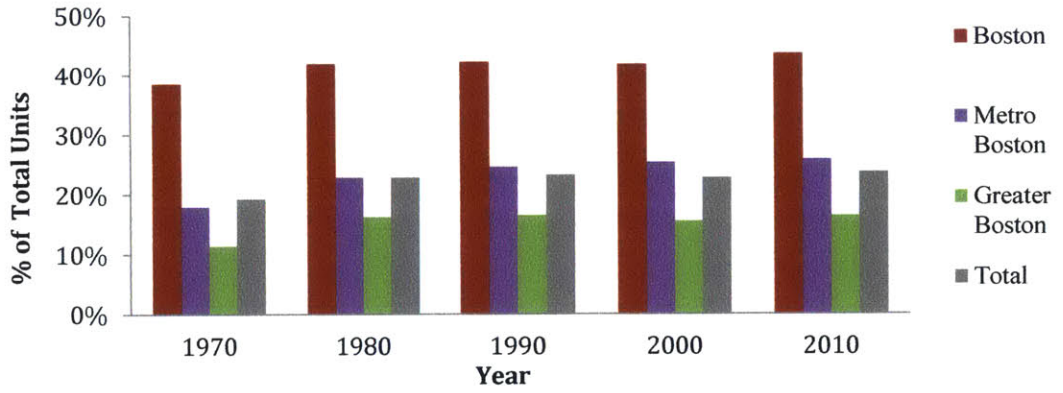

*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

The number of multifamily buildings has been increasing in the region as a whole. The city of Boston has the highest percentage of housing structures with 5 or more residential units, around 43% in 2010. However, this number has increased more rapidly for the rest of the study area,

growing by 8% in the Metro Area and 5% in Greater Boston over the past four decades (Figure 9).

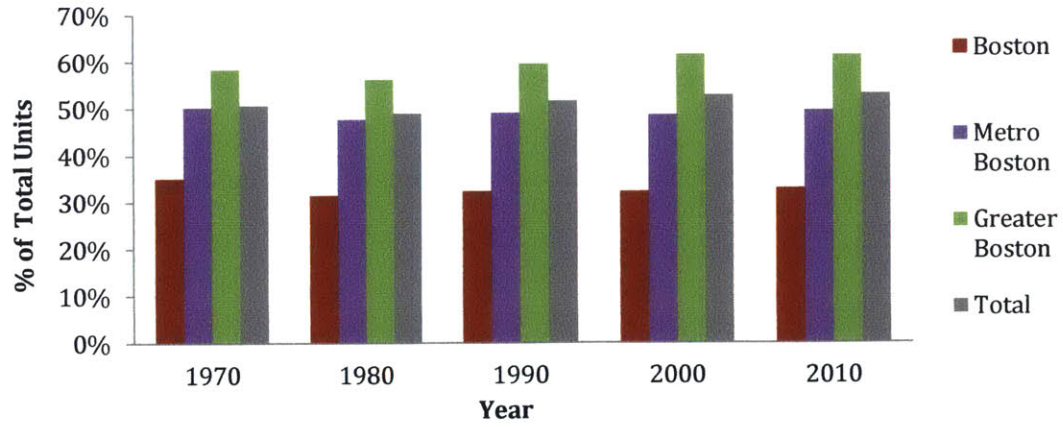
The size of residential units, measured by the number or rooms, has changed little since 1970. The percentage of units with 3 or more rooms has decreased in the city of Boston, from 33% to 32%. In the outer towns, this share has modestly increased, by 3%, in the same period (Figure 10). Note that this modest change in the number of units does not necessarily correlate with unit area.

**Figure 9 Size of buildings. Building with 5 or more units**




*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

**Figure 10 Size of units. Units with 3 of more rooms**




*Source: US Decennial Census from 1970 to 2000 and American Community Survey for 2010*

According to The Greater Boston Housing Report Card 2014-2015, housing costs for both renters and owners have increased since 2000 by around 15% after adjusting for inflation. Median gross rent in nominal terms doubled from 1990 to 2010, from $642 to $1,226. The rising housing costs coupled with stagnating household incomes have led to housing becoming increasingly less affordable. The share of renter households paying more than 30% of their income on rent grew from 41% in 1990 to 50.6% in 2010. And the share of households paying more than 50% of their income on rent grew from 19.6% to 26.4% in the same period. For homeowners, the percentage households paying more than 30% of their gross income in mortgages and taxes increased from 27% to more than 38% in the same period.

To summarize the main population and housing dynamics:

- Total population in the study area has been increasing steadily.
- Population in the region is concentrated in Boston and Metro Boston, with almost 50% of total population.
- Median age has increased, a trend expected to continue as baby-boomers get older.
- Racial diversity has increased.
- Income inequality has increased.
- The majority of households in Boston are renters (slightly decreasing in the last years) while the majority (and growing) in Greater Boston are owners. The Metro Boston subarea has a 50-50 split between renters and owners.
- The share of multifamily buildings is growing, with the Metro Boston sub region presenting the fastest growth.
- The Greater Boston sub region has more large units (3+ rooms) followed by Metro Boston and the City of Boston. The number of large units has decreased in Boston and increased in Greater Boston.
- Housing cost as a percentage of income has increased for both renters and owners.

## 3.2. Residential Location Model Estimation

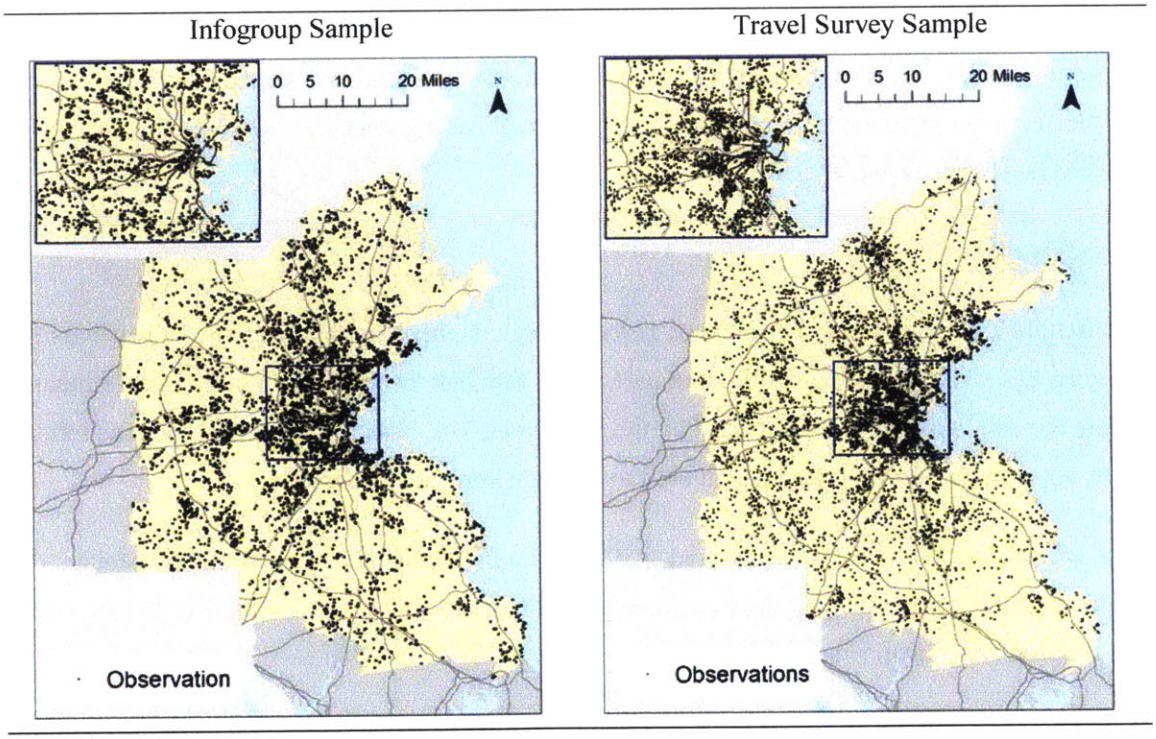
### 3.2.1. Data description

As mentioned in section 2, I estimate residential location models with two different data sets. One is based on the 2012 Massachusetts Travel Survey records and their corresponding parcels obtained from the MassGIS Leve 3 Assessors' Parcel Mapping data. The other comes from the Infogroup customers database. After filtering out the records form the Level 3 Parcel data that

were blank, the matched Travel Survey-L3 Parcel dataset contains 5528 observations in the study area, 743 of which are for households in rent-occupied units. Given the likely differences in location preferences between renters and owners, which arise from the difference in their socio-economic characteristics, lifestages, and/or lifestyles, these two submarkets should be modeled separately. In this thesis I focus on location models for owners. I do not include location choice models for renter-occupied units, which account for almost 40% of total units in the study area, due to time limitations.

The Infogroup dataset contains 1,718,162 homeowners in the study area from which I selected a random sample of 5000 observations for modeling. Figure 11 presents the spatial distribution of the two data sets.

**Figure 11 Sample distribution by value of room**




*Source: Infogroup, Massachusetts Travel Survey 2012*

The Infogroup data contain individual customers and have less information on households' attributes than the Travel Survey data, which include attributes such as number of workers, students, and car. The household attributes selected for the analysis are shown in Table 3 in grey.

**Table 3 Household attributes**

| | Travel Survey | Infogroup |
|---|---|---|
| Income | x | x |
| Household Size | x | x |
| Household Head Age | x | x |
| Vehicle | x | |
| Workers | x | |
| Students | x | |
| Children | | x |
| Household Type | x | x |

For the Infogroup database, the variables household size and household type were constructed based on other available variables. For household type, I used the information on whether the customer is married or single (married=family household and single=non-family households). I calculated household size as one (the customer), plus one if married, plus the number of children in the household. This method may lead to misrepresenting some households, such as a household of four (4) unmarried students over age of 18 (although such a household would more likely be located in a renter-occupied unit).

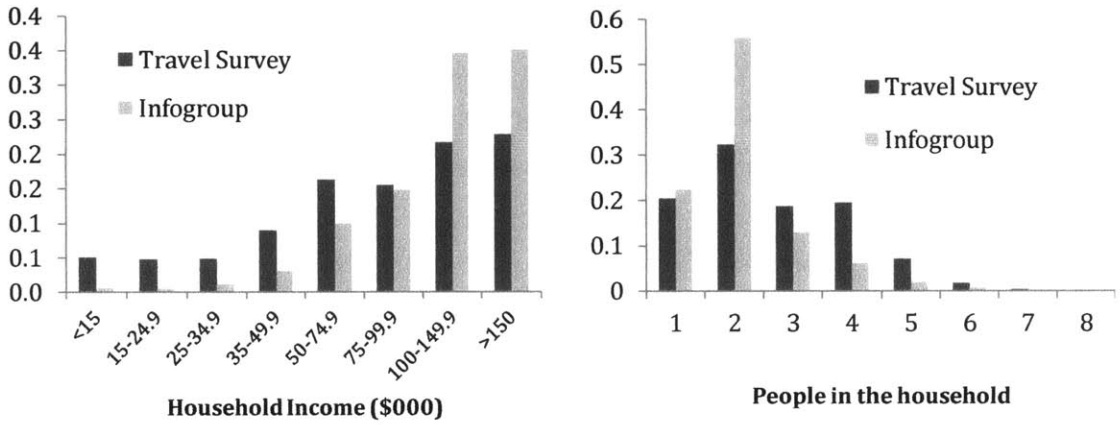
Agent Attributes

Household income in the Infogroup data is, on average, higher than that from the Travel Survey (Figure 12). This may be due, at least in part, to the fact that income in Infogroup is an estimate while the one in the survey is reported by the household. The available Infogroup documentation does not include details on the method used to estimate household income.

The variable I constructed for household size from the Infogroup data has a higher percentage of 2-person households. As mentioned previously, by definition this is either a married couple with no children or a single person with one child.

The distribution of the age of the head of the household is relatively similar between the two datasets (Figure 13). Infogroup reports the maximum age category as 65 years or older, so the larger bar for the 60-70 year age range for the Infogroup data includes older age categories.

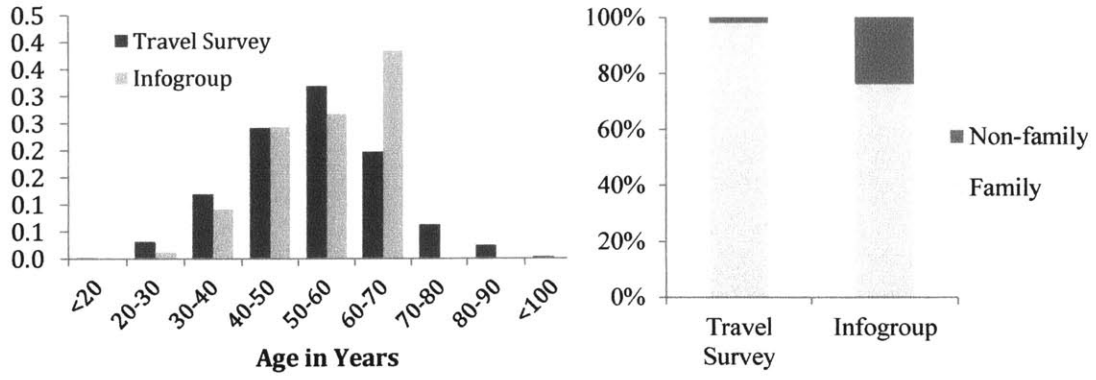
**Figure 12 Household income and household size histograms**




*Source: Infogroup, 2012 Massachusetts Travel Survey*

The number of non-family households is larger in the Infogroup data (23.7%) than in the Travel Survey data (0.02%). Again, this may be at least partly due to the way I constructed the variable for the Infogroup data. For example, according to the categorization criteria state previously, single adult customers who live with their parents (or other type of family other than children) are categorized as non-family, even though the household is a family household (Figure 13).

**Figure 13 Histogram of the age of the head of the household and household type distribution**




*Source: Infogroup, 2012 Massachusetts Travel Survey*

Unit Attributes

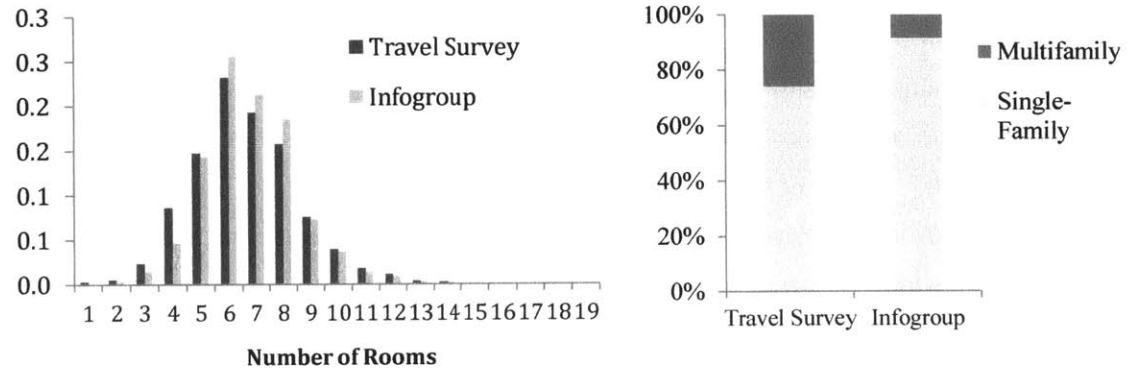
Similar to the household attributes, the residential unit attributes vary by dataset. The Infogroup dataset has information about the specific residential unit, while for the Travel Survey data I construct these variables based on the information available in Level 3 Parcel data. The unit attributes selected for the analysis are shown in Table 4 in grey. The number of rooms for the Travel Survey data was approximated as average rooms per unit for the parcel (the total number of rooms in the building divided by the number of units). This might not be accurate if, for example, the building is a large multifamily complex with multiple communal rooms (e.g. storage rooms). In these cases, the average number of rooms per unit would be overestimated. The FAR (floor-area-ratio) attribute for the Infogroup dataset was approximated as the average FAR of the zone.

**Table 4 Residential unit attributes**

| | Travel Survey | Infogroup |
|---|---|---|
| Building Age | | x |
| Num. Rooms | x | x |
| Num. Baths | | x |
| Num. Bedrooms | | x |
| Unit area | | x |
| Unit Type | x | x |
| Building FAR | x | x |

The distribution of the number of rooms in the unit is relatively similar between the two samples, validating the method used to approximate this variable for the Travel Survey data (Figure 14).

**Figure 14 Histogram of number of rooms in units. Unit type distribution**




*Source: Infogroup, 2012 Massachusetts Travel Survey*

Zonal Attributes

The zonal attributes capture characteristics of the immediate area in which the residential unit is located, as well as its spatial relationship with the rest of the study area. The zone structure used for the analysis is the 2727 Transport Analysis Zones (TAZ) from CTPS. The different zonal attributes were transformed from their original spatial units (e.g. block groups or Town) into this TAZ structure through aggregation and/or disaggregation based on area (spatial split). The zonal attributes evaluated through model estimation and their corresponding sources are:

- Racial composition of the area implemented as ratio of white population to total population (Census data by block group).
- School quality implemented as the SAT scores by school district. The variables are normalized to values between zero and one (Massachusetts Department of Elementary and Secondary Education).
- Median annual income in thousands of dollars (Census data by block group).
- Weighted crime rate per capita by town. Violent crimes are assigned a 0.8 weight and property crimes a 0.2 weight (FBI)
- Property tax rate by town (Massachusetts Department of Revenue)
- Job density (jobs/ha) by different industries in (e.g. retail or amenities) (CTPP by block group)
- Population density (persons/hectare) (Census data by block group).

- Accessibility to different opportunities such as jobs (service, retail, base, total) and built area of specific uses (retail, amenities), by different modes (auto, transit, walk), and at different time periods (morning peak and mid-day). I calculated the built area based on the Level 3 Parcel data. I measure accessibility of a given zone $i$ using a gravity measure of the sum of the opportunities $k$ in all the other zones $j$ divided by the corresponding travel time between zones $t_{ij}$. The travel time comes from a 4-step transportation model developed by Mikel Murga at MIT. I selected this type of accessibility function was over an impedance function in order to avoid the propagation possible errors from the impedance function estimation in the location choice model

$$Accesibility\ to\ Opportunity: ACC_i^k = \sum_j \frac{Opportunity_j^k}{t_{ij}} \quad (3.1)$$

### 3.2.2. Model estimation by categories of agents

Agent categorization

I evaluated several different agent type categorizations, limited in some cases by the number of observations in a given category. The final agent types are cross-categorizations of three main household characteristics: income, age of the head of the households, and household size. A categorization based on these attributes is consistent with the literature and aims to best capture, with the available data, the effect of household lifecycles in location preferences. I divided the households in the sample into low, middle, and high income, depending on the household size (Table 5)

**Table 5 Income levels by total household income ($000) and household size**

| Income level | Household size | | | |
|---|---|---|---|---|
| | 1-person | 2-person | 3-person | 4+ person |
| Low | <$15 | <$35 | <$50 | <$50 |
| Mid | $15-$75 | $35-$150 | $50-$150 | $50-$150 |
| High | $75+ | $150+ | $150+ | $150+ |

In addition, I specified household size categories as small (two persons or less) and large (3 or more persons); and, the age of the head of the household categories as young (less than 35), mid-

age (between 35 and 65), and senior (65 or older). Ultimately, there were not enough observations for a full cross-categorization (18 categories). Table 6 presents the final agent categorization.

**Table 6 Agent categorization**

| Agent Type | Income | Age of Head | Size | Infogroup Obs. | Travel Survey Obs. |
|---|---|---|---|---|---|
| 1 | Low | all | all | 47 | 430 |
| 2 | Mid, High | <35 | 3+ | 40 | 143 |
| 3 | Mid, High | <35 | <3 | 173 | 99 |
| 4 | Mid | 35-64 | 3+ | 450 | 1274 |
| 5 | Mid | 35-64 | <3 | 1299 | 1011 |
| 6 | High | 35-64 | 3+ | 505 | 699 |
| 7 | High | 35-64 | <3 | 1173 | 468 |
| 8 | Mid, High | 65+ | 3+ | 85 | 86 |
| 9 | Mid, High | 65+ | <3 | 1228 | 573 |
| | | | | 5000 | 4783 |

The cross-categorization in Table 6 resulted in the best estimation results. The fact that the best estimation results were obtained when grouping all low-income households together deserves further analysis. This may be due to a limitation in the number of observations in this category (not enough to further sub-divide the group) or to missing variables in the formulation. It may also be due to more complex socio-economic phenomena that limit this population segments' ability to exercise location choice. Low-income households may be *captive* households. That is, they do not really have a choice of where to locate, but rather depend on availability and the location choices of other agents.

Model specification

I estimate the location choice model by groups using Ellickson's formulation (1981) with the adjustment by group size proposed by Lerman and Kern (1983). Agent Type 1 (Low-income households) was taken as the reference. The coefficients estimated reflect how agents value a certain attribute relative to the reference agent. So, for example, a negative coefficient means that the agent places less value on the corresponding location attribute compared to the reference agent (it does not necessarily mean that the agent values the attribute negatively).

Table 7 presents a summary of the variables used in the final specification:

**Table 7 Summary of variables used in the estimation of residential location models**

| Name | Description | Unit of observation | InfoGroup | | | | | | Travel Survey | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| ROOMS | Number of rooms in residential unit | Residential Unit | 2.00 | 6.00 | 7.00 | 6.83 | 8.00 | 20.00 | 1.00 | 6.00 | 7.00 | 6.90 | 8.00 | 53.65 |
| IS_MF | Dummy for unit in multifamily buildings | Residential Unit | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 1.00 |
| DIFF_INC_ LOWER | · log(mid income of zone - income of agent) ; if mid income of zone > income of agent<br>· 0 ; if mid income of zone < income of agent | Zone | -3.85 | 0.00 | 0.00 | 0.27 | 0.00 | 4.32 | -3.62 | 0.00 | 0.00 | 1.07 | 2.65 | 4.93 |
| DIFF_INC_ HIGHER | · log(income of agent - mid income of zone) ; if income of agent > mid income of zone<br>· 0 ; if income of the agent < mid income of zone | Zone | -1.92 | 2.62 | 3.70 | 3.36 | 4.73 | 5.45 | -4.36 | 0.00 | 2.65 | 2.18 | 3.88 | 5.11 |
| SAT | School quality index based on total SAT scores of school district | Zone | 0.00 | 0.78 | 0.82 | 0.83 | 0.89 | 1.00 | 0.00 | 0.76 | 0.82 | 0.82 | 0.89 | 1.00 |
| SHARE_WHI TE | Ratio of white population to total population in zone | Zone | 0.02 | 0.83 | 0.90 | 0.86 | 0.95 | 1.00 | 0.00 | 0.78 | 0.88 | 0.83 | 0.95 | 1.00 |
| RETAIL_DEN | Density of jobs in the retail and amenities sector [jobs/HA] | Zone | 0.00 | 0.05 | 0.16 | 1.08 | 0.50 | 124.7 | 0.00 | 0.06 | 0.23 | 1.26 | 0.76 | 294.8 |
| ACC_SERVE MP_CAR | Auto accessibility to service jobs in the AM peak [jobs/minutes] | Zone | 0.07 | 0.13 | 0.17 | 0.19 | 0.24 | 0.72 | 0.07 | 0.14 | 0.18 | 0.21 | 0.26 | 0.86 |
| POP_DEN | Population density [persons/HA] | Zone | 0.33 | 3.50 | 8.10 | 18.61 | 18.69 | 537.5 | 0.33 | 4.24 | 10.88 | 24.42 | 27.6 | 537.5 |

Estimation results

I specify the same model for the two datasets and conduct a likelihood ratio test to compare the individual estimations with an estimation using the pooled dataset, assuming the coefficients and scale parameter are the same for both individual datasets (fully constrained). This test determines if the coefficients (preferences) change depending on the dataset used. The null hypothesis of this test is that the estimations with the different datasets are the same. Additionally, I conducted a likelihood ratio test between the pooled fully restricted models and a pooled model with different scale parameters for each dataset. The estimation of a scale parameter allows the identification of differences between the variances of the two data sets. Table 8 presents the results of the individual estimations (Model 1 and 2), a model with and the pooled- fully constrained model (Model 3), and the pooled model with different scales (Model 4).

