Title

A Behavioral Framework for Measuring Walkability and its Impact on Home Values and Residential Location Choices

Author

Foti, Fletcher

Publication Date

2014

A Behavioral Framework for Measuring Walkability
and its Impact on Home Values and Residential Location Choices

by

Fletcher Scott Foti

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

City and Regional Planning

in the

Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Paul Waddell, Chair
Professor Elizabeth Deakin
Professor Joan Walker

Spring 2014

A Behavioral Framework for Measuring Walkability
and its Impact on Home Values and Residential Location Choices

Abstract

A Behavioral Framework for Measuring Walkability

and its Impact on Home Values and Residential Location Choices

by

Fletcher Scott Foti
Doctor of Philosophy in City and Regional Planning

University of California, Berkeley

Professor Paul Waddell, Chair

Walking is underrepresented in large area models of urban behavior, largely due to difficulty in obtaining data and computational issues in representing land use at such a small scale. Recent advances in data availability, like the ubiquitous point-of-interest data collected by many private companies, as well as a worldwide dataset of local streets in OpenStreetMap, a standard format for obtaining transit schedules in GTFS, etc, provide the potential to build a scalable methodology to understand travel behavior at a pedestrian scale which can be applied wherever these datasets are available.

In addition, the recent invention of fast network algorithms like Contraction Hierarchies greatly reduce related computational issues, as most network computations in this work are computable in less than a second. This thesis is a presentation of such a scalable methodology, which uses widely available datasets wherever possible, with computations that run quickly to encourage exploration of nuance in urban behavior and transparency of outcomes.

Additionally, indexes like WalkScore have been widely studied in the literature recently, both to predict walking behavior and real estate home values. This dissertation takes the position that WalkScore does not sufficiently support the set of destinations it includes, the weights that are applied, the distance decay function, and most importantly does not account for variation in behavior based on the demographics of the traveler. It is also likely that the use of destinations like coffee shops and bookstores in the index measures a specific kind of walkability that embeds a certain kind of neighborhood into its definition.

This dissertation improves on similar indexes like WalkScore by estimating a model that represents the substitution of destinations around a location and between the modes of walking, automobile, and transit. This model is estimated using the San Francisco Bay Area portion of the 2012 California Household Travel Survey to capture observed transportation behavior, and accounts for the demographics included in the survey. These representations of travel behavior can then be used as right-hand side variables in other urban models: for instance, to create a residential location choice model where measures

of accessibility and available demographics are used to understand why people choose to live where they do.

In all cases, location choice models - both destination choice and residential location choice - use a level of detail not common in the literature in order to accurately represent walkability. This dissertation proposes the concept of "street node geography" which uses the local street network to define the geography with which to perform aggregations in the city. In this conceptualization, land uses and other urban data are mapped to their nearest street intersections, and overlapping aggregations are performed along the street network up to a given horizon distance. This representation of urban space is equivalent to a voronoi diagram around the intersections of the local street network, and can be thought of as having automatically generated set of 226,000 micro-zones in the San Francisco Bay Area. Street node geography thus provides a novel compromise between detail and performance for the kinds of computations performed here.

This dissertation is organized into four topics, one for each of chapters 2-5. The first topic establishes a framework for measuring the network of destination opportunities in the city for each of the walking, transit, and auto transportation modes. Destinations in the form of parcels and buildings, businesses, population, and points of interest are tied to each network so that the distance from each location to every destination can be computed by mode. The use of a points-of-interest dataset as the set of public-facing destinations is novel in the context of a traditional travel demand destination model.

This chapter also creates a case study model of trip generation for home-based walking trips is the 2012 California Household Travel Survey. This model finds that WalkScore is predictive of walking trips, that residential density and 4-way intersections have an additional but small impact, and that regional access by the transit network has a synergistic effect on walking, but regional access by auto has no impact when controlling for regional access by transit.

The second topic engages with the question of the impact of accessibility to local amenities on home values. Although early research has found that the composite index WalkScore is positively correlated with home values, this dissertation unpacks the impact of each category of destination used in WalkScore (as well as several others) on home values. The model shows that some amenities are far more predictive of home values in the datasets used here; in particular, cafes and coffee shops tend to be the indicator of neighborhood-scale urban fabric that has the largest positive relationship with home values, where a one standard deviation increase in access to cafes is associated with a 15% increase in home values.

Although the previous topic provides some evidence that walkable amenities are related to increased home values with the datasets analyzed here, it does not prove that households are valuing walking to these amenities; it is equally plausible that households are capitalizing short driving trips into increased home values. The third topic thus creates a nested mode-destination model for each trip purpose (with destinations nested into modes) so that the logsums of the lower nest give an absolute measure of the accessibility by mode for each purpose for each location in the region.

These logsums are then weighted by the number of trips made for each purpose, and segmented by income and weighted by the incomes of the people that live at each location in the city. The result is an index based only on empirically observed behavior (in this case, the primary dataset is the 2012 CHTS) which is an absolute measure of walking behavior, not just of walkability. The methodology from this chapter yields an index for all three modes, and all indexes are included in the hedonic model described above. The model shows that a one standard deviation change in the auto index has the largest impact on home values, but that the walking index is positive, statistically significant, and almost as large. Although part of the reason for this finding might be that these neighborhoods are undersupplied, where they exist they are clearly in high demand.

The fourth topic then engages with the question of how many people actually value walking when making the residential location choice decision. In this section, latent class choice models are used so that coefficients on the three mode-specific indexes (and other neighborhood descriptors) are allowed to change based on selection into unobserved classes. This can be thought of as a form of consumer preference segmentation for mode-specific accessibility.

The model shows that there are three large segments present in the Bay Area. One that is young and moderately high-income that selects into the walkable neighborhoods of San Francisco, Oakland, and Berkeley (13% of households), one that is transit-oriented and selects into the relatively less-expensive neighborhoods near BART but outside the urban core (37% of households), and one that is composed of middle class families that prefers the idyllic suburbs outside San Francisco (50% of households). Apparently about 50% of Bay Area households value transit access, likely because BART allows commute access to the thriving labor market in the urban core (e.g. the SOMA neighborhood which is the target of so much venture capital in the region).

The main research question explored by this methodology is the question of the size of the segment of the population that is positively affected by walking accessibility for the residential location choice and the results show that this segment exists but is of modest size. However, a major finding of this research is that for planning interventions that seek to increase travel by active modes, members of the transit-oriented segment might have the most latent potential to change their behavior. Perhaps creating denser and more walkable environments around the less expensive neighborhoods near BART stations in the region could relieve pressure on the San Francisco housing market as well as create walkable environments for the lower middle class that appear to be a major component of residential demand in the region.

A ripe area for future research is to perform a gap analysis that compares neighborhoods that are high probability areas for each of the three classes presented here to test for the impact of increases in transit service and pedestrian infrastructure on both the residential location choice and travel behavior. Taking into account the heterogeneity of preferences explored here, the result of such a study would target the locations that could have the highest impact on sustainable behavior for the smallest amount of public investment.

For Kerry, Eva, and my folks: sherpas on Mount Berkeley.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation has been directly supported by the Dean's Normative Time Fellowship, as well as with funds from various research projects through the Metropolitan Transportation Commission. Additional funds have been received in my time at Berkeley from the University of California Transportation Center and the Graduate Division at Berkeley. Without the support of these programs I would never have made it this far.

Many people have helped guide me on the path of knowledge which has so far culminated in this research. First and foremost I must thank my primary advisor, Paul Waddell, for answering my email within 10 minutes so many years ago, finding a way that we could work together, and having patience while I made it up to speed on the depths of the software that is UrbanSim. I would also thank Joan Walker for always being a positive voice of support and for teaching me everything I know about choice modeling. To Betty I owe my conscience for why analytical models and the writing of software might matter to the world. Finally I much appreciate learning from Michael Southworth, who gave me the urban design and artistic slant which will inform much future work, and to John Gliebe, my Masters advisor at Portland State University who first set me upon this path.

Other direct contributions to this dissertation include Karen Frick, who took my ramblings and made them well formed, Ian Carlton who taught me how to navigate the byzantine pathways of this university, Angelo Guevara who helped with the nested logit estimation in Chapter 4, and Akshay Vij, who was a great help with the latent class estimation in Chapter 5. I would also like to thank the folks at MTC, David Ory and Mike Reilly in particular, for their contribution to my understanding of applied planning and particularly for how regional planning functions in the Bay Area. Finally, I would like to give special thanks to Dennis Luxen, who gave to the world an open source version of Contraction Hierarchies and added a small twist for this work, without which I would still be waiting for the many networked queries contained here to finish running.

I must thank many excellent professors from whom whom I have learned - Dan Chatman first and foremost for being the voice of transportation and land use, and Nicholas De Monchaux, Ananya Roy, and Kelly Clifton and Jennifer Dill at Portland State for their excellent courses. I must thank many of my classmates, some for intellectual support - Erick Guerra, Aksel Olsen, Jesus Barajas, Liming Wang, Regina Clewlow, Brian Cavagnolo, Nicola Szibbo, and Carrie Makarewicz and others for their emotional support - Julie Gamble, Eva Hagberg, Geoff Boeing, Yizhen Gu and many others.

Finally I must acknowledge the incredible technical frameworks upon which I built the models contained in this research, and other intellectual works I have created not directly tied to this dissertation. I owe a debt of gratitude to Wes McKinney and his Pandas project upon which I built the new UrbanSim software, and Mike Bostock and D3 for creating web visualizations. Additionally, Leaflet and OpenLayers have been endlessly useful for doing web maps to communicate spatial data to the public. And last but not least, I must thank Guido van Rossum for writing the Python programming language,

without which I would have no career as I would have given up software long ago.

I owe an enormous debt of gratitude to my friends at Synthicity - Carlos Vanegas, Federico Fernandez, Eddie Janowicz, Conor Henley, Jason Oliveira, both for developing the GeoCanvas software, which enabled me (and the world) to see a map of a probability distribution function with 200K points for the first time ever, and for listening to my rants of every good and bad idea I've ever had. All screenshots of maps in this work are taken from the GeoCanvas software.

I must also ackowledge grant NSF IIS-0964412: Integrating Behavioral, Geometrical, and Graphical Modeling to Simulate and Visualize Urban Areas for partial funding provided to complete this dissertation.

I must also thank Kerry and my parents, Beth and Tom, for helping me into, through, and out of this process - for them, this is only the beginning. Much love. Finally, I must thank Little Rock, Princeton, Tampa, Washington DC, Portland, Oakland, and the once and future cities I may ever inhabit, for they shape my every thought of every day. That goes double again for Portland and the people who live there, for being a shining example to which urban life can aspire.

Chapter 1

Introduction

1.1 Motivation

Walking is underrepresented in large area models of urban behavior, largely due to difficulty in obtaining data and computational issues in representing land use at such a small scale [TRB, 2007]. Recent advances in data availability, like the ubiquitous point-of-interest data collected by many private companies, as well as a worldwide dataset of local streets in OpenStreetMap [Zielstra and Hochmair, 2012], a standard format for obtaining transit schedules in GTFS [Catala et al., 2011], etc, provide the potential to build a scalable methodology to understand travel behavior at a pedestrian scale which can be applied wherever these datasets are available.

In addition, the recent invention of fast network algorithms like Contraction Hierarchies [Geisberger et al., 2008] greatly reduce related computational issues, as most network computations in this work are computable in less than a second. This thesis is a presentation of such a scalable methodology, which uses widely available datasets wherever possible, with computations that run quickly to encourage exploration of nuance in urban behavior and transparency of outcomes.

Additionally, indexes like WalkScore [WalkScore, 2011] have been widely studied in the literature recently, both to predict walking behavior and real estate home values [Cortright, 2009, Leinberger and Alfonzo, 2005, Manaugh and El-Geneidy, 2011, Rauterkus and Miller, 2011, Weinberger and Sweet, 2012]. This dissertation takes the position that WalkScore does not sufficiently support the set of destinations it includes, the weights that are applied, the distance decay function, and most importantly does not account for variation in behavior based on the demographics of the traveler. It is also likely that the use of destinations like coffee shops and bookstores in the index measures a specific kind of walkability that embeds a certain kind of neighborhood into its definition.

This dissertation mproves on similar indexes like WalkScore by estimating a model that represents the substitution of destinations around a location and between the modes of walking, automobile, and transit. This model is estimated using the San Francisco Bay Area portion of the 2012 California Household Travel Survey to capture observed transportation behavior, and accounts for the demographics included in the survey. These representations of travel behavior can then be used as right-hand side variables in other urban models: for instance, to create a residential location choice model where measures of accessibility and available demographics are used to understand why people choose to live where they do.

In all cases, location choice models - both destination choice and residential location choice - uses a level of detail not common in the literature in order to accurately represent walkability. This dissertation proposes the concept of "street node geography" which uses the local street network to define the geography with which to perform aggregations in the city. In this conceptualization, land uses and other urban data are mapped to their nearest street intersections, and overlapping aggregations are performed along the street network up to a given horizon distance. This representation of urban space is equivalent to a voronoi diagram around the intersections of the local street network, and can be thought of as having 226,000 micro-zones in the San Francisco Bay Area. Street node geography thus provides a novel compromise between detail and performance for the kinds of computations performed here.

To be clear, numerous other indexes of walking attempt to account for various qualitative aspects of the built environment, like enclosure, street trees, safety, architectural aesthetics and others [Parks and Schofer, 2006, Southworth, 2005, Ewing and Handy, 2009, Clifton et al., 2007, Greenwald and Boarnet, 2001]. Although these studies are generally well regarded, and do capture something which is missing from the framework proposed in this thesis, it is important to have a methodology that is scalable so that it can be broadly applied at relatively low cost to the researcher, and this is an important motivation for the methodology proposed here. For instance, an analysis of statewide data in California would be a simple extension to the work performed here, with no additional data collection effort.

1.2 Context of the Study

> "Partly from historical inheritance and partly from the work of activists who chose to make the city the focus of their activism, [San Francisco] remained a walkable, urban paradise compared to most of America." Gabriel Metcalf in The Atlantic Cities [Metcalf, 2013]

San Francisco is a contemporary lightning rod for urban policy in the United States. At the nexus of the issue is its status as a center for technological innovation, combined with its relatively strict density zoning which controls the increase in supply of housing in the city of San Francisco. Although the growth of technology firms in the Bay Area and Silicon Valley in particular has been well documented [Saxenian, 1996], there appears to be a move by businesses back to the city that has not yet been documented in the academic literature.

According to data from Dow Jones, 13.5 billion dollars in venture capital were invested in the San Francisco Bay region in 2011, more than 4 times the capital invested in Boston or New York [Metcalf, 2013], of which the two zip codes in the neighborhood of SOMA (South of Market in San Francisco) received more investment than any other zip code. Mountain View, home of Google and near San Jose, was a distant third. In fact, zip code 94107 near Potrero Hill received 1.8 billion dollars while Mountain View received only 660 million dollars.

Despite the massive job growth, the housing market has failed to keep pace. San Francisco has produced roughly 1,500 units a year over the past two decades while the roughly comparable (but significantly smaller) city of Seattle has produced more than 3,000 units a year over the same period [Metcalf, 2013]. San Francisco is currently undergoing a boom in new housing - more than 4,220 units are under contract to build in the year 2012 (in only the first 6 months; current data from the city are over two years out-of-date), and an additional 32,120 units have been approved by the planning department for future years, and applications for another 6,940 units have been filed for review [SPUR, 2012]. In dollar amounts, over 3.4 billion dollars were spent on new construction from late 2011 to early 2012 in San Francisco.

According to San Francisco Planning and Urban Research (SPUR), the 1,500 units per year built over the last 20 years is less than half what would have been necessary to meet demand [SPUR, 2012]. This has helped fuel an increase in residential rent levels - 7.6% in 2011 according to SPUR, but nearly 34% over the 2 year period between Nov 2011 to Nov 2013 according to zillow.com). The increase in rents is a signal to build more housing, and the cause of the recent boom in development. Although the signal of increased rents is strong, and apparently has been received by developers as can be seen in the recent building spate, a boom-bust cycle is not an efficient housing market.

> "A dramatic supply-demand imbalance that fuels rent spikes and a corresponding building boom is not a good strategy for a healthy, responsive housing market. It reflects a broken housing market where supply is unable to anticipate and react to rapid demand changes. For decades, San Francisco's political environment has hindered new development. As a result, we have underbuilt housing, creating a longterm structural imbalance of supply and demand that the current building cycle does not come close to addressing." [SPUR, 2012]

This is the context for the research in this dissertation, which uses the CHTS regional travel survey which was administered in 2012, as well as a dataset of home values in the same year. Clearly there is much circumstantial evidence for the "return to the city" in both the non-residential and residential sectors, seen in both the investment of venture capital and the increase in rents and prices. This dissertation addresses the questions of what factors best explain current home prices, and why people are electing to live where they do.

Although the methodology described above is widely applicable, the empirical results for the San Francisco Region are highly localized due to these particular planning and economic issues. Much future work has to be done to replicate this work in other locations and compare behavior amongst urban regions, and that is precisely the motivation for building a general framework of this kind.

1.3 Organization and Contribution by Chapter

This dissertation is organized into four substantive chapters which explore this question in separate parts.

1.3.1 Chapter 2 - Destination framework

Chapter 2 puts forward a new methodology for exploring the question of how accessibility impacts decisions in the city. It is primarily motivated by the lack of a general framework that allows analyzing destination choices for different travel modes within the same methodology. Challenges include 1) the computationally intense nature of representing the entire local street network via OpenStreetMap, since there are over 226K local street intersections in the Bay Area, each of which is used as a possible destination in this dissertation, 2) the schedule-dependent nature of the transit network, which has no counterpart in the on-demand auto and pedestrian networks, as well as the difficulty in obtaining and processing accurate data for those schedules, and 3) the near impossibility of getting accurate travel times by auto for every arterial in the region, where time-dependent congested travel times would be ideal.

Chapter 2 addresses these issues to the degree that current data and methods allow, creating a multi-modal network of transportation options so that urban models can be estimated with variables for accessibility by mode for each of the pedestrian, transit, and auto networks. Numerous land use datasets are linked to these networks, including the built environment (parcels and buildings), businesses (via the NETS dataset), population (synthesized disaggregate population from the census), and point-of-interest datasets of specific destinations. This is one of the first bodies of research to use point-of-interest datasets as a measure of attractiveness of possible destinations in a travel demand framework. As points-of-interest data are available for this research, composite indices like WalkScore are also computable in this framework.

Chapter 2 also uses these aggregation variables in a case study application of the framework in predicting home-based walking trips in the Bay Area portion of the 2012 California Household Travel Survey (CHTS). Results reinforce previous findings that WalkScore is predictive of walking, that residential density has an additional contribution not accounted for in WalkScore and cumulative accessibility indices, and that regional transit access has a synergistic effect with walking but regional access by auto has no impact on walking trip generation when controlling for regional access by transit.

1.3.2 Chapter 3 - Impact of local accessibility on home values

Chapter 3 takes cumulative opportunity accessibility measures and uses them as right-hand side variables in a residential price hedonic model using a large dataset of home prices from the Bay Area collected in 2012. Although initial research indicates that WalkScore is predictive of home values, the specific destinations which are most descriptive of increased home values are not clear when regressing against a composite index like WalkScore. This research "unpacks" WalkScore into accessibility to the component destinations, as well as including a number of other possible destinations (there are over 370 categories in the Factual dataset used here), to measure relative impacts of accessibility to amenities on home values.

Results from Chapter 3 indicate that walking scale access to amenities and disamenities has a clear impact on home values, mostly in theoretically expected directions. Interestingly, some WalkScore destinations, like groceries and restaurants, have a negative impact on home values when controlling for the others, while destinations like cafes and coffee shops have a positive impact on home values. This indicates that some destinations are more structurally descriptive of increased home values than other destinations, with a standard deviation increase in accessibility to cafes associated with an increase in home values of almost 15%.

1.3.3 Chapter 4 - Empirically estimated destination model and associated impact on home values

Although Chapter 3 provides evidence that walkable amenities are correlated with increased home values for the datasets analyzed here, this does not prove that households are actually valuing walking to these amenities; it is equally possible that households are capitalizing short driving trips into increased home values. This chapter leverages the networks from Chapter 2 to estimate a nested mode-destination model in which the modes are the top level of the nests while destinations are the bottom level of the nests. Thus destinations can be combined into a logsum which is an aggregate, absolute measure of accessibility by mode for a household at a given location in the city. Consistent with the long-standing use of nested destination models in travel modeling, these logsums actually represent the probability of choosing a mode for travel from the home location.

Destination choice models are designed to include positive traits of the destination, negative traits of travel to the destination, and attributes of the decision maker, which is an ideal framework to allow highly disparate travel preferences by different income classes to be adequately represented. Logsums are computed for each purpose, for each income group, and for each street intersection in the city, yielding 68M total logsums. These logsums are then combined by weighting each purpose-specific logsum by the number of trips made for that purpose and then weighting each location in the city by the income classes of the people who live at each location.

The result is a more theoretically grounded WalkScore - an estimate of walking which is based on empirical data of how people actually travel, in contrast to the weights used in WalkScore which are chosen because they are empirically plausible. The weighted logsums are an empirically estimated index for walking, transit, and auto for each location in the city that takes into account the number of trips made for each purpose and different travel behavior by people of different income classes.

These indexes by mode are used in the hedonic model from Chapter 3 to test the value of actual walking to local amenities in the valuation of home prices. Accessibility logsums for all three modes are positive, and although the impact of a one standard deviation change in the index for auto is the largest, the index for walking is positive, statistically significant, and almost as large. This is a major contribution to the literature as it moves from the previous correlational relationship of WalkScore and home values toward a more robust behavioral explanation that people are valuing the ability to walk to nearby destinations in their home purchasing decision.

1.3.4 Chapter 5 - Testing variability of accessibility preferences for residential location choice

Most choice models are used to regress a set of independent variables against a categorical dependent variable - which usually represents the choice among available alternatives, like travel modes - to estimate a single set of coefficients that represents the *average* effect of the the independent variables on the dependent variable in the estimation dataset. Of course not every decision maker has the same preferences; in fact, in residential choice preferences for the home location can vary widely according to wealth, cultural background, age and lifestyle stage, presence of children in the household, and other factors. Chapter 5 uses latent class choice models (LCCMs) to allow coefficients of the independent variables to vary based on selection of each household into unobserved classes. These classes can be thought of as consumer segments for the residential housing market, and in fact most early research using LCCMs came from consumer segmentation studies.

The findings from this chapter show that there are three large segments of accessibility preferences in the Bay Area datasets used here. The first seems to be the segment of young professionals discussed in the motivation section. This segment values accessibility primarily by walking and transit, with the largest coefficient for walking, and is the smallest of the three segments, comprising only 13% of households in the Bay Area. The second segment is the lowest income and selects into neighborhoods near the subset of BART stations that are in less expensive neighborhoods, and has a positive and statistically significant coefficient for transit logsums while the coefficient on walking logsums is not statistically significant. This segment comprises 37% of the total population of households. The final segment has the highest average income and selects into the idyllic suburbs throughout the Bay Area. This segment does not appear to have a positive relationship with accessibility by any mode (when controlling for the other variables used in the model) as coefficients on logsums for all three modes are negative and statistically significant. This segment is by far the largest and comprises 50% of Bay Area households.

The main research topic explored in Chapter 4 is the question of the size of the segment of the population that is positively affected by walking accessibility for the residential location choice and the results show that this segment exists but is of modest size. However, a major finding of this research is that for planning interventions that seek to increase travel by active modes, members of the transit-oriented segment might have the most latent potential to change their behavior. Perhaps creating denser and more walkable environments around the less expensive neighborhoods near BART stations in the region could relieve pressure on the San Francisco housing market as well as create walkable environments for the lower middle class that appear to be a major component of residential demand in the region.

A ripe area for future research is to perform a gap analysis that compares neighborhoods that are high probability areas for each of the three classes presented here to test for the impact of increases in transit service and pedestrian infrastructure on both the residential location choice and travel behavior. Taking into account the heterogeneity of preferences explored here, the result of such a study would target the locations that could have the highest impact on sustainable behavior for the smallest amount of public investment.

1.3.5 Chapter 6 - Conclusions

Chapter 6 provides some brief conclusions, contributions to the literature, policy implications, and a section on future work which has been suggested by this research. To summarize, it is likely that walkable amenities are being capitalized into increased home values, and the actual behavior of walking is also associated with increased home values. Although San Francisco is a very expensive and constrained housing market, a large proportion of households are transit-oriented (especially BART-oriented) but quite price sensitive. It is likely that creating dense walkable neighborhoods around BART stations in less expensive areas would both reduce pressure on the San Francisco housing market and create attractive environments for the lower middle class that appear to be the backbone of demand for pedestrian- and transit-oriented neighborhoods in the region.

Chapter 2

Substituting Modes: Accessibility to Destinations via the Multi-modal Transportation Network

2.1 Introduction

This chapter develops and applies an accessibility framework that maps activity destinations and other features in the built environment to a multi-modal transportation network that extends to the local street level of detail, allowing the summation of destinations from the local scale to the regional scale by transportation mode. This chapter's contribution is to allow the comparison of different types of accessibility measures in predicting a given empirical outcome, bringing together the following three threads in the literature.

First, this framework is designed to accurately represent pedestrian-scale accessibility, which continues to be a secondary consideration in current travel modeling practice [Waddell, 2009, TRB, 2007]. Second, this framework is extended to be multi-modal, since it has been theorized that pedestrian demand cannot be measured accurately without also measuring the relative accessibility of auto travel [Chatman, 2008, Krizek, 2003, Crane, 1996]. Finally, this framework implements 3Ds variables that have been widely used in sketch planning, but have been criticized on the grounds that these measures do not propose a behavioral explanation for travel [Crane, 2000, Boarnet and Crane, 2001]. This framework is motivated by the principle that travel is a derived demand most heavily influenced by access to destinations, traits of the routes to those destinations, and mediated by attributes of the decision maker [Cervero, 2002, Guo et al., 2007].

This research makes several methodological advances that allow representation of the full set of local streets for pedestrian-scale accessibility, a hierarchical multi-modal graph to represent the tradeoff between modes, and integration of micro-scale land use data to measure the full population of alternative destinations in the city. This chapter focuses on a model of home-based non-work pedestrian trip generation using the the Bay Area potion of the 2012 California Household Travel Survey (CHTS) as a case study of an application of this framework.

This research is motivated in part by the need for more effective methods to assess the potential impact of policies that encourage compact development in order to reduce VMT and concomitant greenhouse gas (GHG) emissions [Boarnet et al., 2011, Brownstone, 2008] by promoting transportation alternatives such as transit and walking. This topic is of particular relevance in the state of California, where Senate Bill 375 (SB375) mandates that each of its MPOs creates a Sustainable Community Strategy (SCS) that must analyze potential GHG reduction through coordinated land use and transportation policies [Barbour and Deakin, 2012], and must analyze the impact of policies which increase residential density on reductions in automobile use and increases in travel by sustainable modes such as walking, bicycling, and public transit.

2.2 Literature Review

"Urban form must be evaluated in terms of the set of choices it provides – what kinds of destinations are found where and served by what transportation modes – and the characteristics of those choices, including cost and comfort of travel, the amount and quality of the activity at the destinations, etc. In this framework, for example, it is the set of choices correlated with density – not density itself – that shape travel behavior" [Handy, 1996b].

For decades the transportation-land use literature has attempted to relate the demand for travel to aggregate measures of the built environment. Models of travel behavior analyze an outcome variable – often trip generation, mode choice, or vehicle miles traveled (VMT) as a function of demographics and measurements of land use. Land use variables are frequently measured using some variation or extension of the 3Ds Kockelman and Cervero [1997], which capture in turn the density, diversity, and design of a geographic area. Variables commonly include residential density, diversity of land uses within the nearby area, and design of the street grid, and counts of destinations within a constant time distance (known as an isochrone).

Despite limited microdata available at the time, Ewing and Cervero [Ewing and Cervero, 2001] developed broad conclusions based on a wide breadth of empirical literature that trip generation is largely determined by demographics, trip distance varies primarily with the built environment, and mode choice depends on both demographics and the built environment, but predominantly on demographics. Although these early studies were a major contribution to our understanding of the influence of land use on the demand for travel, this type of study was quickly criticized for its lack of behavioral foundation Crane [2000], Boarnet and Crane [2001]. Boarnet [Boarnet, 2011] identified 3Ds-style studies as "reduced form" models and argued for the move to "structural models" which explain *why* residential density, for instance, might influence travel.

2.2.1 Accessibility Measures

Travel is a "derived demand," meaning most travel is performed to reach destinations rather than for the sake of the travel itself, thus the most common causal explanation for different travel patterns is the relative attraction of available destinations and the cost of the travel in order to reach those destinations. Two different frames are used to understand the ability to travel: the first is "mobility," which measures a person's freedom to travel more quickly and reach destinations that are ever further afield, and "accessibility," in which land uses are brought closer to the origin of the trip and more destinations can be reached even if travel conditions are highly congested.