In Model 4 the scale parameter for the InfoGroup data is fixed to 1 and the scale parameter is estimated. The t-statistic of the scale parameter estimated for the Travel Survey data evaluates if parameter is different than 1, that is, different than the scale parameter of the InfoGroup data. In this case, the estimated parameter is significant and, therefore, the null hypothesis of both being equal can be rejected.

Table 9 presents the likelihood ratio test to see if the results change across models.

**Table 8 Residential location model. Estimation results by individual (unconstrained) models, pooled fully constrained model, and pooled model with different scale parameters**

| Variable | AGENT TYPE | | | Model 1: InfoGroup | | | Model 2: Travel Survey | | | Model 3: Pooled | | | Model 4: Pooled - Scaled | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | INC | AGE | SIZE | Beta | t-test | | Beta | t-test | | Beta | t-test | | Beta | t-test | |
| ASC | M,H | <35 | 3+ | -12 | -4.03 | ** | -6.11 | -4.12 | ** | -6.86 | -5.65 | ** | -13.6 | -6.66 | ** |
| | M,H | <35 | <3 | -9.75 | -4.6 | ** | -2.06 | -1.49 | | -7.82 | -6.23 | ** | -16.2 | -7.93 | ** |
| | M | 35-64 | 3+ | -7.06 | -5.36 | ** | -2.35 | -3.83 | ** | -3.55 | -6.71 | ** | -5.48 | -7.37 | ** |
| | M | 35-64 | <3 | -4.77 | -4.46 | ** | -1.84 | -2.93 | ** | -3.25 | -6.56 | ** | -4.34 | -6.48 | ** |
| | H | 35-64 | 3+ | -7.14 | -2.05 | ** | -7.90 | -6.39 | ** | -12.5 | -13.31 | ** | -17.7 | -13.67 | ** |
| | H | 35-64 | <3 | -4.89 | -1.47 | | -4.47 | -3.78 | ** | -16.9 | -16.22 | ** | -22.1 | -16.06 | ** |
| | M,H | 65+ | 3+ | -4.6 | -1.88 | * | -7.14 | -4.37 | ** | -6.82 | -5.37 | ** | -12.8 | -5.25 | ** |
| | M,H | 65+ | <3 | -8.35 | -5.67 | ** | -2.90 | -3.40 | ** | -7.8 | -10.7 | ** | -10.8 | -11.03 | ** |
| SAT | M,H | <35 | 3+ | 4.86 | 2.11 | ** | 1.43 | 1.32 | | 1.84 | 2.4 | ** | 4.78 | 3.28 | ** |
| | M,H | <35 | <3 | 4.76 | 2.98 | ** | -1.18 | -1.15 | | 1.89 | 1.81 | * | 5.28 | 3.13 | ** |
| | M | 35-64 | 3+ | 4.66 | 4.54 | ** | 1.12 | 1.76 | * | 2.22 | 4.78 | ** | 3.27 | 5.41 | ** |
| | M | 35-64 | <3 | 4.64 | 4.78 | ** | 1.27 | 1.97 | ** | 3.17 | 6.73 | ** | 4.26 | 6.83 | ** |
| | H | 35-64 | 3+ | 10.2 | 7.03 | ** | 5.57 | 5.74 | ** | 7.36 | 10.1 | ** | 10 | 10.42 | ** |
| | H | 35-64 | <3 | 9.41 | 6.7 | ** | 2.73 | 3.19 | ** | 8.41 | 12.38 | ** | 11 | 11.97 | ** |
| | M,H | 65+ | 3+ | 1.22 | 0.76 | | 1.57 | 1.26 | | 1.14 | 1.12 | | 2.98 | 1.33 | |
| | M,H | 65+ | <3 | 5.68 | 4.41 | ** | 0.60 | 0.71 | | 3.98 | 5.66 | ** | 5.51 | 6.11 | ** |
| IS_MF | M,H | <35 | 3+ | 4.01 | 2.78 | ** | 0.66 | 1.99 | ** | 1.04 | 3.38 | ** | 1.53 | 3.57 | ** |
| | M,H | <35 | <3 | 4.44 | 3.64 | ** | 0.49 | 1.60 | | 0.674 | 2.88 | ** | 1.11 | 3.48 | ** |
| | M | 35-64 | 3+ | 0.398 | 0.25 | | -0.34 | -1.64 | * | -0.225 | -1.14 | | -0.139 | -0.5 | |
| | M | 35-64 | <3 | 3.16 | 2.63 | ** | 0.32 | 1.64 | * | 0.0715 | 0.4 | | 0.387 | 1.55 | |
| | H | 35-64 | 3+ | 4.31 | 3.49 | ** | -0.60 | -2.22 | ** | 0.0777 | 0.33 | | 0.504 | 1.53 | |
| | H | 35-64 | <3 | 5.19 | 4.27 | ** | 0.48 | 2.18 | ** | 0.993 | 5.18 | ** | 1.72 | 6.24 | ** |
| | M,H | 65+ | 3+ | 3.72 | 2.85 | ** | -0.07 | -0.15 | | 0.217 | 0.6 | | 0.374 | 0.72 | |
| | M,H | 65+ | <3 | 3.7 | 3.06 | ** | -0.03 | -0.14 | | -0.013 | -0.07 | | 0.387 | 1.44 | |

**Table 8 (continued)**

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **RETAIL_DEN** | M,H | <35 | 3+ | 0.01 | 0.07 | | -0.08 | -1.34 | | -0.07 | -1.46 | | -0.11 | -1.35 | |
| | M,H | <35 | <3 | 0.03 | 0.85 | | 0.05 | 1.50 | | 0.05 | 2.32 | ** | 0.07 | 2.28 | ** |
| | M | 35-64 | 3+ | -0.02 | -0.23 | | 0.04 | 1.40 | | 0.05 | 2.15 | ** | 0.07 | 2.06 | ** |
| | M | 35-64 | <3 | 0.04 | 1.20 | | 0.06 | 2.06 | ** | 0.06 | 2.82 | ** | 0.08 | 2.56 | ** |
| | H | 35-64 | 3+ | 0.04 | 1.29 | | 0.07 | 2.48 | ** | 0.07 | 3.37 | ** | 0.10 | 3.14 | ** |
| | H | 35-64 | <3 | 0.03 | 1.10 | | 0.06 | 2.06 | ** | 0.06 | 2.99 | ** | 0.08 | 2.67 | ** |
| | M,H | 65+ | 3+ | 0.01 | 0.24 | | -0.01 | -0.20 | | 0.01 | 0.14 | | -0.02 | -0.22 | |
| | M,H | 65+ | <3 | 0.03 | 0.89 | | 0.05 | 1.74 | * | 0.05 | 2.43 | ** | 0.07 | 2.26 | ** |
| **SHARE_WHITE** | M,H | <35 | 3+ | 3.91 | 2.43 | ** | 1.61 | 2.90 | ** | 2.00 | 3.89 | ** | 3.52 | 4.49 | ** |
| | M,H | <35 | <3 | 2.94 | 4.08 | ** | 1.96 | 2.98 | ** | 2.58 | 5.80 | ** | 4.02 | 6.31 | ** |
| | M | 35-64 | 3+ | 2.11 | 2.86 | ** | 1.49 | 4.59 | ** | 1.82 | 6.30 | ** | 2.55 | 6.32 | ** |
| | M | 35-64 | <3 | 1.79 | 3.60 | ** | 0.98 | 3.14 | ** | 1.58 | 6.31 | ** | 1.93 | 5.80 | ** |
| | H | 35-64 | 3+ | 4.65 | 5.22 | ** | 2.10 | 4.80 | ** | 2.94 | 7.40 | ** | 4.46 | 7.98 | ** |
| | H | 35-64 | <3 | 4.06 | 6.12 | ** | 1.65 | 4.00 | ** | 3.18 | 8.83 | ** | 4.65 | 9.46 | ** |
| | M,H | 65+ | 3+ | 0.50 | 0.48 | | 1.20 | 1.51 | | 1.28 | 2.03 | ** | 2.01 | 2.15 | ** |
| | M,H | 65+ | <3 | 3.15 | 5.60 | ** | 1.70 | 4.29 | ** | 2.51 | 8.41 | ** | 3.46 | 8.70 | ** |
| **ROOMS** | M,H | <35 | 3+ | 0.54 | 4.04 | ** | 0.23 | 2.51 | ** | 0.29 | 3.09 | ** | 0.45 | 3.74 | ** |
| | M,H | <35 | <3 | 0.24 | 2.43 | ** | -0.25 | -3.32 | ** | -0.07 | -1.25 | | 0.02 | 0.25 | |
| | M | 35-64 | 3+ | 0.40 | 4.65 | ** | 0.20 | 5.21 | ** | 0.24 | 6.72 | ** | 0.35 | 7.15 | ** |
| | M | 35-64 | <3 | 0.29 | 3.57 | ** | 0.07 | 1.61 | | 0.09 | 2.51 | ** | 0.17 | 3.58 | ** |
| | H | 35-64 | 3+ | 0.73 | 8.53 | ** | 0.34 | 7.21 | ** | 0.43 | 10.75 | ** | 0.62 | 11.75 | ** |
| | H | 35-64 | <3 | 0.64 | 7.55 | ** | 0.19 | 4.10 | ** | 0.34 | 9.44 | ** | 0.50 | 10.06 | ** |
| | M,H | 65+ | 3+ | 0.42 | 3.84 | ** | 0.29 | 5.67 | ** | 0.29 | 5.37 | ** | 0.41 | 5.48 | ** |
| | M,H | 65+ | <3 | 0.38 | 4.50 | ** | 0.05 | 1.07 | | 0.13 | 3.51 | ** | 0.23 | 4.73 | ** |
| **ACC_SERVEMP_ CAR** | M,H | <35 | 3+ | 7.73 | 1.87 | * | 4.24 | 2.41 | ** | 4.49 | 2.77 | ** | 7.33 | 3.13 | ** |
| | M,H | <35 | <3 | 10.40 | 4.04 | ** | 2.93 | 2.04 | ** | 6.05 | 5.30 | ** | 9.97 | 6.24 | ** |
| | M | 35-64 | 3+ | 6.01 | 2.34 | ** | 2.65 | 2.42 | ** | 3.04 | 3.02 | ** | 4.67 | 3.31 | ** |
| | M | 35-64 | <3 | 4.37 | 1.83 | * | 2.80 | 2.63 | ** | 3.18 | 3.35 | ** | 4.40 | 3.31 | ** |
| | H | 35-64 | 3+ | 7.92 | 3.05 | ** | 5.93 | 4.97 | ** | 6.05 | 5.81 | ** | 8.27 | 5.67 | ** |
| | H | 35-64 | <3 | 11.90 | 4.85 | ** | 6.39 | 5.46 | ** | 8.74 | 8.83 | ** | 12.30 | 8.81 | ** |
| | M,H | 65+ | 3+ | 10.60 | 3.20 | ** | 4.68 | 2.21 | ** | 6.44 | 3.82 | ** | 11.20 | 4.70 | ** |
| | M,H | 65+ | <3 | 11.10 | 4.55 | ** | 5.31 | 4.42 | ** | 7.71 | 7.72 | ** | 10.80 | 7.69 | ** |

**Table 8 (continued)**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **DIFF_INC_HIGHER** | M,H | <35 | 3+ | -0.10 | -0.34 | | 0.05 | 0.38 | | -0.02 | -0.28 | | 0.33 | 1.83 | * |
| | M,H | <35 | <3 | 0.17 | 0.79 | | 0.22 | 1.35 | | 0.74 | 5.66 | ** | 1.57 | 7.36 | ** |
| | M | 35-64 | 3+ | 0.09 | 1.56 | | -0.09 | -2.38 | ** | -0.10 | -3.43 | ** | -0.11 | -2.97 | ** |
| | M | 35-64 | <3 | 0.00 | 0.10 | | 0.04 | 1.01 | | 0.12 | 4.08 | ** | 0.07 | 2.03 | ** |
| | H | 35-64 | 3+ | -1.99 | -3.89 | ** | -0.33 | -2.80 | ** | 0.14 | 1.77 | * | 0.23 | 2.02 | ** |
| | H | 35-64 | <3 | -2.09 | -4.18 | ** | -0.35 | -2.64 | ** | 0.91 | 7.82 | ** | 1.04 | 6.70 | ** |
| | M,H | 65+ | 3+ | -0.15 | -0.67 | | 0.20 | 1.10 | | 0.29 | 2.39 | ** | 0.95 | 3.94 | ** |
| | M,H | 65+ | <3 | -0.10 | -0.83 | | 0.02 | 0.41 | | 0.47 | 8.63 | ** | 0.56 | 7.45 | ** |
| **DIFF_INC_LOWER** | M,H | <35 | 3+ | -0.15 | -0.39 | | -0.08 | -0.48 | | -0.24 | -1.85 | * | -0.21 | -0.88 | |
| | M,H | <35 | <3 | 0.61 | 1.91 | * | 0.21 | 1.04 | | 0.46 | 2.87 | ** | 0.95 | 3.76 | ** |
| | M | 35-64 | 3+ | -0.30 | -4.01 | ** | -0.12 | -3.07 | ** | -0.28 | -9.56 | ** | -0.37 | -9.78 | ** |
| | M | 35-64 | <3 | -0.21 | -3.82 | ** | -0.10 | -2.23 | ** | -0.31 | -10.44 | ** | -0.43 | -11.25 | ** |
| | H | 35-64 | 3+ | -14.10 | -2.64 | ** | -0.22 | -1.13 | | 0.39 | 1.97 | ** | 0.87 | 2.61 | ** |
| | H | 35-64 | <3 | -15.20 | -2.98 | ** | 0.00 | 0.00 | | 1.30 | 5.81 | ** | 1.74 | 4.92 | ** |
| | M,H | 65+ | 3+ | 0.23 | 0.57 | | 0.05 | 0.22 | | 0.01 | 0.05 | | 0.33 | 1.15 | |
| | M,H | 65+ | <3 | 0.46 | 4.13 | ** | 0.13 | 1.88 | * | 0.25 | 4.05 | ** | 0.30 | 3.51 | ** |
| **POP_DEN** | M,H | <35 | 3+ | -0.04 | -2.67 | ** | 0.00 | -0.32 | | 0.00 | -0.98 | | -0.01 | -1.51 | |
| | M,H | <35 | <3 | -0.02 | -2.36 | ** | 0.00 | -1.09 | | -0.01 | -3.06 | ** | -0.01 | -3.63 | ** |
| | M | 35-64 | 3+ | -0.04 | -4.29 | ** | -0.01 | -3.06 | ** | -0.01 | -4.33 | ** | -0.02 | -4.66 | ** |
| | M | 35-64 | <3 | -0.01 | -2.38 | ** | -0.01 | -2.47 | ** | -0.01 | -4.17 | ** | -0.01 | -4.09 | ** |
| | H | 35-64 | 3+ | -0.02 | -2.68 | ** | -0.01 | -1.62 | | -0.01 | -3.40 | ** | -0.02 | -3.84 | ** |
| | H | 35-64 | <3 | -0.02 | -2.52 | ** | -0.01 | -2.24 | ** | -0.01 | -4.40 | ** | -0.02 | -4.65 | ** |
| | M,H | 65+ | 3+ | -0.06 | -4.85 | ** | -0.01 | -1.07 | | -0.02 | -3.30 | ** | -0.04 | -4.37 | ** |
| | M,H | 65+ | <3 | -0.02 | -3.53 | ** | -0.01 | -2.56 | ** | -0.02 | -5.75 | ** | -0.02 | -5.90 | ** |
| Scale InfoGroup | | | | 1.00 | | | | | | 1.00 | | | 1.00 | | |
| Scale Travel Survey | | | | | | | 1.000 | | | 1.00 | | | 0.57 | -17.31 | ** |
| Sample | | | | | 5000 | | | 4785 | | | 9785 | | | 9785 | |
| Final log-like: | | | | | -7875.7 | | | -8670.4 | | | -17377.52 | | | -17231.4 | |
| LRT: | | | | | 6220.852 | | | 3686.64 | | | 8244.645 | | | 8536.907 | |
| Rho-sq: | | | | | 0.28 | | | 0.175 | | | 0.19 | | | 0.20 | |
| Adj rho-sq: | | | | | 0.28 | | | 0.168 | | | 0.19 | | | 0.19 | |

**Table 9 Likelihood ratio test for residential model. Pooled (fully constrained) vs. individual (unconstrained) models and Pooled (fully constrained) model vs. Pooled model with different scales**

---

Pooled (constrained) vs. individual (unconstrained) models

$H_0: \hat{\beta}_R = \hat{\beta}_U \quad where \ \hat{\beta}_R: pooled\ model\ , \hat{\beta}_U: individual\ models$

$T = -2[\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U)]$

$T = -2[-17377.52 + (7875.7 + 8670.4)] = 1662.8$

P-value= 0.000

Reject $H_0$. Preferences change with dataset

Pooled (constrained) model vs. Pooled model with different scales

$H_0: \hat{\beta}_R = \hat{\beta}_S \quad where \ \hat{\beta}_R: pooled\ models\ , \hat{\beta}_S: pooled\ model\ with\ different\ scales$

$T = -2[\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_S)]$

$T = -2[-17377.52 + 17231.4] = 292.262$

df = 1

P-value= 0.000

Reject $H_0$. Models are not the same. Variances of the datasets are different

---

For the Model 4, the scale parameter for the Infogroup data was fixed to 1. The estimated scale parameter of 0.57 for the Travel Survey data indicate that the variance of unobserved factors is lower for this dataset compared to the Infogroup dataset.

The likelihood ratio test indicates that the estimation results do vary depending on the dataset used. The signs of the coefficients are similar in the four models. Given the higher rho-square and level of significance of the individual coefficients, the individual (unconstrained) Infogroup models is preferred over the pooled models. The results of the individual models can be interpretation as follows:

ROOMS: This variable proxies for unit size. The variable is not significant for middle age headed, mid-income households of less than 3 persons, and senior, mid- and high-income, small households (less than 3 persons) for the Travel Survey model. All other agents (except for young, mid- and high-income, small households) value the number of rooms higher than the reference group (all else equal, they will bid more than the reference group for units with more rooms). For both models, larger households value larger units more then smaller households, and the preference for larger units increases with income. This is expected, since demand for space tends to increase with income, all else equal.

IS_MF: for the Infogroup data, the coefficients are significant for all agents except large, middle age headed, mid-income households. All other agents value multifamily households higher than the low-income households (reference group). For this dataset, the agents that place more value on multifamily units are young households and middle age, high-income household. For the travel survey data, the coefficients are not significant for young, mid-and high-income households of less than 3 persons, and senior households. For middle age headed households, preference depends on household size: household of more than 2 persons value multifamily units less than the reference group while small households value multifamily units more than the reference group.

DIFF_INC_HIGHER: this variable aims to measure how agents value being in a location where their income is higher than median income of the area. For both datasets, the coefficient is significant and negative for high-income households. That is, these households place less value on being in comparatively poorer areas than the reference group. These results suggest that high-income households are less likely to locate in comparatively poorer areas.

DIFF_INC_LOWER: this variable aims to measure how agents value being in a location where their income is lower than median income of the area. The coefficients are not significant for young and senior, mid- and high-income households of more than 3 persons in the Infogroup dataset. For the Travel Survey data, coefficients are significant only for middle age headed households and senior, mid- high-income, small households. Except for this last agent type, the significant coefficients in both datasets suggest that the corresponding agent types place less value than the reference group on being in a comparatively higher income area. This might suggest a higher aspiration of low-income households of being in higher income areas.

SAT: this variable is a proxy for school quality in the area. For the Infogroup data, coefficients are significant and positive for all agents except for middle age headed, mid-income, large households (coefficient is not significant for this agent). For the Travel Survey data, coefficients are significant and positive only for middle-age headed, mid- and high-income households. Agents with significant coefficient are willing to bid more than low-income households for locations in areas with good school quality. In other words, low-income households might value school quality equally high (or higher) than these agents, but they are 'not willing' (or not capable) to bid more for locations in these areas. For both datasets, the agents willing to bid the most for higher school quality are high-income households. Except for middle age headed, mid-income households in the Travel Survey data, larger households (which are more likely to have children who go to school), place more value in school quality than similar households (in income level and age) of smaller size (less likely to have kinds).

RETAIL_DEN: this variable is a proxy for the amount of retail and other amenities in the area. The coefficients are significant (and positive) for three agent types in the Travel Survey data: middle age headed, high-income households; middle age headed, mid-income, small households; and senior, mid- and high income, small households. These agents are willing to bid more compared to low-income households for locations with retail outlets and amenities like museums or movie theaters.

SHARE_WHITE: coefficients are significant and positive for all agents in both datasets, except senior, mid- and high-income, large households. This suggests that (almost) all agents are willing to bid more compared to low-income households in order to locate in areas with a larger white population. For the Travel Survey data, which includes information on the race of the households, all agent classifications are more than 80% white. This is in itself evidence of the relationship between race and homeownership (model is for owners only). It might also suggest a preference of agents for being among people from the same race. However, additional analysis is required to make this statement.

ACC_SERVEMP_CAR: coefficients are significant and positive for all agents in both datasets. Given that auto accessibility to service jobs in Boston still shows a highly monocentric pattern, the results suggest that, all else equal, all agents place more value than the reference group in being closer to the Boston metro area. Other than the reference agent type, the agent type willing to pay the least for this attribute are mid-age, mid-income households. The willingness to bid

more for this zonal attribute seems to be more related to income level than to the age of the head of the households or the household size.

POP_DEN: Coefficients are significant and negative for all agents in the Infogroup dataset. For the Travel Survey data, coefficients are significant and negative for middle age headed, mid-income households, middle age headed, high-income, small households, and senior, mid- and high-income, small households. All these households place more value compared to low-income households on lower density. Relative to the referent group (low income households), all other households' willingness to bid decreases as neighborhood density increases. As household size increases, the aversion to density also increases.