The definition of accessibility used here is "a measure of an individual's freedom to participate in activities in the environment" [Weibull, 1976, Miller, 1999]. Accessibility has been measured using different methods [Handy and Niemeier, 1997, Dong et al., 2006], including "iscochrones" that sum opportunities within a distance or travel time, "gravity model" measures which discount the opportunities by some measure of the distance to each destination, and logsum measures [Dong et al., 2006] which estimate coefficients on attractors and impedances using a statistical framework, typically using a discrete choice model.

This work uses the point locations for activities in the city as attractors that generate travel, consistent with the concept of transportation as a derived demand. However, because of data limitations 3Ds variables are used to capture traits of the built environment that are not captured directly by accessibility variables. For instance, residential density likely proxies for a more pleasant and enclosed walking environment [Chatman, 2008, Cervero, 2002], and in lieu of having subjective data on the quality of the pedestrian environment along a route which would be more consistent with the theory espoused here, these types of sketch planning variables must continue to be used to capture such additional considerations which are of interest to planners and thus both types of variables are supported in this framework.

2.2.2 Pedestrian Models

Data has become increasingly available at the pedestrian-scale, and a large body of literature in the demand for pedestrian travel has resulted. Walkscore is a commercially successful online service that computes and maps a weighted combination of the fine-grained location of nine types of nearby destinations [WalkScore, 2011], and has grown to service almost six million queries a day for the WalkScore at specified addresses. The index combines the closest grocery store (weight 3.0), closest 10 restaurants (weights varying from .75 to .2, summing to 3.0), 5 retail establishments, 2 coffee shops, and the closest bank, park, school, bookstore, and "entertainment venue," and there is a clear pattern of decreasing returns to additional destinations (i.e. the closest coffee shop is weighted at 1.25 and the second at .75). The current primary application of WalkScore as a commercial product is in selling residential real estate, and research has identified price premiums for locations with high Walkscores [Cortright, 2009].

Recent research also shows that Walkscore is predictive of walking trip generation [Weinberger and Sweet, 2012, Manaugh and El-Geneidy, 2011], but this study uses modeled pedestrian outcomes derived from travel models which themselves misrepresent the walking environment by ignoring local streets and fine-grained land uses. The theoretical framework established for Walkscore [Frank et al., 2008, Moudon et al., 2006] is intuitive and easily applied, but the set of destinations, the weights given to the destinations, and the distance decay function are empirical questions that merit more investigation (and are explored in Chapter 4 of this dissertation). Additionally, the relationship of pedestrian travel to meso- and macro-scale accessibility is largely missing from this line of research, a shortcoming which is addressed in this framework.

2.2.3 Discrete Choice Frameworks

Discrete choice frameworks have long been the standard approach for modeling travel choices, with mode choice being the canonical application of this methodology. Discrete choice modeling [McFadden, 1980] allows the estimation of indirect utility among a number of alternatives subject to a linear in parameters utility function and a given distribution for a random error term. See Ben-Akiva and Lerman [1985] and Train [Train, 2009] for a thorough treatment of the methods and their application to travel behavior. Williams [1977] developed the theory for using logsum measures from such models to measure consumer surplus. In that sense, the use of logsum measures to represent accessibility is also a measure of consumer surplus.

A framework for mode choice is provided in Cervero [2002], which allows for combining discrete choice estimation of relative utility among travel modes with the use of 3Ds metrics as explanatory variables. This approach is used in Guo et al [2007] which studies the substitution of pedestrian and auto modes, finding that pedestrian travel is largely complimentary to automobile travel (in other words, pedestrian access generates additional walking trips that do not tend to substitute for automobile travel). This chapter contributes to this literature by accounting for all destinations for the walk, auto, and transit modes in order to test this substitution pattern more closely. Chapter 4 expands the approach from this chapter by creating an endogenously estimated mode and destination model using the data and networks described here.

2.3 Framework and Research Questions

This work creates an applied computational framework for computing accessibility variables for use by planning academics and practitioners. It includes support for multimodal networks, for associating any type of land use (or other attribute) to locations within each network, and for computing the many types of accessibility variables discussed here. The project is open source and publicly available and can be expanded as new concepts are added and limitations are addressed. In particular, accessibility variables can be used within a framework for running regressions and choice models on disaggregate datasets suitable for modeling urban behavior.

Although this framework is generic and suitable for a wide variety of applications, a simple application of this framework is developed to predict walking trip generation for non-work home-based trips, which is used to test three hypotheses:

1. Does Walkscore have independent predictive power for walking trip generation after controlling for other typical destination accessibility (cumulative opportunity and gravity-model) variables?

2. Do regional automobile accessibility measures have a negative impact on walking trip generation (substitutive effect) after controlling for local accessibility and regional

transit accessibility? In other words, is there evidence that increased automobile accessibility has a negative impact on walking trip generation?

3. Does population density, an important variable in the 3Ds framework though often subject to theoretical criticism, have independent predictive power after controlling for other destination accessibility variables and composite local accessibility indexes like WalkScore?

2.4 The City as a Network Graph

The foundational assumption in this research is that travel is a derived demand for engaging in an activity at a destination separated from the point of origin by the impedance required to reach the destination. This is roughly consistent with the motivation for an aggregate gravity model [Hansen, 1959], which predicts the volume of travel between two zones based on the size of the origin and destination and the impedance between them. Unlike the aggregate gravity model, the methodology used here, combined with the availability of data on the behavior of individual people, supports analysis of the relative importance of attributes that can include 1) the amount or quality of activity at the destination 2) aspects of the multi-modal route including local streets and 3) attributes of the decision maker which modify the influence of the destination and route.

This work proposes a representation of the city as point-specific activity locations situated within the context of a multi-modal transportation network. This conception is not new; Kevin Lynch in Image of the City defines urban geography in the vernacular of networks as “paths, edges, districts, nodes, and landmarks” [Lynch, 1960]. This research operationalizes the concept with a general framework that unifies numerous spatial data sources useful for travel demand models and models of other urban behavior (e.g. location choices), and places the elements of these datasets within the context of the multi-modal transportation network that enables interaction among different parts of the city. By unifying multiple datasets, a large set of variables is available to begin to identify the character of different locations in the city and test for impacts of each variable on an empirical behavior of interest to planners.

Most geographic spatial data is geo-referenced with a latitude and longitude or other coordinate system, but much data of use to transportation planners is associated to parcels of land. In this research, buildings with associated use and size are assigned to parcel shapes maintained in each region, and thus land uses of this sort are given a geographic position through the regional parcel map. In general, destinations can be assigned to the multi-modal transportation network through either latitude and longitude or through regional parcel identifiers.

A typical schema of urban data relationships is shown in Figure 2.1, which depicts data frequently used in urban modeling [Waddell et al., 2005, Waddell, 2002a]: households and businesses are assigned to buildings, which are assigned to parcels, which are in turn

Dataset	Count of Objects
Parcels	2,023,915
Single Family Houses	1,479,511
Non-SF Buildings	456,749
Establishments/Firms	434,302
Jobs	3,395,967
Households	2,608,023
People	6,996,929

Table 2.1: Number of objects per dataset used in this research

placed within the context of the local street network. Other spatial data can be assigned to a node in the network by latitude or longitude or any other geometry representable in a geographic information system (GIS).

2.4.1 Assignment of Land Use

To perform accessibility calculations, land use data must first be efficiently connected to the transportation network. Datasets are large, and the number of objects of each type for the Bay Area implementation is shown in Table 2.1. Ideally the synthesized population and inventory of firms would be assigned to the parcel map, and the parcel map would contain addresses that define the means of access and egress from each parcel to the local street network. Every person represented in the region would have access to every firm and vice versa via these access and egress points, and this complete graph of parcel connections is referred to here as the Parcel Graph. In an even more data rich environment, exact pedestrian infrastructure including sidewalks and street crossings would be represented, and for very short distances pathways inside of buildings might even by important to accurately capture travel times.

In practical applications such as regional land use and transportation models, creating a full Parcel Graph using data on parcels, buildings, businesses, and households is complex and messy, and the interrelationships among these data elements are confounded by error. Although population might be accurate at higher-level census geographies, the assignment to buildings is typically performed by iterative fitting to observed marginals [Beckman et al., 1996] which introduces error in the relationship of agents and buildings. Firm data is yet more problematic: establishments are tracked in a number of datasets, but businesses with multiple locations can be assigned to a single building, and geographic knowledge is often no more specific than assignment to the nearest tract or block group centroid. Building data are maintained by county assessors and contain myriad errors in spatial encoding including repeated stacked or overlapping parcels, misrepresentation of buildings types, and unrecorded informal units.

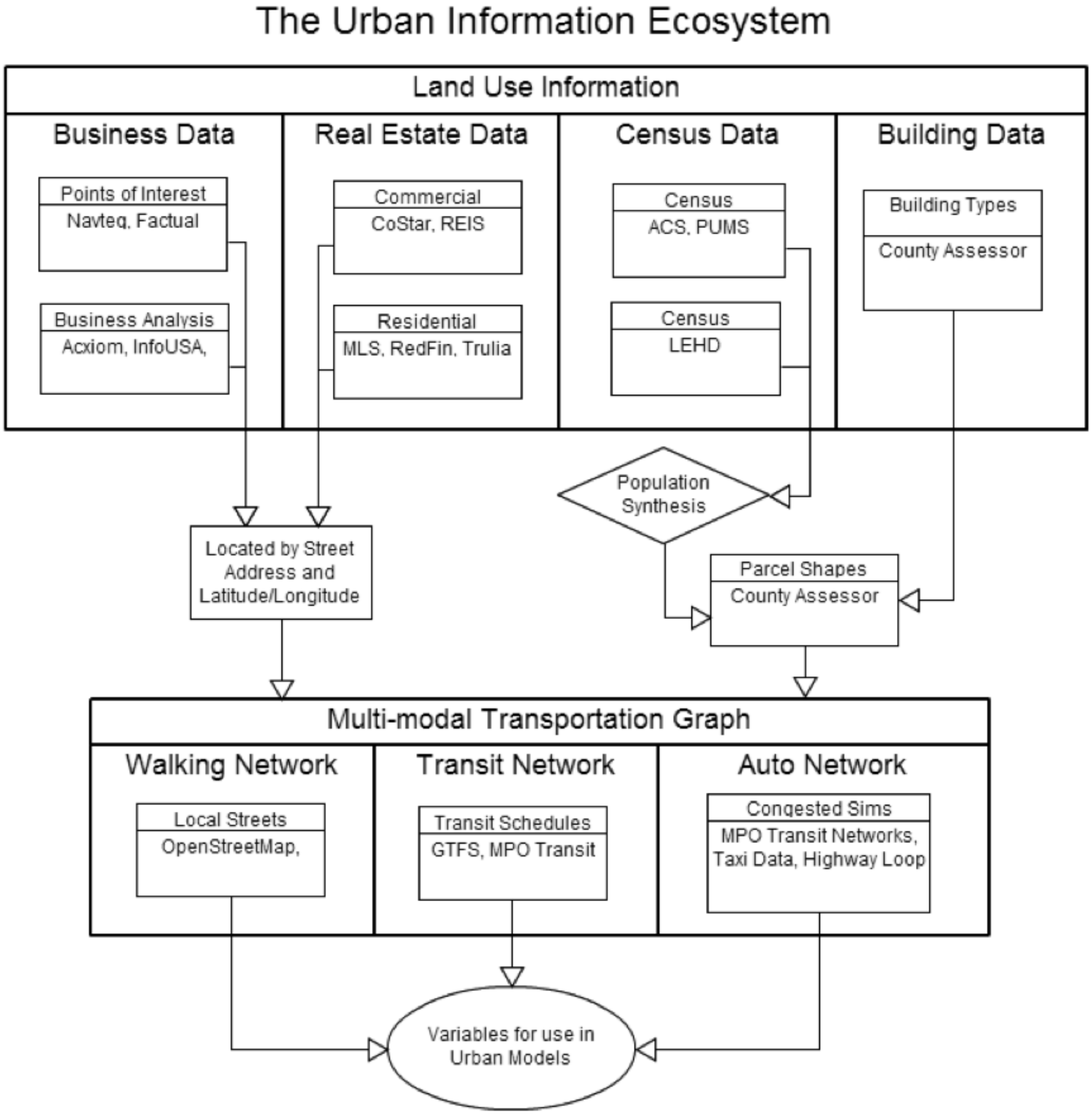


Figure 2.1: The land use and transportation datasets used in this research

2.4.2 Street node geography

An extremely useful simplifying assumption has been made in this research to adopt local street nodes as the primary unit of geography. In this case, each land use is mapped to its nearest street intersection and thus all land uses are connected to the vertices of the network being used, which is referred to here as the Node Graph. Thus the city can be mapped as a Voronoi diagram [Aurenhammer, 1991] of the local street network intersections. Spatial data is assigned first to parcels, then parcel centroids are mapped to the nearest street node, and the relationship between parcels and nodes is used to map land use to the network.

As walking distances are typically significantly larger than the distance from each parcel to its nearest local street intersection, this reduces accuracy of models very modestly within urban areas, while significantly improving computational performance. A dual of this framework would map land uses to the nearest edge (link) on the network rather than to the nearest node (vertex). This is "block face" geography [Clifton et al., 2008] and can be represented by using the "line graph" of the local street network in which every edge is replaced by a vertex and vice versa. In general, this research adopts street node geography as its frame of reference, but this choice is primarily for computational performance and for ease of visualizing inputs and outputs at specific points in space.

Street networks reflect important elements of urban form, such as variability in density across areas of the city. Thus the street network is an important cue that there is less density of land uses in areas that have large distances between nodes, and more density of land uses where there is less distance, and hence using street network geography means space is more accurately represented where it matters the most. From an information theoretic perspective, there is more information where cities are denser and thus street node geography can be used to compress (in the mathematical sense) the city appropriately. The accumulation of land uses to the nearest street node reduces the number of land use elements by almost a factor of ten, which dramatically reduces computational costs while maintaining walking-scale spatial resolution. Almost all computations performed in this analysis can be performed interactively (in less than a second).

This abstraction fails for large parcels like urban parks, university campuses, and corporate office parks. Generally speaking, where location of actual buildings is known, assignment of building to street node directly should be done. Unfortunately, this information is often unavailable and large parcels must be allocated proportionally to all adjacent street nodes. The assignment of land uses can be applied to any network, including networks for other transportation modes discussed next.

2.4.3 The Multi-modal transportation graph

Efficiently measuring region-wide pedestrian-scale access on the local street network is a fundamental goal of this work in order to better represent the opportunities for walking. To this end, all local streets for the entire Bay Area region are used to represent accessibility to destinations via walking. Some have argued that demand for pedestrian travel is not independent of other transportation networks [Crane, 1996, Krizek, 2003], theorizing that bringing destinations closer might reduce the cost of automobile travel at the same time it enables pedestrian travel, possibly inducing more auto travel as the cost of both modes falls. Thus it has been argued that the full set of options in the transportation network must be represented in order to capture these tradeoffs, and for this reason multiple modes are available in this framework.

The multi-modal transportation network (see Figure 2.2) is represented as three separate graphs for walk, auto, and transit modes (biking networks can be included in future work). The walking network is distance based, uses the full set of local streets from the OpenStreetMap project, and includes information on pedestrian infrastructure where available. The transit network is obtained through the Bay Area 511 website via the general transit feed service (GTFS) and is processed into a static network to use in this research. The automobile network is subject to congestion (i.e. traffic) and congested travel times are obtained from the travel model used by the Metropolitan Transportation Commission (MTC) using a regional network that focuses on collector streets and highways.

2.4.3.1 Pedestrian network

The pedestrian network used in this research comes directly from the OpenStreetMap project. OpenStreetMap is derived from the 2007 US Census Tiger line files and has been edited by the general public using the 'crowdsourcing' approach popularized by Wikipedia. The result is a robust and ubiquitous mapping of roads in the United States and is now available for most countries of the world. The user base is massive, with over one million users editing data and over 169 million edges and 176 million nodes currently represented (OpenStreetMap 2013), and the data quality compares favorably to proprietary data sources Zielstra and Hochmair [2012].

2.4.3.2 Automobile network

The use of the MTC travel model for congested skims is necessary for several reasons. First, empirically observed travel times on arterials are simply not available at this time. Although observed travel times are recorded in California PeMS (Performance Measurement System) for all state highways, similar measurements are not available for the arterial network. Thus the calibrated MTC travel model is the best source for high-detail congested travel times. Second, integrating the travel model network allows this framework to be used in scenario planning for the Bay Area SCS and this allows pedestrian-scale sensitivity for non-base year conditions. Finally, although there are 1454 current TAZs in the MTC travel model network, a project is underway to increase the resolution of this network by an order of magnitude. The implementation created as part of this research can interact with any network, and replacement of networks can be made easily where

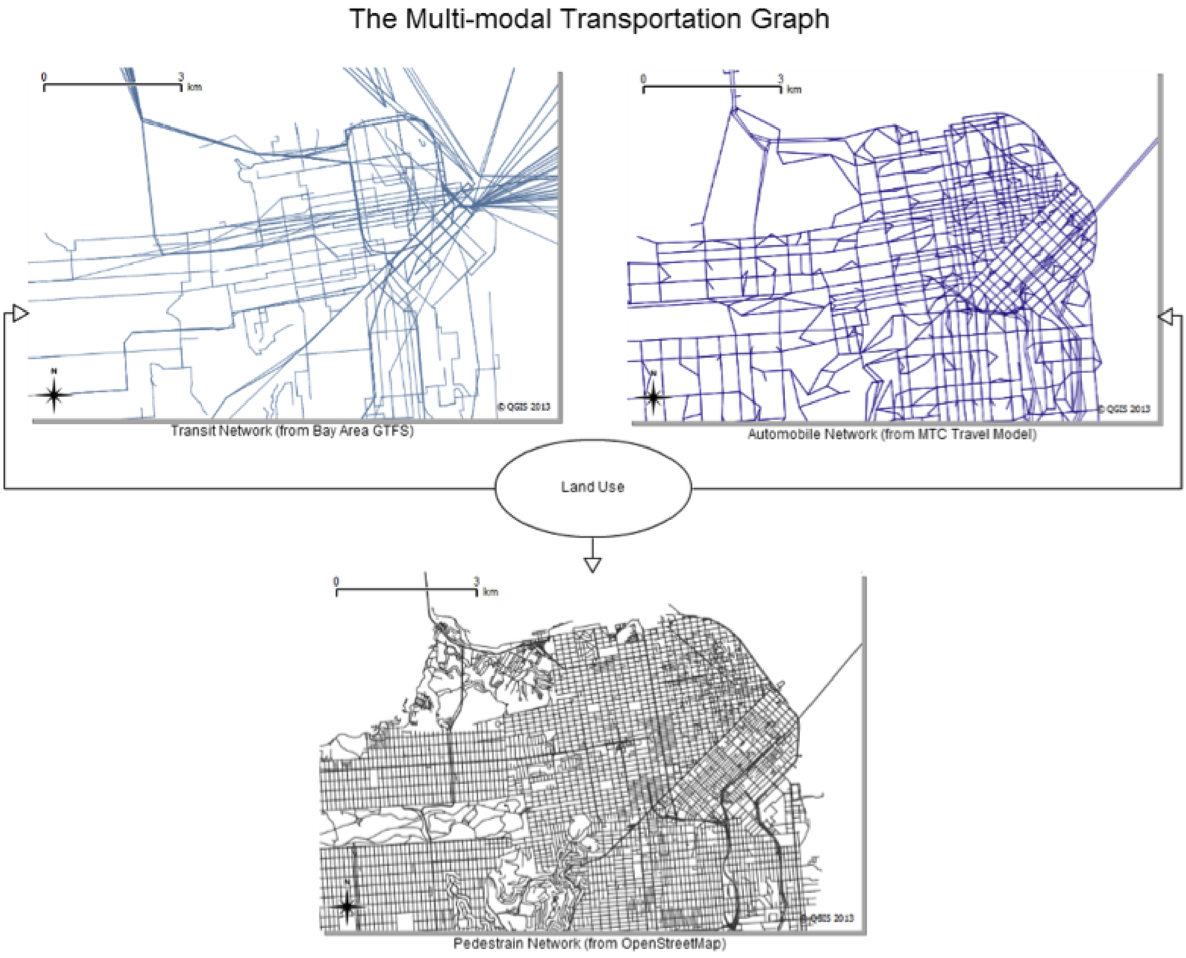

The Multi-modal Transportation Graph
0
3
km
N
© QGIS 2013
Transit Network (from Bay Area GTFS)
0
3
km
N
© QGIS 2013
Automobile Network (from MTC Travel Model)
Land Use
0
3
km
N
© QGIS 2013
Pedestrain Network (from OpenStreetMap)

Figure 2.2: Transit, auto, and walking networks are shown in the clockwise direction for the same geographic area

appropriate.

2.4.3.3 Transit network

The transit network is obtained via the General Transit Feed System (GTFS, originally called the Google Transit Feed System), which is now in widespread use by transit agencies to provide schedules to the long list of routing and mapping services geared towards transit. GTFS provides the complete list of stops, routes, trips, agencies, fares, and the weekly schedule and specification of holidays. Put simply, it is meant to be a representation of every transit vehicle and where each vehicle will be when, and is designed to provide accurate routing information to the public for every time of day. This research is focused on the generation of trips which is largely a result of daily trip planning which would tend to take into account accessibility by transit during a typical service period (e.g. morning peak), and need not accurately represent to-the-minute transit access for each time of the day.

The highly time-dependent transit network is converted to a static network usable for computing accessibility. This is done by taking the first trip from each route after 8AM and placing an 'edge' between adjacent stop locations which has a weight (in seconds) that is taken directly from the GTFS schedule. Transit access is highly dependent on the distance between the origin of a trip and the boarding transit stop and between the destination and the alighting stop, thus the transit network is linked to the local street network in a single unified network.

Travel times for transit come from the transit schedules while travel times on the pedestrian network are taken by dividing the link distance by an average walking speed of three miles per hour. Edges are created between each transit node and the nearest local street node. Edges that link transit to the walking network are currently attributed a constant weight which is a user-specified value for average wait time (typically 3-5 minutes) and these edges represent the time spent waiting for the next transit vehicle. There is significant evidence that wait time and transfers [Taylor and Fink, 2003] are a significantly higher subjective burden than in-vehicle travel time, and the penalty for mode transfers can be modified to fit this theory as appropriate.

2.4.3.4 Unified network

Note that the transportation networks described operate at different geographic scales. Figure 2.2 shows a map at the same scale of the three networks, with the pedestrian network not included in the transit image for the purpose of visual clarity. Note that the pedestrian network is the network of local streets that is ubiquitous where there is development. The automobile network is widespread but not nearly as dense, which enables computation of accessibility for a much wider radius typically reachable by auto. Finally, the transit network is moderately widespread but densely concentrated along corridors. The central location shown in the figure from which many transit lines emanate (lower right section of downtown) is the TransBay Terminal which services all the buses across the Oakland Bay Bridge.

Table 2.2 gives basic descriptive information for each network including the data source, the standard deviation and average number of nodes reachable within fifteen minutes and the average number of nodes within thirty and forty-five minutes. The 45 minute isochrone is used to delineate the typical set of destinations with the count of all reachable nodes in all three mode-specific networks summing to 11 thousand alternatives for the average size of the choice set considered by each person in a destination model. Note that even though the travel model network is less dense than the local street network, the number of nodes reachable within the same amount of time is much larger via this network, thus it's likely the travel model network is scaled with enough detail relative

	Radius (in minutes)	Local Street Network	Automobile Network	Transit Network
Source		OpenStreet-Map	MTC Travel Model	Bay Area GTFS Feed
Count of Nodes	n/a	226,060	11,999	421,491
Count of Edges	n/a	287,161	33,136	660,914
Ave number of Nodes	15	122	1,154	143
Stddev num of Nodes	15	85	678	213
Ave number of Nodes	30	432	3,628	1,063
Ave number of Nodes	45	900	6,565	3,729

Table 2.2: The transportation networks used in the Bay Area SCS implementation, the source of the data, and various characteristics of the network

to the other networks and reduced detail (not including local streets) in the travel model network should not affect the results here.

2.4.3.5 Multiple Impedances

Although travel time is the primary impedance used in this research, it is important to note that this framework is ideal for supporting multiple impedances and measuring the contribution of each to destination choices. For instance, the traits of transit already discussed such as wait time, number of transfers, vehicle technology (e.g. bus vs. train) can be added as additional impedances. For auto, congestion could be theorized to have an independent impact from travel time (for the psychological impact as well as the unreliability of travel time). For walking, a number of additional impedances can be added, including the number of arterials crossed, the presence of sidewalks, and safety along the route, enclosure and 'explorability,' all traits of the route consistent with Southworth's theory of walking [Southworth and Owens, 1993]. These traits can even be interacted with traits of the user to test various hypothesis, e.g. older people might be less willing to walk far or to cross arterials as younger people. This set of hypotheses, though not the direct focus of this chapter, are easily testable in a discrete choice microeconomic framework.

2.5 Data and Methodology

2.5.1 Data

2.5.1.1 CHTS 2012 Travel Survey

This research leverages data in three areas: travel surveys, spatial datasets of land use, and representations of the transportation network. As the networks were discussed at length above, they won't be discussed further here.

Although the Bay Area Transportation Survey (BATS) from 2000 has been used widely in previous research (e.g. Guo et al., 2007, Beckman et al., 1996), the survey did not contain robust data on latitude and longitude which are necessary to capture pedestrian-scale travel decisions. The travel survey used here is the Bay Area portion of the California Household Travel Survey (CHTS) for 2012, which has been recently released and contains accurate latitude and longitude for all locations in the survey.

To emphasize this point, **although CHTS is administered statewide, this research always uses the 9-county Bay Area portion for analysis without exception.** Future work should repeat this analysis elsewhere in California as the data is readily available.

The Bay Area portion of CHTS 2012 is a survey of 9,719 households comprised of 24,030 people, which consists of an activity diary in which participants record locations for all activities and travel modes over the course of a one day period. This survey methodology is limited in that it only tracks a single day of activity rather than the two day period common in many travel surveys. To be clear, this a sample of 24,030 people out of the entire Bay Area population of 7.44 million which means each survey taker represents 310 people in the population. This limits both the detail that can be attained from knowing peoples' habitual behavior as well as the variety of people that might have taken the survey. New methods in data collection, including mobile apps like The Quantified Traveler [Jariyasunant et al., 2013] would go a long way toward rectifying this issue, but for now regional travel surveys are performed in most major regions throughout the world and are considered to be part of the best practices for travel modeling.

CHTS 2012 contains a question not present in BATS 2000 regarding the number of walking trips in the past week, "including trips for exercise." Although this research attempted to use responses to this question as the dependent variable for walking trip generation, model results were not robust, likely due to problems of self-reporting and the mixing of trip purposes. Instead, the number of home-based non-work walking trips from the trip diary portion of the survey are regressed on measures of land use around the survey-taker's place of residence. Figure 2.3 shows the histogram of the number of these walking trips made in the Bay Area portion of the CHTS 2012.

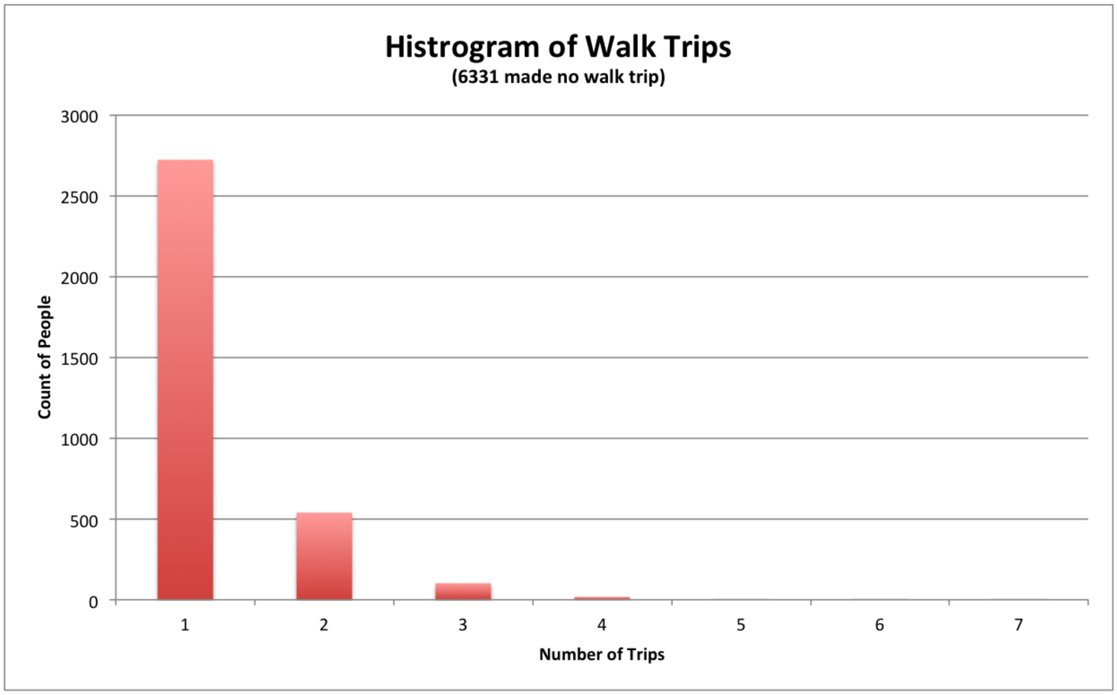


Figure 2.3: Histogram of counts of home-based non-work walking trips present in the California Household Travel Survey 2012

2.5.1.2 Land Use Data

Land use information was collected as part of the UrbanSim implementation for the Bay Area Sustainable Communities Strategy planning effort in 2012 [Waddell, 2013]. This process produced an improved dataset of buildings from county tax assessors that provides non-residential square footage and residential units by building type at the parcel level. The project also generated a synthesized population which represents each person in the Bay Area with associated demographics including age, gender, income, etc. Additional land use data comes from the Factual Places dataset that provides a complete set of destinations by functional category (e.g. restaurants, coffee shops, auto shops, etc). Finally, the California NETS (National Establishment Time Series) dataset contains detailed employment by sector. Many different forms of variables from these datasets have been tested as part of this research (see Table 2.3 for a description of the datasets that have been applied), and significant findings are presented in the results section and in later chapters.