Stability of preferences

As outlined in the introduction, one of my research questions relates to the effects of using different datasets for model estimation (data uncertainty). Here I test the sensitivity of the estimated household bidding behavior across the Infogroup and Travel Survey data, using the stability of preferences approach outlined in the methods section. The $t^*$-statistic is only calculated for coefficients that are significant in both of the individual models.

Table 10 shows the results of the stability of preferences test.

The analysis indicates significant differences in the coefficients between the two models. This might be because of the differences in income distribution between the two data sets (Figure 12). Since the agent categorization is based on income, differences in this household characteristic might results in different coefficients for all agent categories.

**Table 10 Residential location model. Pooled data with unconstrained variables**

| Variable | AGENT TYPE | | | Model 1: InfoGroup | | | Model 2: Travel Survey | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | INC | AGE | SIZE | Beta | t-test | | Beta | t-test | | t* | |
| ASC | M,H | <35 | 3+ | -12 | -4.03 | ** | -3.45 | -4.12 | ** | -2.76 | ** |
| | M,H | <35 | <3 | -9.75 | -4.6 | ** | -1.16 | -1.49 | | | |
| | M | 35-64 | 3+ | -7.06 | -5.36 | ** | -1.33 | -3.83 | ** | -4.21 | ** |
| | M | 35-64 | <3 | -4.77 | -4.46 | ** | -1.04 | -2.93 | ** | -3.31 | ** |
| | H | 35-64 | 3+ | -7.14 | -2.05 | ** | -4.46 | -6.39 | ** | -0.75 | |
| | H | 35-64 | <3 | -4.89 | -1.47 | | -2.53 | -3.78 | ** | | |
| | M,H | 65+ | 3+ | -4.6 | -1.88 | * | -4.03 | -4.37 | ** | -0.22 | |
| | M,H | 65+ | <3 | -8.35 | -5.67 | ** | -1.64 | -3.40 | ** | -4.33 | ** |
| SAT | M,H | <35 | 3+ | 4.86 | 2.11 | ** | 0.81 | 1.32 | | | |
| | M,H | <35 | <3 | 4.76 | 2.98 | ** | -0.67 | -1.15 | | | |
| | M | 35-64 | 3+ | 4.66 | 4.54 | ** | 0.63 | 1.76 | * | 3.70 | ** |
| | M | 35-64 | <3 | 4.64 | 4.78 | ** | 0.72 | 1.97 | ** | 3.78 | ** |
| | H | 35-64 | 3+ | 10.2 | 7.03 | ** | 3.15 | 5.74 | ** | 4.55 | ** |
| | H | 35-64 | <3 | 9.41 | 6.7 | ** | 1.54 | 3.19 | ** | 5.30 | ** |
| | M,H | 65+ | 3+ | 1.22 | 0.76 | | 0.89 | 1.26 | | | |
| | M,H | 65+ | <3 | 5.68 | 4.41 | ** | 0.34 | 0.71 | | | |
| IS_MF | M,H | <35 | 3+ | 4.01 | 2.78 | ** | 0.37 | 1.99 | ** | 2.50 | ** |
| | M,H | <35 | <3 | 4.44 | 3.64 | ** | 0.28 | 1.60 | | | |
| | M | 35-64 | 3+ | 0.398 | 0.25 | | -0.19 | -1.64 | * | | |
| | M | 35-64 | <3 | 3.16 | 2.63 | ** | 0.18 | 1.64 | * | 2.47 | ** |
| | H | 35-64 | 3+ | 4.31 | 3.49 | ** | -0.34 | -2.22 | ** | 3.73 | ** |
| | H | 35-64 | <3 | 5.19 | 4.27 | ** | 0.27 | 2.18 | ** | 4.03 | ** |
| | M,H | 65+ | 3+ | 3.72 | 2.85 | ** | -0.04 | -0.15 | | | |
| | M,H | 65+ | <3 | 3.7 | 3.06 | ** | -0.02 | -0.14 | | | |
| RETAIL_DEN | M,H | <35 | 3+ | 0.0092 | 0.07 | | -0.04 | -1.34 | | | |
| | M,H | <35 | <3 | 0.027 | 0.85 | | 0.03 | 1.50 | | | |
| | M | 35-64 | 3+ | -0.016 | -0.23 | | 0.03 | 1.40 | | | |
| | M | 35-64 | <3 | 0.0387 | 1.2 | | 0.03 | 2.06 | ** | | |
| | H | 35-64 | 3+ | 0.0434 | 1.29 | | 0.04 | 2.48 | ** | | |
| | H | 35-64 | <3 | 0.0341 | 1.1 | | 0.03 | 2.06 | ** | | |
| | M,H | 65+ | 3+ | 0.0114 | 0.24 | | -0.01 | -0.20 | | | |
| | M,H | 65+ | <3 | 0.0287 | 0.89 | | 0.03 | 1.74 | * | | |
| SHARE_WHITE | M,H | <35 | 3+ | 3.91 | 2.43 | ** | 0.91 | 2.90 | ** | 1.83 | * |
| | M,H | <35 | <3 | 2.94 | 4.08 | ** | 1.11 | 2.98 | ** | 2.26 | ** |
| | M | 35-64 | 3+ | 2.11 | 2.86 | ** | 0.84 | 4.59 | ** | 1.67 | * |
| | M | 35-64 | <3 | 1.79 | 3.6 | ** | 0.56 | 3.14 | ** | 2.34 | ** |
| | H | 35-64 | 3+ | 4.65 | 5.22 | ** | 1.19 | 4.80 | ** | 3.75 | ** |
| | H | 35-64 | <3 | 4.06 | 6.12 | ** | 0.93 | 4.00 | ** | 4.45 | ** |
| | M,H | 65+ | 3+ | 0.499 | 0.48 | | 0.68 | 1.51 | | | |
| | M,H | 65+ | <3 | 3.15 | 5.6 | ** | 0.96 | 4.29 | ** | 3.62 | ** |
| ROOMS | M,H | <35 | 3+ | 0.541 | 4.04 | ** | 0.13 | 2.51 | ** | 2.84 | ** |
| | M,H | <35 | <3 | 0.242 | 2.43 | ** | -0.14 | -3.32 | ** | 3.54 | ** |
| | M | 35-64 | 3+ | 0.402 | 4.65 | ** | 0.11 | 5.21 | ** | 3.26 | ** |
| | M | 35-64 | <3 | 0.292 | 3.57 | ** | 0.04 | 1.61 | | | |
| | H | 35-64 | 3+ | 0.731 | 8.53 | ** | 0.19 | 7.21 | ** | 6.02 | ** |
| | H | 35-64 | <3 | 0.639 | 7.55 | ** | 0.10 | 4.10 | ** | 6.05 | ** |
| | M,H | 65+ | 3+ | 0.418 | 3.84 | ** | 0.16 | 5.67 | ** | 2.28 | ** |
| | M,H | 65+ | <3 | 0.377 | 4.5 | ** | 0.03 | 1.07 | | | |

**Table 10 (continued)**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ACC_SERVEMP_CAR** | M,H | <35 | 3+ | 7.73 | 1.87 | * | 2.40 | 2.41 | ** | 1.25 | |
| | M,H | <35 | <3 | 10.4 | 4.04 | ** | 1.66 | 2.04 | ** | 3.24 | ** |
| | M | 35-64 | 3+ | 6.01 | 2.34 | ** | 1.50 | 2.42 | ** | 1.71 | * |
| | M | 35-64 | <3 | 4.37 | 1.83 | * | 1.58 | 2.63 | ** | 1.13 | |
| | H | 35-64 | 3+ | 7.92 | 3.05 | ** | 3.35 | 4.97 | ** | 1.70 | * |
| | H | 35-64 | <3 | 11.9 | 4.85 | ** | 3.61 | 5.46 | ** | 3.26 | ** |
| | M,H | 65+ | 3+ | 10.6 | 3.2 | ** | 2.64 | 2.21 | ** | 2.26 | ** |
| | M,H | 65+ | <3 | 11.1 | 4.55 | ** | 3.00 | 4.42 | ** | 3.20 | ** |
| **DIFF_INC_HIGHER** | M,H | <35 | 3+ | -0.096 | -0.34 | | 0.03 | 0.38 | | | |
| | M,H | <35 | <3 | 0.166 | 0.79 | | 0.12 | 1.35 | | | |
| | M | 35-64 | 3+ | 0.0915 | 1.56 | | -0.05 | -2.38 | ** | | |
| | M | 35-64 | <3 | 0.0042 | 0.1 | | 0.02 | 1.01 | | | |
| | H | 35-64 | 3+ | -1.99 | -3.89 | ** | -0.19 | -2.80 | ** | -3.50 | ** |
| | H | 35-64 | <3 | -2.09 | -4.18 | ** | -0.20 | -2.64 | ** | -3.74 | ** |
| | M,H | 65+ | 3+ | -0.146 | -0.67 | | 0.11 | 1.10 | | | |
| | M,H | 65+ | <3 | -0.1 | -0.83 | | 0.01 | 0.41 | | | |
| **DIFF_INC_LOWER** | M,H | <35 | 3+ | -0.154 | -0.39 | | -0.05 | -0.48 | | | |
| | M,H | <35 | <3 | 0.612 | 1.91 | * | 0.12 | 1.04 | | | |
| | M | 35-64 | 3+ | -0.302 | -4.01 | ** | -0.07 | -3.07 | ** | -2.99 | ** |
| | M | 35-64 | <3 | -0.207 | -3.82 | ** | -0.05 | -2.23 | ** | -2.57 | ** |
| | H | 35-64 | 3+ | -14.1 | -2.64 | ** | -0.13 | -1.13 | | | |
| | H | 35-64 | <3 | -15.2 | -2.98 | ** | 0.00 | 0.00 | | | |
| | M,H | 65+ | 3+ | 0.23 | 0.57 | | 0.03 | 0.22 | | | |
| | M,H | 65+ | <3 | 0.458 | 4.13 | ** | 0.07 | 1.88 | * | 3.26 | ** |
| **POP_DEN** | M,H | <35 | 3+ | -0.037 | -2.67 | ** | -0.001 | -0.32 | | | |
| | M,H | <35 | <3 | -0.016 | -2.36 | ** | -0.002 | -1.09 | | | |
| | M | 35-64 | 3+ | -0.035 | -4.29 | ** | -0.006 | -3.06 | ** | -3.55 | ** |
| | M | 35-64 | <3 | -0.015 | -2.38 | ** | -0.004 | -2.47 | ** | -1.68 | * |
| | H | 35-64 | 3+ | -0.020 | -2.68 | ** | -0.003 | -1.62 | | | |
| | H | 35-64 | <3 | -0.016 | -2.52 | ** | -0.004 | -2.24 | ** | -1.84 | * |
| | M,H | 65+ | 3+ | -0.056 | -4.85 | ** | -0.004 | -1.07 | | | |
| | M,H | 65+ | <3 | -0.023 | -3.53 | ** | -0.005 | -2.56 | ** | -2.67 | ** |
| Sample | | | | | 5000 | | | 4785 | | | |
| Final log-like: | | | | | -7875.7 | | | -8670.4 | | | |
| LRT: | | | | | 6220.85 | | | 3686.6 | | | |
| Rho-sq: | | | | | 0.28 | | | 0.18 | | | |
| Adj rho-sq: | | | | | 0.28 | | | 0.17 | | | |

Prediction test

The prediction test seeks to measure how accurately the models predict location choices. The test compares the models' predictions of agent types for each location (that is, which agent type has the highest probability of being located at a given location) to the actual agent location. Aside from determining the accuracy of the different models, this test can shed light on the impact of the differences in coefficients between the models (calculated previously) on model accuracy. Table 11 and 12 present the prediction test results for the individual (unconstrained) models.

The rows in the table indicate the agent types observed in the different locations, and the columns show predicted agents for these locations according to the model. So, for example, there are 106 locations occupied by low –income households in the Infogroup data. The model accurately predicts 1.9% of the times (2 out of 106) that a low-income household is most likely to occupy these locations. For 90.6% of these locations, the model predicts the most likely agent type to be found is mid-income, mid-age household of less than 3 people. And for 0.9% of the locations, the model predicts the most likely agent type to be found is mid-, high-income senior households of less than 3 people. The diagonal of the table indicates the percentage of agent types that were predicted accurately for the locations in question.

The Infogroup model has a higher general accuracy than the Travel Survey model (38.6% vs. 32.3%). Neither model is able to accurately predict the locations of mid-, high-income, young, large households and mid-, high-income, senior, large households. Additionally, the Infogroup model is not able to accurately predict the location of mid-income, mid-age, large households and the Travel Survey the locations of senior, mid- and high-income, large households. The agent types with more observations have the highest prediction accuracy. This may be due to the use of Lerman and Kern's specification, which includes the size of the groups.

To distinguish between differences in accuracy that arise from the size of the categories, and those that arise from differences in model estimation (e.g. differences in specific coefficients), the prediction test is done on the Infogroup model using the Travel Survey data. The results are presented in Table 13.

The total accuracy of the model decreases from 38.6% to 18.2%. The level of prediction accuracy for individual agent types is similar to the prediction test with the Infogroup data. It also seems to be independent of the number of observations per category in the data that used in the test.

**Table 11 Prediction test. Infogroup model with Infogroup data**

| | | Model Prediction | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Obs. | Agent Type (Inc/Age/Size) | L / All / ALL | M,H / <35 / 3+ | M,H / <35 / <3 | M / 35-64 / 3+ | M / 35-64 / <3 | H / 35-64 / 3+ | H / 35-64 / <3 | M,H / 65+ / 3+ | M,H / 65+ / <3 |
| 106 | L / All / ALL | 1.9% | | | | 90.6% | | 0.9% | | 6.6% |
| 40 | M,H / <35 / 3+ | | | | | 62.5% | 2.5% | 25.0% | | 10.0% |
| 173 | M,H / <35 / <3 | 0.6% | | 2.3% | | 57.2% | 0.6% | 24.9% | | 14.5% |
| 431 | M / 35-64 / 3+ | | | | | 82.1% | 0.2% | 5.3% | | 12.3% |
| 1422 | M / 35-64 / <3 | 0.1% | | | | 75.5% | | 8.4% | | 16.0% |
| 505 | H / 35-64 / 3+ | | | | | 26.5% | 2.0% | 56.2% | | 15.2% |
| 1029 | H / 35-64 / <3 | | | 0.3% | | 24.5% | 1.7% | 57.9% | | 15.5% |
| 84 | M,H / 65+ / 3+ | 1.2% | | | | 59.5% | | 19.0% | 1.2% | 19.0% |
| 1210 | M,H / 65+ / <3 | 0.2% | | 0.1% | | 55.0% | 0.1% | 24.3% | | 20.3% |
| 5000 | | | | | | | | | | 38.6% |

**Table 12 Prediction test. Travel Survey model with Travel Survey data**

| | | Model Prediction | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Obs. | Agent Type (Inc/Age/Size) | L / All / ALL | M,H / <35 / 3+ | M,H / <35 / <3 | M / 35-64 / 3+ | M / 35-64 / <3 | H / 35-64 / 3+ | H / 35-64 / <3 | M,H / 65+ / 3+ | M,H / 65+ / <3 |
| 430 | L / All / ALL | 3.0% | | | 53.3% | 36.7% | 5.8% | 1.2% | | |
| 145 | M,H / <35 / 3+ | 2.8% | | | 51.7% | 33.1% | 9.7% | 2.8% | | |
| 99 | M,H / <35 / <3 | | | | 46.5% | 49.5% | 2.0% | 2.0% | | |
| 1274 | M / 35-64 / 3+ | 0.5% | | | 73.5% | 15.2% | 9.4% | 1.3% | | |
| 1011 | M / 35-64 / <3 | 0.6% | | | 55.8% | 34.0% | 7.3% | 1.8% | | 0.5% |
| 699 | H / 35-64 / 3+ | 0.3% | | | 57.8% | 7.2% | 32.0% | 2.7% | | |
| 468 | H / 35-64 / <3 | 0.2% | | | 49.8% | 22.9% | 20.5% | 5.3% | | 1.3% |
| 86 | M,H / 65+ / 3+ | 3.5% | | | 62.8% | 19.8% | 12.8% | 1.2% | | |
| 573 | M,H / 65+ / <3 | 1.2% | | 0.2% | 56.9% | 25.5% | 13.1% | 2.4% | | 0.7% |
| 4785 | | | | | | | | | | 32.3% |

**Table 13 Prediction test. Infogroup model with Travel Survey data**

| | | Model Prediction | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Total Obs. | Agent Type (Inc/Age/Size) | L / All / ALL | M,H / <35 / 3+ | M,H / <35 / <3 | M / 35-64 / 3+ | M / 35-64 / <3 | H / 35-64 / 3+ | H / 35-64 / <3 | M,H / 65+ / 3+ | M,H / 65+ / <3 |
| 430 | L / All / ALL | 1.4% | | | | 56.0% | | 34.4% | | 8.1% |
| 145 | M,H / <35 / 3+ | | | | | 35.9% | 1.4% | 53.8% | | 9.0% |
| 99 | M,H / <35 / <3 | 1.0% | | | | 42.4% | | 47.5% | | 9.1% |
| 1274 | M / 35-64 / 3+ | 0.5% | | | | 39.4% | 1.3% | 49.9% | 0.1% | 8.8% |
| 1011 | M / 35-64 / <3 | 0.3% | | | | 46.0% | 0.6% | 44.9% | | 8.2% |
| 699 | H / 35-64 / 3+ | 0.3% | | | | 12.9% | 1.7% | 78.0% | | 7.2% |
| 468 | H / 35-64 / <3 | | | | | 21.4% | 0.6% | 68.4% | | 9.6% |
| 86 | M,H / 65+ / 3+ | | | | | 45.3% | 1.2% | 43.0% | | 10.5% |
| 573 | M,H / 65+ / <3 | 0.9% | | | | 35.1% | 0.7% | 51.3% | 0.2% | 11.9% |
| 4785 | | | | | | | | | | 18.2% |

Summary

The likelihood ratio test indicates that the estimation results vary depending on the dataset used. Comparing coefficients across the two models suggests that the difference in the income distribution of the households between the two datasets may account for some of the differences in the estimation results. The prediction test shows similar total accuracy for both models when applied to the data used for estimation; however, accuracy by agent type varies significantly. For these tests, predictions' accuracy by agent type seems to be positively correlated with the size of the agent categories. The prediction test for the Infogroup model using the Travel Survey data presents a lower total accuracy but a similar level of accuracy by agent type to the prediction tests with the Infogroup data. This seems to suggest that (1) the model estimation is sensitive to the number of observations per category, and (2) the model accuracy for individual agent types is sensitive to the size of the groups (representation) in the data used to estimate the model, but not in the data used in the prediction test. In other words, the models appear to be good at predicting the location of the agents that were well represented (large sample) in the data used for the estimation.

Given the apparent relationship between number of observations per category in the estimation data and model accuracy, it is difficult to determine the impact of the differences in estimation (i.e. different coefficients identified in the preference stability test) on the prediction accuracy of the models.

One approach to determining which model is better, as well as the full impact of the differences between the models, is to run the prediction test with the entire population from the census. This test was not done due to time limitations.

It is also important to keep in mind that in general (but especially for certain applications) a good model would have a good level of both total accuracy and individual accuracy by agent type. That is, we want a model that represents well the majority of location choices, and also is able to represent well each individual agent type.

In spite of their differences, both models show similar results in terms of relative location preferences between agent types. In terms of the relationship between life stage and location choice, the following conclusions can be drawn from the models:

- In general, income level seems to have a bigger impact on location preferences than the age of the head of the households or the household size. This can be seen in the fact that, for the same variables (i.e., same location attribute) in the same model estimation, the largest differences in magnitude of the coefficients are mostly, first, between low income households and the rest, and second, between mid-age mid-income and mid-age high-income households. For most of the attributes, the coefficients for young households and for seniors (which combine mid- and high income) often fall between those of mid-age mid-income and mid-age high-income households.
- Preference for the size of the unit seems to be driven more by household size than by age of the head of the household. Senior households of 2 or less persons do seem to value larger units less compared to large households. This might suggest that seniors move to smaller units after the children leave the nest. However, it is not clear if these households were ever in a larger unit or if they were living with their children before.

Possible limitations of the models include:

- Heterogeneity within categories. In an estimation by household categories, all the households within a category are assumed to have the same preferences with respect to location attributes in the bid function. This strong assumption simplifies the analysis and the implementation of the models in forecasting exercises (it is easier to forecast households categories than individual households). However, heterogeneity within agent categories might result in poor model fit and insignificant coefficients. An alternative to an estimation by categories of agents is an estimation based on individual households. Subsection 3.2.3 presents this approach.
- Omitted variable bias. A better characterization of the households and the residential units might be needed. Clearly, when choosing a location, households look at more than neighborhood characteristics, the number of rooms, and whether the unit is single or multifamily. I tried to account for some of these characteristics in some specifications, including proximity to amenities to the coast and greenspaces; but these were not significant and omitted from my final models.
- Errors due to the spatial definition of the zone. The zone structure used to implement the zonal characteristics comes from a transportation model, as that model supplies the transportation levels of service and also provides the zone structure for a land use forecasting model. However, this zone structure may add errors in the models given the level of resolution and the size variation from zone to zone. The relatively low level of resolution in

some areas can result in low variability in zonal characteristics between observations that fall in a large zone. The variation in size results in an uneven comparison. For example, the calculation of the retail density in a small zone well characterizes the surrounding area of a household located in that zone. But for a large zone, the average retail density might not be the same as the retail density in the immediate surroundings of a particular residential unit located in that zone. One alternative would be to determine a buffer distance for each residential unit and characterize the area within that buffer. This approach is computationally intensive and also limits the implementation of the location models in aggregate integrated Land Use-Transportation models, which are based on Transport Analysis Zones.

- Errors arising from utilizing outputs from other models, specifically the travel time matrices from the 4-step transportation model. Any errors in travel time estimation from the 4-step model propagate into the location choice model.
- Errors in the functional forms. These might result in biased estimation and poor model fit. Several variable transformations were evaluated, but additional analysis on this regard might be needed. For example, the utility (bid) with respect to certain variables might have a structural break. That is, the relationship between utility and the explanatory variables is not constant. This might be the case for accessibility. Agents might value differently being within a 15-minute, 45-minutes, or 2-hour drive from certain opportunities. In these cases, a piece-wise linear or Box-Cox transformation might be necessary.
- Model structure errors. If the heterogeneity within agent types varies from agent to agent, the variance in the location probability might not be the same for all agents. This violates an assumption on the multinomial logit. It might also be the cases that household' location preferences vary depending on higher-level decisions, such as location in the city vs. location on the suburbs. In this case, a nested logit specification would be more appropriate.

### 3.2.3. Model estimation by individual households

Estimating a location choice model by categories of households assumes that all households within a category have the same location preferences. In reality, characterizing households by a crude income level or size masks greater differences within those categories while also eliminating the possibilities to include other household attributes of relevance, such as vehicle ownership. Capturing such variations within a category-based model would require a large number of groups (cross-categorization of different factors). An alternative approach is to define individual households, and not household groups, as the agents bidding for units. So, in theory, the choice set for a given unit consists of all the individual households in the study region.

Estimating a model with a choice set this large would be too computationally intensive and probably impractical. Instead, a random sample of households can be selected as the choice set for each of the units in the analysis (McFadden, 1978).

In this approach, only one bid function is defined. The specification allows the differences in preferences between households to be evaluated by interacting household attributes with location attributes in the bid function. That is, household attributes are explicitly formulated in the bid function. For the group formulation, these characteristics are implicit in the differences in coefficients by agent type.

I tested this approach using the Travel Survey data since it has more household attributes than the Infogroup data. I include renters in the estimation. For each unit, I randomly sampled twenty "alternatives" (households) for each unit.

I tested different specifications, using the same variables used in the estimation by groups. The best bid function is:

$$\begin{aligned} BID = {} & \beta_{ACC\_TR}(ACC\_TR \cdot NO\_VEH \cdot METRO) \\ & + \beta_{ACC\_CAR}(ACC\_CAR \cdot VEH \cdot WRK \cdot HIGH\_INC) \\ & + \beta_{FAR}(FAR \cdot VEH) + \beta_{INC}(INC\_DIFF) \\ & + \beta_{TYPE}(UNIT\_TYPE \cdot RENTER) \\ & + \beta_{RACE}(SHARE\_WHITE \cdot HH\_RACE) \\ & + \beta_{SCHOOL}(SAT \cdot STUD \cdot HH\_TYPE) + \beta_{SIZE}(SIZE\_DIFF) \\ & + \beta_{AMEN}(LU\_AMEN \cdot HIGH\_INC) \end{aligned} \tag{3.2}$$

where:

$\beta_{ACC\_TR}$: Measures the effect of the transit accessibility of the zone to total jobs ($ACC\_TR$), for a household with no vehicles ($NO\_VEH$) and a unit in the metro area ($METRO$).

$\beta_{ACC\_CAR}$: Measures the effect of car accessibility of the zone to service jobs ($ACC\_CAR$), for a household with cars (VEH), workers (WRK), and high-income ($HIGH\_INC$).

$\beta_{FAR}$: Measures the effect of the built density (Floor-Area-Ratio) for households that own vehicles ($VEH$). The FAR aims at capturing the parking availability in the zone.

$\beta_{INC}$: Measure the effect of differences between the household's income and the median income of the area. The variables $INC\_DIFF$ is defined as the absolute value of the difference between the income of the household and the median income of the zone.

$\beta_{TYPE}$: Measures the effect of units in multifamily buildings ($UNIT\_TYPE$) for households that are renters ($RENTER$).

$\beta_{RACE}$: Measures the effect of the ratio of white population to total population in the zone ($SHARE\_WHITE$) for households that are identified as white ($HH\_RACE$) in the Travel Survey.

$\beta_{SCHOOL}$: Measures the effect of school quality (SAT score index of the zone) for family households ($HH\_TYPE$) with students ($STUD$). The household type was included in order to evaluate the effect of school quality in the zone only for families with children, and not for households of, for example, a group a college students

$\beta_{SIZE}$: Measures the effect of a mismatch between the size of the household and the size of the unit. I define the variable $SIZE\_DIFF$ as the absolute value of the difference between the household size and the number of rooms, divided by the household size. The variable is based on the assumption that households want housing units that match their space requirements, not necessarily space, per se. The difference or mismatch between the size of the unit and the size of the household is divided by the size of the household in order to account for the fact that a one-room mismatch is not as important for large household and units (7 people can accommodate in a 6 room unit), as it is for small unit and households.

$\beta_{AMEN}$: Measures the effect of amenities (e.g. restaurants, theaters, museums) for households of high-income level. Amenities in the zone are measured as the sum of the total built area of amenities in the zone based on the Level 3 Parcel data. This variable was normalized to values between 0 and 1 to provide an index of built area amenities.

The estimation results are presented in Table 14.

**Table 14 Estimation results by individual households**

| Beta | Value | Std. err | t-test | |
|---|---|---|---|---|
| $\beta_{ACC\_TR}$ | 0.276 | 0.205 | 1.35 | |
| $\beta_{ACC\_CAR}$ | 4.03 | 0.221 | 18.23 | ** |
| $\beta_{FAR}$ | -0.703 | 0.102 | -6.86 | ** |
| $\beta_{INC}$ | -0.986 | 0.0432 | -22.81 | ** |
| $\beta_{TYPE}$ | 1.55 | 0.0613 | 25.29 | ** |
| $\beta_{SHARE\_WHITE}$ | 0.575 | 0.0623 | 9.22 | ** |
| $\beta_{SCHOOL}$ | 0.132 | 0.0354 | 3.73 | ** |
| $\beta_{SIZE}$ | -0.531 | 0.0499 | -10.62 | ** |
| $\beta_{AMEN}$ | 0.00876 | 0.0072 | 1.22 | |
| Rho-square: | | 0.048 | | |
| Adjusted rho-square: | | 0.048 | | |

Most of the coefficients are significant and their signs are reasonable. The model can be interpreted as follows:

$\beta_{ACC\_TR}$: Not Significant. If significant, the positive value would indicate that households with no cars value transit accessibility positively for locations in the metro area (where there is a denser transit network).

$\beta_{ACC\_CAR}$: High-income households, with cars, and where someone in the household works, value car accessibility to service employment positively.

$\beta_{FAR}$: Households that have vehicles prefer low FAR, likely due to better parking availability in areas with low FAR

$\beta_{INC}$: Households negatively value a mismatch between their income and the median income of the zone. This suggests that household want to live among people with the same income level.

$\beta_{SHARE\_WHITE}$: White households positively value being in areas with high ratios of white populations.

$\beta_{TYPE}$: Renters prefer multifamily units.

$\beta_{SCHOOL}$: Family households with students value school quality positively

$\beta_{SIZE}$: Households negatively value a mismatch between household size and unit size.

$\beta_{AMEN}$: High-income households value proximity to amenities positively.

The low rho-square indicates that the selected variables have a relatively low explanatory power. The model provides behavioral insights, but would be problematic for forecasting exercises. Even with a high rho-square, implementation of these models in forecasting exercises in a land use model is a general limitation of this approach. To use this formulation to predict where future population would locate, would require a microsimulation approach, requiring a synthetic population with all the relevant attributes of the households and units being forecast.

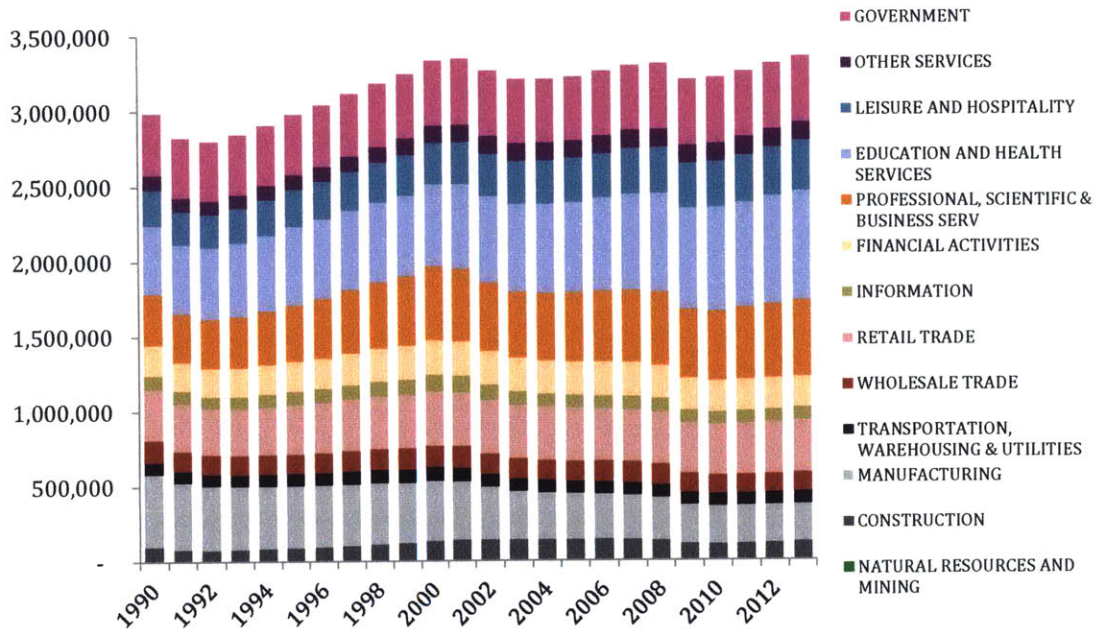
## 4. NON RESIDENTIAL LOCATION CHOICE – FIRMS

We now turn to the question of firm location choice in Greater Boston. Similar to the case of households, in this chapter I first paint a general picture of the urban dynamics over time, then describe the data used in the models and finally present the model specifications, results and interpretations.

### 4.1. Employment Dynamics

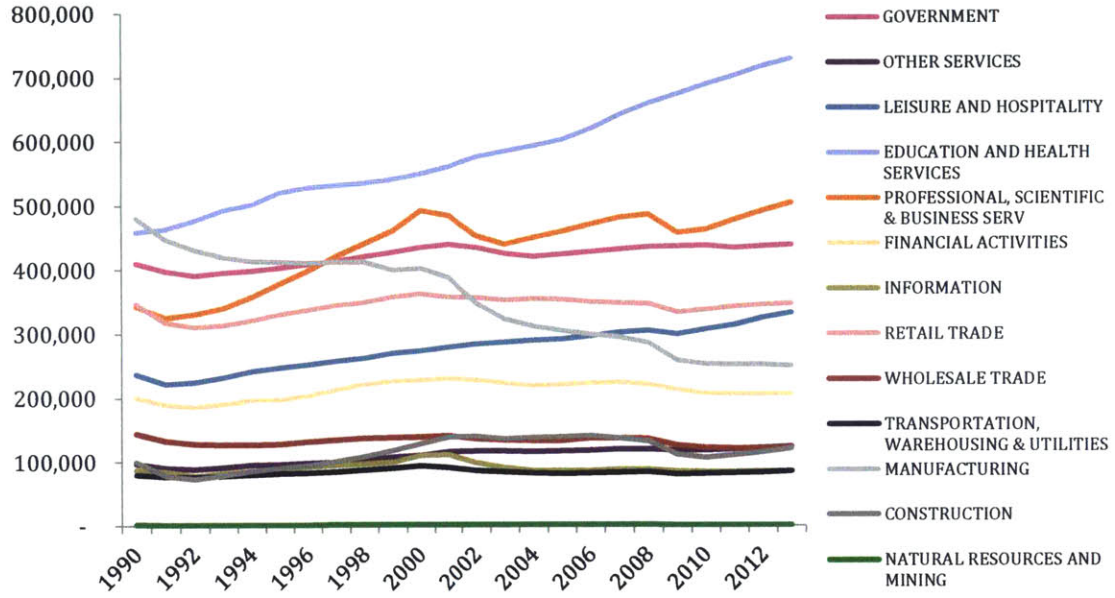
Statewide, employment in Massachusetts grew rapidly from 1990 to 2000 and then more slowly from 2000 to today. Job recovery was faster after the financial crises at the end of 2008 than after the *dot.com* crises in 2002 (Figure 15). Employment has increased rapidly in the service sector, especially in the education, health services, and amenity (arts, entertainment, leisure, hospitality) industries, while it has steadily declined in manufacturing. Professional, scientific, and business related service jobs had a large increase between 1990 and 2000, after which there seems to be stabilization. Jobs in this sector present the highest volatility and sensitivity to economic shocks (Figure 16).

**Figure 15 Employment evolution in Massachusetts, Total by Industry**




*Source: Massachusetts Executive Office of Labor and Workforce Development*

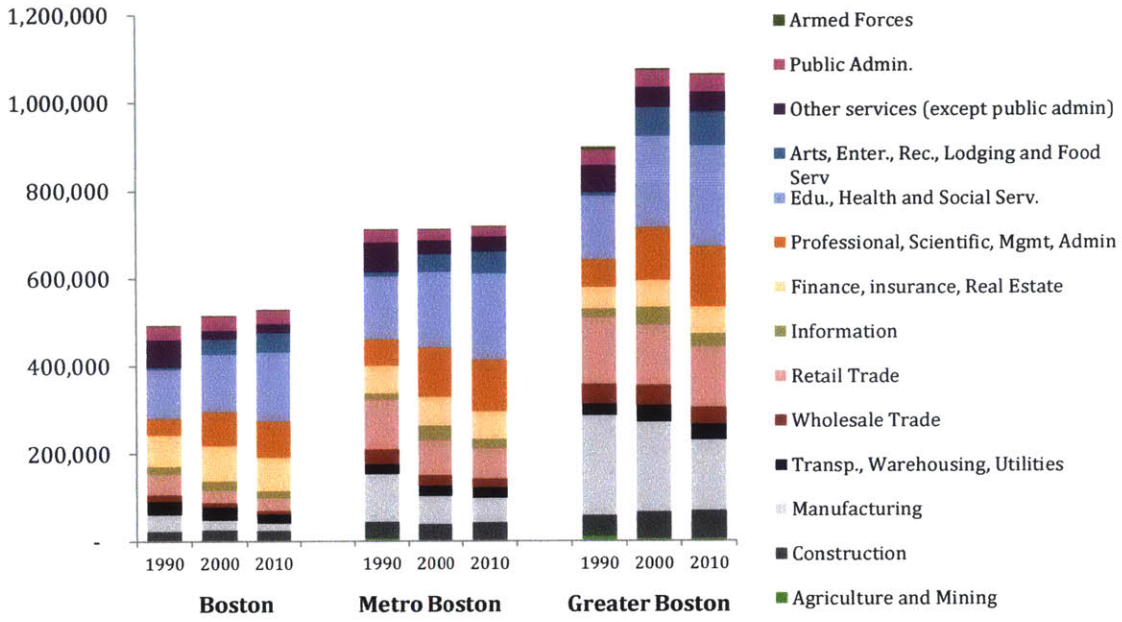
**Figure 16 Employment evolution in Massachusetts by Industry**




*Source: Massachusetts Executive Office of Labor and Workforce Development*

Following the same subdivision of the Boston Metropolitan area used in the residential sector, we can observe that the largest increase in jobs took place in the Greater Boston region followed by the city of Boston (Figure 14). Within each sub-area, the employment changes by industry follow similar trends. That is, neither the increase in service jobs nor the decrease in manufacturing jobs was concentrated in any of the tree sub areas. Services such as maintenance and repairs (*Other Services* category) decreased more in Boston and the Metro Area than in the Greater Boston region.

**Figure 17 Employment evolution in Greater Boston**

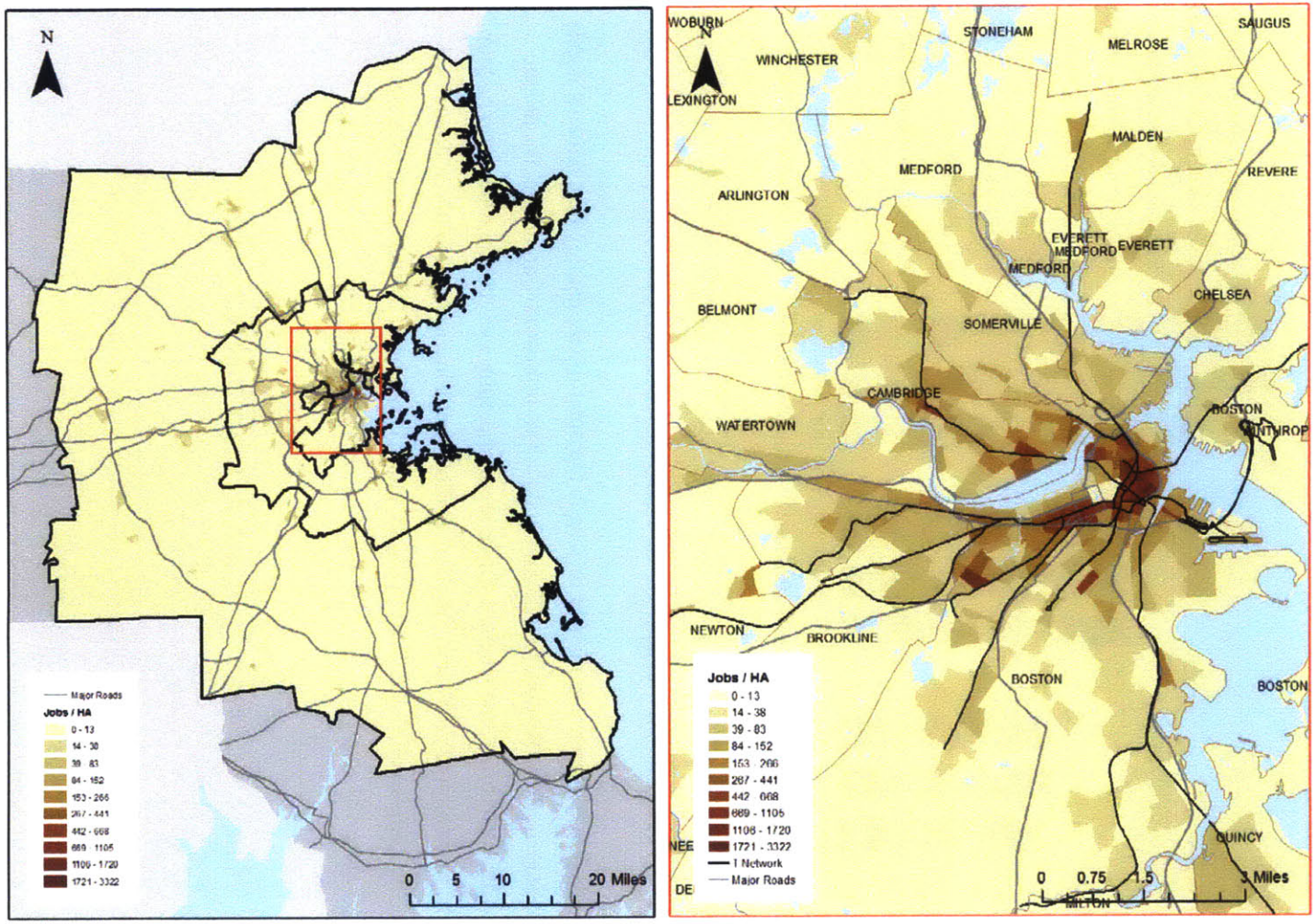



*Source: CTPP*

Spatial distribution of employment

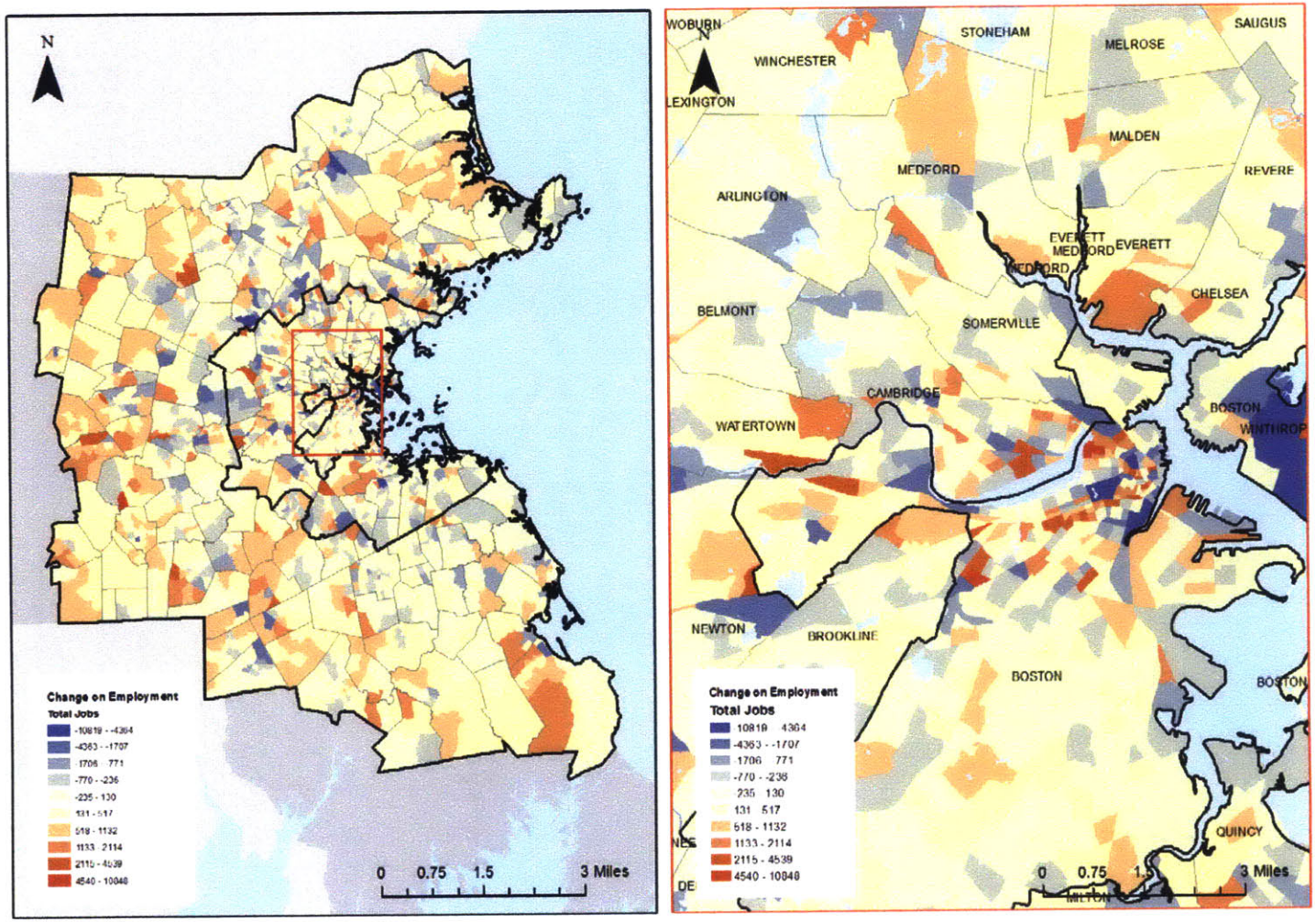
Despite the larger increase in number of jobs in the outer towns, employment remains highly concentrated in Metro Boston, specifically in the cities of Boston, Cambridge, Brookline, Somerville, Watertown, and Chelsea. Other areas of employment concentration can be found along the main highways. Within the core, employment concentrates mainly in downtown Boston, the Longwood Medical Area (LMA), along Boylston St. (Prudential Center), and Kendall Square (Figure 18).