Dataset	Source	Columns	Examples
Firms	Dun and Bradstreet, Infogroup	NAICS category, number of employees, sales volume	Sum of employment or sales by category
Population	Synthesized or Anonymized	Demographics (age, race, sex, household income, employment status)	Sum of population by income, or average income
Points-of-interest	Online places datasets	Categorical definition of location (cafe, bookstore, liquor store)	Accessibility to a certain category
Land Uses	MPOs, County Assessors	Building size (sqft and stories) by land use or type (single family detached, attached, multifamily, strip retail, big box, etc)	Mixing of land uses by square footage, sometimes with a multiple (e.g. residential sqft is 5x of retail)
Other	Various	Pollution, Health Metrics, Crime, etc	Count of property or violent crime locations

Table 2.3: Typical urban datasets and how they can be applied with the methodology described here

2.5.2 Methodology

2.5.2.1 Accessibility

Variables in this research project include 1) traits of the decision maker taken from the travel survey 2) traits of the routes taken from the transportation network (typically impedances) and 3) aggregations of the built environment around a point in urban space. This approach enables the efficient computation of 3Ds and accessibility metrics for use as righthand-side variables in urban models. Existing GIS technology is not flexible or efficient enough to integrate directly into UrbanSim, R, and other statistical packages, therefore a generic accessibility library was coded in the C++ language in order to create and test new variables. The most common analysis using this framework is the ‘buffer query,’ which takes four parameters:

- the network to use
- the ‘range’ to use – this is the maximum distance to nodes that is to be included in this computation
- the ‘aggregation’ type to use – this is typically sum, but can be average, standard deviation, min, max, etc
- the ‘decay’ to use – this is consistent with the gravity model in that items further away affect the point of interest less than objects nearby. Decays can be linear, exponential, or flat (which applies no decay).

The summation for a buffer query takes the same basic form as the gravity model, but is implemented at a local street node level. For the object x being aggregated within a set of nodes R with a range defined by a maximum impedance, the buffer query (B) is defined for each location:

$$B = agg(x * decay(t)) \; \forall \; node \, \epsilon \, R \tag{2.1}$$

where t is the generalized impedance (typically time or distance) to each node in the summation. In the common case, the aggregation is simply a sum and the decay is linear which applies a coefficient of 1.0 at impedance 0 and 0.0 at the maximum distance ($maxt$):

$$B = \sum (x * \frac{maxt - t}{maxt}) \; \forall \; node \, \epsilon \, R \tag{2.2}$$

‘Density’ and ‘Destination’ variables can both be computed using the above equation. ‘Design’ variables typically include the number of intersections, block length, the connectivity to the regional network, and the percent of four way intersections [Southworth and Owens, 1993], which are all traits of the network and can be computed with special functions built into this framework. ‘Diversity’ variables require two or more density variables and are calculated using Shannon Entropy [Shannon, 1948]:

$$D = P(x) \log \frac{1}{P(x)} \; \forall \, x \, \varepsilon \, X \tag{2.3}$$

where D represents diversity and X is the set of variables for which to compute the entropy. For instance, jobs-housing balance is computed by performing a buffer query on jobs and one on housing and using the two values to compute an entropy. The same can be done for mixing of land uses and other possible entropy measures.

2.5.2.2 Accessibility in the context of spatial analysis

Table 2.4 describes the possible metrics that can be computed using the network-based framework described here. Computations range from simple summary statistics, of which cumulative opportunity measures are a special case that use a summation aggregation. Logsum measures are computed within the same framework, but combine multiple input variables with coefficients that are estimated using discrete choice models. Disaggregate measures like those used to compute WalkScore measure the nearest or Nth nearest point location of a category in a point-of-interest dataset.

Metrics describing the graph of streets have also become popular recently (e.g. Sevtsuk, 2010). For instance, "centrality" measures the degree to which nodes are included in the shortest path between other nodes and can indicate high-value locations for land uses like retail. Finally, 'mixing' measures like jobs-housing balance are described as 'diversity' measures in the original 3Ds. All of these measures fit within the same framework of network-based aggregations, which are in contrast to traditional GIS techniques that use aggregations within polygons (TAZs, census geographies, cities, and states) and point-based statistics or point processes [Snyder, 1975, Daley and Vere-Jones, 2007].

Logsums warrant further explanation as they are the only statistically estimated measure of accessibility presented here. The definition and derivation of logsums is left to the methodology section of Chapter 4, though several high-quality discussions exist in the literature already [Ben-Akiva and Lerman, 1985, Dong et al., 2006]. As used in this chapter, logsums are a statistically rigorous measure of accessibility to all the destinations available to a decision maker. Where cumulative opportunity measures simply count the amount of an object within a radius, logsums estimate the willingness of a certain kind of person to travel a certain distance to a destination with certain traits, as observed in an estimation dataset (e.g. the travel survey described above).

The logsums built as part of Chapter 4 use a nested construction that places destination below mode so that the logsum of each lower nest is an aggregate measure of accessibility for a given mode. By definition, normalizing logsums across modes gives a prediction of the probability of choosing each mode, thus this particular kind of logsum is an absolute measure of the accessibility for each mode and can be used in the estimation of other urban models. Logsums are not used in the model in this chapter, but is critical to Chapters 4 and 5.

Type	Description	Parameters	Notes
Summary Statistic	Mean, Median, Std Dev, Max, Min	Radius	Ranges can overlap
Cumulative Opportunity	Special case of the above which performs a sum	Radius, Decay	Ranges can overlap, similar to smoothing function
Logsum	Coefficients on measures of multiple land uses are included	Radius, Estimated Coefficients	The same as above but with estimated coefficients of variables including distance
Disaggregate	Distance to closest point-of-interest rather than sum	Which point-of-interest to use, decay function	Can be combined into index like WalkScore
Graph Metrics	Often used as ‘design’ variables, but are really measures of graph density, connectedness, etc	Custom computation	Also related to Space Syntax [Hillier and Tzortzi, 1976] and centrality [Sevtsuk, 2010]
Mixing	Diversity measures like entropy combine multiple measures of the above	Custom computation	Jobs-housing balance could be used as two summary stats, but a mixing function like entropy would be large even when jobs and housing are both small

Table 2.4: The types of aggregations that can be performed using the accessibility framework on the above land use datasets

2.5.2.3 Statistical Methodology

The framework presented in the previous section is a methodology for computing urban accessibility variables, and these variables can be mapped directly to aid in visual understanding of data, or can be used as predictors in a variety of statistical models. Typical methods used in urban modeling include linear regressions (e.g. hedonic models), poisson and negative binomial regressions (e.g. trip generation), and discrete choice (e.g. mode choice or destination/location choice). Future applications of this framework could use these variables in basic machine learning algorithms including clustering, classification, neural networks, and recommender systems. Generally, machine learning algorithms focus on predictive power rather than interpretability of coefficients and causal inference which makes conventional statistical techniques more common in planning and related social sciences.

A case study for the application and testing of the accessibility framework described above is created by regressing the number of home-based non-work walking trips from CHTS 2012 on various aggregations of land use for different modes of transportation. Non-work trips are used from the home location as these are frequently short discretionary trips that are likely to have the walk mode substitute for the automobile mode as built environment density increases. Trip generation could be performed for each trip purpose independently, but as the vast majority of households make no walking trips, trip purposes are merged to gain greater statistical significance for coefficients. Regressing destination choice for specific trip purposes on attributes of destinations is performed in Chapter 4.

Ordinary least squares regression (OLS) is suitable for normally distributed continuous outcome variables, of which trip generation is neither. As such, trip generation is frequently estimated using “count models,” which are based on the poisson and negative binomial distributions, and the number of trips is modeled using explanatory variables that include demographics of the person making the trips as well as measures of the built environment. This work closely follows Ma and Goulias [1999] and uses a Poisson regression to predict the number of trips made, and several other studies have used similar models [Wallace et al., 1999, Jang, 2005, Wootton and Pick, 1967].

The Poisson model, in which the number of trips is generated from a Poisson process, is formulated as

$$P(y_i = j \mid X_i) = \frac{e^{-\tau_i \lambda_i}(T_i \lambda_i)^{y_i}}{y_i!} \tag{2.4}$$

$$log\lambda_i = log\lambda(X_i\beta) = BX_i \tag{2.5}$$

where j is the number of activities, X_i is the set of observed variables for individual i, β is the set of coefficients for the variables X_i, and λ_i is the rate of occurrence of a trip per unit time, here set to a single day.

The model is estimated using maximum likelihood with the Python statsmodels library, and simulated using the same. Negative binomial count models are also tested, but

Variable	**Coefficient**	**Z-score**	**Category**
Residential Units	0.03	2.9	Built Env
Walkscore	0.37	12.2	Built Env
Median Year Built	-0.12	-2.8	Built Env
Percent of 4-way intersection	0.05	3.6	Street Design
Ratio of Cars to Population	-0.13	-2.8	Demo/Built
Median Income	0.27	7.9	Demo
% Low Inc HH of Total HH	0.15	4.4	Demo
Ratio of Workers to Population	0.16	4.1	Demo
Jobs by Transit within 45 mins	0.08	4.4	Regional
Older	0.40	3.2	Individual
Employed	-0.32	-9.1	Individual
Flex Hours	0.21	3.8	Individual
Constant	-1.90	62.0	Constant
Null Log-likelihood	-11158		
Log-likelihood	-10531		
Pseudo R-squared	.056		

Table 2.5: Coefficients for the preferred model of pedestrian trip generation around the home location using the Bay Area portion of the 2012 CHTS

Poisson is chosen for better goodness of fit. Explanatory variables include demographic variables that exist in the estimation dataset (CHTS 2012) such as gender, age, income, employment status, flexible work hours, and presence of children in the household. Numerous measures of the built environment and demographics in the local neighborhood are used as explanatory variables in the model, as well as regional accessibility measures by transit and by auto to capture accessibility at a wider scale. A thorough set of variables that could be theorized to describe urban neighborhoods is provided in Appendix A to this dissertation, with a marker for those variables which have been tested in this model estimation (many variables are not available due to data limitations).

2.6 Results

The preferred model coefficients and z-scores are provided in Table 2.5. The model includes all the variables that are significant among the variables that have been tested, and a list of variables that are theorized as descriptive of neighborhood demographics and built environment characteristics is provided as Appendix A to this dissertation. In the model shown here, residential unit density and WalkScore are both significant and positively related to walking trip generation around the home location, with WalkScore having both a much larger coefficient and much higher significance (independent variables are divided by their standard deviation in order to enable rough comparison of coefficients). Year built is inversely related to walking (i.e. newer neighborhoods have less walking), which likely is related to the historic older neighborhoods in highly walkable San Francisco.

The only design variable that is significant after controlling for the other variables shown is the percent of 4-way intersections within a 500 meter radius, which has been theorized as indicative of highly-connected street networks, and might also be related to the gridded networks in highly walkable San Francisco. Other design variables tested include the number of intersections and street length within a radii, average edge curvature (ratio of geometric distance divided by airline distance), number of cul-de-sacs, and the number of connection points to the larger street network at the .5km boundary, none of which are significant when controlling for the variables shown here.

The ratio of cars to population has the expected negative relationship to walking. Median income in the neighborhood has a positive relationship relationship to walking, likely indicating safe and pleasant neighborhoods, but the percent of low income households also has a positive relationship to walking, probably indicating increased likelihood of walking by lower income people. Although employment status at the individual level is negatively related to walking, likely due to time constraints, the ratio of employed persons to total persons in the neighborhood has a positive relationship to walking, which is likely indicative of the correlation between higher income neighborhoods and safe and pleasant pedestrian environments.

The only regional accessibility variable that is significant when controlling for the other variables included in this model is accessibility to jobs (for all sectors) by transit within 30 and 45 minutes. Here the 45 minute radius is used for its slightly higher explanatory power. Accessibility to jobs by automobile at radii of 15, 30, and 45 minutes and to a number of individual employment sectors are tested for inclusion but none are significant.

Individual demographics are also included, with older people (age > 50) tending to walk more, employed people walking less, and people with flexible work schedules walking more. Income is tested for inclusion in the model and is not found to be significant. This somewhat surprising result is robust over many model specifications and likely shows that walking in the San Francisco area is somewhat common across all income classes.

2.7 Discussion

The model presented in this chapter is intended as a case study application of this framework, but it does provide some insight into several possible correlates of walking as represented in the Bay Area region of the new California Household Travel Survey. Of note is the statistical significance of Walkscore in the model, which confirms recent findings as to WalkScore's positive correlation with walking behavior [Weinberger and Sweet, 2012, Manaugh and El-Geneidy, 2011]. Although WalkScore has the higher coefficient and much higher significance, nonetheless residential density maintains a separate positive and significant impact, despite the argument that density is not a causal factor in walking [Crane, 2000, Boarnet and Crane, 2001]. Note that Chatman [2008] discusses the issue and proposes that density might be an easily measured proxy for walkable environments. Other measures of destination accessibility were tested in the model - e.g. access to retail jobs within 500 meters - but none were as significant as WalkScore.

On the other hand, many of the street design variables that have been discussed as being correlates of walking [Ewing and Cervero, 2001, Southworth and Owens, 1993] are not found to be statistically significant in this model. The only street design variable which is significant is percent of 4-way intersections, which is likely related to the gridded historic street networks in highly walkable San Francisco. Although street design is likely to be an important factor in walking - see, for instance, the renovation and widening of the Valencia Street corridor and resulting increase in pedestrian activity - it is likely that currently available data is insufficient to capture properly these design characteristics.

Finally, it is interesting to note that very few metrics of regional accessibility have a significant relationship with home-based walking behavior, when controlling for the other variables included here. The only significant regional accessibility variable among those tested as part of this research is access to all jobs within 45 minutes via the the transit network. Other variables tested include: access to all jobs, retail jobs, information sectors jobs, and FIRE jobs (finance, insurance, and real estate) within 15, 30, and 45 minutes drive. There appears to be little indication in this travel survey that increased accessibility at the regional scale by automobile has any negative impact on walking [Handy, 1996a, Crane, 1996], which is contrary to the finding from Guo et al. [2007].

The statistical significance of regional access by transit does indicate that having access to destinations via the transit network has a positive and synergistic effect on waking, which is a promising area of future research. While previous work in this area typically measures transit access as the distance to the nearest transit stop, this work makes a clear contribution by measuring transit access to destinations along the transit network, considering the speed of the vehicle, access and egress times, and transfer time. When accounting for these additional factors, increased regional access by transit has a positive impact on the generation of walking trips around the home. This model formulation also allows the differentiation of regional accessibility by transit and by automobile: when controlling for regional access by the transit system, regional access by the automobile network is not significant in predicting home-based walking trips.

2.8 Conclusion

This chapter has presented a novel new compromise between speed, flexibility, and understanding in modeling preferences for urban travel. It proposes a concept of land uses positioned relative to a hierarchical multi-modal graph of transportation options in the city that allows the computation of accessibility variables by mode to any other destinations in the city. These measures can then be used as independent variables that characterize the quality of neighborhoods in models of urban behavior, such as price hedonic models, location choice models, and travel demand models. The use of street node geography is the enabling abstraction that allows land uses to be quickly mapped to their nearest location on each of the mode-dependent networks and then acted upon in a parameterized aggregation as described in the methodology section.

Numerous data sources have been unified in order to perform the research presented here, including OpenStreetMap to represent the local streets, the regional travel model for capturing congested automobile travel times, and the General Transit Feed Specification for representing transit vehicle schedules. Land use data includes a synthesized population, datasets of businesses and points-of-interest, as well as governmental datasets of buildings and parcels in the region. The unification of multiple datasets allows the inclusion of a large number of variables computed to represent different aspects of the character of a neighborhood's built environment and demographic makeup.

As a case study application of this framework, a model of pedestrian trip generation using the Bay Area portion of the new CHTS 2012 survey is developed. The coefficient on WalkScore is both large and highly significant, showing that destination accessibility is very important in understanding the generation of walking trips, but residential unit density maintains a separate small but significant coefficient, perhaps serving as a proxy to an attractive pedestrian built environment. Interestingly, the only regional variable significant in the model is accessibility to jobs via the transit network, which shows both the lack of negative impact of increased auto accessibility on walking as well as demonstrates the strong positive and synergistic relationship of having a strong regional transit network on walking trips.

Chapter 3

Measuring the Impact of Accessibility to Local Amenities on Home Values

3.1 Introduction

Hedonic models have long been the primary method used to identify the contributions of individual aspects of a property to its sales price or rent, but little research has been performed to date that uses individual neighborhood amenities as explanatory variables in such models to understand their contribution to home values. High quality point-of-interest (POI) datasets (datasets that identify destinations by category like restaurants, coffee shops, dry cleaners, city parks, etc) are now widely available and are used, for instance, in the WalkScore algorithm to measure walkability [WalkScore, 2011] and recent work has found that the WalkScore index is correlated with increased home values [Cortright, 2009].

This is taken as evidence that walkability is valued as a positive trait when purchasing a house, but this relationship is not necessarily causal. More thorough research must be done in order to provide convincing evidence that people are actually walking to these destinations, and that a preference for the choice to walk is being capitalized into increased home values for homes than enable more of this behavior. This chapter explores the relationship between local amenities and home values more systematically by testing the impact of accessibility to individual amenities on home values, and Chapter 4 addresses the question of whether a preference for walking is being capitalized into home values.

Typical hedonic models [Kain and Quigley, 1970, Rosen, 1974, Kain and Quigley, 1975] include variables that are theoretically related to home values but which are often limited by data availability. Common variables to include are traits of the unit like square footage, lot size, number of bedrooms and bathrooms, and view, as well as traits of the neighborhood, like school districts, crime, architectural quality, and access to amenities via the multi-modal transportation network. In practice, one of the most predictive independent variables for home values is average income in the neighborhood [Ramírez et al., 2008], but this is a fairly tautological relationship. It does not answer the question of how this neighborhood became preferred by high-income households originally, which resulted in increasing home values as the residents' purchasing power increased.

This research proposes and tests the theory that some high income households choose to live in high income neighborhoods not strictly due to a preference to be around people like them, known as homophily [Lazarsfeld and Merton, 1954], but because of a preference to have increased access to amenities they and their neighbors support by patronizing together. This shared “amenity infrastructure” is easily testable with a point-of-interest dataset using neighborhood income as a control variable. If the theory is true, the predictive power of neighborhood income should decline as amenities are included in the hedonic model.

This relationship is also bi-causal as high price amenities do seek to locate near high-income households, which is the primary motivating principle behind agent-based urban models like UrbanSim [Waddell, 2002b], which model both residential and commercial agents. Future work will explore this relationship from the perspective of shops and restaurants seeking to locate near their clientele.

3.2 Additional Literature

3.2.1 Hedonic Models

Core theoretical building blocks for this research include bid-rent theory, put forward in the early development of urban economics as a field [Alonso, 1964, Muth, 1969], and hedonic regression [Boyle and Kiel, 2001, Cheshire and Sheppard, 1995, 1998], a methodology to estimate the implicit prices of amenities in bundled goods, such as housing [Rosen, 1974]. Combining these two building blocks, extensive research has been performed to analyze how locational amenities such as accessibility are capitalized into residential property values [Nelson, 1977, Edmonds Jr, 1983, Waddell et al., 1993, Waddell and Nourzad, 2002, Lee et al., 2010a].

The logic behind this theoretical approach is straightforward: agents that value specific amenities such as travel time savings will bid more in terms of rent or purchase price at those locations that have higher values of such amenities, and in so doing, they are more likely to outbid other agents for the right to occupy those sites. A further consequence of this logic is that higher competition for advantageous sites results in higher land values and subsequently translates to a higher development intensity on such sites, as a result of substitution from increasingly expensive land costs to relatively less expensive capital costs – in the form of taller buildings – through capital-land substitution.

3.2.2 Commercial Land Uses and Home Values

Bartholomew and Ewing [2011] review the literature on the impact of pedestrian- and transit-oriented development on home prices, presenting results from numerous studies on the impact of transit investments and urban design on home values, but with few recent studies of the impact of mixing of land uses and local amenities on the same. A notable exception is Matthews Matthews and Turnbull [2007], which presents conflicting results, depending on the auto-oriented or pedestrian-oriented nature of the neighborhood (i.e. local amenities only increase value in pedestrian-oriented neighborhoods), a result also seen in Rauterkus and Miller [2011] when relating WalkScores to home values. Song and Knaap [2004] investigated the topic using a Portland dataset and found less equivocal results. They state, “our fundamental conclusion is that mixing certain types of land uses with single family residential housing has the effect of increasing residential property values,” although they do note that larger commercial areas can have a negative impact on home values.

Despite the lack of empirical studies, the last decade has produced numerous theories on the positive impact of local amenities on inter- and intra-regional location choices. Different authors give this theory different names, of which The Consumer City [Glaeser et al., 2001, Gottlieb and Glaeser, 2006] and The City as an Entertainment Machine [Clark, 2003] are two of the most common. In this theory, scarce and high-value goods like fine restaurants, architectural aesthetics, and natural beauty are seen as drivers of economic growth, as high-income (and presumably high-productivity) workers migrate to regions and neighborhoods replete with these amenities. There are a number of studies which relate these traits to various measures of economic output, but few that test to see if intra-regional variation in these amenities is related to the variation in residential home prices.

The methodology proposed in this work is also supported by the theory of "scenes" proposed by Terry Nichols Clark, in which available amenities help define the character of a neighborhood [Silver et al., 2010, 2011]. "These constellations of amenities define the scene by making available an array of meaningful experiences to residents and visitors" [Silver et al., 2011]. In their work, Silver and Clark define clusters of amenities into different axes (e.g. Transgression) and these get clustered into a typology of neighborhoods (e.g. Bohemian). The approach used here is to test the hypothesis that if such amenities are valued by citizens of the Consumer City, a residential hedonic model should reflect this with significant positive coefficients on accessibility to amenities used as independent variables.

3.3 Research Objectives

This work seeks to contribute to the following research questions:

- Does WalkScore have a positive correlation with home values in this dataset? As WalkScore is a weighted combination of access to 9 different amenity categories, is access to some of the categories more correlated with home values than others? Are there categories of amenities that are not included in WalkScore that are also strongly correlated with home values?

- Does adding accessibility to nearby amenities to an hedonic model that includes the average neighborhood income reduce the significance and size of the coefficient on the latter, supporting the theory that people sort into neighborhoods of similar incomes partially to access a shared set of amenities rather than simply to be near households of similar demographic makeup?

3.4 Data and Methodology

3.4.1 Data

This research relies on a large number of residential home listings from 2012 in the San Francisco Bay Area with a limited number of associated attributes. The particular estimation dataset used contains 209,075 listings with a mean value of 306 dollars per square foot and a median of 266 dollars per square foot. Throughout this research prices per square foot is used rather than absolute prices so as to remove rebound effects in which households consume more housing when home prices are cheaper, and given that

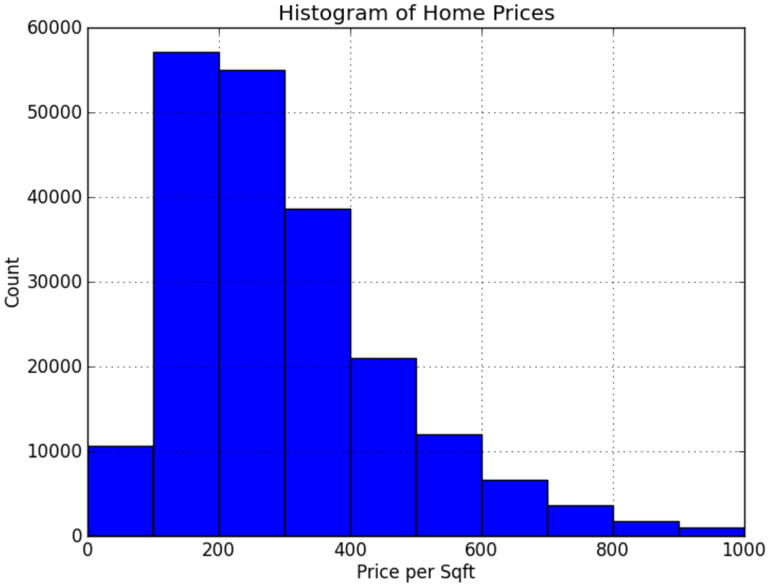


Figure 3.1: Histogram of home prices per sqft in the residential sales listings

price per sqft is the metric most consistent with bid-rent curves described in the literature review. The histogram of home values from the dataset is shown in Figure 3.1, which shows the large positive skew in the data: there is a long tail of high priced homes.

The San Francisco Bay Area is one of the most economically rich and scenically beautiful locations in the United States and the observed prices reflect this status. During the time period this data was collected - spring of 2012 - the median home price in the United States was roughly 160,000 while the median in the Bay Area is fully 66% higher than the national average at 275,000. Although home prices throughout the Bay Area remain high, San Francisco is an especially expensive real estate market with a median almost four times the national average.

3.4.2 Hedonic Models

Real estate prices are modeled using hedonic regression of the log-transformed property value per square foot on attributes of the parcel and the neighbood surrounding each parcel. The hedonic regression equation encapsulates interactions between market demand and supply, revealing an envelope of implicit valuations for location and structural characteristics [DiPasquale and Wheaton, 1996]. The model was estimated from residential listings with Ordinary Least Squares (OLS), using a standard semi-log specification in which the dependent variable is log-transformed:

$$log(P_i) = \alpha_i + \beta(X_i) + \epsilon_i \tag{3.1}$$

where i indexes observations in the estimation dataset described above, and ϵ is the error term. All attributes of the residential properties are used in the estimation to attempt to control for building quality variation, and the variables include: *historic* buildings (pre-1940), *modern* buildings (post-2000), natural log of *unit size* in square feet, and natural log of *lot size* in square feet. Two characteristics of the neighborhood are used as control variables and these are *regional accessibility* which sums the number of jobs accessible by automobile within 30 minutes and the *average income* of households in the local area, defined in this case as households within half a kilometer. All distances are network distances and are computed using the framework described in Chapter 2. Note that because the dependent variable is the price per square foot rather than an absolute price, the coefficient on any transformation of the *unit size* variable is expected to be negative in the hedonic model due to decreasing returns of additional square footage to homebuyers. Ceteris paribus, larger units have a smaller price per sqft.

3.4.3 Point-of-interest Variables

Point-of-interest datasets are now nearly ubiquitous for use in internet applications, and give latitude-longitude locations for categorized amenities, which include businesses like restaurants, cafes, groceries, and shopping, as well as personal services like medical centers, law firms, and accountants, public institutions like city parks and government buildings, and natural amenities like historic and tourist locales. The quality of these datasets is also very high compared to typical urban datasets like population datasets synthesized from aggregate census geographies and datasets of businesses which often fail to separate interrelationships between different branches of a business. A simple point-of-interest dataset has value as it contains the “public facing” aspect of a business, and thus would be the destination for individual travel and the spatial location which would have value to a home buyer for its proximity to the property. In the simplest terms, these datasets can be thought of as an online yellow pages, and are used as such by many websites.

The application of these data in this chapter is to characterize neighborhoods by the POIs located nearby to every property listing in the estimation dataset. The POI dataset used in this research was provided by Factual, Inc. and contains 370 different categories of destinations, and these categories are used without modification as independent variables in the hedonic model of prices. Although POI datasets have excellent geographic coverage, one shortcoming of this dataset is that there are no additional descriptive columns in the data associated with a destination. Thus the flawed assumption that each destination is equivalent to all others in a given category must be made. Future research can correct this assumption by including quality data as it becomes available. Most disaggregate indices, like WalkScore, use a form of destination accessibility that gives the distance to the nearest (or second or third nearest, etc) destination in a given category directly. This work uses a cumulative opportunity or gravity measure to sum the available destinations along the local street network within a parameterized distance. The equation for cumulative opportunity [Dong et al., 2006] is :

$$A_i(R) = \sum_j a_j W_j \tag{3.2}$$

where a_j is a continuous variable describing an attribute of the environment at each location in the city (in this research the variable is a binary 0/1 indicating the presence of a destination at a given locale), and W_j is equal to 1.0 where the local street network distance is less then the parameterized distance r and 0.0 otherwise. The generalized gravity model can be expressed as:

$$A_i(R) = \sum_j a_j f(r_{ij}) \tag{3.3}$$

The function $f(r_{ij})$ is called the *decay* and is also parameterized. The decay used everywhere in this work that gravity measures are discussed is linear for ease of interpretation. Thus $f(r_{ij}) = 1 - r_{ij}/R$ so at $r_{ij} = 0$, W_j equals 1.0 and where $r_{ij} \geqslant R$, W_j equals 0.0. More simply, the impact of the destination on the measure is reduced according to the proportion of distance traversed from the location of measurement to the user-defined *horizon*. Although these variables can be expressed mathematically, they are more intuitively understood when mapped, and this research uses a new software platform to enable detailed mapping at the parcel-scale to provide intuitive visual understanding of these variables.

The POI variables included fall into three categories according to how they are theoretically expected to impact home values. The variables and categories are shown in Table 3.1. The *common* category contains items that are frequent walking destinations, and most are used in the WalkScore index. The *high-income* category contains destinations that might be expected to indicate amenities valued by high-income households [Silver et al., 2011], and locally undesirable land uses (LULUs) [Been, 1993] which are expected to indicate depressed neighborhoods or reduced home values.

3.5 Results

Table 3.2 shows a chloropleth map of destination accessibility for all of the destination categories used in this analysis (Figure 3.2 shows the region represented in the images without coloring). These maps show only a portion of the Bay Area including most of San Francisco, Oakland, and Berkeley, but each variable is computed for the entirety of the Bay Area when used in the hedonic models. The maps can be interpreted as heat maps, with darker colors indicating more destinations of that type nearby (all maps are represented with an equal interval legend so colors and values are evenly spaced from min to max). The maps are also color-coded such that green represents common destinations, purple represents high-income destinations, and blue represents the LULU category.

Name	Category	Count	Sign/Magnitude	WalkScore
Restaurants	Common	36,441	+, large	included
Shopping	Common	85,631	+, large	included
Groceries	Common	5,618	+, large	included
Cafes	Common	3,918	+, large	included
Bookstores	Common	1,702	+, small	included
Entertainment Venues	Common	12,415	+, small	included
Parks, Outdoor Sites	Common	3,555	+, small	included
Florists	High-income	2,559	+, small	n/a
Yoga	High-income	395	+, small	n/a
Sushi	High-income	330	+, small	n/a
Farmer's Markets	High-income	41	+, small	n/a
Health Spas	High-income	11,742	+, small	n/a
Tourist Sites	High-income	164	+, small	n/a
Historic Sites	High-income	44	+, small	n/a
Bars	LULU	2,235	-, small	included
Fast Food	LULU	3,409	-, large	n/a
Tattoo Parlors	LULU	396	-, large	n/a
Liquor Stores	LULU	4,122	-, large	n/a
Pawn Shops	LULU	112	-, large	n/a
Concerts	LULU	1,031	-,large	n/a

Table 3.1: Categories of destinations to be used in home price hedonic models with expected signs and magnitudes

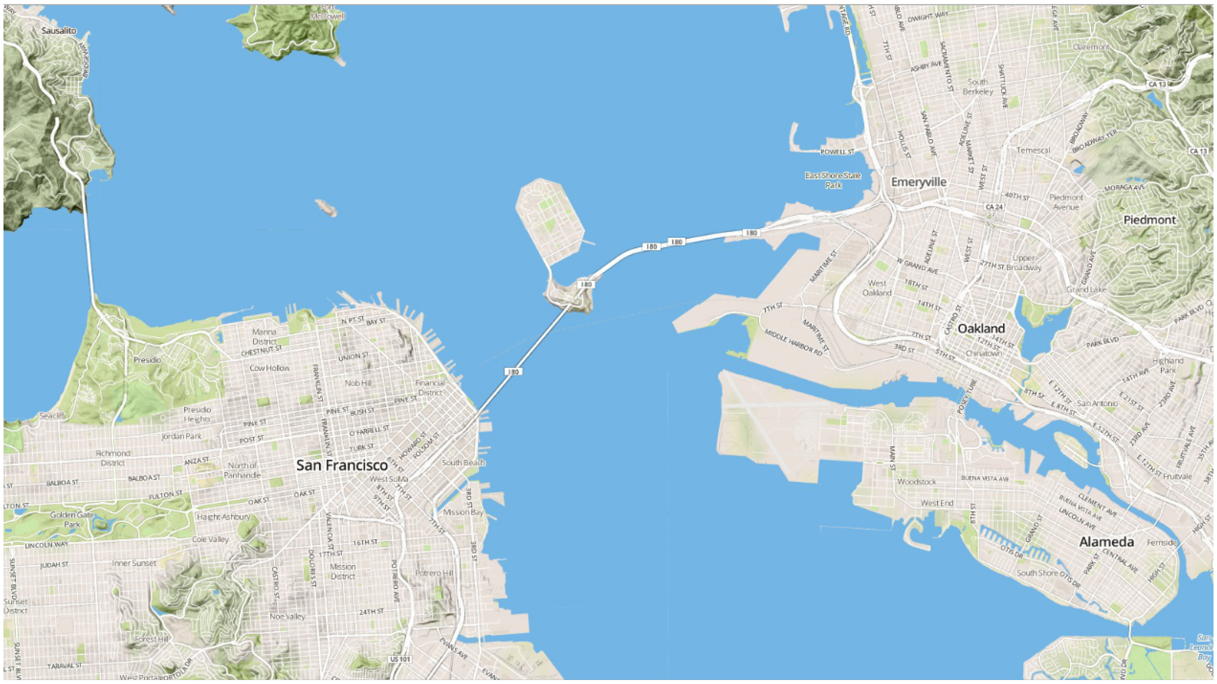
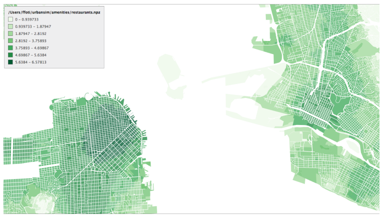


Figure 3.2: Map of the Bay Area region shown in the cloropleth maps

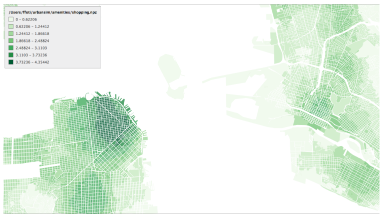
There is significant overlap in high accessibility locations among amenities in the *common* (green) category, with downtown San Francisco being the most common location for almost all of the destination categories. Berkeley has high accessibility to cafes and Downtown Oakland's Lake Merritt area has high accessibility to outdoor space and parks. The *high-income* (purple) category has much greater spatial variation, with yoga locations concentrated away from the region's CBDs mostly in the affluent neighborhoods. Sushi is prevalent in the Chinatown and Richmond neighborhoods, as both neighborhoods have high East Asian populations. Farmer's Markets occur at Fisherman's Wharf and West Berkeley, and tourist sites occur mostly in The Mission neighborhood of San Francisco. In the *LULU* category, bars occur more frequently downtown and in the Mission, Fast Food near Market Street, Tattoo Parlors on Haight Street, and Pawn Shops in The Tenderloin. Thus there are subtle but important differences in access to each category of destination that provides some evidence in support of the theory of *scenes* [Silver et al., 2010]. In short, concentrations of destinations appear to at least partially explain one's experience of the city as a diverse urban fabric.

Table 3.2: Maps of accessibility to destinations by category. Stronger colors indicate more nearby destinations

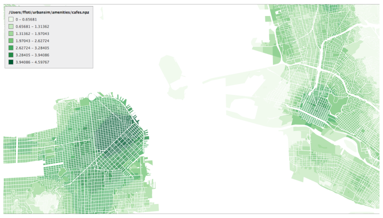
Restaurants



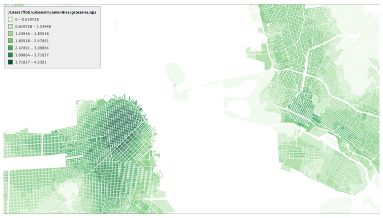
Shopping



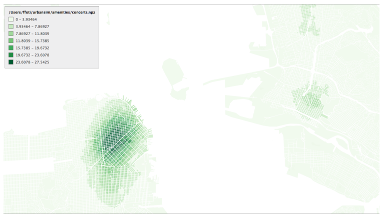
Cafes



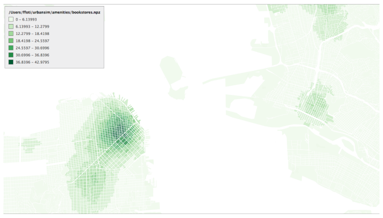
Groceries



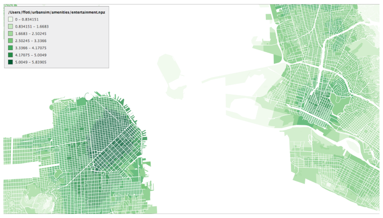
Concert Halls



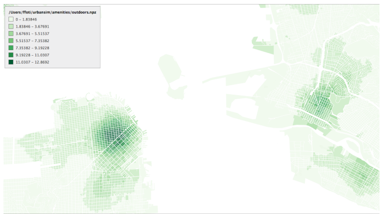
Bookstores



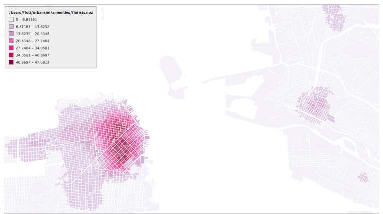
Entertainment



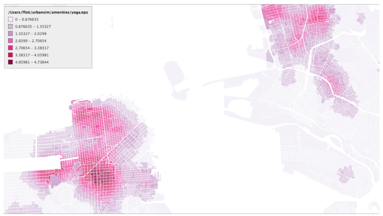
Recreation



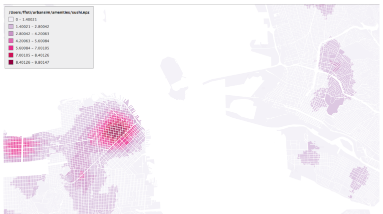
Florists



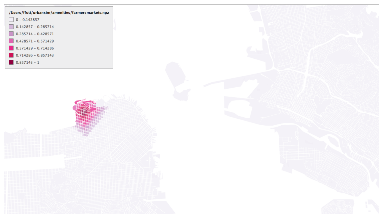
Yoga



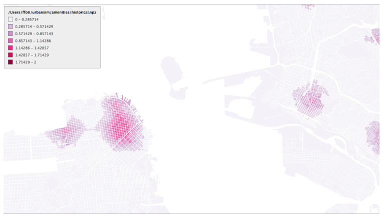
Sushi



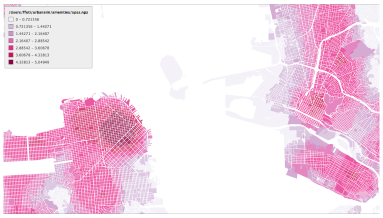
Farmer's Markets



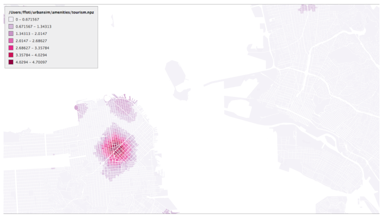
Historic Sites



Health Spas



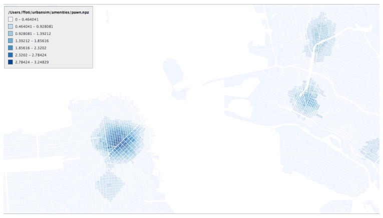
Tourist Sites

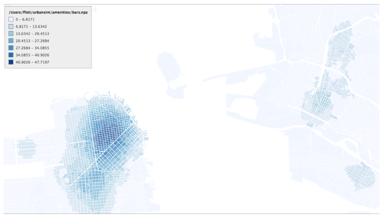


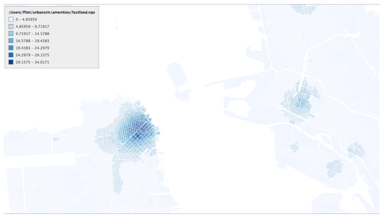
Pawn Shops

Bars

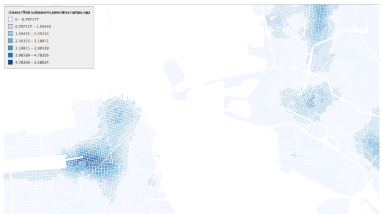
Fast Food



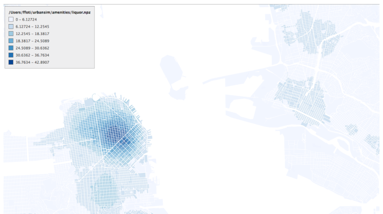




Tattoo Parlors



Liquor Stores



Once variables are generated for all of the possible destination categories, they can be used as independent variables in an hedonic model of sales price per square foot. Figure 3.3 shows a map of *average* rent in San Francisco for the estimation dataset, but *individual* sales prices (and characteristics of individual properties) were used in the model estimation.

The three estimated models are shown in Table 3.3. Model 1 is the traditional residential home price hedonic model, which uses all unit-level attributes that are available in the dataset, as well as a regional accessibility metric and the average income within the neighborhood. There are positive coefficients on historic and modern properties, and on regional accessibility and average income in the neighborhood. There is a negative coefficient on unit square footage, indicating that larger units are cheaper per sqft, which is a pattern that holds in Models 2 and 3. There is a small negative coefficient on lot size which does not conform to theory and which doesn't occur in Models 2 and 3. The r-squared for Model 1 is the smallest among the three models. Coefficients are roughly comparable in magnitude since they are divided by the standard deviation of the variable, and the coefficient on average income is the largest which is typical of residential price hedonic models.

Model 2 tests the amenity value of those destinations which are components of WalkScore in addition to the variables used in Model 1 (which have similar signs and magnitudes in both models). Positive coefficients occur for cafes, entertainment, and bookstores, while negative values occur for groceries, shopping, and restaurants. This is somewhat misleading as all WalkScore destinations are positively correlated with residential home values when used in isolation. There are correlations between different variables in this analysis due to the colocation of non-residential uses in commercial corridors and centers which cause some variables to have negative coefficients when controlling for others. Future work could use factor analysis to eliminate the correlations among related destinations, but this would not allow the disambiguation of accessibility impacts that is the purpose of this research. The r-squared for Model 2 is .504, which is a significant improvement over Model 1.

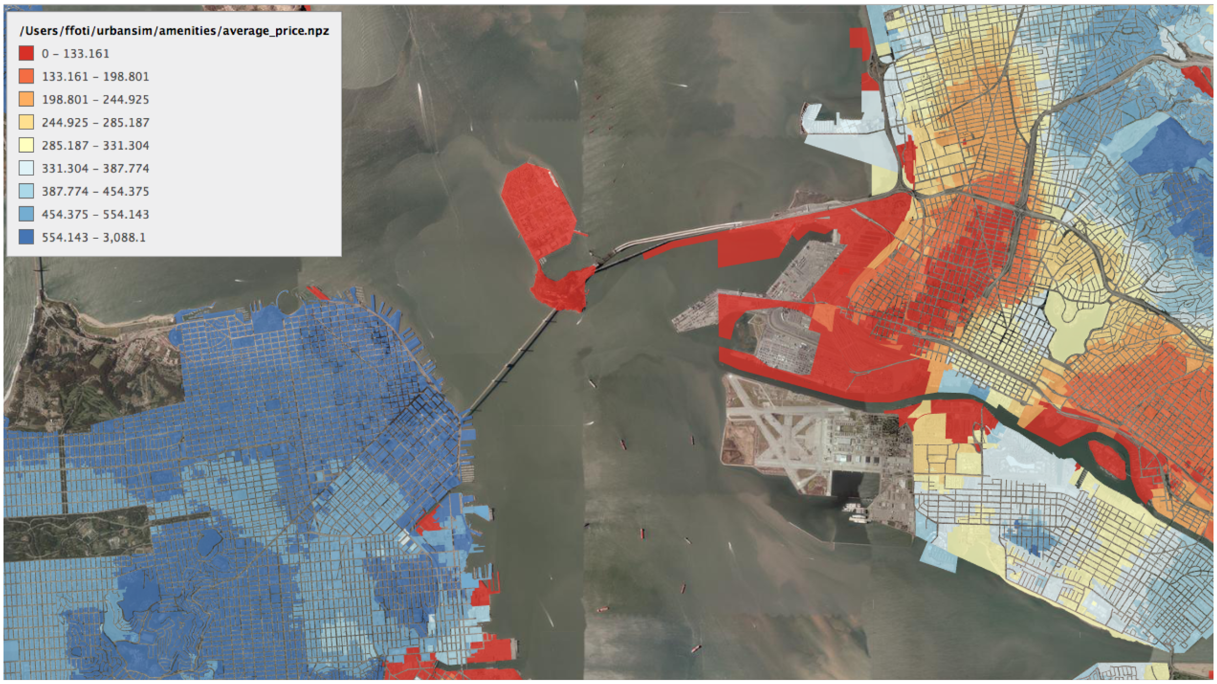


Figure 3.3: Average residential rent per square foot

Model 3 includes all the amenity variables used in this study, again using unit variables from Model 1 as controls, and all controls have similar magnitudes and signs as Model 1. Model 2 shows that high-income amenities are related to high-value home prices, and Model 3 extends this hypothesis by including additional high-income amenities like sushi, yoga, and florists - in addition to the typical WalkScore destinations, as well as a few LULUs like tattoo parlors and pawn shops. Accessibility to cafes is still the strongest predictor of the amenity accessibility variables, with large positive coefficient and statistical significance.

The additional amenities have theoretically expected signs, with sushi, yoga, florists, spas, and tourist destinations all having positive and significant coefficients, while fast food, tattoo parlors, pawn shops, concert venues, and bars have significant and negative coefficients. Somewhat counterintuitively, liquor stores have a small positive and significant coefficient, likely because many urban liquor stores double as groceries and provide a desired service to the surrounding neighborhood. Outdoor recreation is positive as hypothesized, but is not significant. Shopping, restaurants, and groceries all maintain the negative and significant coefficient present in Model 2. Model 3 has an r-squared of .52 which is a small but significant improvement over Model 2.

To increase the interpretability of coefficients, Table 3.5 shows the actual price impact (in dollars per square foot) of a one standard deviation change of each of the variables used in Model 3. The table is also sorted from largest negative impact to largest positive impact to aid legibility. In other words, an increase in local accessibility to fast food restaurants of

	MODEL 1		**MODEL2**		**MODEL3**	
	Traditional Vars		WalkScore Vars		All Vars	
R-squared	0.42		0.504		0.52	
Variables	Coefficient	T-score	Coefficient	T-score	Coefficient	T-score
Historic	0.26	82.59	0.09	27.7	0.05	16.5
Modern	0.18	36.45	0.17	37.4	0.18	39.6
Accessibility	0.16	140.46	.14	117.7	.14	122.4
Ave Income	1.03	294.03	1.03	289.1	.96	262.0
ln of Unit Sqft	-0.32	-120.08	-0.36	-142.0	-0.37	-146.6
ln of Lot Size	-0.01	-17.79	0.02	33.4	0.02	37.3
Constant	-5.55	-148.39	-5.2	-139.3	-4.5	-115.4
Restaurants			-0.10	-57.3	-0.08	-39.7
Groceries			-0.09	-36.5	-0.11	-40.9
Cafes			0.26	87.5	0.25	80.9
Shopping			-0.02	-7.56	-0.04	-14.6
Entertainment			0.13	60.01	0.11	46.6
Bookstores			.02	29.1	0.02	18.75
Sushi					0.05	21.2
Yoga					0.06	19.52
Florists					0.03	35.92
Spas					.013	6.2
Tourism					0.076	3.17
Recreation					.0013	.94
Concerts					-0.004	-2.43
Fast Food					-0.044	-43.2
Tattoo Parlors					-0.043	-15.8
Pawn Shops					-0.07	-11.36
Bars / Pubs					-0.02	-22.98
Liquor Stores					.013	18.28

Table 3.3: Comparison of three hedonic models using 1) traditional accessibility variables 2) destinations used in WalkScore 3) the complete set of destination categories

Variables	**Coefficient**	**Median**	**Sigma**	**Price Equiv**
Fast Food	-0.044	0.56	1.89	-14.62
Groceries	-0.11	0.58	0.75	-14.50
Bars / Pubs	-0.02	0.08	3.34	-11.69
Restaurants	-0.03	2.04	1.29	-6.56
Shopping	-.04	0.24	0.61	-3.90
Tattoo Parlors	-.043	0.00	0.47	-3.11
Pawn Shops	-0.07	0.00	0.20	-1.93
Concerts	0.004	0.00	1.62	-.50
Recreation	.0013	.36	.98	.99
Spas	.013	1.07	1.02	3.31
Tourism	0.076	0.00	0.20	3.69
Historic	0.05	0.00	0.33	3.94
Yoga	0.06	0.00	0.51	6.72
Sushi	0.05	0.00	0.64	7.00
Modern	0.18	0.00	0.20	7.80
ln of Lot Size	0.02	8.64	2.31	9.84
Bookstores	0.02	0.04	2.42	10.29
Florists	0.03	0.34	3.17	19.95
Entertainment	0.11	0.93	0.98	22.65
Cafes	0.25	0.44	0.73	39.28
Constant	1.00	6.86		
Sum at Median		5.26		
Median Price		191.74		

Table 3.4: The variables used in the model with the associated impact of an increase of one standard deviation of that variable on price per sqft

one standard deviation is associated with a $14.62 drop in residential sales price per square foot, while an increase in accessibility to cafes of one standard deviation is associated with a $39.28 increase in residential sales price per square foot. The median sales price in the dataset is $266, so these changes are significant, yielding a -5.5% and +15% change in price respectively (although most variables are within a range of -5% to +8%). It should be emphasized that because of the correlated nature of variables included, the dollar impact of cafes, for instance, is likely to be counterbalanced by a negative impact of restaurant accessibility since these variables are correlated. Again, signs and magnitudes coincide with theoretical expectations other than the negative impact of groceries, restaurants, and shopping. This counterintuitive effect is discussed further in the next section.

3.6 Discussion

The results presented in the last section provide reasonable evidence as to the value of individual amenities being capitalized into home sales prices. As the dataset contains a very large count of observations, a larger than normal set of explanatory variables can be investigated. In Model 3, accessibility to 22 different categories of amenities is tested, and all but one yields high statistical significance. Although most variables have theoretically expected signs, Model 2 shows that access to cafes is the variable most explanatory of higher home values, and when controlling for this and the other amenities in WalkScore, groceries, restaurants, and shopping have a counterintuitive negative sign. It should be noted that groceries only occur as a broad category in this dataset, with local "corner stores" occurring in the same category as large format warehouse groceries. Although different categories of restaurants and shopping are available, destinations are not identified as neighborhood or auto-oriented in format.

One interpretation of these results is that different aspects of WalkScore are more positively related to higher home values than others. In this dataset, high-income amenities like cafes, bookstores, and certain entertainment destinations are most correlated with increased nearby home values. Past research which relates the composite WalkScore index to residential home values is accurate but is somewhat misleading; some destinations in the index are more positively correlated with residential home values than others. It might provide more insight into the underlying behavior that drives home values to disambiguate the impact of different categories of destinations.

Additionally, future work should attempt to differentiate not only the category of amenity, but also the size of amenity, as some destinations are neighborhood scale and others are regional scale. It is possible that access to cafes is the destination with the strongest predictive power largely because it is an accurate indicator of neighborhood-scale urban form. This could be due to the fact that groceries, shopping, and to some degree restaurants are built with either a regional-scale or neighborhood-scale format, but there is no regional-scale format for cafes. It is possible that the regional-scale format for destinations might not increase nearby home values, and this aspect of the data might confound the results presented here.

3.6.1 The Substitution of Amenity Accessibility for Household Income Sorting

The final topic of investigation for this chapter is to analyze the degree to which accessibility to amenities explains home values in a way that substitutes for the explanatory power of household income sorting as proposed in the introduction and research questions. Figure 3.4 varies the radius of analysis (for the amenity variables only) from 0 meters (i.e. no amenity variables) to 20 kilometers. The y-axis contains both the r-squared outcome at each radius as well as the coefficient on average income at each radius. It is hypothesized that the r-squared metric should increase to a certain point and then lose explanatory power in a non-monocentric city (in a purely monocentric city the r-squared would increase indefinitely as the radius expands). Additionally, if access to amenities is a valid substitute for household income sorting in the prediction of increased home values, the coefficient on income should go down as the explanatory power of amenity access increases.

First, it is clear that access to amenities adds explanatory power to Model 1 in Table 3.3 which has an r-squared of .42; the highest r-squared with amenity accessibility variables added is .614, which is a significant improvement and is very high for an hedonic model where price per square foot is the dependent variable. Second, the r-squared goes up significantly and consistently as the radius is expanded up to 9 kilometers, at which point it does not change significantly out to a range of 20 kilometers. Third, over the same range, the coefficient on average income in the neighborhood around each house drops from 1.02 to a low of .64 at a radius of 17km.

Together these results paint a clear picture that access to amenities does explain some proportion of residential prices that in Model 1 is attributed to household income sorting, but with a large remainder that is not explained by the addition of amenity accessibility variables. Additionally, the radius of analysis which provides the largest explanatory power to the model is larger than is typically considered in the literature on the relationship of walkable amenities to home values. For this dataset, a radius of 9km to nearby amenities provides the highest explanatory power of residential prices, using similar destinations to those that are often used to justify a preference for walkability. Although walkability is not ruled out as a factor by these results, it is not clear that access at larger than walkable distances is controlled for accurately in previous studies.

3.7 Conclusion

The results presented in this chapter show that access to specific amenities can be capitalized into home values with both positive and negative impacts. This conclusion supports the hypothesis that WalkScore is correlated with higher home values, but shows that not all destinations in a composite measure like WalkScore are equally predictive of home values. Some amenities tested here, like cafes, florists, and entertainments venues are correlated with large increases in home value, while access to land uses like fast food, tattoo parlors, and pawn shops has a negative correlation with home values consistent with the concept of locally undesirable land uses.