**Figure 18 Employment density in 2010**



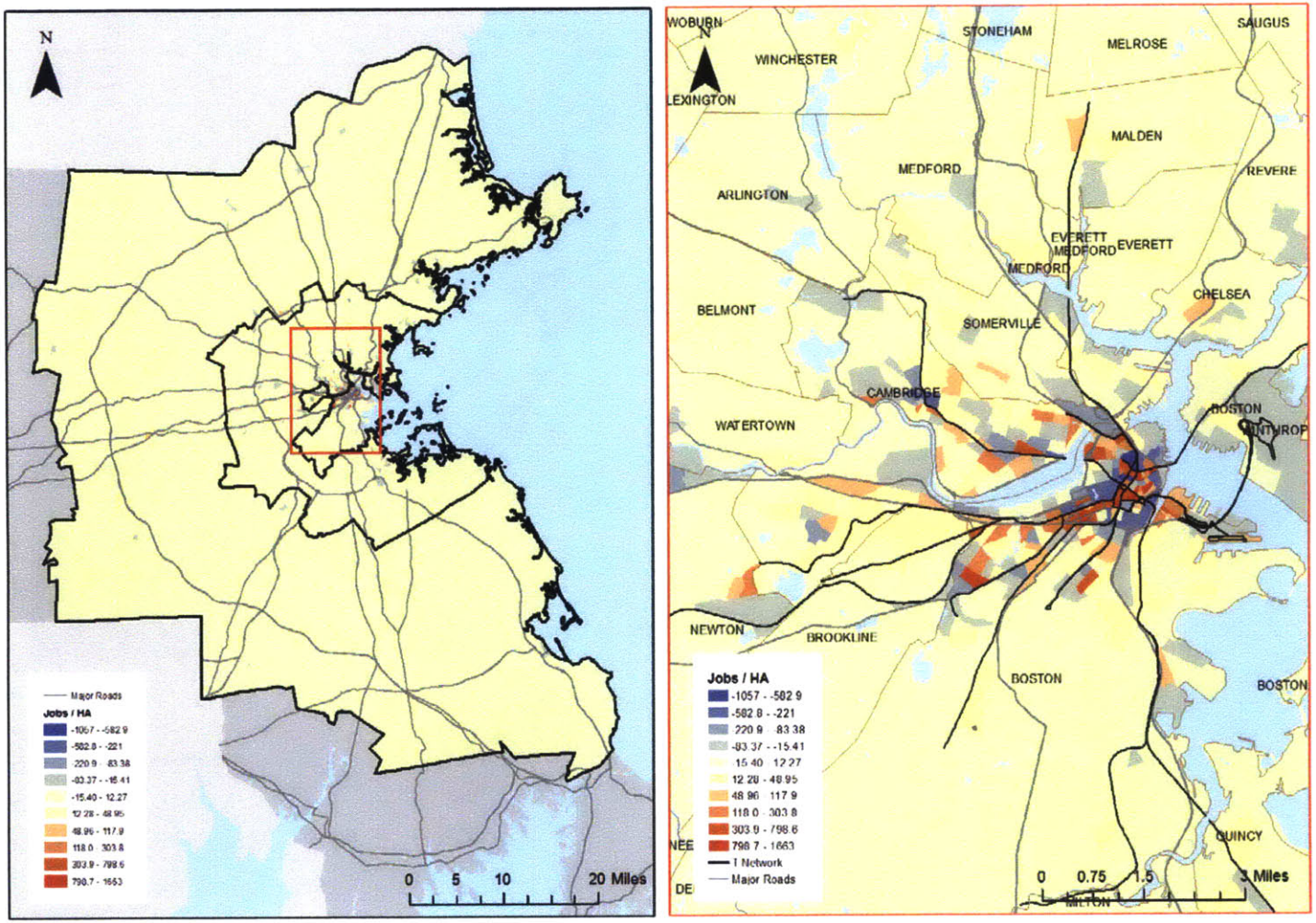

*Source: CTPP*

**Figure 19 Change in employment from 1990 to 2010**




*Source: CTPP*

**Figure 20 Change in employment density between 1990 and 2010**




*Source: CTPP*

Outside of the Metro area, the increase in employment has been concentrated along the highway 495 and the corridor between Boston and Providence. The change in total employment has been spread throughout the region, however the employment distribution structure in Greater Boston has not changed much, except in the core (mainly Boston and Cambridge) where employment density has increased in specific areas such as the Longwood Medical Area (LMA), Kendall Square, in the Charles MGH area, and along Boylston Street (Figure 20).

Employment spatial relationships

As a first approach to analyzing firms' preferences for clustering and agglomeration economies, I analyzed the spatial relationship of jobs by industry. From a production function perspective, firms need access to labor (of different types) and to suppliers and customers. A firm might value access to each of these production factors differently, depending on the nature of its business or the firm's industry. The aggregated analysis of the spatial relationship between jobs provides some initial insights on firm preferences to locate close to one another. To analyze which industry sectors prefer to be closer together, I focus on intra-sector spatial relationships. To analyze the spatial relationship between employment by industry, I define the following metrics.

*Intra-Sector Employment Accessibility (ISEA):*

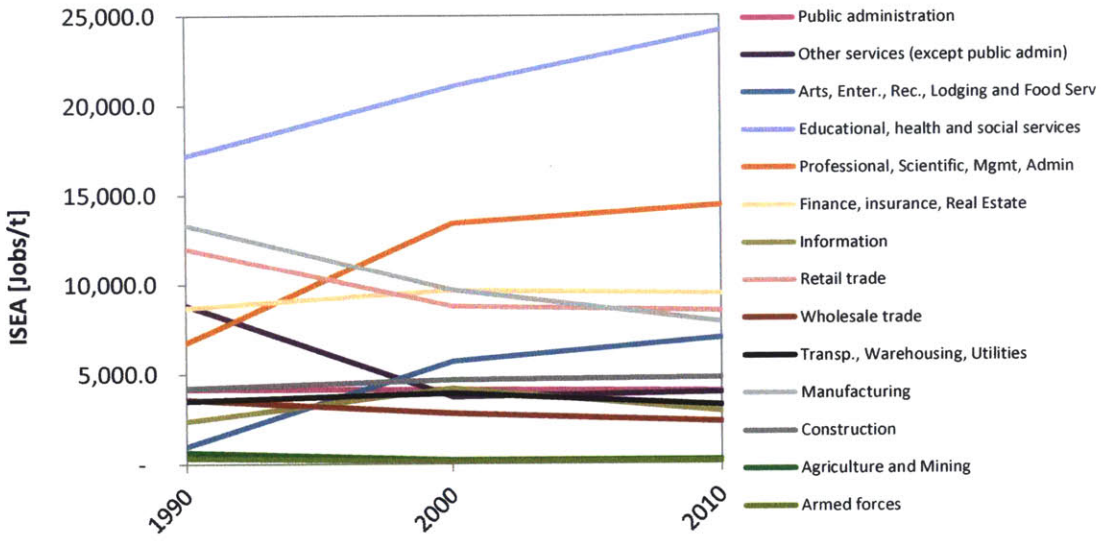
This metric represents the relative ease of jobs within the same sector of accessing each other, measured in Jobs/Minutes. The ISEA takes the form of a gravity measure: for jobs in a specific industry $k$ in a given area $i$ ISEA is defined as the sum of the jobs in industry $k$ in all areas $j$ (including zone those in area $i$) divided by their corresponding travel time $t_{ij}$.

$$ISEA_i^k = \sum_j \frac{Jobs_j^k}{t_{ij}}. \quad (4.1)$$

Thus, ISEA can increase either by in increase in the total number of jobs, or a decrease in travel time (jobs coming closer together), or both. I calculate this metric using inter-zonal travel time matrices for auto travel in the AM period, which come from the Boston area 4-Step

Transportation Model of 1990, 2000, and 2010.[4] Not surprisingly, the employment sectors with the largest increases in ISEA are the ones that presented the largest increase in total number of jobs: Education and Health Services; Leisure and Hospitality Services; and Professional, Scientific Business, Services. (Figure 21)

**Figure 21 Average ISEA by industry**



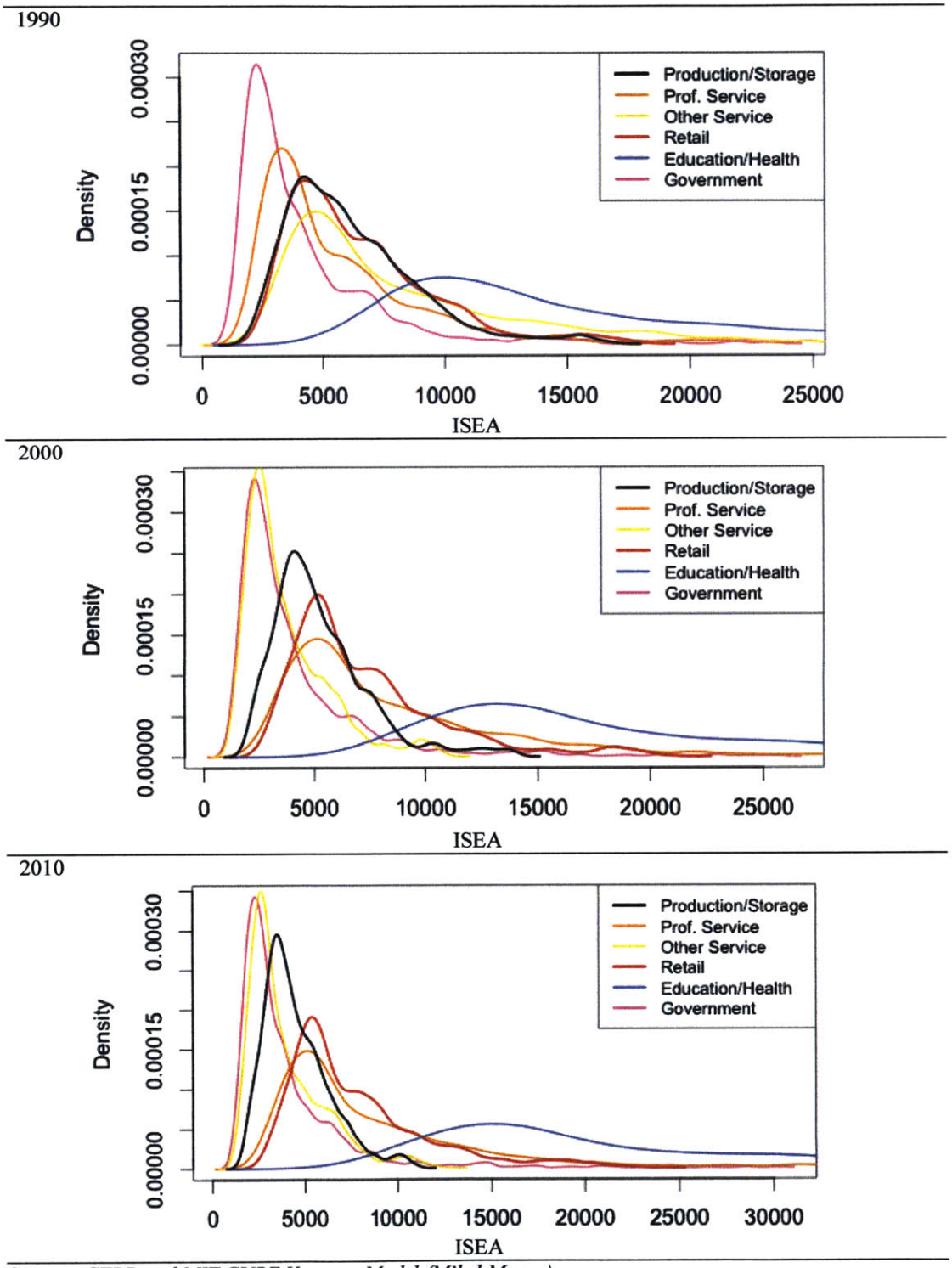

*Source: CTPP and MIT CUBE Voyager Model (Mikel Murga)*

The evolution of the distribution of the average ISEA by industry super-sectors, defined as Retail (retail trade, leisure, hospitality), Professional Service (professional, scientific, business, information), Other Services, Education and Health Services, Government, and Production and Storage (construction, manufacturing, transportation, warehousing, utilities, wholesale), suggests a general increase in ISEA (a shift to the right in the distribution) in professional service, retail, and education/health sectors. Employment in the Professional Service super-sector also present an increase in the spread of the distribution. On the other hand, production and the other services sectors show a decrease in mean ISEA (shift to the left) as well as a decrease in the spread of their distributions. (Figure 22)

---

[4] *Travel time matrices come from a 4-step transport model developed by Mikel Murga at MIT*

**Figure 22 ISEA by industry super-sector by time period**




*Source: CTPP and MIT CUBE Voyager Model (Mikel Murga)*

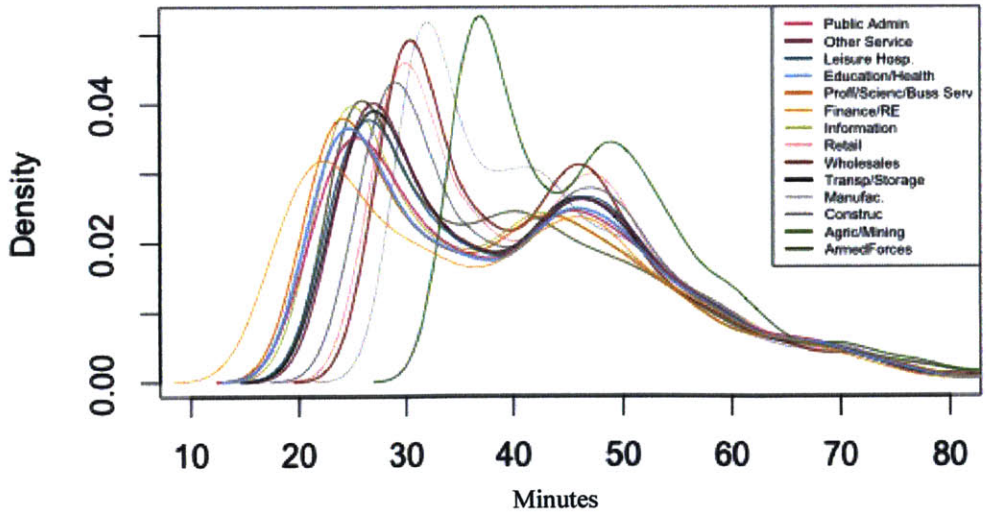
*Weighted average travel time (WATT):*

The objective of defining a weighted average travel time (WATT) is to distinguish between an increase in total number of jobs and the change in the spatial distribution of jobs over time, which cannot be done with the ISEA measure. WATT of employment in a given sector $k$ for a given area $i$ is defined as the sum of travel time to all other areas $j$, adjusted by the proportion of jobs in the industry $k$ in the area $j$ to total number of jobs in that industry in the whole region. In other words, travel time to zones with a high number of jobs weighs more than that to zones with a small number of jobs:

$$WATT_i^k = \frac{1}{Total\ Jobs^k} \sum_j t_{ij} \cdot Jobs_j^k. \tag{4.2}$$

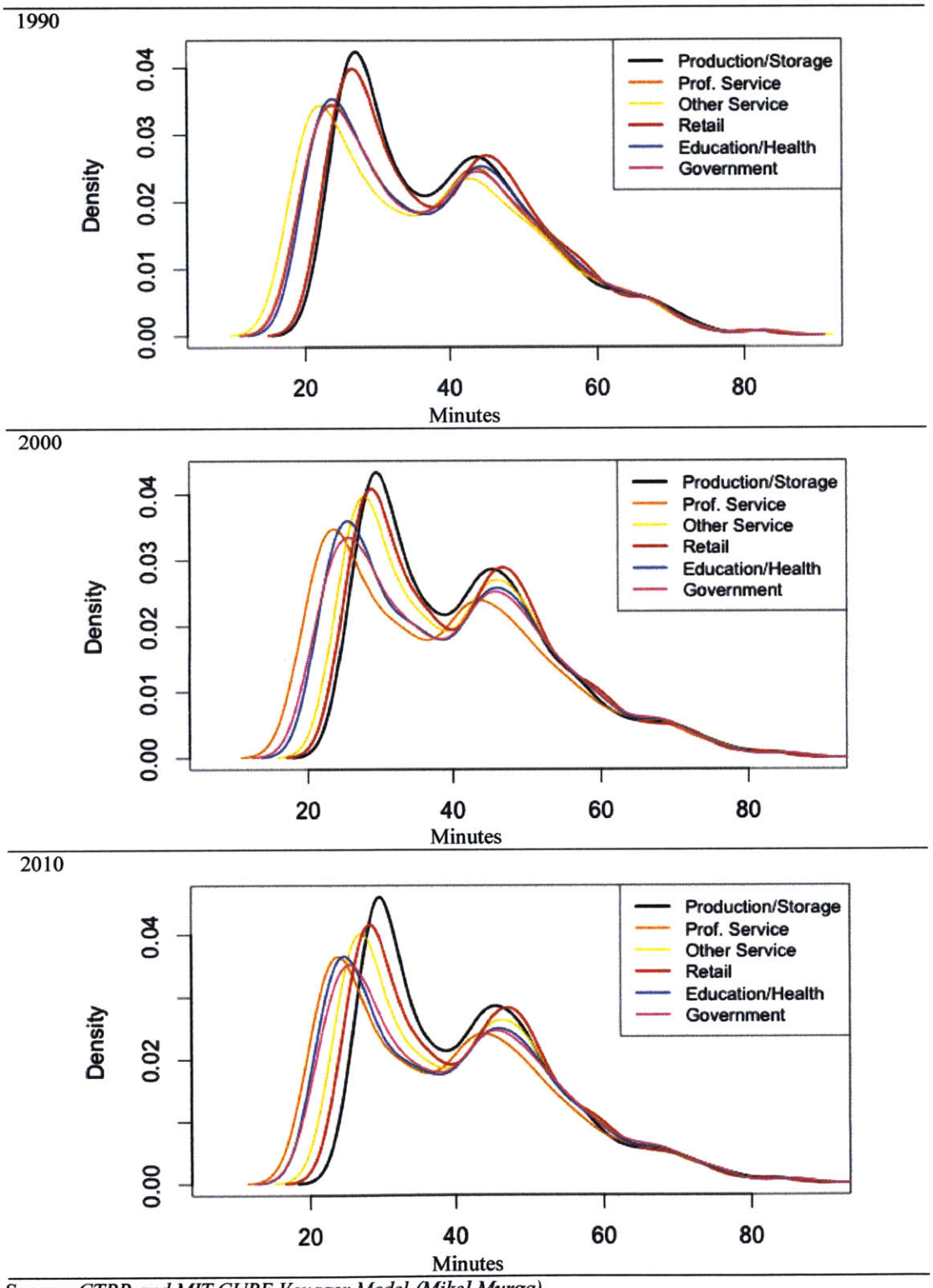
The shape of the density distribution of WATT reflects the spatial structure of the city: the downtown area has the highest concentration of employment, then sub-centers exist with lower, but still high employment concentrations. In general, the service sectors (e.g. finance, professional, science, business, education and health, and government services) present higher spread in the WATT distribution, which would indicate a more even spatial distribution of jobs in those sectors compared to jobs in the manufacturing, wholesales, storage, retail, and agriculture sectors (Figure 23)

**Figure 23 WATT distribution by industry in 2010**




*Source: CTPP and MIT CUBE Voyager Model (Mikel Murga)*

**Figure 24 WATT by industry super-sector by time period**

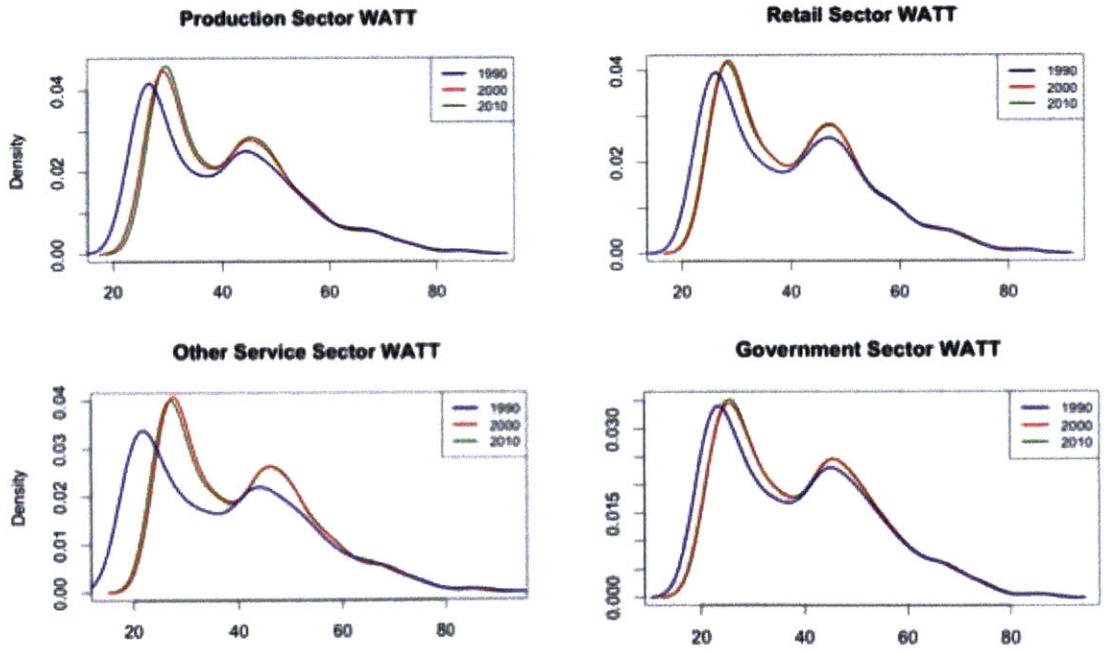


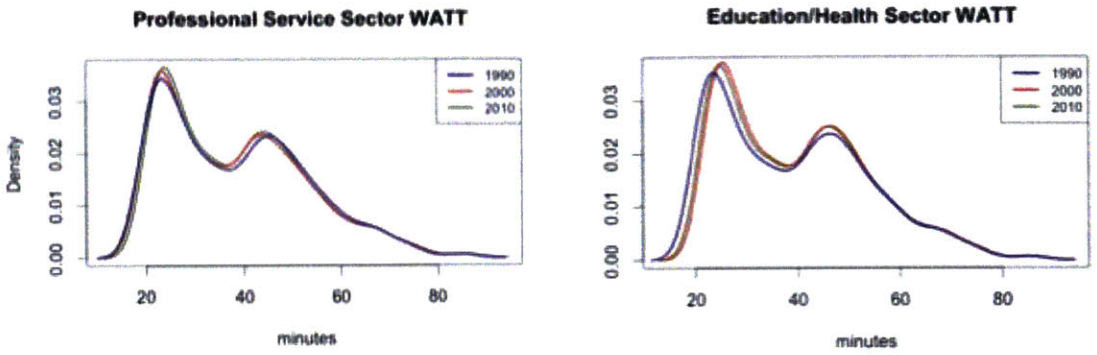

*Source: CTPP and MIT CUBE Voyager Model (Mikel Murga)*

An analysis across the different years, using the corresponding travel time matrix, indicates a relatively stable trend of lower but more spread WATT distribution for service sectors and higher and more concentrated WATT distribution for retail and production and storage sectors. The other services sector present an increase in WATT between 1990 and 2000 (Figure 24).

The travel time used for the ISEA and the WATT calculations come from a 4-Step transportation model and therefore include the estimated effects of congestion. In order to compare the evolution of WATT due exclusively to changes in the spatial distribution activities via the transport network (and not due to congestion), I recalculated WATTs for the three years but using only the 2010 travel time matrix. In this case, travel time can be interpreted as a network distance between zones. This analysis confirms an increase in WATT for the production and storage, retail, and other services sector (Figure 25). The highest increase was in the other services sector. This suggests an increase in relative distance (suburbanization) between jobs in these sectors, and/or core area-located businesses (with lower WATT) with these sector jobs closing (probably more likely given the decrease in total number of service sector jobs; see Figure 16). The remaining businesses with these sector jobs are located in the suburbs, with greater distances between them.

**Figure 25 WATT evolution by industry super-super sector**






*Source: CTPP and MIT CUBE Voyager Model (Mikel Murga)*

## 4.2. Firms Location Model Estimation

I now turn to models aiming to explain firm location decisions. I used data from Infogroup's database of individual businesses in 2010 and 2000. The geo-located records contain information on each firm's operational characteristics (e.g. 8-digit NAICS code, employee size, sales volume information) as well as some information on the physical space the firm occupies (categories of square footage). In order to be able to characterize the firms' space in more detail, I matched the Infogroup records to individual parcels from the MassGIS Level 3 Assessors' Parcel Mapping data. This dataset provides information on the building in which the Infogroup records are located such as FAR, the total building coverage area, land use code, and building age. For the Infogroup data for the year 2000, only records from the MassGIS data that matched to buildings built before 2000 were used. Zonal variables were constructed for both time periods based on census information on population and employment, tax information and aggregated parcel data.

### 4.2.1. Data description and model specification

Not all the records in the parcel data have information on all the variables of interest for the analysis. I only used records with complete information.

Since the bid-choice method assumes that locations are the outcome of a bidding process between utility maximizing agents, I removed Infogroup records for businesses, institutions, or agencies that do not fit this underlying location logic – essentially agents which do not compete with other agents for location. Such agents include universities and colleges, police and firefighter stations, judicial courts, or zoos. The Appendix contains a complete list of the records filtered for the analysis. I also excluded natural resource-related business from the analysis because they represent 0.3% of total employment in the study area and, given the nature of their industry,

might have unique location requirements (e.g. quarries). Historic preservation buildings (based on the building classification code from the Level 3 Parcel Data), which usually have specific use restrictions, were also not included. In order to reduce possible noise in the analysis, businesses with less than 10 employees at the location were also excluded.

After being thusly filtered, the total number of observation were 3,535 for 2010 and 3,367 for 2000. The lower number of observations for 2000 is due to more incomplete records from Infogroup that were left out (e.g. businesses with no information on number of employees), business records matched to incomplete parcel records, and the application of the building age filter (building built before 2000). Figure 26 show the distribution of the samples by firm industry and employee size.

**Figure 26 Sample distribution by firm industry and employees size**

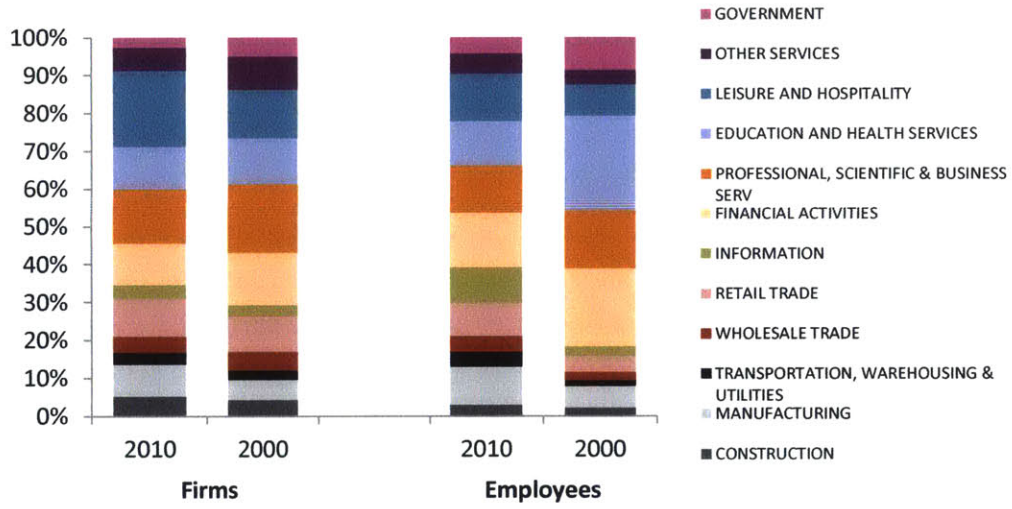
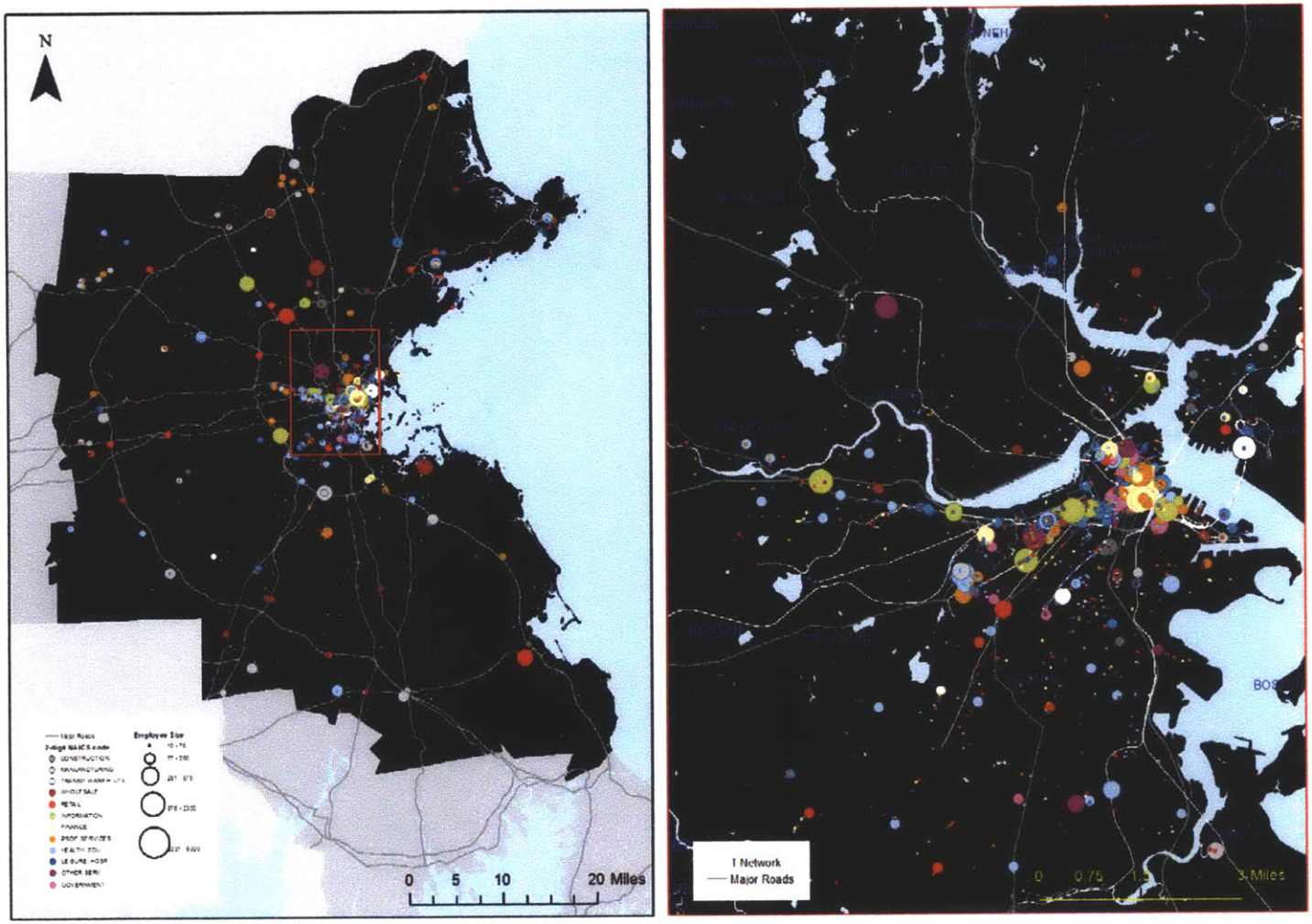


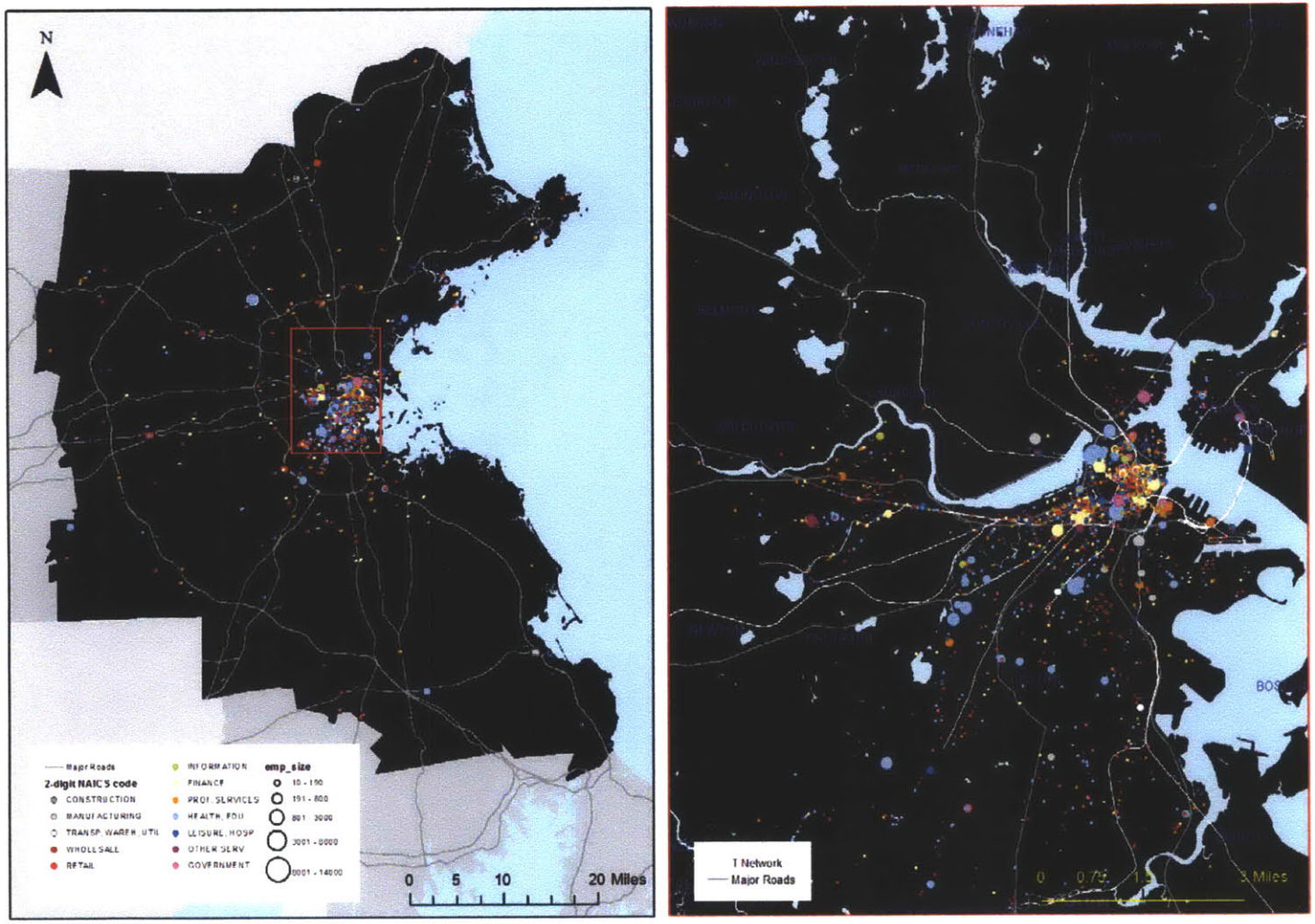

Figure 27 and Figure 28 shows the spatial distribution by industry and employment size of the observations in the final samples used in the estimations. Consistent with the spatial distribution of total jobs, the samples are more concentrated in Boston and its surrounding towns and along the main highways. Within Boston, the main employment concentrations such as the financial district, Boylston Street, or LMA, can also be identified. The new commercial developments around Kendall Square are underrepresented in the sample. This may be due to the fact that many of the new buildings in this area were (and, in fact, still are) under construction in 2010. The sample for 2000 has fewer large firms along the main highways. This is mainly due to the lower level of detail for the 2000 Infogroup records, which might cause some observations to be filtered out of the analysis, according to the filtering criteria mentioned previously.

**Figure 27 Sample for 2010 firm location model estimation**




*Source: Infogroup*

**Figure 28 Sample for 2000 firm location model estimation**




*Source: Infogroup*

### 4.2.2. Model estimation

Model specification

Similar to the residential case, the firm model was specified as a multinomial logit with a linear utility (bid) function. The model specification represents the industry types as different agents. This approach is supported by the literature and is consistent with the way in which cities and states frame their economic development objectives. In the case of Boston, for example, cities and towns may be particularly interested in bolstering a knowledge-based economy with a strong emphasis on high tech, health, and associated education services. Based on the aggregated analysis presented earlier, and after evaluating different agent categorization alternatives through model estimations, I settled on an approach which grouped firms by their 2-digit NAICS codes into 6 types of agents. Table 15presents these groups.

**Table 15 Agent type description for firm location model**

| 2-Digit NAICS | 2-Digit NAICS Description | Agent Type | Agent Description |
|---|---|---|---|
| 23 | CONSTRUCTION | 1 | Production and Storage |
| 31, 32, 33 | MANUFACTURING | | |
| 22, 48, 49 | TRANSPORTATION, WAREHOUSING & UTILITIES | | |
| 42 | WHOLESALE TRADE | | |
| 44, 45 | RETAIL TRADE | 2 | Retail/Leisure |
| 71, 72 | LEISURE AND HOSPITALITY | | |
| 81 | OTHER SERVICES | 3 | Other Services |
| 92 | GOVERNMENT | 4 | Government |
| 61, 62 | EDUCATION AND HEALTH SERVICES | 5 | Education and Health |
| 51 | INFORMATION | 6 | Professional Services |
| 52, 53 | FINANCIAL ACTIVITIES | | |
| 54, 55, 56 | PROFESSIONAL, SCIENTIFIC & BUSINESS SERV | | |

As with the residential model, the variables used for the firm location choice model estimation can be group into three categories: agent-specific attributes, unit-specific attributes, and zonal attributes. Since the estimation was done by groups instead of individual firms, many of the individual characteristics of firms such as the sales volume, number of employees, and whether it is a headquarter or branch, cannot be included in the analysis. To capture the differences between agents, other than the nature of their industry, I calculated an average employee size (based on the

individual observations) and an average salary or wage for each agent type. The unit-specific characteristics are based on the Infogroup data on space used by each firm, and from the parcel data. The zonal attributes capture the immediate area in which the firms are located, as well as their spatial relationship with the rest of the study area. This allows distinguishing between firms' preferences for immediate surroundings (e.g. job density, population density) and for accessibility at the metropolitan level (e.g. accessibility to population, accessibility to employment). This is a way of measuring the spatial limits in which firms value agglomeration. As with the residential location model, the zone structure used for the analysis is the 2727 Transport Analysis Zones (TAZ) from CTPS. The different zonal attributes were transformed form their original spatial units (e.g. block groups or Town) into this TAZ structure through aggregation and spatial splits. As in the residential choice case, the accessibility is calculated as a gravity function, using auto travel time matrix matrices for the AM peak from the MIT Cube Voyager 4-step model. The measures were normalized to values between 0 and 1.

$$Accesibility\ to\ Opportunity: ACC_i^k = \sum_j \frac{Opportunity_j^k}{t_{ij}} \tag{4.3}$$

$$i,j: Zone\ indexs\ ;\ t: travel\ time\ in\ minutes\ ;\ k: type\ of\ opportunity$$

Given the broad range of firms within the agent types specified, it is difficult to restrict the choice set for a given location. For example, under the firm agent type Production and Storage, fall all the firms in the manufacturing industry category, which would include both a large chemical production plant as well as a small jewelry manufacturer. Consequently, I could not assure that a specific agent type may not be allowed in a particular location (Back Bay, for example). This illustrates the problems of the heterogeneity within agent types, which is discussed in the analysis of results. For this reason, I assume all agent types are available for all locations.

Table 16 provides a summary of the variables used in the final specification.

**Table 16 Summary of variables used in the estimation of firm location model**

| Variable | Description | Unit of observation | 2010 | | | | | | 2000 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
| logAREA | log of the area of the parcel covered by building | Parcel | 1.800 | 6.519 | 7.792 | 7.908 | 9.056 | 12.910 | 2.027 | 6.187 | 7.209 | 7.262 | 8.306 | 12.910 |
| MBTA | Dummy for parcels located within 400m of a subway station | Parcel | 0.000 | 0.000 | 0.000 | 0.451 | 1.000 | 1.000 | 0.000 | 0.000 | 1.000 | 0.616 | 1.000 | 1.000 |
| Prox_UNI | Inverse of distance in meters to nearest university | Parcel | 0.000 | 0.000 | 0.001 | 0.020 | 0.004 | 1.000 | 0.000 | 0.001 | 0.002 | 0.008 | 0.005 | 1.000 |
| LU_IND | Dummy if parcel's property type classification code is industrial, light-industrial, utility, vacant, or warehouse | Parcel | 0.000 | 0.000 | 0.000 | 0.131 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.099 | 0.000 | 1.000 |
| LU_TAX | Dummy if parcel's property type classification code is that of tax exempt building | Parcel | 0.000 | 0.000 | 0.000 | 0.110 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.121 | 0.000 | 1.000 |
| LU_MIX | Dummy if parcel's property type classification code is retail, lodging, or mix-use | Parcel | 0.000 | 0.000 | 0.000 | 0.228 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.220 | 0.000 | 1.000 |
| LU_OFF | Dummy if parcel's property type classification code is office, or mix-use | Parcel | 0.000 | 0.000 | 0.000 | 0.362 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.375 | 1.000 | 1.000 |
| FAR | Average floor-area-ratio of the zone | Zone | 0.100 | 0.140 | 0.660 | 2.630 | 3.480 | 22.840 | 0.100 | 0.605 | 2.530 | 4.001 | 6.000 | 22.840 |
| IND_DEN | Density of employments in the same industry | Zone | 0.000 | 0.002 | 0.012 | 0.109 | 0.091 | 1.000 | 0.000 | 0.010 | 0.069 | 0.175 | 0.262 | 1.000 |
| RETAIL_DEN | Retail employments density | Zone | 0.000 | 0.002 | 0.017 | 0.120 | 0.126 | 1.000 | 0.000 | 0.018 | 0.092 | 0.223 | 0.378 | 1.000 |

**Table 16 (continued)**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACC_HWY | Network distance index to access point to the interstate roadways and to Rt 24, Rt 128 (N. of I-95), and to limited sections of Rt. 2 and Rt. 3. | Zone | 0.014 | 0.046 | 0.074 | 0.112 | 0.140 | 1.000 | 0.014 | 0.046 | 0.074 | 0.112 | 0.140 | 1.000 |
| ACC_IND | Intra-sector employment accessibility (ISEA) | Zone | 0.043 | 0.255 | 0.450 | 0.473 | 0.667 | 1.000 | 0.047 | 0.362 | 0.578 | 0.570 | 0.794 | 1.000 |
| ACC_EMP_1 | Accessibility index to workforce in professional, technical, executive, manager, and administration occupation | Zone | 0.190 | 0.410 | 0.710 | 0.659 | 0.880 | 1.000 | 0.200 | 0.690 | 0.860 | 0.780 | 0.900 | 1.000 |
| ACC_EMP_2 | Accessibility index to workforce in sales, admin. Support, and clerical occupations | Zone | 0.220 | 0.470 | 0.800 | 0.717 | 0.950 | 1.000 | 0.240 | 0.770 | 0.900 | 0.829 | 0.940 | 1.000 |
| ACC_EMP_3 | Accessibility index to workforce in production, transportation, and material moving occupations | Zone | 0.260 | 0.570 | 0.820 | 0.745 | 0.920 | 0.980 | 0.280 | 0.810 | 0.920 | 0.856 | 0.960 | 1.000 |

Estimation results and analysis

I evaluated numerous specifications, testing different combinations of variables. A challenge with the formulation is the high correlation between several of the different accessibility measures due to the urban structure of the study area (high concentration of jobs and population) coupled with the relatively low resolution of the travel time matrices. For this reason, not all the accessibility measures could be included in the same estimation.

In the models, I use the production/Storage agent type as the reference group. When analyzing the results, one must remember that, given the under-defined nature of the multinomial logit, only the difference in preferences between groups can be assessed. So the correct interpretation of the magnitude and the sign of a given coefficient is how much more (or less) a given agent type values a change in a given variable in comparison to the reference group.

Table 17 presents the best estimation in terms of the coherence of the signs of the coefficients, their level of significance, and the goodness of fit for individual models for each time period (Models 1 and 2), a pooled fully constrained models (Model 3), and a model with pooled data but different scale parameters (Model 4).