More generally, the analysis in Figure 3.4 which relates an increasing radius of accessibility measures to increasing explanatory power of the hedonic model casts serious doubt on the idea that homebuyers are capitalizing walkability directly into home values. It is true that access to amenities at the local scale is correlated with increased home values, but it is also true that access to amenities at a distance larger than typical walking distances has even greater explanatory power of increased home values, thus the evidence that there is a specific contribution to home values from a preference for walkable neighborhoods is incomplete. Nonetheless, the optimal radius of 9km with a linear decay (with

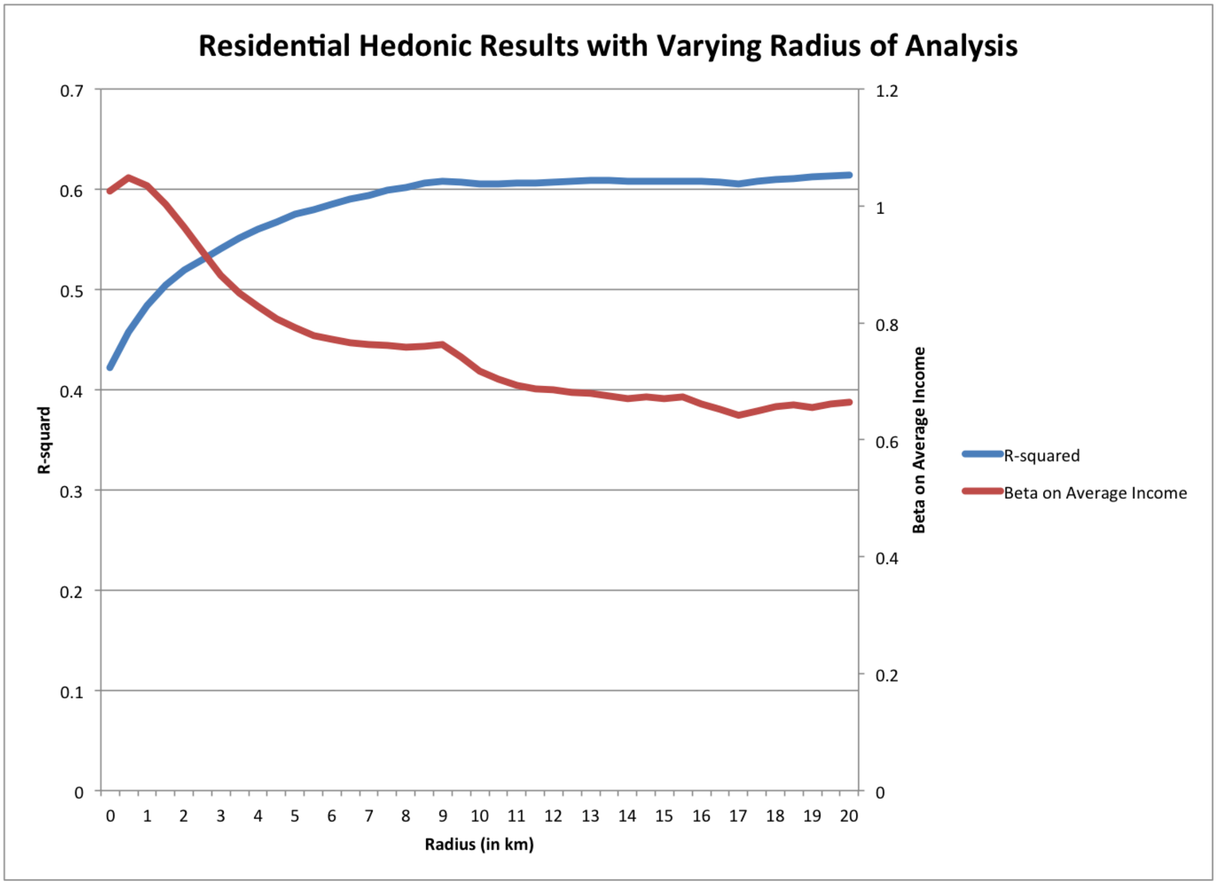


Figure 3.4: The increasing r-squared and decreasing coefficient on average income with increasing radius up to 9 km

destinations 4.5 km away already discounted 50%), shows a clear preference for dense living, and homes with amenities within a short drive or an easy bike ride are often sold at a price premium. The question of whether a preference for actual walking behavior is translated into increased home values is the topic of the next chapter.

It is also likely that different segments of the population of a city value access to amenities in different ways, for which this research does not account. For instance, it is possible there is a significant population which values walkable amenities and drives the demand for the housing market in San Francisco, and another segment of households which is willing to access regional amenities by transit and short automobile trips but does not require walkable access to amenities. Identifying the relative size and demographic makeup of these market segments is the topic of Chapter 5.

Finally, this work makes the assumption that preference for travel is directly related to distance - clearly the choice of some routes and destinations is affected by the safety and aesthetic quality of a route, and some boundaries are not present in network-based definitions of accessibility at all, like jurisdictional borders, school districts, physical boundaries like highways and railways, and neighborhood boundaries between adjacent social communities. A more general framework of accessibility is necessary to account for these factors, and the outline of a proposal for this framework is explored in the next chapter.

Chapter 4

Creating an Empirically Estimated Walking Index using a Nested Mode-Destination Choice Model

4.1 Introduction

Much recent interest has been given to walking indexes like the commercial product WalkScore [WalkScore, 2011], which has been validated as a measure of walking [Weinberger and Sweet, 2012, Manaugh and El-Geneidy, 2011] and as a positive correlate of residential real estate values [Cortright, 2009, Rauterkus and Miller, 2011]. Although the theory behind a walking index of this sort is established in Frank et al. [2008] and Moudon et al. [2006], the empirical basis for the destinations chosen, the weights assigned to destination categories, and the distance decay function should be tested more thoroughly. Of additional importance is the inclusion of aspects of the decision maker (e.g. household income) in the framework used here, an aspect completely missing from the WalkScore algorithm, and which has been shown to have a large impact on the decision to walk [Manaugh and El-Geneidy, 2011].

This chapter leverages the accessibility framework from Chapter 2 in representing walking-scale accessibility and creates a 2-level nested mode-destination model where choice of destination is conditional on the choice of mode, so that the logsum of the lower level creates a general measure of accessibility by mode which can be used in other models [Dong et al., 2006]. The travel modes used here are walking, automobile, and transit, and future work will include estimation and application of bike accessibility. Additionally, this study is one of the first to use the new California Household Travel Survey (CHTS), conducted in 2012, to estimate current trends in travel preferences, which the literature shows are changing rapidly [Dutzik and Baxandall, 2013].

This research also measures preferences at a more precise geographic scale than previous travel demand models. The local street network is utilized fully to measure walking accessibility, with nearly 226,000 street intersections representing possible destinations of trips. The specific trip purposes from the travel survey are not aggregated to more general trip purposes so that precise locations from a point-of-interest dataset can be used as the attractors of trips. For instance, 'eat out' trips are estimated to 'restaurant' destinations, 'indoor exercise' trips to 'recreation' destinations, and so forth. The resulting logsums by specific trip purpose for home-based non-work trip purposes are then combined into a composite index using the number of trips from a trip generation model as weights. Thus a single index with value from 1 to 100 for each mode is produced, which should address the shortcomings described above by taking the destinations, trip generation, and distance decay directly from observed behavior in a regional travel survey.

In addition, the currently established relationship of WalkScore to residential property values is largely correlational rather than causal; accessibility measures that comprise WalkScore use the distance to the nearest destinations in nine different categories, and assume that bringing destinations closer to the home causes more frequent walking trips, and that a preference for walking is being translated into increased home values. However it is just as plausible, using much the same reasoning as Crane [1996], that the reductions in travel distances are enabling shorter automobile trips and that this convenience is being capitalized into increased home values. Although the relationship of local accessibility and home values has been established, this is not the same as providing evidence that a preference for the behavior of walking is being translated into increased home values.

The indexes created in this work are ideal to measure the impact of walking behavior on home values, using the composite utilities for travel by mode that result from the methodology described above in an hedonic regression of home values to test whether the utility for walking, when controlling for the utility for travel by automobile and transit, is positively correlated with home values. Hence this chapter includes the indexes described above in the hedonic model of residential home prices which was created and explored in Chapter 3 to test this hypothesis.

4.2 Additional Literature

4.2.1 Nested mode-destination models

When alternatives in a discrete choice model have unobserved correlation among alternatives the IIA (Independence of Irrelevant Attributes) assumption of standard multinomial logit models is violated. The nested logit formulation is used to address this limitation by placing alternatives which have unobserved correlation into different nests and assuming that IIA only applies within each nest. The nested logit technique was first presented in McFadden [1978] and is derived mathematically in the methodology section of this chapter. The technique has been applied to mode choice [Sobel, 1979], in which certain travel modes are nested - e.g. public transit modes or non-motorized modes share a nest, and has also been applied to create hierarchies that capture spatial correlation in location choice or destination choice models [Kitamura et al., 1979].

Although sequential models of mode and destination are standard in 4-step travel modeling practice [de Dios Ortúzar et al., 2001] and are frequently used in activity-based models [Jonnalagadda et al., 2001], the choices of destination and mode are generally interrelated and thus the choice of destination conditional on mode (or vice versa) requires a nested model to capture this correlation. Some recent attempts have been made to estimate joint mode-destination models [Yagi and Mohammadian, 2008, Richards and Ben-Akiva, 1974], and at least one nested mode-destination model has been created that nests destination below mode - as this study does - with good results [Newman and Bernardin Jr, 2010].

4.2.2 Activity-based travel models

Mode and destination models have long been a central component of activity based travel modeling [Ishaq et al., 2013]. Bowman [1998] and Ben-Akiva et al. [1998] introduced the concept of the activity-based travel model (ABM), a completely disaggregate model in which all activities for all people are synthesized by representing each person moving through his/her simulated day. The Portland implementation of the Ben-Akiva and

Bowman framework has five levels of hierarchical choices: activity-patterns, time-of-day, mode-destination, sub-tours and intermediate stops. The San Francisco ABM [Jonnalagadda et al., 2001] and Florida ABM [Pendyala, 2004, Pendyala et al., 2005] both keep the same basic structure and hierarchy of models [Ishaq et al., 2013].

More recent models, like the Sacramento ABM (SACSIM), keep the same basic form but increase spatial and temporal resolution [Bradley et al., 2009, 2010]. SACSIM begins to model destinations at the parcel level, but does so by sampling first the TAZ and then a parcel within the TAZ during both estimation and simulation. In the work presented in this chapter, destinations are not required to be sampled during simulation, and the full population of 226 thousand alternatives is assigned a probability during simulation. This is not only computationally efficient but corrects sampling bias present when sampling during simulation.

The nested model used in this chapter focuses on the third step in the Bowman and Ben-Akiva choice hierarchy and does not attempt to model people as they move through their day. Instead, logsums are computed without accounting for time-of-day substitution by using travel times during the morning commute for automobile and transit in order to represent an accessibility measure that might have the largest impact on a long term choice like residential location choice. Future work should include logsums at multiple times of day in order to test the hypothesis that accessibility at different times of day can affect residential location choice.

4.3 Research Objectives

This work seeks to answer the following research questions:

- Does creating a nested mode-destination model representing purpose-specific destinations at the appropriate scale by mode yield significant coefficients with theoretically expected signs? To what degree do demographics of the choice maker, in particular income, affect these coefficients?

- Are logsums by mode significant in trip generation when measured at the pedestrian scale, or does the result from Ewing and Cervero [2001] that demographics are more predictive of trip generation than accessibility hold with this methodology?

- Are logsums by purpose weighted by the number of trips per purpose significant in an hedonic model of residential home prices? In other words, do people value walking accessibility, as measured using the mode-specific logsums created in this chapter, in their home purchase when controlling for transit and auto accessibility measured analogously?

4.4 Data and Methodology

4.4.1 Data

There are a number of datasets that are used in this research, including mode-specific travel networks, point-of-interest datasets to describe possible destinations, and the California Household Travel Survey (CHTS) 2012 travel survey. A dataset of residential home prices in the Bay Area is also used to apply this methodology in an hedonic model of home prices. Most of these datasets are described in more detail in Chapters 2 and 3, but is described briefly here with appropriate citations. The region of study is the San Francisco Bay Area, which is comprised of 6.78 million people, 9 counties, and is one of the most economically rich and scenically beautiful metropolitan areas in the United States. The regional government in the Bay Area is the Metropolitan Transportation Commission (MTC).

Mode-specific travel networks describe the travel time to different areas in the city by mode (for more information, see Chapter 2). Walking networks use the entire set of local streets provided by OpenStreetMap, the automobile network and simulated travel times per time period are included in the output of the MTC travel model, and the transit network combines the walking network with transit schedules provided by the Bay Area 511 GTFS (General Transit Feed Specification) service. All data was collected during 2011-2012 in order to coincide with the behavior observed in CHTS 2012.

The most recent travel survey available in the Bay Area region is the CHTS from 2012, which provides a long awaited update to the 2000 Bay Area Travel Survey (BATS) which has been used in numerous studies in the intervening years [Bhat and Guo, 2007, Guo et al., 2007, Eluru et al., 2010, Pinjari et al., 2007, Cervero and Duncan, 2003]. This is one of the first studies to use CHTS 2012 to update estimates of travel behavior, and some direct comparisons of travel behavior in the two surveys are available in Clelow et al. [2014]. CHTS 2012 is a travel diary for a single day for 9,719 households with 24,031 people in those households and contains a total 108,184 activities and 84,259 trips in the Bay Area portion for the Bay Area portion of the survey. Locations are address-matched and provided as precise latitude and longitude coordinates. Trip purposes and modes are recorded with detailed categories. This study combines specific modes into more general modes as shown in Table 4.1, but does not combine specific purposes into general purposes as much previous research does.

Specific purposes are linked with specific destinations in the point-of-interest dataset which provides exact geographic coordinates of destinations by category in the city. In simplest terms, the point-of-interest data can be thought of as an online yellow pages which might represent the consumer facing aspect of a business that a person would see if conducting an online search looking for nearby destinations. As such the dataset is ideal for representing the attractors of trips, although no quality information is present in the data and all destinations of the same type must be assumed to be identically attractive. Future work can utilize quality as an additional characteristic of attraction using this

General Mode	Specific Mode	Count of Trips
Walk	Walk	9,896
Auto	Driver or Carpool, Rental Car	68,028
Transit	All buses, trains, paratransit, public shuttle, and ferry	3,991

Table 4.1: Mapping of specific modes provided in the CHTS survey to general modes for which a separate logsum is computed

methodology.

Purposes and associated destinations are provided in Table 4.2; purposes are drawn from those available in the travel survey and nested mode-destination models for all non-work home-based purposes are estimated in this study. The available trip purposes overlap significantly with those used in creating the WalkScore index, and an additional column is provided in the table to indicate whether a destination is available in the WalkScore index.

Finally, a dataset of home prices was collected by saving publicly-available home listings for a three month time period in 2012 for homes throughout the Bay Area. The particular estimation dataset used contains 209,075 listings with a mean value of 306 dollars per square foot and a median of 266 dollars per square foot. During the time period this data was collected - spring of 2012 - the median home price in the United States was roughly 160,000 while the median in the Bay Area was fully 66% higher than the national average at 275,000. More detail on this dataset is provided in Chapter 3, as well as a description of the hedonic methodology and hedonic model results that are directly comparable to those provided in this chapter.

4.4.2 Nested mode-destination model

Multinomial logit (MNL) models are a method used to regress outcome variables on independent predictors where the outcome variable is discrete rather than continuous. The methodology was originally developed in McFadden [1980] and has been frequently applied to travel modeling [Ben-Akiva and Lerman, 1985]. Travel modeling is a natural application of the discrete choice methodology as the choice of, for example, mode of transportation is categorical rather than continuous and can be attributed to traits of the destination, attributes of the trip, and characteristics of the decision maker. The basic form of the MNL model is shown below which takes this form when the error term is Gumbel distributed.

$$P(i) = \frac{e^{V_i}}{\sum_i e^{V_i}} \tag{4.1}$$

Purpose	Description (from survey)	Survey ID	Destination Category	Total Count of Destinations	Available in WalkScore
Routine Shopping	Groceries, Clothing, Convenience Store, Household Maintenance	27	Groceries, Convenience Stores, Clothing and Accessories, Hardware, etc	5,618; 5,450; 7,510; 1,114	Yes
Shopping for Major Purchases	Appliance, Electronics, New Vehicle, Major Household Repairs	28	Home Improvement, Automotive, Electronics, Home Appliances, etc	24,906; 1,133; 5,146; 2,145	No
Household Errands	Bank, Dry Cleaning, etc.	29	ATM, Banking, Dry Cleaning, Hardware, etc	949; 8,367; 3,438; 1,114	Yes
Personal Business	Visit Government Office, Attorney, Accountant	30	Business and Professional Services	34,349	No
Eat Out	Eat meal at restaurant/diner	31	Food and Beverage, Restaurants	20,407	Yes
Health Care	Doctor, Dentist, Eye Care, Chiropractor, Veterinarian	32	Health and Medicine	7,903	No
Civic/Religious	Civic/Religious Activities	33	Community and Government	6,849	No
Outdoor Exercise	Playing Sports, Jogging, Bicycling, Walking, Walking the dog, etc.	34	Sports and Recreation	737	No
Indoor Exercise	Gym, Yoga, etc.	35	Sports and Recreation	737	No
Entertainment	Movies, Watch Sports, etc.	36	Arts, Entertainment, and Nightlife	3,260	Yes
Social	Visit friends/relatives	37	Arts, Entertainment, and Nightlife	3,260	No

Table 4.2: Specific trip purposes provided in BATS 2012, their description, and the set of destinations used as "attractors" in the point-of-interest dataset for estimating the model

MNL models make the assumption that alternatives have the trait of IIA (Independence of Irrelevant Alternatives), which states that adding or removing an alternative should affect all other alternatives proportionately to their initial probabilities. In the case of mode-destination it's likely that some alternatives substitute for others at greater rates; for instance, a person with a preference for walking might prefer to substitute a different destination while maintaining the choice of walking, or alternatively, if a person has a preference for a given destination, she might substitute a different mode at greater rates rather than switching to a new destination. In short, the choice of mode and destination is highly interrelated and nested logit should be used to relax the IIA assumption.

The "nests" involved are often described as a sequential choice. In the case of the choice of mode and destination, the process of decision making can be described as "mode then destination" or "destination then mode." Thus either could be assigned to the upper level of the nest while the other is assigned to the lower level [Newman and Bernardin Jr, 2010]. In truth, the nested logit formulation makes no presumption as to the sequential nature of the choices, rather it is an econometric structure which relaxes the IIA assumption and allows correlation of some alternatives to others due to unobserved factors.

This chapter always uses a nested structure with mode in the upper nest and destination in the lower nest (see Figure 4.1) so as to create a logsum factor which describes accessibility to all available destinations for a specific mode. As described in Dong et al. [2006], this creates a measure of accessibility which is empirically estimated - in this case using CHTS 2012 and the point-of-interest dataset to describe land use - the result of which is a numeric value which describes the accessibility at each location in the city.

Nested logit (following Lee and Waddell, 2010) computes the probability of choosing a destination d in the city as

$$P(d) = P(d|m) \cdot P(m) \tag{4.2}$$

where $P(d|m)$ is the conditional probability of choosing destination d given the selection of mode m and $P(m)$ is the probability of choosing mode m. The lower level choice of destination conditional on mode takes the standard MNL form

$$P(d|m) = \frac{e^{V_d \, \mu_m}}{\sum_{d' \, \epsilon \, D_m} e^{V_{d'} \, \mu_m}} \tag{4.3}$$

where V_d is the utility for the destination in question and μ_m is the nest parameter for the nest associated with that alternative. As with MNL, the probability is the exponentiated utility of the alternative divided by the sum of exponentiated utilities of alternatives in the nest; unlike MNL the NL version also contains μ_m which is the nesting parameter associated with each nest. The marginal probability of choosing mode m is

$$P(m) = \frac{e^{V'_m \, \mu}}{\sum_{m'' \, \epsilon \, M} e^{V''_m \, \mu}} \tag{4.4}$$

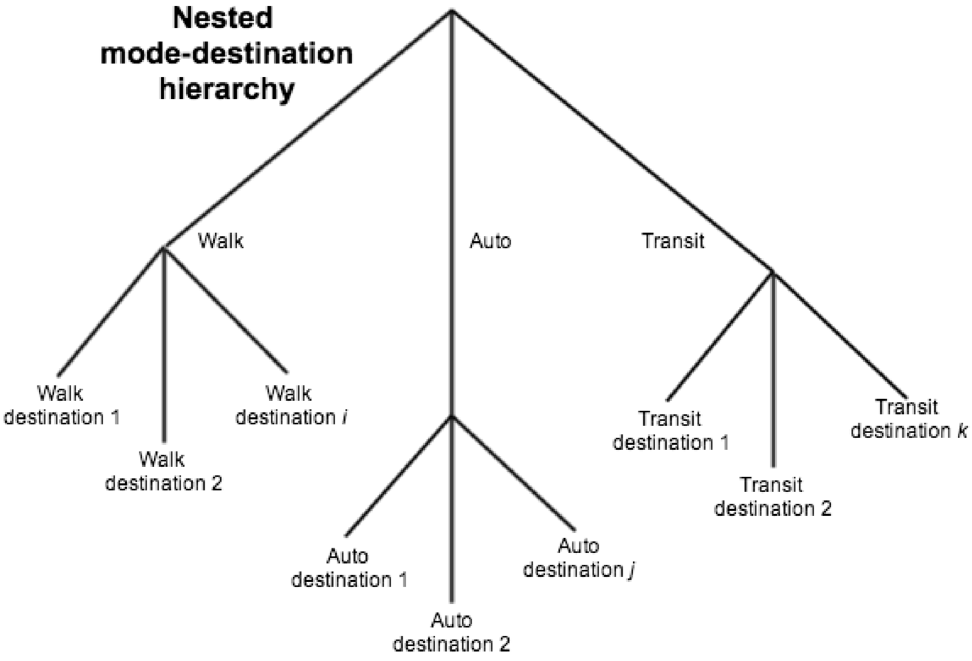


Figure 4.1: Two-tier nested structure mode mode choice and then destination choice

where V'_m is the logsum associated with nest m and μ is the top level nesting parameter (which is usually set to 1.0). The logsum represents the expected value of utilities for all alternatives in the nest. The logsums computed here use all available alternatives and do not sample during simulation, even if there are several thousand alternatives. This simplifies the computation of the logsum (as it does not need a correction for sampling) and the resulting formula for the logsum is

$$V'_m = \frac{1}{\mu_m} \ln(\sum_{d' \,\epsilon\, D_m} e^{V_d \, \mu_m}) \tag{4.5}$$

During estimation, 100 alternatives are sampled per nest (99 alternatives are sampled and the 1 chosen alternative is always included). Although McFadden showed that estimation is consistent when sampling, in NL a correction factor is necessary. This research follows the work of Guevara and Ben-Akiva [2013] and uses the correction the authors describe as “1_0” in which logsums are scaled proportional to the number of alternatives in the nest divided by the number of alternatives that are sampled. The authors show that, with non-emperical data, the 1_0 method performs as well as the other methods and requires a substantially simpler implementation.

4.4.3 Trip Generation model

This chapter uses a Poission model for trip generation which is described in detail in Chapter 2. Explanatory variables include demographic variables that exist in both the estimation dataset (Bay Area portion of the CHTS 2012) as well as the simulation dataset (the synthesized population provided by MTC), which includes gender, income, employment status, student status, and household. Measures of the built environment are logsums by mode for the appropriate income and household location of the decision maker. Additional variables are used to capture the age of the decision maker, where "younger" is defined as age less than 18, "older" is age greater than 70, and dummies for age in the "20s" and "30s" are also used. Trip generation models are estimated for each specific purpose provided in the travel survey.

4.4.4 Methodology for simulation

The purpose of this research is to create a set of logsums which are appropriate for use in other estimated models, such as a residential price hedonic model. The process is most easily described as a series of steps that must be executed in sequence:

1. Nested mode-destination choice models are estimated for each specific purpose which yield estimated parameters on travel times by mode, alternative specific coefficients by mode, and household income interacted with mode. All non-work home-based purposes are included and travel is assumed to start at the home location (only home-based trips are included). This creates a logsum by purpose and by mode that is appropriate for using in an hedonic model or location choice model.

2. Logsums are simulated for each purpose and for each income class. This implementation uses ten income classes that are the same as the ten income classes present in the survey (shown in Table 4.3). The model is not segmented when estimating, but for the purposes of creating a probability distribution function (PDF) with all 491,025 alternatives (summed across modes, from Chapter 2), choosers are discretized into ten representative people, one per income class in the survey. After this step in the process, logsums exist for each of 10 income classes, for each of 10 purposes, for each of 3 modes, for each of up to 226,060 alternatives per mode, for a total of roughly 68 million logsums.

3. Logsums by mode and demographics are used to estimate trip generation for each of the 10 trip purposes using a Poisson regression.

4. The number of trips is simulated for each person in the simulated population of the Bay Area. The population is synthesized from aggregate census geographies for use in the MTC travel model and contains 5.27 million people above the age of 18 and below the age of 75. As a result of this step, the number of simulated trips by specific purpose is predicted for each person in the synthesized population.

Income Class	1	2	3	4	5	6	7	8	9	10
Household Income (in thousands)	0-10	10-25	25-35	35-50	50-75	75-100	100-150	150-200	200-250	250+

d

Table 4.3: The ten income classes used for simulation (income classes are those specified in the CHTS 2012 survey)

5. A purpose-weighted and person-weighted logsum is computed as shown in Equation 4.6, where i is the index of each person that is located in the city, and j is the index of each of the ten trip purposes. In words, each person has an associated logsum which is weighted by the number of trips he is predicted to make, so that trips he makes more frequently are weighted higher. These purpose-weighted logsums are then averaged for all the people at a given location in the city. Locations used in this research are the street intersections of the local street network.

   Thus a relative accessibility is now assigned to every location in the city which is weighted by the people that live there and the number of trips that each person is predicted to take. This is called the PPW (purpose and person weighted) logsum when referred to here. Each PPW logsum is then divided by its maximum value and multiplied by 100 so as to create an index comparable to WalkScore, although the indexing step is primarily for visualization and ease of interpretation and is entirely optional. Note that every step in the process is based on an empirical model of how people substitute destinations and modes, the number of trips they make by purpose, and demographics are included so that travel behavior which varies strongly by traits of the person is accounted for.

$$PPW_n = \frac{\sum_i \sum_j logsum_{i,j} * trips_j}{size(I)} \; \forall \, n \epsilon N \tag{4.6}$$

4.5 Results

4.5.1 Estimation of nested mode-destination model

Nested mode-destination models for each purpose are estimated as described in the methodology section and the results are shown in Table 4.4. All ten purposes take the same simple form to aid in interpretability, with all coefficients being highly significant in all models. Travel time is negative in all models, with coefficients between .04 and .08. Travel time interacted with walking destinations is also negative, showing that walking trips are typically shorter than the other modes, while travel time interacted with transit is positive, as transit trips are longer on average than the other modes.

The attracting element (from Table 4.2) is always positive, and is allowed to vary by mode as the magnitude of the attraction variable varies by mode (the travel model network is sparse so there are typically more destinations per location). Household income is interacted with each mode and is negative when interacted with walking and even more negative when interacted with transit, indicating that households are increasingly likely to drive as their income rises. Note that outdoor exercise is dropped from these results as hypothetical attractors of outdoor exercise (like parks) are not significant in the model - the generators of outdoor exercise is an active research topic but won't be explored further here.

The nesting scale factors - μ's - are expected to be greater than one by the construction of nested logit. This is the case in each model for the auto and transit nests, but is not always the case for the walk nest. These values are thus constrained to be at least 1.0 during estimation to match the construction of the model, with a 1.0 nesting parameter being equivalent to having no nest at all. Model fit is reported as log-likelihoods, and likelihood ratios are high for all purposes.

Note that household income is the only interaction term used here. Although other interaction terms could be used, the small number of interactions keeps the model parsimonious and allows for the generation of maps for each income level which are intuitive and easy to understand (as shown in the next section). In the next chapter, latent classes are used to identify correlated sets of demographics which identify different lifestyle clusters.

4.5.2 Simulation of nested mode-destination model

Once models are estimated they can be simulated and mapped for a discrete set of purposes and decision makers. Here logsums are estimated by purpose and by income class for each mode and for each alternative, resulting in about 68 million logsums. Note that these logsums can be normalized to create mode choice probabilities, or mode splits, by dividing by the sum across modes for each purpose, income class, and destination. For the purpose of this research, a logsum is more accurate than mode split as a measure of absolute accessibility, but for other applications mode split might be more applicable.

In travel modeling, it is typical to simulate actual destination choices, which is an additional step beyond what is done here, but for the purpose of computing an aggregate measure of accessibility, logsums are sufficient. An informative byproduct of simulating a logsum for every possible location in the city is that these logsums can be mapped with no missing data. In this case there are 300 maps (one for each purpose, income, and mode). Two of these maps are shown in Figure 4.2, which show the accessibility of walking for routine shopping for a person of the lowest income classification and for the highest income classification; logsums for the lowest income households are higher indicating a greater utility for walking (which can be due to preferences or constraints), although the

	Routine Shopping		**Major Purchases**		**Household Errands**		**Personal Business**	
Null Loglik		-13536		-1021		-4913		-3664
Converged Loglik		-6503		-603		-2187		-2234
Loglik Ratio		0.52		0.41		0.55		0.39
Variables	Coeff	T-score	Coeff	T-score	Coeff	T-score	Coeff	T-score
mu (walk)	1	104.74	1	17.15	1	69.1	1	46.86
mu (auto)	5	61.13	2.29	20.32	5	35.74	2.77	39.42
mu (transit)	1.19	61.74	1.22	16.73	1.39	28.96	1.16	42.97
travel_time*walk	-0.08	-22.53	0	-0.05	-0.07	-13.58	-0.1	-10.76
travel_time*transit	0.07	22.44	0.09	8.56	0.09	15.03	0.11	25
travel_time	-0.06	-62.89	-0.08	-18.52	-0.07	-37.44	-0.07	-35.05
attractor*walk	0.38	11.38	0.62	3.11	0.18	3.2	0.36	5.28
attractor*auto	0.08	15.49	0.14	4.78	0.03	3.28	0.08	6.26
attractor*transit	0.64	13.3	1.02	8.25	0.36	3.71	0.9	17.45
hhincome*walk	-0.42	-79.38	-0.51	-14.69	-0.4	-49.99	-0.35	-30.61
hhincome*transit	-0.94	-84.44	-0.82	-22.81	-0.87	-42.23	-0.98	-61.95

	Social		**Entertainment**		**Indoor Exercise**		**Eat Out**	
Null Loglik		-9051		-3782		-4860		-7963
Converged Loglik		-6164		-2745		-2250		-4772
Loglik Ratio		0.32		0.27		0.54		0.40
Variables	Coeff	T-score	Coeff	T-score	Coeff	T-score	Coeff	T-score
mu (walk)	2.4	33.76	1	51.02	1	57.77	1	77.85
mu (auto)	3.03	59.82	2.82	39.98	3.73	38.61	4.95	50.55
mu (transit)	1.51	46.24	2.33	26.74	1.43	25.34	1.9	40.29
travel_time*walk	-0.05	-11.31	-0.07	-10.2	-0.05	-7.95	-0.09	-18.93
travel_time*transit	0.12	39.52	0.16	53.52	0.08	11.29	0.11	39.53
travel_time	-0.05	-49.37	-0.05	-30.88	-0.08	-38.32	-0.04	-49.42
attractor*walk	0.12	2.69	0.75	5.63	0.96	5.67	0.27	4.56
attractor*auto	-0.01	-2.48	0.12	10.02	0.11	7.61	0.08	13.01
attractor*transit	0.52	11.49	0.66	10.23	1.06	4.05	0.54	13.83
hhincome*walk	-0.13	-19.81	-0.38	-34.2	-0.43	-44.99	-0.43	-59.93
hhincome*transit	-0.89	-78.83	-0.73	-62.5	-0.73	-33.91	-0.76	-77.14

Table 4.4: Listing of all nested mode-destination models by specific trip purpose in the CHTS 2012 survey. Attractors vary by purpose and are listed in Table 4.2.

relatively high accessibility locations in the city are higher regardless of income.