**Table 17 Firm location model. Estimation results by individual (unconstrained) models, pooled (fully constrained) model, and pooled model with different scales**

| | | Model 1: 2010 | | | Model 2: 2000 | | | Model 3: Pooled | | | Model 4: Pooled - Scaled | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Agent** | **Value** | **t-test** | | **Value** | **t-test** | | **Value** | **t-test** | | **Value** | **t-test** | |
| ASC | Retail/Amen | 4.73 | 10.11 | ** | 7.00 | 11.56 | ** | 4.20 | 13.74 | ** | 3.85 | 13.58 | ** |
| | Other Serv. | 5.07 | 8.50 | ** | 12.20 | 12.33 | ** | 6.23 | 15.2 | ** | 5.63 | 14.50 | ** |
| | Gov. | 8.85 | 11.14 | ** | 7.70 | 7.39 | ** | 4.28 | 8.6 | ** | 3.64 | 7.87 | ** |
| | Edu/Health | 7.23 | 13.27 | ** | 4.35 | 6.28 | ** | 4.37 | 12.72 | ** | 3.82 | 11.79 | ** |
| | Prof. Serv | 9.39 | 16.56 | ** | 6.83 | 11.07 | ** | 5.35 | 16.56 | ** | 4.80 | 15.75 | ** |
| logAREA | Retail/Amen | -0.06 | -1.82 | * | -0.21 | -4.69 | ** | -0.11 | -4.28 | ** | -0.11 | -4.49 | ** |
| | Other Serv. | -0.25 | -5.26 | ** | -0.48 | -5.52 | ** | -0.33 | -8.29 | ** | -0.30 | -8.34 | ** |
| | Gov. | -0.09 | -1.32 | | -0.03 | -0.37 | | -0.04 | -0.89 | | -0.04 | -0.89 | |
| | Edu/Health | -0.12 | -2.79 | ** | -0.13 | -2.60 | ** | -0.15 | -4.74 | ** | -0.14 | -4.87 | ** |
| | Prof. Serv | -0.04 | -0.87 | | -0.08 | -1.81 | * | -0.09 | -3.1 | ** | -0.08 | -3.24 | ** |
| MBTA | Retail/Amen | 1.59 | 7.16 | ** | 1.72 | 7.81 | ** | 1.67 | 10.91 | ** | 1.50 | 10.82 | ** |
| | Other Serv. | 1.28 | 4.45 | ** | 3.93 | 10.20 | ** | 2.56 | 12.41 | ** | 2.36 | 12.65 | ** |
| | Gov. | 4.89 | 9.56 | ** | 3.72 | 9.20 | ** | 3.25 | 11.81 | ** | 2.88 | 11.64 | ** |
| | Edu/Health | 1.52 | 5.47 | ** | 1.69 | 7.02 | ** | 1.51 | 8.61 | ** | 1.38 | 8.87 | ** |
| | Prof. Serv | 3.31 | 11.08 | ** | 1.94 | 8.41 | ** | 2.14 | 12.63 | ** | 1.90 | 12.35 | ** |
| Prox UNI | Edu/Health | 1.12 | 2.59 | ** | 0.02 | 0.02 | | 0.78 | 1.93 | ** | 0.59 | 1.51 | |
| LU_IND | Retail/Amen | -1.50 | -7.23 | ** | -1.44 | -6.06 | ** | -1.46 | -9.6 | ** | -1.34 | -9.63 | ** |
| | Other Serv. | -1.09 | -3.70 | ** | -1.64 | -3.30 | ** | -1.37 | -5.61 | ** | -1.28 | -5.76 | ** |
| | Gov. | -2.47 | -4.17 | ** | -3.76 | -3.55 | ** | -2.78 | -5.69 | ** | -2.60 | -5.69 | ** |
| | Edu/Health | -2.27 | -7.08 | ** | -2.58 | -7.08 | ** | -2.31 | -9.93 | ** | -2.10 | -9.89 | ** |
| | Prof. Serv | -1.32 | -4.63 | ** | -1.53 | -6.19 | ** | -1.28 | -7.83 | ** | -1.19 | -8.04 | ** |
| LU_TAX | Retail/Amen | 0.96 | 3.96 | ** | 0.43 | 1.53 | | 0.68 | 3.82 | ** | 0.57 | 3.57 | ** |
| | Other Serv. | 1.33 | 4.30 | ** | 0.20 | 0.44 | | 0.89 | 3.85 | ** | 0.73 | 3.50 | ** |
| | Gov. | 0.60 | 1.44 | | 0.14 | 0.32 | | 0.58 | 2.24 | ** | 0.46 | 1.98 | ** |
| | Edu/Health | 1.56 | 5.67 | ** | 1.45 | 5.28 | ** | 1.53 | 8.33 | ** | 1.35 | 8.13 | ** |
| | Prof. Serv | 0.58 | 1.88 | * | 0.22 | 0.78 | | 0.43 | 2.24 | ** | 0.36 | 2.14 | ** |
| LU_MIX | Retail/Amen | 1.44 | 8.73 | ** | 1.27 | 6.72 | ** | 1.36 | 11.24 | ** | 1.20 | 10.79 | ** |
| | Other Serv. | 0.60 | 2.65 | ** | 0.50 | 1.54 | | 0.38 | 2.22 | ** | 0.32 | 2.09 | ** |
| | Gov. | -1.56 | -3.21 | ** | 0.19 | 0.60 | | -0.24 | -1.05 | | -0.13 | -0.65 | |
| | Edu/Health | -0.13 | -0.59 | | 0.21 | 0.95 | | 0.07 | 0.5 | | 0.08 | 0.62 | |
| | Prof. Serv | -0.45 | -1.97 | ** | -0.17 | -0.85 | | -0.23 | -1.65 | * | -0.20 | -1.63 | * |

**Table 17 (continued)**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LU_OFF | Retail/Amen | 0.11 | 0.67 | | -0.27 | -1.44 | | 0.04 | 0.31 | | 0.00 | 0.04 | |
| | Other Serv. | 0.10 | 0.45 | | -1.05 | -3.20 | ** | -0.15 | -0.92 | | -0.17 | -1.13 | |
| | Gov. | -0.67 | -1.91 | * | -1.37 | -3.98 | ** | -0.69 | -3.1 | ** | -0.65 | -3.31 | ** |
| | Edu/Health | 0.07 | 0.35 | | -0.43 | -1.97 | ** | 0.01 | 0.08 | | -0.02 | -0.14 | |
| | Prof. Serv | 0.61 | 2.92 | ** | 0.20 | 1.11 | | 0.54 | 4.38 | ** | 0.46 | 4.16 | ** |
| FAR | Retail/Amen | 0.15 | 3.74 | ** | 0.20 | 5.46 | ** | 0.16 | 5.97 | ** | 0.14 | 5.77 | ** |
| | Other Serv. | 0.21 | 4.29 | ** | 0.95 | 14.92 | ** | 0.50 | 15.03 | ** | 0.45 | 14.89 | ** |
| | Gov. | 0.67 | 10.06 | ** | 0.72 | 11.64 | ** | 0.48 | 11.74 | ** | 0.42 | 11.24 | ** |
| | Edu/Health | 0.29 | 5.68 | ** | 0.29 | 7.08 | ** | 0.27 | 8.39 | ** | 0.23 | 8.20 | ** |
| | Prof. Serv | 0.67 | 14.29 | ** | 0.48 | 13.92 | ** | 0.53 | 19.19 | ** | 0.46 | 17.20 | ** |
| IND_DEN | Retail/Amen | 11.10 | 10.04 | ** | 11.30 | 11.99 | ** | 10.10 | 15.35 | ** | 9.10 | 15.26 | ** |
| | Other Serv. | 9.47 | 7.72 | ** | 15.50 | 14.07 | ** | 12.00 | 17.08 | ** | 10.80 | 16.85 | ** |
| | Gov. | 19.70 | 13.46 | ** | 15.80 | 14.82 | ** | 13.60 | 18.67 | ** | 12.10 | 18.10 | ** |
| | Edu/Health | 12.40 | 10.49 | ** | 9.14 | 9.33 | ** | 8.76 | 12.58 | ** | 7.84 | 12.56 | ** |
| | Prof. Serv | 17.40 | 14.77 | ** | 2.13 | 2.19 | ** | 6.71 | 9.95 | ** | 5.75 | 9.50 | ** |
| RETAIL_DEN | Prof. Serv | -1.69 | -5.25 | ** | 5.21 | 12.85 | ** | 2.09 | 10.03 | ** | 2.02 | 11.03 | ** |
| ACC_HWY | Retail/Amen | -3.05 | -4.16 | ** | -0.75 | -8.25 | ** | -0.43 | -8.11 | ** | -0.38 | -8.26 | ** |
| | Other Serv. | -3.89 | -3.69 | ** | -1.17 | -8.42 | ** | -0.43 | -6.35 | ** | -0.35 | -5.90 | ** |
| | Gov. | -7.92 | -4.67 | ** | -1.18 | -7.02 | ** | -0.42 | -4.52 | ** | -0.28 | -3.66 | ** |
| | Edu/Health | -5.75 | -6.12 | ** | -0.28 | -3.18 | ** | -0.19 | -3.65 | ** | -0.14 | -3.07 | ** |
| | Prof. Serv | -9.12 | -8.78 | ** | -0.82 | -8.76 | ** | -0.45 | -8.04 | ** | -0.41 | -8.41 | ** |
| ACC_IND | Retail/Amen | -16.70 | -18.08 | ** | -17.50 | -17.52 | ** | -14.60 | -24.07 | ** | -13.10 | -22.40 | ** |
| | Other Serv. | -18.20 | -15.34 | ** | -52.70 | -27.10 | ** | -28.90 | -38.38 | ** | -26.50 | -34.26 | ** |
| | Gov. | -47.60 | -23.59 | ** | -42.90 | -24.24 | ** | -30.50 | -35.93 | ** | -27.20 | -30.34 | ** |
| | Edu/Health | -27.90 | -22.76 | ** | -20.00 | -18.70 | ** | -18.80 | -27.85 | ** | -16.70 | -24.38 | ** |
| | Prof. Serv | -42.70 | -31.93 | ** | -21.70 | -21.42 | ** | -23.10 | -35.43 | ** | -20.30 | -28.48 | ** |
| ACC_EMP_3 | Prod/Stge | 6.00 | 5.89 | ** | -4.14 | -4.64 | ** | -1.07 | -1.85 | * | -1.39 | -2.70 | ** |
| ACC_EMP_2 | Retail/Amen | 11.60 | 10.67 | ** | 0.45 | 0.50 | | 3.70 | 6.22 | ** | 2.91 | 5.40 | ** |
| ACC_EMP_2 | Other Serv. | 12.80 | 11.35 | ** | 9.85 | 9.46 | ** | 8.85 | 14.05 | ** | 7.74 | 13.21 | ** |
| ACC_EMP_1 | Gov. | 20.10 | 14.33 | ** | 8.28 | 7.53 | ** | 8.61 | 11.96 | ** | 7.45 | 11.33 | ** |
| ACC_EMP_1 | Edu/Health | 16.90 | 13.45 | ** | 4.01 | 4.19 | ** | 6.38 | 9.8 | ** | 5.28 | 8.85 | ** |
| ACC_EMP_1 | Prof. Serv | 20.80 | 16.33 | ** | 2.53 | 2.69 | ** | 7.16 | 11.18 | ** | 5.86 | 9.83 | ** |
| Scale 2000 | | | | | 1.00 | | | 1.00 | | | 1.26 | 5.85 | ** |
| n: | | | 3535 | | | 3367 | | | 6902 | | | 6902 | |
| Final log-like: | | | -3424.4 | | | -3098.0 | | | -7645.8 | | | -7624.5 | |
| LRT: | | | 5818.9 | | | 5869.7 | | | 9441.9 | | | 9484.5 | |
| Rho-sq: | | | 0.459 | | | 0.486 | | | 0.382 | | | 0.383 | |
| Adj rho-sq: | | | 0.449 | | | 0.476 | | | 0.377 | | | 0.378 | |

The scale parameter of the 2010 was fixed to 1. The scale parameter estimated in Model 4 indicated that the variance of unobserved factor in greater in the 2000 data than in the 2010 data.

Table 18 presents the likelihood ratio test to see if the results change across models.

**Table 18 Likelihood ratio test for firm model. Pooled (constrained) vs. individual (unconstrained) models and Pooled (constrained) model vs. Pooled model with different scales**

---

Pooled (constrained) vs. individual (unconstrained) models

$H_0: \hat{\beta}_R = \hat{\beta}_U \quad where \ \hat{\beta}_R: pooled\ models\ , \hat{\beta}_U: individual\ models$

$T = -2[\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U)]$

$T = -2[-7645.8 + (3424.4 + 3098.0)] = 2246.69$

P-value= 0.000

Reject $H_0$. Preferences have change over time

Pooled (constrained model vs. Pooled model with different scales

$H_0: \hat{\beta}_R = \hat{\beta}_S \quad where \ \hat{\beta}_R: pooled\ models\ , \hat{\beta}_S: pooled\ model\ with\ different\ scales$

$T = -2[\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_S)]$

$T = -2[-7645.8 + 7625{,}5] = 42.57$

df = 1

P-value= 0.000

Reject $H_0$. models are not the same. Variances of the datasets are different

---

The likelihood ratio tests indicate that the estimation results do vary depending on the dataset that is used, which indicates that location preferences have changed over time. The signs and significance of the coefficients are similar in the four models. Given their higher rho-square, the individual (unconstrained) models are proffered over the pooled models. The results of the individual models can be interpretation as follows:

The results of the individual models can be interpreted as follows:

ASC: Alternative specific constant. All else equal, in 2010 a firm in the professional service sector is more likely to win a bid for a property than firms in other sectors. In 2000, a firm in the other service sectors is more likely to win a bid for a property than other firms. The ASCs, to some extent, capture the effect of missing variables.

Log_AREA: Log of the building coverage, which is a proxy for the floor-plate area. Units with larger floor plates are valued more highly by firms in the production/storage sector. Results are not significant for firms in Retail and Government sector in 2010 and 2000 as well as in Professional Services in 2010

MBTA: Within 400m of subway station. Urban rail access is significant and positive for all agents in both periods. All agent types value proximity to the T more highly than firms in the production/storage sector. Firms that value proximity to the T the highest are those in the government and professional service sectors in 2010, and in government and other service sectors in 2000.

Prox_UNI: Proximity to universities and colleges. Only firms in the Education/Health sector in 2010 appear to value university proximity positively, relative to the rest of the agents. In 2000 it is not significant for any agents.

LU_IND, LU_MIX, LU_TAX, LU_COM: Variables related to the property classification code of the building. Due to time limitations and data availability, the same classification codes were used for both time periods. That is, the parcels that were used in the 2000 dataset (i.e. buildings built before 2000) are assumed to have the same classification code in 2010 and 2000. For both time periods, firms in the production/storage sector will bid more for space in properties classified as industrial. Firms in Retail/Leisure will bid more for space in properties classified as retail, lodging or mix-use. Firms in education/health and other services will bid more for space in tax-exempt properties. Firms in professional services are more likely to bid for space in properties classified as office building in 2010. Alternatively, and more technically correct, these results may indicate what type of agent a landlord prefers as a tenant given the land use classification of her building.

FAR: Floor-area-ratio. Coefficients on FAR are significant and positive for all agents in both time periods. All other agent types value FAR more highly than firms in the production/storage sector. The firms that value FAR the most are those in the government and professional service sector in 2010, and government and other service sector in 2000.

IND_DEN: Job density in the same industry. This variable aims to capture clustering preferences. Coefficients are significant and positive for all agents in both time periods. All other agent types value proximity to employment in their same industry more highly than firms in the production/storage sector. The firms that value this type of agglomeration the highest are those in the government and professional service sector in 2010, and government and other service sector in 2000.

IND_DEN_RETAIL: Density of retail jobs in the area. For firms in the retail sector, this is the IND_DEN variable. For all agent types except professional services, retail density was insignificant. In 2010 firms in the professional service sector value proximity to retail more negatively compared to other agents, while in the 2000 data they value it more positively.

ACC_HWY: Accessibility to point of access to main highways. Coefficients are significant and positive for all agents in both time periods. In both cases, firms in the production/storage sector value being close to the main highways more than other agents.

ACC_IND: Accessibility to firms in the same industry, which can be interpreted as a measure of proximity at the metropolitan level (different than the immediate proximity captured by the IND_DEN variable). Coefficients are significant and positive for all agents in both time periods. Firms in the production/storage sector value this type of accessibility higher than other firms in both time periods.

ACC_EMP3, ACC_EMP2, ACC_EMP1: accessibility to employees' place of residence by employee type within the metro area. These variables were highly correlated with ACC_IND when evaluated for the whole study area, but not when evaluated inside Metro Boston. Therefore, I include it as an explanatory variable only for locations within that sub-region. They were also highly correlated with each other. For this reason, I matched agents with work force type, which resulted in the best goodness of fit while also making intuitive sense. Firms in the production/storage sector value accessibility to employees in production, transportation, and material moving occupations more highly than other agent types in the 2010 dataset, but less than other agents in the 2000 dataset. Firms in the retail/leisure and other services sector value accessibility to employees in sales, administration, support, and clerical occupations more than other sectors. Finally, firms in the government, education/health, and professional services value accessibility to employees in professional, technical, executive, manager, and administration occupations more than other sectors.

Stability of preferences

As outlined in the introduction, one of my research questions relates to the temporal transferability of location choice models. If location preferences change over time, a model estimated for one time period might not represent the location behavior of agents in a different time period. This is a critical issue considering that these types of models are often used to forecast future urban development scenarios. Here I test the sensitivity of the estimated firms' bidding behavior across the 2010 and 2000 data, using the stability of preferences approach outlined in the methods section.

To compare the preferences for individual location attributes, the coefficients of the 2000 model are adjusted based on the scale parameter estimated in Mode 4. After the adjustment, the coefficients of the two models are compared using the t-statistic described in the methods section, which is only calculated for coefficients that are significant in both of the individual models. The results are presented in Table 19.

**Table 19 Firm location model. Preferences stability test**

| | | 2010 | | | 2000 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Name** | **Agent** | **Value** | **t-test** | | **Value** | **t-test** | | **t*** | |
| ASC | Retail/Amen | 4.73 | 10.11 | ** | 8.82 | 11.56 | ** | -4.57 | ** |
| | Other Serv. | 5.07 | 8.50 | ** | 15.37 | 12.33 | ** | -7.45 | ** |
| | Gov. | 8.85 | 11.14 | ** | 9.70 | 7.39 | ** | -0.56 | |
| | Edu/Health | 7.23 | 13.27 | ** | 5.48 | 6.28 | ** | 1.70 | * |
| | Prof. Serv | 9.39 | 16.56 | ** | 8.61 | 11.07 | ** | 0.81 | |
| logAREA | Retail/Amen | -0.06 | -1.82 | * | -0.27 | -4.69 | ** | 3.13 | ** |
| | Other Serv. | -0.25 | -5.26 | ** | -0.61 | -5.52 | ** | 2.97 | ** |
| | Gov. | -0.09 | -1.32 | | -0.04 | -0.37 | | | |
| | Edu/Health | -0.12 | -2.79 | ** | -0.17 | -2.60 | ** | 0.64 | |
| | Prof. Serv | -0.04 | -0.87 | | -0.11 | -1.81 | * | 0.93 | |
| MBTA | Retail/Amen | 1.59 | 7.16 | ** | 2.17 | 7.81 | ** | -1.62 | * |
| | Other Serv. | 1.28 | 4.45 | ** | 4.95 | 10.20 | ** | -6.51 | ** |
| | Gov. | 4.89 | 9.56 | ** | 4.69 | 9.20 | ** | 0.28 | |
| | Edu/Health | 1.52 | 5.47 | ** | 2.13 | 7.02 | ** | -1.48 | |
| | Prof. Serv | 3.31 | 11.08 | ** | 2.44 | 8.41 | ** | 2.08 | ** |
| Prox UNI | Edu/Health | 1.12 | 2.59 | ** | 0.03 | 0.02 | | 0.68 | |
| LU_IND | Retail/Amen | -1.50 | -7.23 | ** | -1.81 | -6.06 | ** | 0.86 | |
| | Other Serv. | -1.09 | -3.70 | ** | -2.07 | -3.30 | ** | 1.41 | |
| | Gov. | -2.47 | -4.17 | ** | -4.74 | -3.55 | ** | 1.55 | |
| | Edu/Health | -2.27 | -7.08 | ** | -3.25 | -7.08 | ** | 1.75 | * |
| | Prof. Serv | -1.32 | -4.63 | ** | -1.93 | -6.19 | ** | 1.44 | |
| LU_TAX | Retail/Amen | 0.96 | 3.96 | ** | 0.55 | 1.53 | | | |
| | Other Serv. | 1.33 | 4.30 | ** | 0.25 | 0.44 | | | |
| | Gov. | 0.60 | 1.44 | | 0.18 | 0.32 | | | |
| | Edu/Health | 1.56 | 5.67 | ** | 1.83 | 5.28 | ** | -0.60 | |
| | Prof. Serv | 0.58 | 1.88 | * | 0.28 | 0.78 | | | |

**Table 19 (continued)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| LU_MIX | Retail/Amen | 1.44 | 8.73 | ** | 1.60 | 6.72 | ** | -0.55 | |
| | Other Serv. | 0.60 | 2.65 | ** | 0.63 | 1.54 | | | |
| | Gov. | -1.56 | -3.21 | ** | 0.24 | 0.60 | | | |
| | Edu/Health | -0.13 | -0.59 | | 0.27 | 0.95 | | | |
| | Prof. Serv | -0.45 | -1.97 | ** | -0.22 | -0.85 | | | |
| LU_OFF | Retail/Amen | 0.11 | 0.67 | | -0.34 | -1.44 | | | |
| | Other Serv. | 0.10 | 0.45 | | -1.32 | -3.20 | ** | 3.03 | ** |
| | Gov. | -0.67 | -1.91 | * | -1.73 | -3.98 | ** | 1.91 | * |
| | Edu/Health | 0.07 | 0.35 | | -0.54 | -1.97 | ** | | |
| | Prof. Serv | 0.61 | 2.92 | ** | 0.26 | 1.11 | | | |
| FAR | Retail/Amen | 0.15 | 3.74 | ** | 0.25 | 5.46 | ** | -1.58 | |
| | Other Serv. | 0.21 | 4.29 | ** | 1.19 | 14.92 | ** | -10.38 | ** |
| | Gov. | 0.67 | 10.06 | ** | 0.90 | 11.64 | ** | -2.29 | ** |
| | Edu/Health | 0.29 | 5.68 | ** | 0.37 | 7.08 | ** | -1.13 | |
| | Prof. Serv | 0.67 | 14.29 | ** | 0.61 | 13.92 | ** | 1.01 | |
| IND_DEN | Retail/Amen | 11.10 | 10.04 | ** | 14.24 | 11.99 | ** | -1.93 | * |
| | Other Serv. | 9.47 | 7.72 | ** | 19.53 | 14.07 | ** | -5.43 | ** |
| | Gov. | 19.70 | 13.46 | ** | 19.91 | 14.82 | ** | -0.10 | |
| | Edu/Health | 12.40 | 10.49 | ** | 11.52 | 9.33 | ** | 0.52 | |
| | Prof. Serv | 17.40 | 14.77 | ** | 2.68 | 2.19 | ** | 8.66 | ** |
| RETAIL_ DEN | Prof. Serv | -1.69 | -5.25 | ** | 6.56 | 12.85 | ** | -13.67 | ** |
| ACC_HWY | Retail/Amen | -3.05 | -4.16 | ** | -0.94 | -8.25 | ** | -2.84 | ** |
| | Other Serv. | -3.89 | -3.69 | ** | -1.47 | -8.42 | ** | -2.26 | ** |
| | Gov. | -7.92 | -4.67 | ** | -1.49 | -7.02 | ** | -3.76 | ** |
| | Edu/Health | -5.75 | -6.12 | ** | -0.36 | -3.18 | ** | -5.70 | ** |
| | Prof. Serv | -9.12 | -8.78 | ** | -1.04 | -8.76 | ** | -7.73 | ** |
| ACC_IND | Retail/Amen | -16.70 | -18.08 | ** | -22.05 | -17.52 | ** | 3.43 | ** |
| | Other Serv. | -18.20 | -15.34 | ** | -66.40 | -27.10 | ** | 17.71 | ** |
| | Gov. | -47.60 | -23.59 | ** | -54.05 | -24.24 | ** | 2.15 | ** |
| | Edu/Health | -27.90 | -22.76 | ** | -25.20 | -18.70 | ** | -1.48 | |
| | Prof. Serv | -42.70 | -31.93 | ** | -27.34 | -21.42 | ** | -8.31 | ** |
| ACC_EMP_3 | Prod/Stge | 6.00 | 5.89 | ** | -5.22 | -4.64 | ** | 7.39 | ** |
| ACC_EMP_2 | Retail/Amen | 11.60 | 10.67 | ** | 0.57 | 0.50 | | | |
| ACC_EMP_2 | Other Serv. | 12.80 | 11.35 | ** | 12.41 | 9.46 | ** | 0.22 | |
| ACC_EMP_1 | Gov. | 20.10 | 14.33 | ** | 10.43 | 7.53 | ** | 4.90 | ** |
| ACC_EMP_1 | Edu/Health | 16.90 | 13.45 | ** | 5.05 | 4.19 | ** | 6.80 | ** |
| ACC_EMP_1 | Prof. Serv | 20.80 | 16.33 | ** | 3.19 | 2.69 | ** | 10.12 | ** |
| n: | | | 3535 | | | 3367 | | | |
| Final log-like: | | | -3424.4 | | | -3098.0 | | | |
| LRT: | | | 5818.9 | | | 5869.7 | | | |
| Rho-sq: | | | 0.459 | | | 0.486 | | | |
| Adj rho-sq: | | | 0.449 | | | 0.476 | | | |

The estimation suggests changes in preference for the following location attributes:

- Preference for accessibility to employees has increased for all firms expect for those in the other service sector. This suggests that, except for the the other service sector, employers increasingly value proximity to employees.
- Preferences for proximity to jobs in the same sector (IND_DEN) decreased for firms in the other service sector and increased for firms in the professional service sector. The decrease in

preferences for the other service sector is consistent with the aggregated analysis that showed an increase in WATT for this sector.