4.5.3 Estimation of trip generation

The next step in creating logsums weighted by purpose is to estimate a trip generation model to use as weights for each purpose-specific logsum, making the assumption that the value of accessibility to the destinations of a trip would be in proportion to the frequency of the trip. A poisson count model is estimated for each purpose in CHTS 2012, using the total count of home-based trips for each household for the day recorded in the travel diary as the dependent variable, and using demographics from the travel survey and logsum accessibilities for the correct income and the correct geographic location computed in the step above as independent variables. Although there are more specific demographic traits present in the travel survey than are used in the estimation, the variables must be limited to those that are also present in the simulated population for use in the next step. The definition of the variables used here is described in the methodology section of this chapter.

The significant coefficients are presented in Table 4.5. All variables are dropped that have a z-score less than 1.64 (p-value greater than .1). R-squared values for all models are very low, but the coefficients for demographics and accessibility variables match theoretical expectations and previous results in the literature, therefore low r-squared's are likely due to the fact that CHTS 2012 records a single day of travel and prediction of trips on a single day is an event that is highly subject to random fluctuation.

Demographics have many significant impacts on trip generation. Employed people make fewer trips of every purpose, students make fewer shopping trips but more exercise trips, females make more social, exercise, civic, health, and routine shopping trips, but fewer eat out trips. Age has numerous impacts, with people in their 20s and 30s making more social trips, but fewer shopping tips. Household size has a negative impact on most trip purposes but with a positive impact on civic/religious trip purposes. Higher education levels have a positive impact on most trip generation purposes, except for social, health care, and major shopping purchases.

Accessibility logsums are also significant for several trip purposes. Greater walk accessibility is correlated with more exercise, entertainment, eat out, and routine shopping trips, but fewer home-based personal business trips. Increased transit accessibility has a positive impact on errands, personal business, eat out, and entertainment trips, but with a negative association with major purchases. Auto accessibility is positively related to major purchases but negatively with household errands, possibly indicating a substitution of these two trip purposes in suburban built environments.

These results seem to indicate that demographics and accessibility measures both have moderate impacts on trip generation, although the results are highly dependent on the specific trip purpose - for instance accessibility has no significant impact on health, civic/religious, and social trip purposes. This result seems consistent withEwing and Cervero [2001] which finds that both demographics and accessibility impact trip genera-

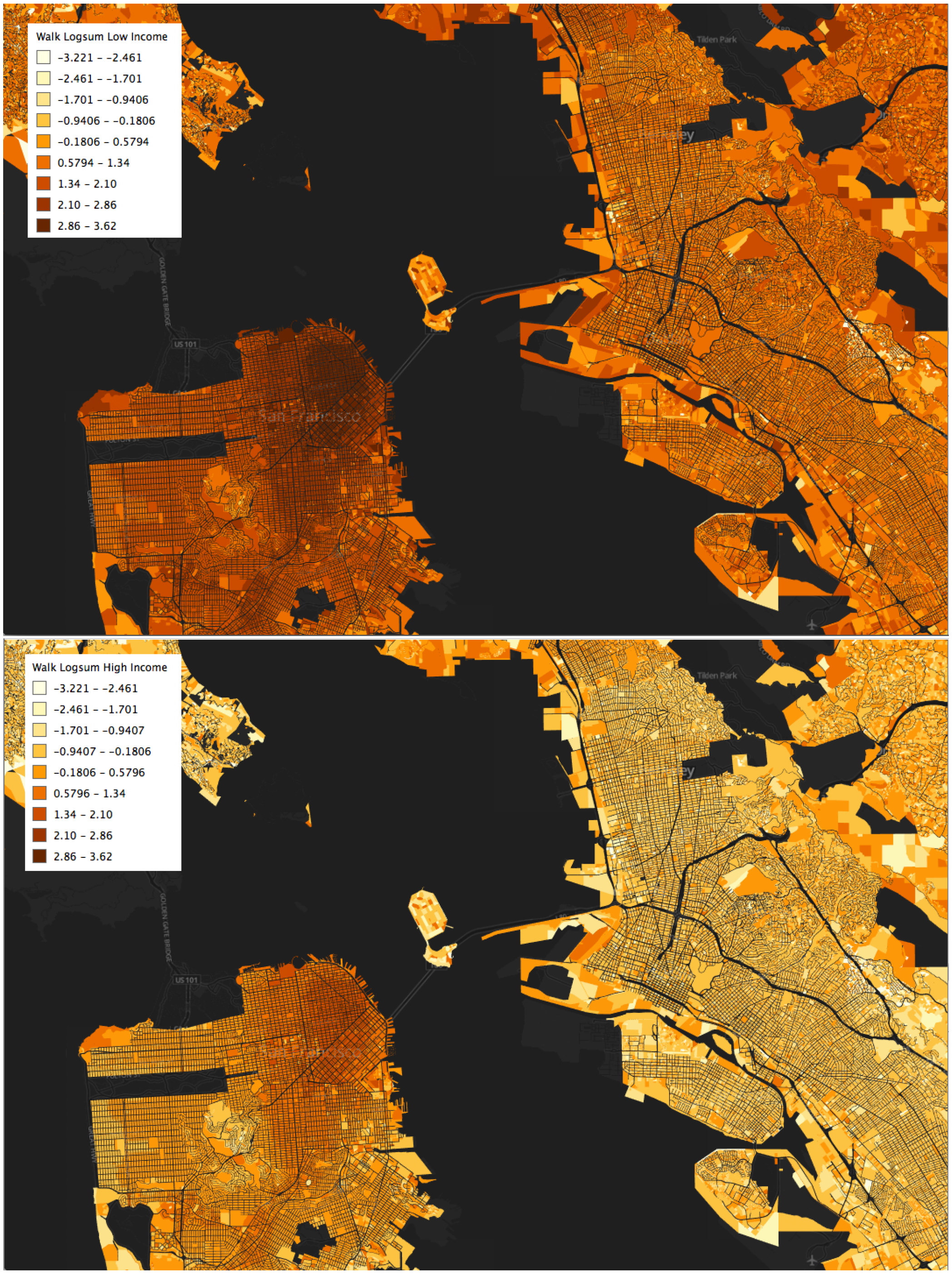


Figure 4.2: The logsums are higher for low income (above) then high income (below) for the purpose of routine shopping indicating an increased likelihood for walking

tion, although perhaps this research provides some evidence as to the increased importance of accessibility for trip generation. For the purpose of this work, these trip generation models are sufficient to create a rough estimate of the number of trips per person in order to weight the highly correlated set of logsums computed above.

4.5.4 Simulated number of trips

In this step, the trip generation models from the previous step, which were estimated using the CHTS 2012 survey, are simulated on the synthesized population from the nine county Bay Area. The synthesized population used here is provided by MTC and is the same one the regional agency uses for travel modeling, and contains 5.27 million people with age greater than 18 and less than 75. The mean and standard deviation of the simulated number of trips by purpose for the entire synthesized population is presented in Table 4.6. The number of trips per person and purpose are used in the next step.

4.5.5 Purpose and person weighted logsums

Logsums have now been generated for each person, purpose, and mode, and the number of trips has been simulated for the synthesized population. The next step is to average over purpose and person in order to create a single weighted logsum by mode for each location in the city. This is described in detail in the methodology section of this chapter; logsums are weighted by the number of simulated trips and summed for each person and these weighted logsums are averaged among people for each location in the city. The result is a logsum measure that is weighted by the frequency of travel by each purpose and adjusted for the demographics of the people that live in that location.

This creates purpose and purpose weighted (PPW) logsums that are ideal for use in a model which does not contain the same disaggregate information by person. In this research, these logsums are used as independent variables in a residential price hedonic model for which the demographics of the person that actually purchased the house are not observed. The best estimate of a person-based accessibility measure for each home purchase is the weighted logsum computed here per mode and per location in the city (although no logsum can be computed for a location where there are no residents). These logsums are mapped in Figures 4.3, 4.4, and 4.5, which show the PPW logsum for each of the walking, automobile, and transit modes. Note that locations without any residents have no index value as the index is necessarily based on the demographics of the people that live in a location; parcels without an associated index value are colored white in the maps.

4.5.6 Using mode-specific indexes in residential hedonic models

This section provides a sample application of the purpose and person weighted logsums in an empirical model, in this case a residential price hedonic model. Although much

	Routine		**Major**		**Household**		**Personal**	
	Shopping		**Purchases**		**Errands**		**Business**	
R-squared		0.02		.02		.04		.01
Variables	Coeff	Z-score	Coeff	Z-score	Coeff	Z-score	Coeff	Z-score
Female	.13	3.36						
Younger			-1.43	-5.27	-.95	-6.2		
Older	-.14	-1.9						
20s	-.44	-4.1			-.57	-3.1		
30s			-.62	-1.9	-.47	-3.2	-.46	-3.0
Employed	-.38	-8.6	-.45	-3.1	-.60	-8.5	-.47	-6.2
Student	-.30	-3.9			-.25	-1.7	-.28	-2.7
HH Size	-.05	-3.0			-.10	-3.5		
Income					.07	3.0		
Education	.06	4.3			.13	5.2	.10	4.30
Auto Lsum			.22	2.6	-.11	-2.3		
Transit Lsum			-.10	-2.3	.14	4.0	.08	2.9
Walk Lsum	.09	4.2					-.17	-2.7
Constant			-6.0	-8.8	-1.7	-4.6	-2.9	-22.0

	Social		**Entertainment**		**Indoor**		**Eat Out**	
					Exercise			
R-squared		0.01		0.01		0.03		.03
Variables	Coeff	Z-score	Coeff	Z-score	Coeff	Z-score	Coeff	Z-score
Female	.15	3.7			.17	2.7	-.15	-3.0
Younger								
Older	-.18	-2.1						
20s	.31	3.7						
30s	.17	2.2	-.38	-2.5				
Empl Ratio	-.33	-7.0	-.33	-4.3	-.47	-6.5	-.34	-6.3
Stud Ratio	.15	2.8			.18	1.8		
HH Siz	-.09	-5.5			-.1	-3.6	-.14	-6.3
Income			.12	4.5	.16	9.2	.16	11.0
Education			.10	4.9	.16	7.1	.13	8.0
Auto Lsum			-.10	-2.3				
Trans Lsum			.06	1.8			.08	3.1
Walk Lsum			.11	1.6	.12	3.0	.13	2.5
Constant	-1.9	-28.4	-2.9	-7.6	-4.2	-26.6	-3.0	-21.6

Table 4.5: Listing of Poisson models of trip generation described here

	Routine Shopping	Major Purchases	Household Errands	Personal Business	Eat Out	Health Care	Civic / Religious	Indoor Exercise	Entertainment	Social
Mean	.289	.024	.087	.059	.151	.053	.045	.047	.050	.177
Std Dev	.087	.005	.032	.015	.027	.017	.029	.019	.010	.038

Table 4.6: The mean and standard deviation of the number of simulated trips for the synthesized population

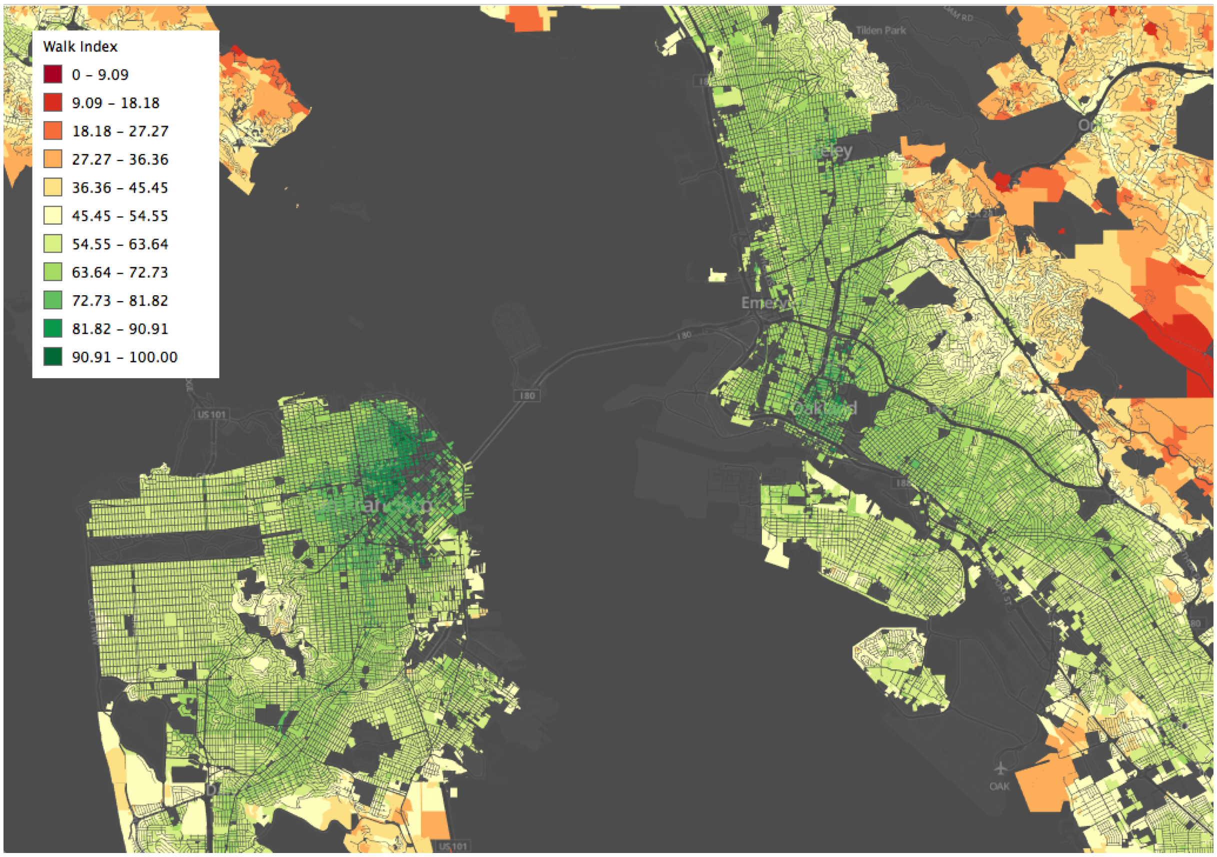


Figure 4.3: Equal interval map of the walking index for the entirety of San Francisco

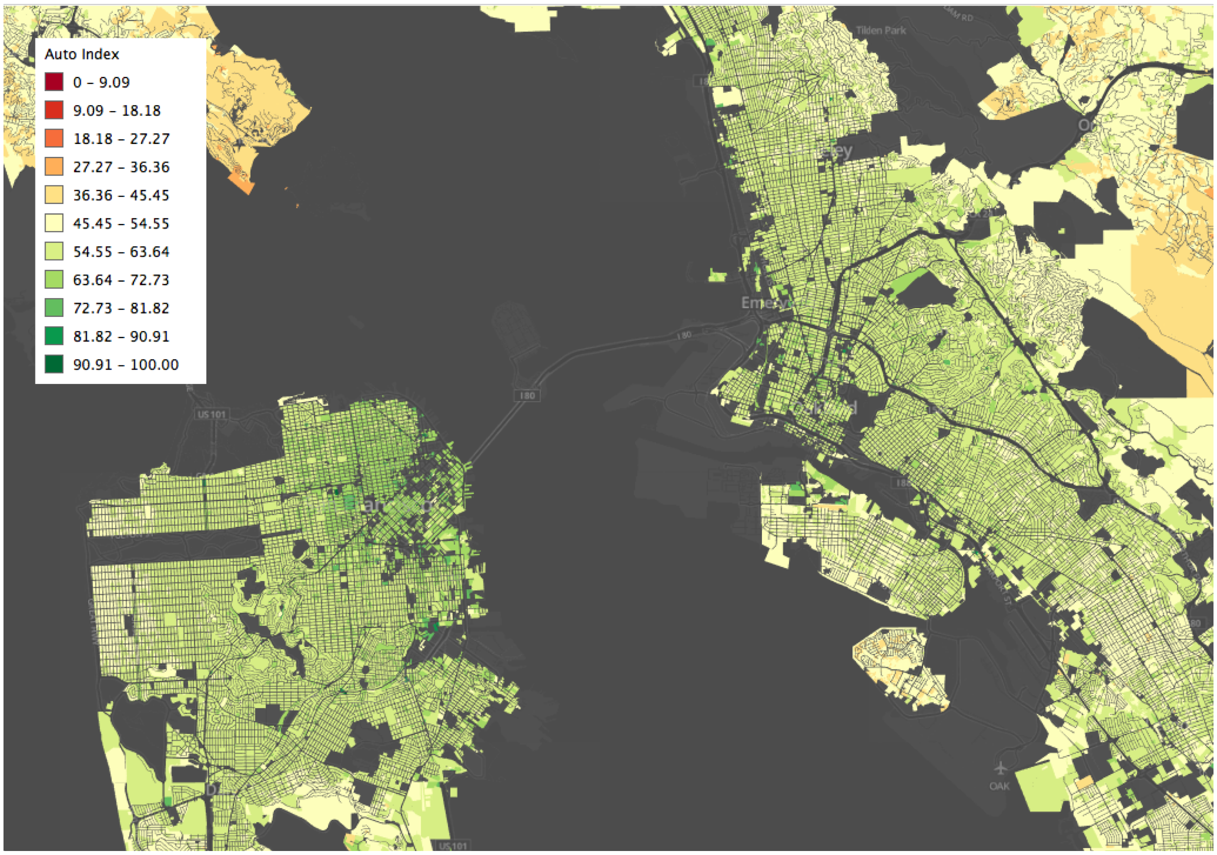


Figure 4.4: Equal interval map of the auto index for the entirety of San Francisco

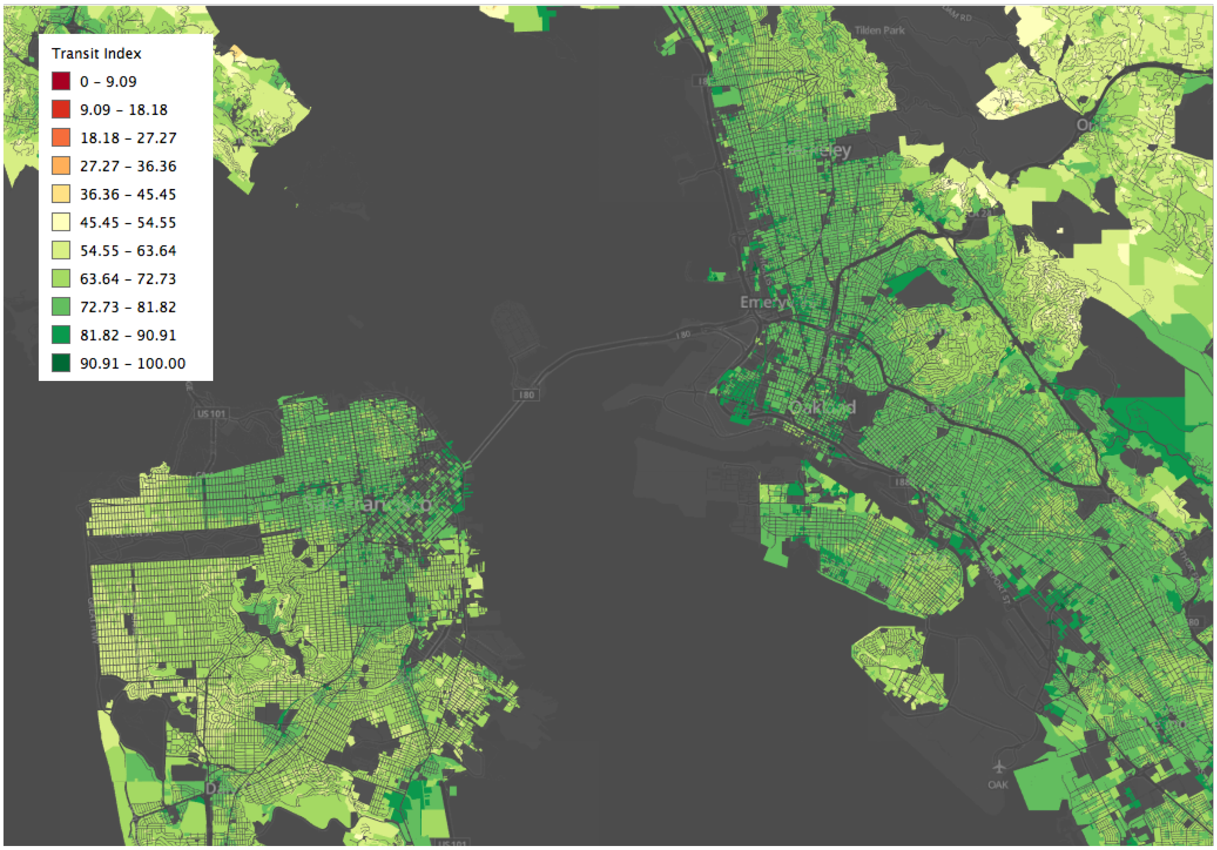


Figure 4.5: Equal interval map of the transit index for the entirety of San Francisco

computation must take place to create the weighted logsums, the output is intuitive. Logsums are an empirically estimated measure of how people in the Bay Area travel, which are weighted by the frequency of trips people make, and then averaged over the people that live in a specific location to account, for example, for how different income classes sort into neighborhoods and travel in different ways due to constraints and preferences. As a final step to aid in interpretability, the logsums are normalized and multiplied by 100 to create an index for each mode. This walking index is roughly comparable to WalkScore, but based on empirical data.

The empirically estimated walk, transit, and auto indexes are included in the residential price hedonic model used in Chapter 3. As in the previous chapter, the available characteristics of the unit are included as control variables, and instead of the cumulative opportunity accessibility measures to individual amenities used before, this model includes the PPW indexes by mode generated as part of this research. The coefficients from the estimated model are shown in Table 4.7. As before, neighborhood income is a positive correlate of home value, increased square footage reduces price per square foot, and lot size, historic units, and new units are all positive correlates of home price.

Interestingly, all three accessibility indexes are positive when accounting for the others. This makes intuitive sense; accessibility for each mode should be a positive trait for some home purchasers, although this was difficult to disentangle with the highly correlated accessibility metrics used in the previous chapter. Correlation matrices of the indexes are shown in Table 4.8, with the most correlated indexes being auto and transit with a value of .46. Future work could use factor analysis to remove the correlation of accessibility by mode, but factor analysis makes interpretability of coefficients more difficult so logsums are used directly here.

With the entire population of mode-specific indexes and relevant coefficients from the residential price hedonic model, the dollar value of a one standard deviation change in the mode-specific index can be computed, which is shown in Table 4.9. Not surprisingly, the highest dollar value is attributable to a one standard deviation change in automobile accessibility, with a dollar value of $41.99 per square foot on a mean value in the dataset of $306 dollar per sqft (14% of the mean square foot price). A one standard deviation change in walking accessibility is also worth $36.62, or 12% of the sales price per square foot, which indicates a significant and large price premium paid for the ability to walk to nearby destinations. This is perhaps not surprising given the incredible price premiums people pay for excellent walking accessibility in the city of San Francisco. A one standard deviation change in transit access is valued at $34.69 per square foot. These findings are discussed further in the next section.

4.6 Discussion

The methodology presented in this chapter first computes logsum accessibility measures for each specific purpose and income level, demonstrating that disaggregate point-

R-squared		.484
	Coefficient	T-score
Ave Neighborhood Income	1.34	379
Square Footage	-.375	-146.1
Lot Size	.0047	9.9
Historic Unit	.2403	79.12
New Unit	.0901	17.55
Constant	-8.98	-188.6
Walk Index	.0174	84.44
Auto Index	.0239	143.17
Transit Index	.0228	96.49

Table 4.7: Residential price hedonic model which include accessibility indexes by mode

	walk	auto	transit
walk	1.0		
auto	.21	1.0	
transit	.19	.46	1.0

Table 4.8: Correlation matrix for mode-specific indexes

	Std Dev	Coeff	Price Impact per Square Foot
Walk Index	6.49	.0174	$36.62
Auto Index	5.38	.0239	$41.99
Transit Index	4.71	.0228	$34.69

Table 4.9: The standard deviation of the accessibility index and coefficient from Table 4.7 combine to estimate the dollar value of a one standard deviation change in accessibility by each mode

of-interest datasets can be used as attractors in destination models, rather than the more typical use of jobs by employment sector. Next, the logsums by purpose are weighted by the simulated number of trips by purpose for each person and then averaged over the people that live in a given location. The result is then normalized and multiplied by 100 to provide an index of accessibility by mode in all locations of the city which takes into account the travel preferences and constraints of the people that live there.

These mode-specific metrics provide ideal measures of accessibility to use in other models, such as a residential price hedonic model in which we do not observe the person who purchases each house. As presented in the last section, these mode-specific indexes are all strongly predictive of real estate home values when controlling for traits of the unit. Theory would suggest that all absolute measures of accessibility should be weighted positively when purchasing a home - the result presented here, but this has not always been a given in America's suburban-oriented history of development.

Perhaps not surprising is that the dollar equivalent of a change in automobile index of one standard deviation is larger than the dollar equivalent of a one standard deviation change in walking index, although both of the walking and transit indexes are positive in the hedonic price model. A number of qualifications should be made here: first, these results are specific to the Bay Area. The methodology should be tested in other locations, as the Bay Area is a unique region with very high home values in the amenity-rich and very dense city of San Francisco. However these results are clearly supported by the enormous price premiums paid to live in the city of San Francisco, even if San Francisco might still be a special case.

Second, it should be noted that the variance in the walking index is the highest while the variance of the transit index is the lowest. Comparing a one standard deviation change between modes might thus bias the results toward increased value of walking accessibility, but at the same time is indicative of the limited supply of walkable neighborhoods. Although there is greater supply of high accessibility auto neighborhoods, this does indicate that, where present, high walking and high transit accessibility neighborhoods are valued.

Third, it should be noted that the way the models are constructed does not penalize the ability to drive for short trips in dense areas like urban San Francisco. Because only travel_time is used as an impedance, and not, for instance, the ability to park easily, this means that by construction any area that is high in walkability index will also be high in auto index. This is almost certainly a flawed assumption and should be corrected to take into account the difficulty of parking, and thus of driving for short trips, in dense urban areas. Once this correction is made, it's likely that impact of walking and auto accessibility is more separable in the model. However, the small correlation between the walking index and auto index indicates that this might not have a huge impact on the results presented here.

Finally, it is tempting to conclude that more high walking and transit accessibility neighborhoods should be created, which could possibly accrue environmental and social benefits, in addition to providing profit to real estate developers and homeowners. Indeed, the recent spate of development in the city of San Francisco - with more than 4,220 units currently being developed in the year 2012 SPUR [2012] - seems to support this. Nonetheless, caution is advised due to the issue of self-selection. Although it's probably true that walkable neighborhoods are highly valued in the Bay Area, and that more walkable neighborhoods can be provided without subsidy to real estate developers, it is certainly not true that everyone prefers walkable neighborhoods. Future research must investigate this research more precisely, with one possible method being to allow coefficients to vary among people using the latent class modeling technique Walker and Li [2007], and to measure the relative size of different segments. This is precisely the topic of the next chapter.

4.7 Conclusion

This chapter adds to previous research by first regressing disaggregate amenities on peoples' observed travel behavior using a nested mode-destination model estimated using the Bay Area portion of the CHTS 2012, computing a weighted logsum measure which accounts both for the frequencies of trips by purpose as well as the preferences and constraints of the people that are traveling, and averaging this measure for each location in the city. The resulting accessibility indexes by mode are shown to be highly correlated with residential real estate prices.