- Preference for density has increased for firms in the professional service sector.
- Preferences for accessibility to the main highways (ACC_HWY) have decreased for firms in the professional service, education and health, and government sectors.

Prediction test

A prediction test was conducted for each individual model, using their corresponding dataset. Additionally, the prediction accuracy of the 2010 model was evaluated using the dataset of 2000. This test aims to measure the forecast (or in this case *backcast*) capability of the model. That is, how accurate would the 2010 model predict the locations of firms in a different time period. The prediction tests are presented in Table 20 – Table 22.

**Table 20 Prediction tests. 2010 model with 2010 data**

| Agent Type | Obs. | Model Prediction | | | | | |
|---|---|---|---|---|---|---|---|
| | | Prod /Storage | Retail /Amen | Other Serv. | Gov. | Edu /Health | Prof. Serv |
| Prod/Storage | 740 | 76.8% | 20.7% | | | 0.9% | 1.6% |
| Retail/Amen | 1064 | 11.3% | 74.2% | | | 4.1% | 10.3% |
| Other Serv. | 218 | 14.2% | 60.6% | | | 12.8% | 12.4% |
| Gov. | 89 | | 15.7% | | | 4.5% | 79.8% |
| Edu/Health | 397 | 4.8% | 40.3% | | | 19.1% | 35.8% |
| Prof. Serv | 1027 | 1.9% | 11.2% | | | 0.9% | 86.1% |
| | 3535 | | | | | | 65.6% |

**Table 21 Prediction test. 2000 model with 2000 data**

| Agent Type | Obs. | Model Prediction | | | | | |
|---|---|---|---|---|---|---|---|
| | | Prod /Storage | Retail /Amen | Other Serv. | Gov. | Edu /Health | Prof. Serv |
| Prod/Storage | 572 | 75.5% | 12.8% | | | 1.9% | 9.8% |
| Retail/Amen | 723 | 12.7% | 55.6% | 0.3% | 0.3% | 8.4% | 22.7% |
| Other Serv. | 301 | 2.0% | 3.3% | 82.7% | 5.3% | 1.0% | 5.6% |
| Gov. | 172 | 0.6% | 20.3% | 32.6% | 35.5% | 1.7% | 9.3% |
| Edu/Health | 409 | 7.3% | 14.9% | 0.7% | 4.9% | 31.5% | 40.6% |
| Prof. Serv. | 1190 | 5.3% | 9.5% | 1.6% | 0.2% | 4.2% | 79.2% |
| | 3367 | | | | | | 65.8% |

**Table 22 Prediction test. 2010 model with 2000 data**

| Agent Type | Obs. | Model Prediction | | | | | |
|---|---|---|---|---|---|---|---|
| | | Prod /Storage | Retail /Amen | Other Serv. | Gov. | Edu /Health | Prof. Serv |
| Prod/Storage | 572 | 94.9% | 5.1% | | | | |
| Retail/Amen | 723 | 55.6% | 39.6% | | 0.3% | 1.0% | 3.6% |
| Other Serv. | 301 | 20.3% | 22.6% | | 25.2% | 4.7% | 27.2% |
| Gov. | 172 | 11.0% | 33.1% | | 26.2% | 2.3% | 27.3% |
| Edu/Health | 409 | 72.9% | 26.2% | | 0.2% | 0.2% | 0.5% |
| Prof. Serv. | 1190 | 61.1% | 31.3% | | | 1.4% | 6.1% |
| | 3367 | | | | | | 28.2% |

When evaluated with the dataset that was used for its estimation, the 2010 model has high location prediction accuracy for firms in the professional services (86.1% accuracy), retail/leisure (74.2% accuracy), and production/storage (76.8% accuracy) sectors. However, it has no (0%) accuracy when predicting firms in the government and other services category. The most common mistake when predicting location of firms in the government sector is mistaking them for firms in the professional service sector (79.8%). For firms in the other-sector services, the most common error is mistaking them for firms in the retail sectors (60.6%).

When evaluated with the dataset that was used for its estimation, the 2000 model has similar total accuracy (65.8%) to the 2010 model (65.6%). However, unlike the 2010 model, the 2000 model is able to accurately predict some locations for all types of agents (no zeros in the diagonal).

When evaluated with the 2000 dataset, the 2010 model's total accuracy decreases from 65.6% to 28.2%. That is, the 2010 model is able to accurately predict 28.2% of the locations in 2000. The largest decrease in location prediction accuracy is for firms in the professional service sector (from 86.1% to 6.1%).

McFadden Omitted Variables Test

Using the McFadden omitted variables test described in the methods section, I test possible violation of the IIA assumption. The test design aims to analyze if the location preferences are the same for firms in the Metro Area and firms in the Greater Boston area (or core vs. suburbs). The 2010 model was estimated adding a METRO variable for observations in the Metro Area. The variable is defined according to equation 2.31. The coefficients of the auxiliary variables were not significant.

Summary

The firms' location choice models for 2010 and 2000 show similar results in terms of significance, relative willingness to pay between agent categories (i.e., who is willing to pay more for a specific attribute), goodness of fit, and total accuracy level. The 2000 model accurately predicts the locations of more types of individual agents, while the 2010 model does not accurately predict any location for firms in the other sector and government sector industries. As with the residential models, the firms' location models seem to be more accurate at predicting the location of the agent types that were well represented (i.e. large number of observations) in the data used for model estimation.

The results of the likelihood ratio test indicate that the estimation results vary depending on the data set that is used. That is, firms' location preferences have apparently changed over time. The stability of preferences test indicates that firms in the other service sector have lower willingness to pay for locations that are close to their employee base. Their willingness to pay for being close to jobs in the same sector has also decreased. For this industry sector, the aggregated analysis shows a decrease in total jobs in Boston and Metro Boston sub regions, as well as an increase in WATT. Both the results on location preferences and the aggregated analysis suggest a suburbanization of the other service sector. That is, jobs in this industry sector are now less concentrated in the metro area (they have likely been priced out of these areas), and more spread out into the Greater Boston sub area. Consequently, the WATT for this sector has increased (jobs are now more far apart). This is reflected in the location choice model as a decrease in other service sector firms' willingness to pay to be close to one another. Moving out of the metro area also means being farther away from places of high population density, hence the decrease in willingness to pay for being close to employee residences. Since the location choice models show relative preferences between firms, a decrease in willingness to pay does not necessarily mean that firms in this sector value this location attribute less per se. It just mean that, compared to 2000, in 2010 there are firms in other sectors who are willing to pay more for these location attributes. This seems to be the case for firms in the professional service sector, which have increased their willingness to pay for proximity to jobs in the same sector. Additionally, these types of firms show higher preference for population density and lower preference for proximity to the main highways, relative to 2000. This suggests an apparent centralization of firms in this sector.

The prediction test shows a low level of accuracy for the 2010 model evaluated with the 2000 dataset (28.2% total accuracy). That is, the 2010 model does not explain very well the location of firms in 2000. As with the residential location choice models, it is not clear to what extent the loss in accuracy is due to changes in preferences (determined in the stability of preference tests) or to the difference in number of observations per agent category of the two samples used for estimation.

Possible limitations of the models are:

- Heterogeneity within agent categories. This issue goes beyond not being able restrict the choice set (mentioned previously). It relates to the very core of the analysis. Does a large chemical producing firm and a small jewelry manufacturer have the same preferences for location? That is, should one group them in the same category? The answer is, probably not. Preliminary alternatives to bypass this problem such as estimating individual models by firm size categories were evaluated with no major improvement in estimation results.
- Omitted variable bias. A better characterization of the commercial space that firms occupy is needed. For example, a large law firm might be located in the top floor of a class A office building in downtown Boston, and a drug store might be located in the first floor of the same building. These two firms will probably have different location preferences. Even though they are located in the same building, they are not likely to compete for the same commercial space. Without a better characterization of commercial space, this type of analysis cannot capture those preference differences.
- As already discussed in the residential model cases, several other errors may afflict these firm location models: the spatial definition of zones, error propagation from the transportation model, errors in functional form, and model structure error. In the latter case, I conducted a preliminary McFadden omitted variables test to identify possible IIA violations. Even though the coefficients estimated in the test were not significant, additional analysis is recommended.

# 5. CONCLUSIONS

The spatial distribution of activities across an urban area is often the result of historical urban processes and a reflection of differences in opportunities, rather than the outcome of preferences. A superficial analysis of observed agent distribution patterns might lead to wrong conclusions and wrong decisions. An in depth understanding of what city agents look for in a location can, and should, inform planning policies and intervention that seek to match preferences with opportunities.

Location choice models based on discrete choice theory can help identify the location preferences that explain the spatial distribution of agents in an urban area. These models require detailed data about the characteristics of agents and their corresponding location. In this thesis I explored such models for households and firms in Greater Boston. The objective of the analysis is to get insight on the relationship between residential location choices and life stages, and firms' clustering preferences by industry. These topics are important given (1) the demographic changes forecasted for Greater Boston, specifically baby boomers aging, and (2) the continuing move from a manufacturing-based economy to a service- and knowledge-based economy. These changes in population and economy will likely require a change in housing stock in order to better match supply with demand, and changes in stock of commercial space in order to continue boosting the firms that drive the economy of the region.

I estimated location models using multiple datasets from different sources and different time periods in order to analyze (1) the sensitivity of model estimation results to different data sources, and (2) changes in location preferences over time. To analyze the models' data source-sensitivity, I estimated residential location choice models for the same population in the same time period (households in Greater Boston in 2010) with two different datasets: one based on Infogroup data on costumers and one based on the 2012 Massachusetts Travel Survey. As is often the case for modeling, each dataset was processed and complemented with other data in order to construct a dataset suitable for model estimation. In order to analyze if location preferences have changed over time, I estimated location choice models for firms in 2010 and 2000.

An aggregated analysis of population and housing in Greater Boston shows a greater increase in population across the broader Greater Boston sub region compared to the more inner Boston Metro Area over the last 40 years. On the other hand, the city of Boston's share of total housing has increased in the same period. This suggests a change in family structure and household size

between the different sub regions (small households moving into the Metro Area and larger households moving into the suburbs). The median age has increased in the region as a whole, and this trend is expected to continue, as baby-boomers age. The location choice analysis suggests that income has a bigger impact on willingness to pay for location attributes than age of the head of the household or household size. This is critical given that housing affordability has decreased over time and income inequalities have increased. Housing cost as percentage of income has increased for both renters and owners. This means that increasingly, location preferences are constrained by housing costs. The preferences for size of the unit seem to be driven more by household size than by age of the head of the household, which is in line with the notion of aging baby boomers seeking smaller housing units. Keep in mind, however, that the number of rooms indicates unit size in the models, not floor area.

The location choice analysis presented in this thesis is for homeowners only. Renters represent close to 40% of the households in the study area and over 60% for the city of Boston. In order to get a better understanding of housing dynamics, the analysis presented in this thesis needs to be complemented with a location choice analysis for renters.

The analysis indicates that the models are sensitive to the specific dataset that is used in the estimation. That is, two data sets that represent the same population in the same period of time can result in two different model estimations. The accuracy of the different models' estimations seems to be correlated to the category size (number of observations of the individual agent categories) in the sample data that was used for the estimation.

The aggregated analysis of employment shows a decrease in jobs in the manufacturing and other services industries and an increase in jobs in professional service and education and health sectors. The location choice models suggest that willingness to pay for clustering has changed between 2000 and 2010. In 2000 the firms that valued proximity to jobs in the same industry were those in the government and other service sector, while in 2010 it was firms in the professional services, government, and education and health sectors. These changes in willingness to pay, coupled with the change in total number of jobs, seem to have resulted in (1) a move to the suburbs for firms in the other service sector and (2) a higher concentration of firms in the professional service, education and health, and government sector in the inner Metro Area and the city of Boston.

This thesis presents a first approach to developing location choice models for Greater Boston. It should serve as a starting point to future models. Next steps in this direction should include a better characterization of both the residential units and the commercial space occupied by households and firms respectively, and the development of residential location models for renters. Additionally, a homogeneous spatial unit of analysis for the zonal characteristics (e.g. a 400 meter buffer) should be tested.

## 6. APPENDIX

**Table 23 Businesses not included in the firm location model estimation**

| NAICS 2-digite codes | Description |
|---|---|
| 21 | Mining, Quarrying, and oil and Gas Extraction |
| 11 | Agriculture, Forestry, Fishing and Hunting |
| **NAICS 3-digit codes** | **Description** |
| 928 | National Security and International Affairs |
| 922 | Justice, Public Order, and Safety Activities |
| **NAICS 8-digit codes** | **Description** |
| 61121002 | Junior Colleges |
| 61131008 | Colleges & Universities |
| 61131009 | |
| 61131010 | |
| 61111007 | Elementary and Secondary Schools |
| 71121101 | Sports Teams & Clubs |
| 71121102 | |
| 71121103 | |
| 71121104 | |
| 71211001 | Museums |
| 71211004 | |
| 71213006 | Zoos & Botanical Gardens |
| 71219003 | Nature Parks & Other Similar Institutions |
| 71219004 | |
| 71219006 | |
| 71219007 | |
| 71311001 | Amusement & Theme Parks |
| 71311002 | |
| 71311003 | |
| 71312001 | Amusement Arcades |
| 71312003 | |
| 71391002 | Golf Courses & Country Clubs |
| 71393003 | Marinas |
| 71393004 | |
| 71393005 | |
| 71393006 | |
| 71393007 | |
| 71393008 | |
| 71393013 | |

**Table 23 (continued)**

| | |
|---|---|
| 71399002 | All Other Amusement & Recreation Industries |
| 71399005 | |
| 71399006 | |
| 71399009 | |
| 71399014 | |
| 71399019 | |
| 71399020 | |
| 71399021 | |
| 71399024 | |
| 71399028 | |
| 71399031 | |
| 71399034 | |
| 71399044 | |
| 71399050 | |
| 71399059 | |

## 7. REFERENCES


Alonso, W. (1964). Location and Land Use: Toward a General Theory of Land Rent, Harvard University Press, Cambridge,Massachusetts.

Anas A (1982) Residential location markets and urban transportation: economic theory, econometrics and public policy analysis. Academic Press, New York.

Baum-Snow N. (2014), Urban Transport Expansions, Employment Decentralization, and the Spatial Scope of Agglomeration Economies.

Ben-Akiva M.E. and Lerman, S. R. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand, The MIT press, Cambridge MA.

Cahill D ., (2006). Lifestyle Market Segmentation. Haworth Press, New York.

Carlton D. (1979). Why new firms locate where they do: An econometric model. In Interregional Movements and Regional Growth, e.d W. Wheaton. Washingtone, D.C.: The Urban Institute pp. 13 50.

Chattopadhyay, S. (1998). An Empirical Investigation into the Performance of Ellickson's random Bidding Model, with an Application to Air Quality Valuation, Journal of Urban Economics 43(2): 292 – 314.

Clark W. A. V., Deurloo M. C. (2006). Aging in place and over-consumption, Journal of Housing and Built Environment 21 257–270.

Clark, W.A.V., and Huang, Y. (2003). The Life Course and Residential Mobility in British Housing Markets, Environment and Planning A, Vol. 35, pp. 323-339.

Ellickson, B. (1981). An Alternative Test of The Hedonic Theory of Housing Markets, Journal of Urban Economy 9, 56-79.

Florida, R. (2002). The economic geography of talent. Annals of the Association of American Geographers, 92, 743–755.

Frenkel, A., Bendit, E., & Kaplan, S. (2012). The linkage between the lifestyle of knowledge workers and their intra-metropolitan residential choice: A clustering approach based on self-organizing maps. Computers, Environment and Urban Systems, 39, 151–161.

Frenkel, A., Bendit, E., & Kaplan, S. (2013). Residential location choice of knowledge-workers: The role of amenities, workplace and lifestyle. CITIES, 35, 33–41.

Gibbs R. M., Bernat G. A., Jr. (1997). Rural Industry Clusters Raise Local Earnings. Rural Development Perspectives 12(3):18-25.

Glaeser E. L. (2004). Reinventing Boston: 1630-2003. Journal of Economic Geography 5 pp. 119-153.

Gross, D. J. (1988). Estimating Willingness To Pay For Housing Characteristics: An Application of the Ellickson Bid-Rent Model, Journal of Urban Economics 24(1): 95 – 112.

Gross, D. J., Sirmans, C. and Benjamin, J. D. (1990). An Empirical Evaluation of The Probabilistic Bid-Rent Model: The case of homogenous households, Regional Science and Urban Economics 20(1): 103 – 110.

Hansen, E (1987). Industrial Location Choice in Sao Paulo, Brazil: A Nested Logit Model. Regional Science and Urban Economics 17. 89-108.

Hurtubia, R. and Bierlaire, M. (2012). Estimation of Bid Functions for Location Choice and Price Modeling With a Latent Variable Approach, Technical Report TRANSP-OR 120206, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Jara-Díaz, S. R. andMartínez, F. J. (1999). On The Specification of Indirect Utility and Willingness to Pay for Discrete Residential Location Models, Journal of Regional Science 39(4): 675–688.

Krugman, P. (1991), "Geography and Trade", Leuven: Leuven University Press and Cambridge (MA), London: the MIT Press.

Kunzmann, K. R. (2009). The strategic dimensions of knowledge industries in urban development. DISP – The Planning Review, 177, 40–47.

Lawton, P., Murphy, E., & Redmond, D. (2013). Residential preferences of the 'creative class'? Cities, 31, 47–56.

Lee K.S. (1982). A model of intra-urban employment location: An application to Bogota, Colombia. Journal of Urban Economics 12.

Lee, B.H.Y, Waddell, P. (2010). Residential Mobility and Location Choice: a Nested Logit Model with Sampling of Alternatives, Transportation 37:587-601.

Lerman, S.R. and Kern, C. R. (1983). Hedonic Theory, Bid Rents, and Willingness to Pay: Some Extensions of Ellickson's Results, Journal of Urban Economy: 13, 358-363.

MAPC, Donoso P. Citi labs, (2013). Boston Region Land Use Model.

Marshall, A. (1920). Principles of Economics. London: Macmillan.

Martínez, F. (1996). Mussa: Land use model for santiago city, Transportation Research Record: Journal of the Transportation Research Board 1552(1): 126–134.

McCarthy, K.F. (1976). The Household Life Cycle and Housing Choices, Papers of the Regional Science Association, Vol. 37, pp. 55-80.

McFadden, D. (1978). Modeling the Choice of Residential Location, A. Karlqvist (ed.), Spatial Interaction Theory and Residential Location, North-Holland, Amsterdam, pp. 75–96.

McGranahan, D., & Wojan, T. (2007). Recasting the creative class to examine growth processes in rural and urban counties. Regional Studies, 41, 197–216.

Mokhtarian P. L., Cao X. (2008), Examining the impacts of residential self-selection on travel

behavior:a focus on methodologies, Transportation Research B 42 204–228.

Myung-Jin Jun (2013), The Effect of Housing Preferences for an Apartment on Residential Location Choice in Seoul: A Random Bidding Land Use Simulation Approach, Land Use Policy 35 1-426

Raspe, O., & Van Oort, F. G. (2006). The knowledge economy and urban economic growth. European Planning Studies, 14, 1209–1234.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, The Journal of Political Economy 82(1): 34 – 55.

Rossi, P.H. [1955] (1980) Why Families Move, Sage Publications, Beverly Hills and London.

Schwanen T., Mokhtarian P. L. (2005), What if you live in the wrong neighborhood? The impact of residential neighborhood type dissonance on distance travelled, Transportation Research D 10 127–151.

Shukla V., Waddell P. (1991). Firm location and land use in discrete urban space. A study of the spatial structure of Dallas-Forth Worth, Regional Science and Urban Economics 21 225 253.

Smith B., Olaru D. (2012), Lifecycles stages and residential location choice in the presence of latent preference heterogeneity, Environment and Plannin A 2013, volume 45, 2495-2514

Ström S. (2010), Housing and first births in Sweden, 1972–2005, Housing Studies 25 509–526

The Kitty and Michael Dukakis Center for Urban and Regional Policy Northeastern University (2015), The Grater Boston Housing Report Card 2014-2015 Fixing an Out-of-Sync Housing Market.

Tomaney, J., & Bradley, D. (2007). The economic role of mobile professional and creative workers and their housing and residential preferences. Town Planning Review, 78, 511 529.

United States Census Bureau. Americal Comunity Survey 2008-2013.

United States Census Bureau. Decenial Census 1970, 1980, 1990, 2000.

Van Wee B. (2009), Self-selection: a key to a better understanding of location choices, travel behaviour and transport externalities, Transport Reviews: A Transnational Transdisciplinary Journal 29 279–292.

Veal A. J. (2001), Leisure, culture, and lifestyle, Society and Leisure 24 359–376.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M. and Ulfarsson, G. (2003). Microsimulation of Urban Development and Location Choices: Design and Implementation of UrbanSim, Preprint of an article that appeared in Networks and Spatial Economics, Vol. 3, No.1, pp. 43-67.

Wedemeier, J. (2010). The impact of the creative sector on growth in German regions. European Planning Studies, 18, 505–520.

Yigitcanlar, T., Baum, S., & Horton, S. (2007). Attracting and retaining knowledgeworkers in knowledge cities. Journal of Knowledge Management, 11, 6–17.