In the analysis performed here, the dollar value of a single standard deviation increase in the walking index is only slightly lower than the dollar value attributable to a single standard deviation of the auto index. Although there are some qualifications to the results presented in this chapter, this is strong evidence that people do in fact highly value the ability to walk to nearby amenities, even when accounting for income constraints and accessibility by other modes. This is a major contribution to the literature as it moves from the previous correlational relationship of WalkScore and home values toward a more robust behavioral explanation that people are valuing the ability to walk to nearby destinations in their home purchasing decision.

Chapter 5

Using Latent Class Models to Explore the Heterogeneous Impact of Accessibility by Mode on Residential Location Choices

5.1 Introduction

Residential location choice is one of the key determinants in how cities are shaped. The preferences of households making location decisions and how they change over time is one of the most important areas of research for city planning, as these decisions can have enormous impacts on traffic congestion, social equity considerations, and environmental consequences. Although many characteristics of neighborhoods and dwelling units affect the residential location choice including crime, school quality, and size and quality of the dwelling unit and lot [McFadden, 1978, Lerman, 1976, Quigley, 1985], a common research topic in transportation planning is the impact of accessibility to nearby destinations on the choice of residence [Lee et al., 2010a, Chatman, 2009, Ben-Akiva and Bowman, 1998].

This impact is particularly important to the planning literature, as many planning interventions hope to create more dense residential neighborhoods in order to reduce greenhouse gas emissions (GHGs) from driving, improve transit accessibility for households which can't afford automobiles, and increase the prevalence of active transportation to positively impact public health outcomes. In fact, many recent policies like Senate Bill 375 in California have legislated that land use impacts on GHG emissions must be accounted for in long-term regional plans [Barbour and Deakin, 2012], although many researchers have questioned the efficacy of such policies on the basis that many households are unwilling to live in dense neighborhoods or travel using transit and nonmotorized modes [Boarnet, 2011, Boarnet and Sarmiento, 1998, Boarnet et al., 2011, Crane, 1996].

Although previous research establishes a moderate but significant impact of accessibility on the residential location decision, very little research has tested for heterogeneous impacts of accessibility by mode on the residential location choice and that is the topic explored in this chapter. This topic is of vital importance as research which investigates average behavior in a cross-sectional dataset almost certainly loses the heterogeneity of preferences which must be accounted for in understanding the impact of the choice of residence on the willingness to travel by the transit and walking modes. This chapter thus uses latent class choice models to cluster preferences for mode-specific accessibility on residential choice to answer the following questions.

Are there significant clusters of households exhibiting different location choice preferences among locational attributes such as accessibility by walking, transit and drive modes? What number of latent classes most effectively represents clustering of preferences? Are some clusters positively influenced by walking and transit accessibility, and if so, how large are they and what demographics predict membership into the clusters?

5.2 Background

Residential location choice is the process by which a household makes tradeoffs between characteristics of different dwelling units and neighborhoods. Numerous studies have analyzed this decision making process from both a qualitative and quantitative perspective

[Chatman, 2009, Giuliano and Small, 1991, Tiebout, 1956, Bhat and Guo, 2007, Cao et al., 2008, Cervero and Duncan, 2002, Lee et al., 2010b, Muth, 1969].

Most previous residential location choice models use Lerman's grouped alternatives choice (GOC) approach (Lerman, 1976 in Guo and Bhat, 2007) where alternatives are collections of residential units and represent large contiguous geographic shapes in the city, for instance census tracts or transportation analysis zones [Lerman, 1976, Waddell, 1996, 2000, Deng et al., 2003]. This makes analysis of the impacts of accessibility by walking difficult, as the built environment can change significantly within these large geographies. Recent studies have begun to model residential choice where alternatives are specific buildings or parcels [Waddell et al., 2010, Lee et al., 2010a, Lee and Waddell, 2010], but none of these studies model the impact of local accessibility at the parcel scale.

This chapter adds to previous research by testing the impact of statistically estimated measures of accessibility by the walk, transit, and automobile modes derived in the previous chapter on the residential location choice. In that chapter, a nested mode-destination model is estimated using the observed behavior in the regional travel survey in the San Francisco Bay Area to predict the indirect utility of each destination for each mode and for each income class of the decision maker (as income is a highly significant factor in walking behavior Manaugh and El-Geneidy, 2011, Pucher and Renne, 2003). The destinations are summed in a statistically principled way to create a logsum, which is a measure of accessibility for each mode and income class. These logsums are then matched to the income class of each household in the travel survey and a residential choice model is estimated in which the appropriate logsums by walk, transit, and automobile modes are used as independent variables.

Additionally, and perhaps more importantly, standard discrete choice techniques find one set of coefficients for the entire estimation dataset, which results in the average behavior for the set of choices that are observed. It is certain that all people do not exhibit this average behavior, rather people have a distribution of preferences which is difficult to estimate econometrically. For instance, in the location choice decision the relevant question is not "does the average person value walking?" but rather, "what is the relative size of the population that values walking highly enough for it to affect the home buying decision?" There are a small number of advanced discrete choice techniques which allow investigation of the heterogeneity of preferences in choice making, including manual segmentation, mixed logit, and latent class models. This last is the methodology used here as it is ideal for discovering clusters of preferences and was originally developed for the purpose of consumer market segmentation.

This chapter adds to previous research by applying the latent class choice methodology developed in Walker and Li [2007] and Vij et al. [2011] to a residential location choice model where alternatives are small geographies. This methodology allows for multiple sets of coefficients for the residential location choice, based on an endogenously estimated set of "classes" where membership in classes is regressed on available demographic variables. In other words, based on income, presence of children, and other demographic predictors, the coefficients on the value of walking, transit, and driving accessibility and other variables are allowed to change. Although there are a small number of studies which apply latent class models to residential location choice (described below), this is the first to incorporate walking-scale accessibility, the intent of which is to understand the relative size of segments that value waking and transit and what demographics predict membership into these clusters.

5.3 Data and Methodology

5.3.1 Discrete Choice and Latent Class Models

Discrete choice modeling [McFadden, 1980] allows the estimation of indirect utility among a discrete number of alternatives subject to a linear in parameters utility function and a given distribution for a random error term. See Ben-Akiva and Lerman [1985] or Train [Train, 2009] for a thorough treatment of the methods and their application to travel behavior. Williams [1977] developed the theory for using logsum measures from such models to measure consumer surplus. In that sense, the use of logsum measures to represent accessibility is also a measure of consumer surplus.

A major shortcoming of the straightforward implementation of discrete choice models is that it does not allow for heterogeneity of preferences when estimating coefficients on the explanatory variables, although interaction variables and segmentation are techniques can begin to address this limitation. Traditional discrete choice models estimate coefficients which are the average value of a coefficient from the estimation dataset. To allow for heterogeneity in preferences, Latent Class Choice Modes (LCCMs) were first developed in the marketing sciences to identify relatively homogenous clusters of consumers that differ significantly from each other in their consumer behaviors (Kamakura and Russell, 1989 in Vij, 2013).

LCCMs have a class membership model and a manually-specified number of class-specific choice models conditional on the choice of a class in the class membership model. Heterogeneity is captured by allowing coefficients in each class-specific model to be independent. The class membership model typically incorporates traits of the decision maker - in this case demographics of the household making the residential location choice - while the class specific models typically include traits of the alternatives, in this case accessibility measures by mode, price, and attributes of the neighborhoods around each possible residential location.

Early applications of discrete choice models tended to focus on transportation behavior like mode choice, but also included work on residential location choice [Lerman, 1976, Quigley, 1976, McFadden, 1978] and on residential mobility [Clark and Van Lierop, 1987]. LCCMs have become increasingly common in the field of travel demand and have been applied to travel mode choice [Atasoy et al., 2011, Vij et al., 2011], vehicle ownership [Train, 2008, Hidrue, 2011], and residential location [Walker and Li, 2007, Olaru et al., 2011].

This last is particularly important as these are the only two previous studies which combine LCCMs with residential choice. Walker and Li find three segments in the residential location choice which they name suburban dwellers, urban dwellers, and transit riders, although they find a prima facia contradiction that suburban dwellers have preferences to travel by transit while the urban dwellers often prefer to travel by auto. This chapter adds to this research by using a logsum measure of the accessibility of destinations by mode in order to investigate closely the question of how people make tradeoffs in mode-specific accessibility when choosing the residential location.

5.3.2 Data and Methodology for the Residential LCCM

As described in the literature review, LCCMs have a class membership model and class-specific models which are conditional on selection into a class by the class membership model. The number of classes is a user-specified parameter to the estimation process and is chosen based on goodness of fit and interpretability of coefficients. In the residential choice model specified here, location choice alternatives are the 226,000 street intersections in the San Francisco Bay Area. Since the number of alternatives is large, sampling must be used for estimation, and the sample size which produces robust results, defined as producing consistent results for any random sample, is 50 (the chosen alternative plus 49 randomly sampled alternatives). When simulating the entire population of alternatives is used, so the probability density function is created for all 226,000 alternatives for use in mapping high probability areas in the city.

5.3.2.1 Class-membership Model

The class-membership model is used to predict the probability of membership into a specific latent class. The form of the class-membership model is the same as that of the standard MNL, shown below which takes this form when the error term is Gumbel distributed. Here j is the index of the latent class, while i is used to index specific alternatives.

$$p(j) = \frac{e^{V_j}}{\sum_j e^{V_j}} \tag{5.1}$$

There is a rich set of demographic predictors available in the new CHTS household survey which can be used as explanatory variables in the class membership model, including age of the head of household, household size, household income (in ten categories), presence of children, number of employed workers, number of students, related/non-related status, residential building type (e.g. single family detached), rent/own, and number of years in the current residence. Each of these variables can be tested in the class-membership model for statistical significance and theoretical sign.

5.3.2.2 Class-specific Models

The class-specific model is also a straightforward MNL model when conditioned on selection into a specific class. Here the probability of choosing alternative i conditional on class j is given by:

$$p(i|j) = \frac{e^{V_{i,j}}}{\sum_i e^{V_{i,j}}} \tag{5.2}$$

The probability of alternative i and class j is the standard formula for joint probabilities using the equations above.

$$p_{i,j} = \frac{p_{i|j} * p_j}{\sum_j p_{i|j} * p_j} \tag{5.3}$$

Log-likelihood is maximized using the Expectation Maximization (EM) method now common in machine learning. The intuition for this process is that log-likelihood increases as the probability of the chosen alternative increases for the higher probability classes for each household. Thus log-likelihood increases when either the probability of the alternative conditional on the class increases or the probability of selection into the class increases consistent with the concepts of joint and conditional probabilities. Although the definition of the latent class model is fairly easy to understand, the selection into a class is an abstract concept and is thus never actually observed, which makes the analytical gradients used for estimating the coefficients of the LCCM somewhat more involved and the reader is referred to Vij [2013] pps. 17-22 for the derivation.

The class-specific models have a specification which includes variables for logsums for travel by walking, transit, and auto, as well as price, income interacted with price, average income in the neighborhood, and residential units to act as a supply variable. The prices are taken from MLS residential listings from 2012 so are coeval with the travel survey - the median price within 1000 meters on the street network is computed so as to fill in areas for which there are no observations and to smooth over outliers. These are the same prices used as the dependent variable in the hedonic model in Chapter 3.

5.3.2.3 Estimation of the LCCM

Estimation of the LCCM is performed in Python using custom Numpy code developed by the author, and is incorporated into a new implementation of UrbanSim [Waddell, 2002b] as open source software available at *https://github.com/synthicity/urbansim*. Estimation is performed using the Expectation Maximization algorithm until the difference in the log-likelihood for successive iterations is smaller than a certain threshold (here 1e-6 is used as the threshold for convergence). The estimation process is thus an expectation step in which the NC+1 models are simulated and the resulting probabilities are used to compute a log-likelihood for the overall model followed by a maximization step in which NC+1 models are independently estimated to local convergence where NC is the number of classes specified by the modeler. A loop is then executed around each EM step until the model reaches global convergence. The reader is referred to the code in the UrbanSim project for a reference implementation, and to the methodology in Vij [2013] which is followed precisely in this work.

5.3.3 Data and Methodology for Estimating Logsums

This chapter builds on Chapter 4 of this dissertation which computes logsums of accessibility to destinations. Although interested readers are referred to the previous chapter for detailed information, the methodology is summarized briefly here.

The population of all possible destinations is known by using a clean and thorough point-of-interest dataset provided by Factual, and the regional travel survey (CHTS, described above) provides an estimation dataset that describes how people tradeoff modes and destinations. The specific model used is a nested mode-destination choice model where destinations are street intersections which have attractors (the number of destinations at the intersection) and impedances (the travel times by mode) which are used as explanatory variables.

Specific trip purposes in the survey are linked to specific destinations so that micro-scale land use can be linked to its associated travel behavior. In some cases, the correspondence is perfect - e.g. eat out trips with restaurant destinations - in other cases, like "social" trip purposes, the set of destinations is less well defined (see Chapter 4 for the complete list). The logsums are then weighted by a trip generation model so that the result is an indexed logsum that is a weighted average of purpose-specific logsums where the weights are the number of trips that are made by that purpose. The assumption implicit in this methodology is that accessibility matters more for trips that are made more frequently, in direct proportion to the frequency of trips made for each purpose.

The result of the previous phase of the analysis is a complete set of logsums by mode for each income category (see Chapter 4 for the income category definitions, which are the same as those from CHTS) and for each possible home location, where street intersections are also used to represent the population of home locations as consistent with "street node geography" described in Chapter 2.

5.4 Results

For comparison purposes, Table 5.1 shows coefficients for a non-segmented MNL model. Although in a typical choice situation attributes of the user would be used as interaction terms in this model, here the specification is used that is exactly the same as the class-specific models described below for comparison purposes. The coefficients for the unsegmented model are positive on walk (although very small and with a p-value of .1) and negative on transit and auto logsums, with the coefficient on the auto logsum being larger and more significant. Price is negative and interacting income with price is

	Mode	Beta	Stderr	T-score	Significance
Unsegmented MNL	Walk Logsum	0.01	0.01	1.20	.
	Transit Logsum	-0.02	0.01	-1.73	*
	Auto Logsum	-0.14	0.01	-9.92	***
	Average Income	-0.05	0.02	-2.22	*
	Residential Units	1.11	0.02	50.65	***
	Price	-0.99	0.03	-34.69	***
	Income x Price	0.24	0.01	275.72	***
	Null Loglik	-34261			
	Final loglik	-32206			
	Loglik Ratio	0.06			

Table 5.1: Unsegmented MNL Estimation Results

positive and both are highly statistically significant. Residential units associated with the street intersection is positive and acts as a supply variable. Finally, average income in the neighborhood is a small negative value which is counter to expectations and this varies significantly among the different classes in the class-specific model.

The number of latent classes is determined by the modeler, and the preferred model in this case uses three classes. The determination of the number of classes is discussed in more detail later, but there appear to only be three large classes in the dataset and the introduction of more classes is not informative behaviorally nor does it improve the fit of the model much.

5.4.1 Class-specific Models

Coefficients for the class-specific models are shown in Table 5.2. Class 1 has positive and significant coefficients on walking and transit logsums but negative on the auto logsums, with a small negative coefficient on price and a small negative coefficient on average income. For Class 1, the coefficient on the walk logsum is much larger than the coefficient on the transit logsum. The magnitude of the coefficient for residential units is much larger than for the other classes, which indicates this class's proclivity for density which is seen in Figure 1. For Class 1, all coefficients are highly statistically significant.

In Class 2, the coefficient on transit is positive but auto is negative and walk is insignificant. Although there are few places in the Bay Area where transit access is high and walk access is not, this group is clearly most driven by transit access. Price is negative and income interacted with price is positive as in the other classes, and residential units is positive but not as large as in Class 1. Average income in the neighborhood is still negative for this class. All coefficients but walk logsums are highly statistically significant for this class.

In Class 3, the coefficients on all accessibility logsums are negative indicating that this

	Mode	Beta	Stderr	T-score	Significance
Class 1	Walk Logsum	0.79	0.07	12.08	***
Urbanites	Transit Logsum	0.31	0.05	6.42	***
	Auto Logsum	-0.37	0.06	-5.87	***
	Average Income	-0.54	0.08	-7.16	***
	Residential Units	2.17	0.08	25.87	***
	Price	-0.74	0.11	-6.58	***
	Income x Price	0.31	0.01	113.86	***
Class 2	Walk Logsum	0.01	0.03	0.03	
Transit-oriented	Transit Logsum	0.08	0.02	3.80	***
	Auto Logsum	-0.18	0.03	-7.20	***
	Average Income	-0.29	0.03	-9.35	***
	Residential Units	1.34	0.04	33.98	***
	Price	-1.32	0.05	-15.12	***
	Income x Price	0.35	0.01	-27.69	***
Class 3	Walk Logsum	-0.05	0.03	-1.97	*
Suburbanites	Transit Logsum	-0.04	0.02	-2.40	**
	Auto Logsum	-0.17	0.02	-8.14	***
	Average Income	1.09	0.03	31.28	***
	Residential Units	0.92	0.03	28.90	***
	Price	-2.99	0.05	-61.31	***
	Income x Price	0.43	0.01	242.09	***

Table 5.2: Class Specific Estimation Results

class is not drawn to high-accessibility areas. Price is still quite negative with a large interaction term, although it should be noted that the income level of this class is higher on average than the other two. For this class the average income of the neighborhood is large, positive, and statistically significant. The coefficient on residential units is the smallest of the three classes. Together this creates a suburban orientation which is clearly seen in Figure 3.

5.4.2 Mapping the PDF

As residential location choice is inherently a spatial process, the preferences of each latent class can be mapped for a far more intuitive understanding of the tradeoffs each class is making. To do this, each class-specific model is configured to simulate the entire probability density function (PDF) so that a probability is computed for each of the 226,000 street intersections that are used as alternatives. These probabilities are then mapped to the nearby parcels shapes which are colored using equal interval coloring and projected on top of aerial imagery for context.

The result is shown in Figures 1, 2, and 3, which map the PDF for each of Class 1, 2, and 3 respectively. In the maps, darker colors indicate higher probability areas for the location preferences of each class. Legends are not provided as the numbers are very small probabilities (due to the large number of alternatives), and don't provide additional understanding. The probabilities are mapped with an equal interval color scheme so that the colors themselves should provide an intuitive understanding of the relative probability of a class choosing the darker, more colorful, areas.

The behavior of the classes is explained in more detail in the discussion section below, but from the maps it is immediately apparent that Class 1 is the set of households that select into the dense and expensive urban core of San Francisco, as well as the urban areas in Downtown Oakland and Downtown Berkeley (thus dubbed the Urbanites), Class 2 is the second largest class and tends to move into the poorer neighborhoods in East Oakland, Richmond, and far out on the Pittsburgh/Bay Point and Dublin/Pleasanton BART lines (thus titled the Transit-oriented), and Class 3 is by far the largest and selects into the expensive suburbs around San Francisco including Orinda and Lafayette, Oakland Hills, Marin County, and the auto-oriented sections of San Francisco (thus named the Suburban Commuters).

5.4.3 Class-membership Model

The class-membership model is given in Table 5.3. As in all discrete choice models, the alternatives (in this case the 3 classes) can only be measured relative to a reference alternative, which in this case is Class 1. Compared to Class 1, Class 2 (the Transit-oriented) has a negative coefficient on income and non-related, and significant positive coefficients are estimated for household size, unit ownership, employment ratio, detached housing, older age of head of household, and presence of children and young children. Compared to Class 1, Class 3 has a negative and significant coefficient on non-related households and positive and significant coefficients on household income, household size, owner occupied, age of head of household, and children and young children.

5.4.4 A Note on the Number of Classes

The number of classes is a configuration parameter, and the choice of preferred model must be based on the log-likelihood measures, the relative class sizes, and the ease of interpretation of the coefficients in both the class specific and class membership models. For this dataset, the three class model is intuitive to understand, the log-likelihood doesn't improve much with additional classes, and any additional classes beyond three are very small. Although there is some variation from run to run due to the random starting coefficients, typical statistics for the three, four, and five class models are shown in Table 5.4. In fact, in many cases a four or five class model tends to converge to the three class model presented here, depending on the initial sample of alternatives and random starting coefficients.

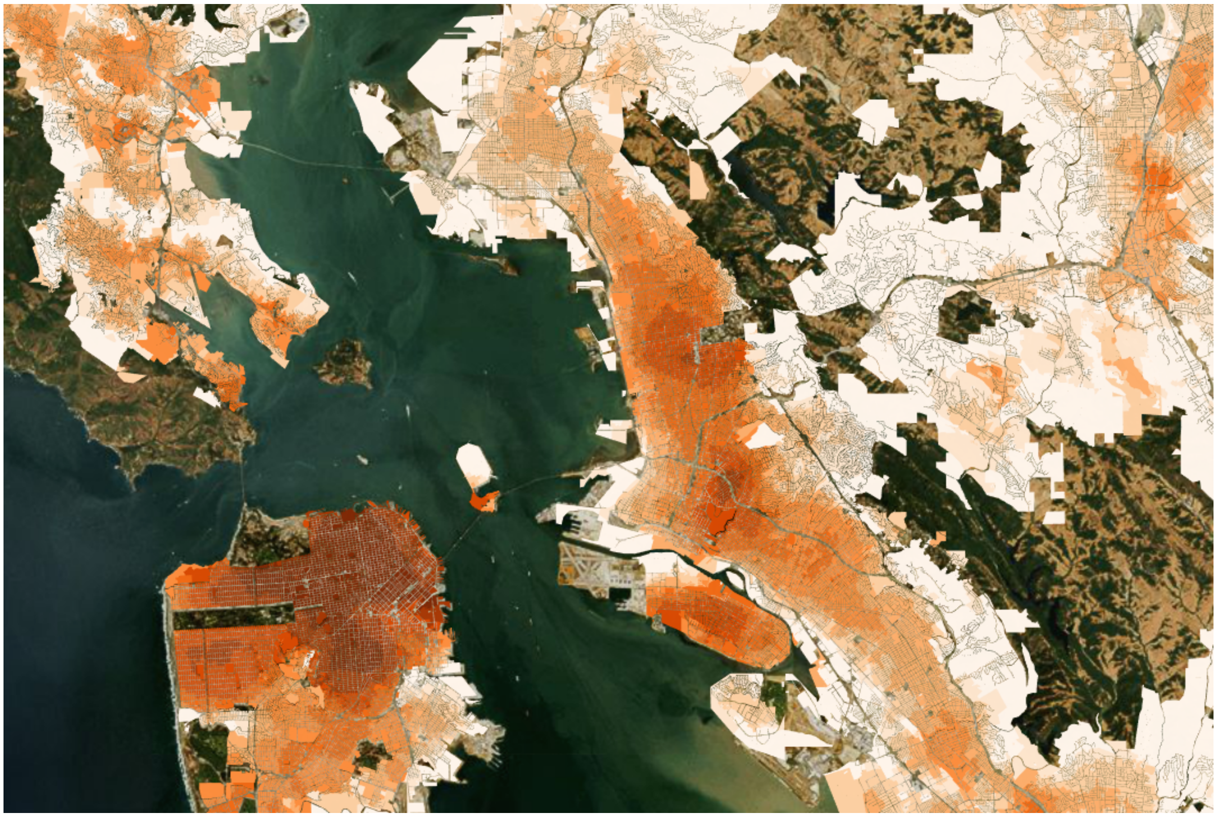


Figure 5.1: A map of the probability density function (PDF) for the residential choice of households in Class 1 (no legend is used because the probabilities are very small; darker colors indicate relatively higher probability areas)

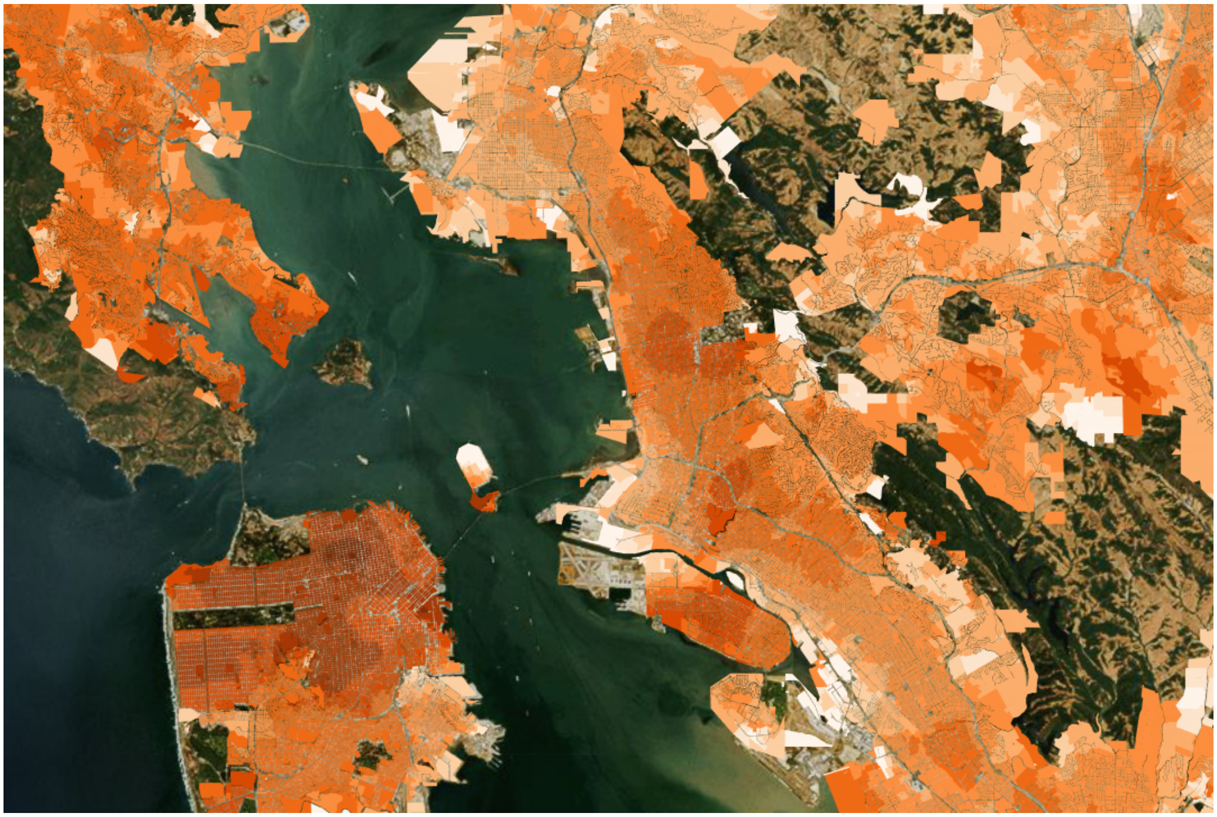


Figure 5.2: A map of the PDF for residential choice of households in Class 2

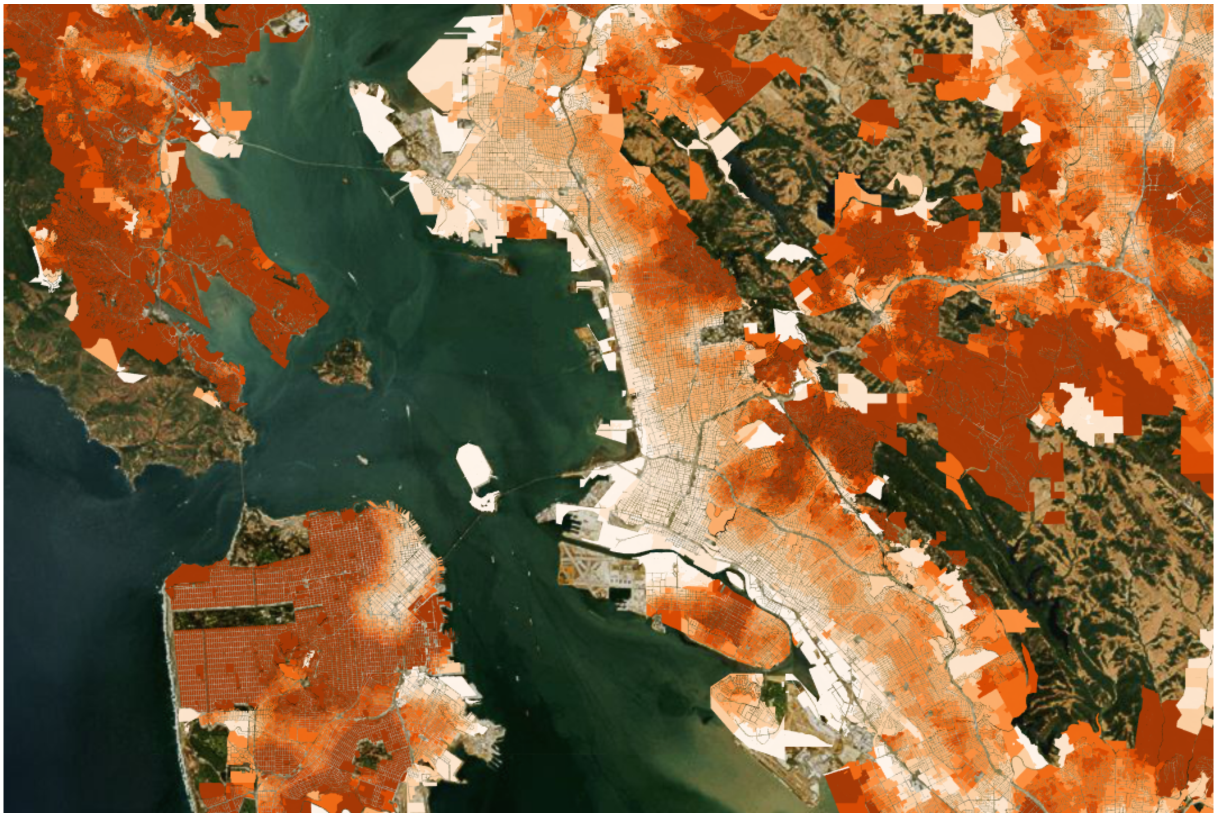


Figure 5.3: A map of the PDF for residential choice of households in Class 3

	Variables	Coefficient	Stderr	T-score	Significance
Class 2	Household Income	-0.79	0.12	-6.60	***
Transit-oriented	Household Size	8.36	0.24	34.96	***
	Owner Occupied	37.05	0.55	67.04	***
	Employed Ratio	1.75	0.46	3.81	***
	Detached Unit	33.56	0.50	66.26	***
	Non-related Hhld	-11.44	0.78	-14.67	***
	Age of head	1.47	0.20	7.22	***
	Young Children	36.23	1.33	27.16	***
	Children	16.86	0.70	24.17	***
	Constant	-18.50	1.20	-15.38	***
Class 3	Household Income	0.52	0.20	2.62	***
Suburbanites	Household Size	9.37	0.27	34.24	***
	Owner Occupied	40.01	0.73	55.06	***
	Employed Ratio	-	-	-	
	Detached Unit	-	-	-	
	Non-related Hhld	-14.44	1.05	-13.78	***
	Age of head	1.40	0.24	5.76	***
	Young Children	38.55	1.39	27.70	***
	Children	16.20	0.73	22.08	***
	Constant	-	-	-	
	Null Loglik	-34261			
	Final loglik	-31103			
	Loglik Ratio	0.09			

Table 5.3: Class Membership Estimation Results

	3 class	4 class	5 class
Null Loglik	-34261	-34261	-34261
Final Loglik	-31103	-31080	-31060
Loglik Ratio	.09	.09	.09
Class 1 Size	1.24M	1.23M	1.23M
Class 2 Size	.67M	.67M	.71M
Class 3 Size	.45M	.45M	.37M
Class 4 Size		4K	7K
Class 5 Size			6K

Table 5.4: Comparison of 3, 4, and 5 class latent class models (class sizes are sorted from largest to smallest)

5.5 Discussion

The demographic makeup of the classes can be easily understood by simulating the class membership model on the households in the survey and making a monte carlo choice based on the probabilities that are derived from the model. Each household also has a weight which is a statistical measure of the number of households that household represents in the entire population (in order to reach demographic marginals known to be true for the population). In this way, simulating for the surveyed population and multiplying by the expansion weights results in a demographic distribution for each class in the entire population. Table 5.5 presents the descriptive statistics of such a sample enumeration, giving estimates of the various statistics in the entire population of the Bay Area. The three classes are described in words below based on these descriptive statistics.

- Class 1 - the Urbanites - live in the highly urban areas of San Francisco, as well as Downtown Oakland and Downtown Berkeley. Since these households live near the urban core, they have a positive coefficient on accessibility by walking and transit, with the largest coefficient on walk accessibility of the three classes. They are the second highest segment in terms of average household income (to Class 3). The class membership model tells us that Urbanites tend to have the smallest household sizes, the largest rate of one person households, the highest ratio of employed household members to total household members, the smallest ratio of vehicles to household members, the smallest percentage of young children and school-age children, with the lowest rate of home ownership and the lowest rate of single family household occupancy. Class 1 accounts for 13% of households but because of their small average household size, members of these households account for only 7% of the population. To describe them in one sentence, they are young professional single person households (56% live alone) or roommates who rent in highly walkable and relatively expensive urban areas.

- Class 2 - the Transit-oriented - are the second largest group of Bay Area households, which are largely defined by the lowest income of the three classes and their associated high sensitivity to price. These households have low rates of car ownership (though the Urbanites are lower still) and so tend to locate near transit stations in East Oakland, Fremont, Richmond, and along the distant stops of the Pittsburgh/Bay Point and Dublin/Pleasanton BART lines. They have the second highest rate of single family housing, the second highest rate of home ownership, the second highest rate of employment, the second highest rate of one person households, and the second highest rate of both young children and school-age children present in these households. They contain 36% of households and 34% of the total population. To describe them in one sentence, they are the lower income households which tend to live near the subset of BART stations which are surrounded by low cost housing; many of the neighborhoods are high crime, but many of these households are raising families and own houses as these are the areas where housing is relatively affordable in the exorbitantly expensive Bay Area.

- Class 3 - the Suburban Commuters - are by far the largest group of Bay Area households, and are the highest income households, with a median income twice that of the other two classes. They tend to locate in the desirable suburbs of the Bay Area, including Marin County, the Oakland Hills, Orinda and Lafayette, Pacific Heights and Twin Peaks in San Francisco, and Silicon Valley. They live in almost exclusively owner occupied single family housing, with a median of almost 14 years in their current residence. These are largely families, with a median household size of three, while 36% contain school-age children and 12% contain young children (only 4% are one person households). Only 11% of these households have an age of head of household which is less than 40 years old. The median household has as many vehicles as people and half as many employees as people (though given the large household size these can still be two earner households). This group comprises 50% of all households and 59% of the Bay Area population. To describe them in one sentence, these are middle class and high-income households raising children in the idyllic suburbs of the Bay Area.

There are a number of interesting conclusions that can be drawn from analyzing the makeup of the three latent classes and their location preferences. In general, Class 1 is the young professionals selecting into high rent areas in the urban core, and appear to have traded off home ownership and single family housing for walkable neighborhoods and a greater ability to foster nascent careers and social connections. Class 3 is the middle class and high income families that have chosen to live in the idyllic suburbs around San Francisco; this class appears to have traded off accessibility for bigger houses, larger lots, and safer neighborhoods. Class 2 is the lowest income of the three, and is predominately located in the lower-cost neighborhoods near BART stations. This class appears to have traded off walk accessibility (which has an insignificant coefficient) for lower cost housing, while maintaining access to the larger region through the public transit network.

The three classes do appear to have heterogeneous impacts of accessibility. For the first, smallest class walking accessibility is a driving force in the residential location decision, for the second class transit accessibility has the largest and most significant coefficient, and the third class tends to have a generally negative relationship with accessibility. The class which values walking accessibility the most - Class 1 - is composed of young, moderately high income professionals with a large proportion of single member households, a demographic class which research indicates has been growing considerably in recent years Nelson [2009].

Although Walker and Li [2007] also found three classes when applying similar methodology to the residential location choice, the previous study found an urban class with auto tendencies and a suburban class with transit tendencies. Both of these behaviors might be present in Class 3 of this study, with auto-oriented neighborhoods in San Francisco being as likely as BART-accessible neighborhoods in the suburbs for this class, the unifying factor being the high-value housing, high-income neighborhoods, and access to high-paying

Class	1	2	3
Size (number of households)	324,189	859,656	1,164,581
Percent of households	13%	37%	50%
Total Population	428,282	1,998,452	3,532,020
Percent of population	7%	34%	59%
Percent Single Family Housing	2%	74%	99%
Percent Owner Occupied	31%	78%	96%
Years Lived in Residence (median)	9	12	14
Income Class Category (median)	50K-75K	50K-75K	150K-200K
Ratio of Employed to Total People (median)	1.0	.75	.50
Ratio of Vehicles to Total People (median)	.50	.75	1.0
Household Size (median)	1	2	3
Percent One Person Households	56%	14%	4%
Age of Head of Household (median)	53	55	58
Percent Young HH (age of head < 40)	23%	16%	11%
Percent Young Children	1%	9%	12%
Percent School-age Children	2%	23%	36%

Table 5.5: Descriptive statistics for the 3 latent classes

jobs. It seems possible that the previous study, which was performed on a 1994 dataset from Portland, OR, did not have a truly walking-oriented class like Class 1 in the present study.

5.6 Conclusions

This chapter has described a methodology for using latent class choice models to allow for heterogeneous preferences in the residential location choice decision as estimated using the Bay Area portion of the 2012 California Household Travel Survey. LCCMs are comprised of a class membership model which uses the diverse set of demographic descriptors present in CHTS to assign classes of behavior to households, and a number of class-specific models which with independent coefficients to allow for differing preferences among classes. This chapter uses a specification for location choice with accessibility by auto, walk, and transit, supply of residential units, average income in the neighborhood, and average price as independent variables. Alternatives are the roughly 220 thousand street intersections in the Bay Area.

The preferred model reveals that there are three large classes of behavior in the travel survey, and that additional classes are relatively small. Of the three classes, one is comprised of young professionals which locate primarily in the urban core and have positive coefficients on transit and walk accessibility with the largest and most significant coefficient for walking, the second is the lowest income and tends to locate in low-cost neighborhoods around BART stations (with a negative coefficient on auto access and walking accessibility is not statistically significant) and the third is the highest income and least price sensitive which locates in the idyllic suburbs around San Francisco, with a negative relationship to all accessibility measures when controlling for price, residential supply, and average neighborhood income.

It should be noted that schools and crime are not included as covariates in this study, as all data in this study is available regionally except for these two datasets. Future work should perform the methodology presented in this chapter on a subset of the Bay Area where crime and school data are available so as to properly control for these variables. Nonetheless it is likely that the neighborhood income and price variables are highly correlated with crime and school quality, and since the coefficients on income and price are used as controls, it is possible that the impact of crime and schools is controlled for adequately through these proxy variables.

It should also be noted that in the Bay Area the set of variables used here to explain residential location choice is quite correlated. High income neighborhoods, price, residential density, and accessibility by all three modes essentially are different measures of development density. In particular, residential units and the walking logsums are highly correlated, with a Pearson correlation of .65. In light of the large previous literature discussing the relationship of residential density and pedestrian accessibility, it is worth noting that in this dataset and using this methodology, these two variables are significantly correlated, so much so that it's clear that a trivial-to-compute summation of residential units captures almost all of the impact of the extensive methodology presented here which attempts to account for available destinations, trip generation, mode substitution, and income effects.

This has two implications. First, previous studies which use residential density as an explanatory variable in travel behavior are likely capturing walking behavior by proxy, at least for Bay Area studies, which is usually the stated purpose of using residential density in these studies. Second, the residential units measure which in this paper is used to control for availability of residential units (i.e. the supply variable), is also highly correlated with walkability so that some of the explanatory power currently attributed to supply is probably due to the positive correlates of density such as nearby walkable destinations.

To answer the research questions posed in this study, it appears that nearly half the households in the Bay Area are not driven by accessibility considerations at all, and that only 13% of households are truly preferring of walkable neighborhoods. Nonetheless, there is a large segment of households - 37% - which are lower-income and transit-oriented and which are often overlooked by Bay Area planning processes. For planning interventions which seek to increase travel by active modes, members of this segment might have the most latent potential to change their behavior. As Class 1 is probably well served by pedestrian neighborhoods already (and is relatively small) and Class 3 has no apparent proclivity to high-accessibility areas, it should be noted that Class 2 is a large and seemingly underserved population living in neighborhoods around BART stations whose pedestrian environments could be easily upgraded.

A ripe area for future research is to perform a gap analysis which compares neighborhoods that are high probability areas for each of the three classes presented here to test for the impact of increases in transit service and pedestrian infrastructure on both the residential location choice and travel behavior. Taking into account the heterogeneity of preferences explored here, the result of such a study would target the locations which could have the highest impact on sustainable behavior for the smallest amount of public investment.

Chapter 6

Conclusions

6.1 Summary of key findings

In Chapter 2, an accessibility framework is created to allow the computation of the distance to all destinations from all locations in the city. This allows the measurement of accessibility to destinations by mode for each of walking, transit, and auto, as well as WalkScore-like disaggregate indexes, and “3Ds” density, diversity and design variables. A case study application for trip generation for home-based walking trips in the Bay Area portion of the 2012 California Household Travel Survey finds that WalkScore is descriptive of walking trips, but that residential density and 4-way intersections have an additional but small impact, and that regional access by the transit network has a synergistic effect on walking, while regional access by auto has no impact.

In Chapter 3, the relationship between accessibility to local amenities and home values is tested. Although early research has found that the composite index WalkScore is positively related to home values, this research unpacks the impact of each category of destination in WalkScore (as well as several other categories) on home values. The hedonic model presented in Chapter 3 shows that accessibility to certain amenities is far more predictive of home values than others in the datasets used in this analysis; in particular, cafes and coffee shops tend to be the indicator of neighborhood-scale urban fabric that has by far the largest positive impact on home values, where a one standard deviation increase in access to cafes is associated with almost a 15% increase in home values.

Chapter 4 expands on the findings from Chapter 3 by providing evidence that the willingness to walk to these destinations (as opposed to making short driving trips) is valued when purchasing the home. In this chapter, logsums from a nested mode-destination model (using the networks from Chapter 2) are generated that are an accurate measure of the actual choice to walk to the complete set of destinations around each home location. Logsums by purpose are weighted by the number of trips made for each purpose, and the models are segmented by income and then income-specific logsums are averaged over all people that live in each location in the city. The result is an empirically estimated index of walking, transit, and auto accessibility for each location in the city.

These indexes are then used as right-hand side variables in the hedonic model from Chapter 3. Not surprisingly, travel by all 3 modes has a positive relationship to home values, and a one standard deviation change in the weighted logsum for automobile has the greatest impact on home values, although the impact of walking accessibility on home values is positive, statistically significant, and almost as large. Interestingly, the variance in availability of neighborhoods that enable walking is much larger than the variance in neighborhoods that enable driving probably due to the relatively higher supply of drivable neighborhoods, but the results from this chapter show that where walkable neighborhoods are available there is unequivocal evidence that increased walking is related to increased home values for the datasets analyzed here.

Chapter 5 then investigates the residential location choice using latent class choice models (LCCMs) so that coefficients on the three mode-specific indexes, as well as price, residential supply, and average income in the neighborhood, are allowed to change based on selection into unobserved classes. This is a type of consumer preference segmentation aimed at elucidating the impact of mode-specific accessibility to destinations on the residential location decision.

The model shows that there are three large segments present in the Bay Area. The first likely captures the young, moderately high-income professionals that select into the walkable neighborhoods of urban San Francisco, Oakland, and Berkeley, though this segment is only 13% of Bay Area households. The second is transit-oriented and selects into the subset of relatively less expensive neighborhoods near BART stations but outside the urban core, and comprises 37% of all households, and the third which is composed of middle class families that prefer the idyllic suburbs outside San Francisco. This last segment seems to have a generally negative relationship with accessibility for all modes and comprises 50% of all Bay Area households. The main research topic explored in Chapter 5 is the question of the size of the segment of the population that is positively affected by walking accessibility for the residential location choice and the results show that this segment exists but is of modest size.

However, a major finding of this research is that for planning interventions that seek to increase travel by active modes, members of the transit-oriented segment might have the most latent potential to change their behavior. Perhaps creating denser and more walkable environments around the less expensive neighborhoods near BART stations in the region could relieve pressure on the San Francisco housing market as well as create walkable environments for the lower middle class that appear to be a major component of residential demand in the region.

A ripe area for future research is to perform a gap analysis that compares neighborhoods that are high probability areas for each of the three classes presented here to test for the impact of increases in transit service and pedestrian infrastructure on both the residential location choice and travel behavior. Taking into account the heterogeneity of preferences explored here, the result of such a study would target the locations which could have the highest impact on sustainable behavior for the smallest amount of public investment.

6.2 Contributions to the literature

One contribution of this dissertation to the literature has been to address the question of the impact of WalkScore as a predictor of both walking and of increased home values. This dissertation uses an empirical dataset (CHTS 2012) to estimate a nested destination choice model which increases accuracy of WalkScore in five ways: 1) the distance decay function can be estimated based on how far people are observed to travel for each mode, 2) the set of destinations can be matched against the specific trip purposes from CHTS 2012 as well as the set of destinations from the point-of-interest dataset, 3) the relative contribution of each destination can be weighted by the trip generation observed in CHTS

2012, 4) the model can be segmented by income so that different travel preferences for each income class are taken into account, 5) perhaps most importantly, the actual willingness to walk to destinations can be related to the set of destinations that are available by other modes, particularly auto.

The contribution of this work is thus a more rigorous, empirically based index that measures walking rather than walkability. This index (and the index for auto and transit) can then be used in models of home values to determine the impact of accessibility by walking on home values. Accessibility logsums for all three modes are positive, and although the impact of a one standard deviation change in the index for auto is the largest, the index for walking is statistically significant and almost as large as the auto index. This is strong evidence that the ability to walk to destinations is being translated into high property values in the Bay Area.

The final contribution of this work is to create one of the first residential location choice latent class models, which allows the impact of the indexes described above to vary based on selection into unobserved classes. Interestingly, there are only 3 large segments: one that is young and urban, a second that is transit oriented and lives in less expensive areas near BART stations, and a third which has the highest average income and lives in the idyllic suburbs surrounding San Francisco.

The somewhat unexpected finding from this research is that although the first class does appear to have a large and significant coefficient on walking, at 13% of households this class is not large enough to drive the regional housing market. Instead there appears to be a large - 37% of households - segment of the population which is most influenced by access to transit, apparently due to the regional job access provided by the BART rail system. The impact of BART as a unifying factor in the spatial structure of the region for a significant number of households, perhaps even a plurality of households, seems clear.

6.3 Policy implications

This research has clear and broad planning implications for the San Francisco Bay Area, and possibly for other regions although this research will have be carefully executed elsewhere. The broad conclusion that walkability has a strong relationship to home values, and thus to rents and prices, implies that more of these walkable neighborhoods should be built in locations where the market supports development without subsidy. The additional fact that prices have risen so steeply in San Francisco, combined with the historical record of only 1,500 units per year for over a decade suggests that the market has been constrained by zoning policy in the areas of highest demand.

Indeed, the fact that 4,000 units are currently being built while 32,000 more units are being permitted suggests that developers are willing to build in San Francisco, but most development can only take place where there are few residents now - in Mission Bay, Treasure Island, and the Hunter's Point Shipyard. Densification is currently taking place along Market St. as well, but the opportunities for infill in the city as it is currently zoned are few. As most parcels are already nearly built to capacity (within zoning), little redevelopment can occur.

And perhaps it should not, as an argument can be made that development interests, social equity concerns, transit supporters, and historic preservationists all align on this issue. Even given the strong political will against yet more densification in SF, the strong and growing labor market requires that residential development must occur somewhere. In fact prime locations for this development exist, with those in proximity to the many BART stations in Oakland as prime candidates.

Creating walkable communities around the Oakland BART stations would preserve the historic neighborhoods of San Francisco, while providing a strong new tax base for Oakland and built-in ridership for BART, as well as possibly generating more resources for public safety, schools, and beautification in a city that desperately needs it. In addition, this research has found that at least 37% of the Bay Area residential housing market would prefer much cheaper housing that is BART-accessible, and possibly walkable. To date, the surfeit of development in San Francisco hasn't carried across the Bay to Oakland. Perhaps a unification of interests of the sort described here can solve many of the issues in the housing market on both sides of the Bay. In fact, laws such as SB375 mandate the consideration of alternatives like this in order to address the future emissions of GHGs - this research shows that these policy interventions might create neighborhoods that are also preferred by a large number of households in the Bay Area.

6.4 Future research

Much future research will need to occur related to these topics, as many of these issues need more nuanced understanding and empirical evidence from additional datasets. Of primary concern is to identify the locations where residential development is feasible at current rents and prices in the Bay Area; is development profitable only in San Francisco or would it also be profitable in Oakland? If further development is not profitable in Oakland, what would be the subsidy required to incent building there? Given that development clearly is profitable without subsidy in San Francisco (although this is not a given city-wide), how much additional capacity is there in the city of San Francisco within zoning restrictions and without subsidy? In places where more dense development than current height and bulk limits is feasible, what is an estimate of the amount of development (and tax revenue) foregone in San Francisco as a result of the zoning restrictions?

Additionally, the question of which transit lines are associated with residential preferences should be investigated further. Is BART the only transit service that encourages development or is there evidence that CalTrain and MUNI also have an impact on home prices and location choices? Does reliability and level-of-service of transit have an impact? Do investments in bicycle infrastructure have any resultant economic benefits like those found here for transit and walking?

Finally, there is an ever-expanding set of urban data from which to draw for additional research. New data sources like cell phones and mobile apps that automate travel diaries would help to record travel behavior more precisely. Additional quality information associated with the set of destinations (e.g. services like Yelp) would increase detail in understanding variation in land use. The quality of the pedestrian environment or a record of the transit or auto experience would be extremely helpful in understanding the specific quality of a transportation route. Finally, private companies now have access to a great deal of information on consumer segmentation, lifestyle preferences, and attitudes of much of the American population - clearly this data could contribute to the knowledge of how people travel and why they value the neighborhoods they do.

Bibliography

W. Alonso. *Location and land use: toward a general theory of land rent.* Harvard University Press, 1964.

Bilge Atasoy, Aurélie Glerum, and Michel Bierlaire. Mode choice with attitudinal latent class: A swiss case-study. In *2nd International Choice Modelling Conference, Leeds*, 2011.

Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

Elisa Barbour and Elizabeth A. Deakin. Smart growth planning for climate protection. *Journal of the American Planning Association*, 78(1):70–86, 2012.

Keith Bartholomew and Reid Ewing. Hedonic price effects of pedestrian-and transit-oriented development. *Journal of Planning Literature*, 26(1):18–34, 2011.

R. J Beckman, K. A Baggerly, and M. D McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30(6):415–429, 1996.

Vicki Been. Locally undesirable land uses in minority neighborhoods: disproportionate siting or market dynamics. *Yale LJ*, 103:1383, 1993.

M. Ben-Akiva, J. L Bowman, et al. *The day activity schedule approach to travel demand analysis.* PhD thesis, Massachusetts Institute of Technology, 1998.

M. E Ben-Akiva and S. R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. The MIT Press, 1985.

Moshe Ben-Akiva and John L. Bowman. Integration of an activity-based model system and a residential location model. *Urban Studies*, 35(7):1131–1153, 1998.

C. R Bhat and J. Y Guo. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transportation Research Part B: Methodological*, 41(5):506–526, 2007.

M. Boarnet and R. Crane. *Travel by design: The influence of urban form on travel.* Oxford University Press, USA, 2001.

M. G. Boarnet. A broader context for land use and travel behavior, and a research agenda. *Journal of the American Planning Association*, 77(3):197–213, 2011.

M. G Boarnet and S. Sarmiento. Can land-use policy really affect travel behaviour? a study of the link between non-work travel and land-use characteristics. *Urban Studies*, 35(7):1155, 1998. URL `http://usj.sagepub.com/content/35/7/1155.short`.

M. G. Boarnet, D. Houston, G. Ferguson, S. Spears, Y. H. Hong, and G. Ingram. Land use and vehicle miles of travel in the climate change debate: Getting smarter than your average bear. *Climate change and land policies*, page 151–187, 2011.

J. L Bowman. *The day activity schedule approach to travel demand analysis*. PhD thesis, Massachusetts Institute of Technology, 1998.

Melissa A. Boyle and Katherine A. Kiel. A survey of house price hedonic studies of the impact of environmental externalities. *Journal of real estate literature*, 9(2):117–144, 2001.

Mark Bradley, John L. Bowman, and Bruce Griesenbeck. SACSIM: an applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3(1):5–31, 2010.

Mark A. Bradley, John L. Bowman, and Bruce Greisenbeck. Activity-based model for a medium sized city: Sacramento. *Traffic engineering & control*, 50(2):73–79, 2009.

D. Brownstone. Key relationships between the built environment and VMT. *Transportation Research Board*, page 7, 2008.

X. Cao, P. Mokhtarian, and S. Handy. Examining the impacts of residential self-selection on travel behavior: methodologies and empirical findings. 2008.

Martin Catala, Samuel Dowling, and Donald Hayward. Expanding the google transit feed specification to support operations and planning. Technical report, 2011.

R. Cervero. Built environments and mode choice: toward a normative framework. *Transportation Research Part D: Transport and Environment*, 7(4):265–284, 2002.

R. Cervero and M. Duncan. Residential self selection and rail commuting: a nested logit analysis. Technical report, University of California Transportation Center, 2002.

R. Cervero and M. Duncan. Walking, bicycling, and urban landscapes: evidence from the san francisco bay area. *American journal of public health*, 93(9):1478, 2003.

D. G. Chatman. Deconstructing development density: Quality, quantity and price effects on household non-work travel. *Transportation Research Part A: Policy and Practice*, 42(7):1008–1030, 2008.

D. G Chatman. Residential choice, the built environment, and nonwork travel: evidence using new data and methods. *Environment and planning. A*, 41(5):1072, 2009.

Paul Cheshire and Stephen Sheppard. On the price of land and the value of amenities. *Economica*, page 247–267, 1995.

Paul Cheshire and Stephen Sheppard. Estimating the demand for housing, land, and neighbourhood characteristics. *Oxford Bulletin of Economics and Statistics*, 60(3): 357–382, 1998.

Terry Nichols Clark. The city as an entertainment machine (research in urban policy)(v. 9). 2003.

William AV Clark and Walter FJ Van Lierop. Residential mobility and household location modelling. *Handbook of regional and urban economics*, 1:97–132, 1987.

Regina Clelow, Fletcher Foti, and Paul Waddell. The socioeconomics of daily travel: Trends in the san francisco bay area. *Transportation Research Board Yearly Conference*, 2014.

K. Clifton, R. Ewing, G. J Knaap, and Y. Song. Quantitative analysis of urban form: a multidisciplinary review. *Journal of Urbanism*, 1(1):17–45, 2008.

Kelly J. Clifton, Andréa D. Livi Smith, and Daniel Rodriguez. The development and testing of an audit for the pedestrian environment. *Landscape and Urban Planning*, 80 (1):95–110, 2007.

Joe Cortright. Walking the walk: How walkability raises home values in US cities. 2009.

R. Crane. On form versus function: Will the new urbanism reduce traffic, or increase it? *Journal of Planning Education and Research*, 15(2):117–126, 1996.

R. Crane. The influence of urban form on travel: an interpretive review. *Journal of Planning Literature*, 15(1):3–23, 2000.

Daryl J. Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*, volume 2. Springer, 2007.

J. de Dios Ortúzar, L. G Willumsen, et al. *Modelling transport.* Wiley, 2001.

Yongheng Deng, Stephen L. Ross, and Susan M. Wachter. Racial differences in homeownership: the effect of residential location. *Regional Science and Urban Economics*, 33(5):517–556, 2003.

Denise DiPasquale and William C. Wheaton. *Urban economics and real estate markets.* Prentice Hall Englewood Cliffs, NJ, 1996.

X. Dong, M.E. Ben-Akiva, J.L. Bowman, and J.L. Walker. *Moving from trip-based to activity-based measures of accessibility [An article from: Transportation Research Part A]*. Elsevier, 2006.

Tony Dutzik and Phineas Baxandall. A new direction: Our changing relationship with driving and the implications for america's future. 2013.

Radcliffe G. Edmonds Jr. Travel time valuation through hedonic regression. *Southern Economic Journal*, page 83–98, 1983.

Naveen Eluru, Abdul R. Pinjari, Ram M. Pendyala, and Chandra R. Bhat. An econometric multi-dimensional choice model of activity-travel behavior. *Transportation Letters: the international journal of transportation research*, 2(4):217–230, 2010.

R. Ewing and S. Handy. Measuring the unmeasurable: urban design qualities related to walkability. *Journal of Urban Design*, 14(1):65–84, 2009.

Reid Ewing and Robert Cervero. Travel and the built environment: A synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, 1780(-1): 87–114, January 2001.

L. D Frank, J. Kerr, J. F Sallis, R. Miles, and J. Chapman. A hierarchy of sociodemographic and environmental correlates of walking and obesity. *Preventive medicine*, 47 (2):172–178, 2008.

R. Geisberger, P. Sanders, D. Schultes, and D. Delling. Contraction hierarchies: Faster and simpler hierarchical routing in road networks. *Experimental Algorithms*, page 319–333, 2008.

G. Giuliano and K. A Small. Subcenters in the los angeles region. *Regional Science and Urban Economics*, 21(2):163–182, 1991.

Edward L. Glaeser, Jed Kolko, and Albert Saiz. Consumer city. *Journal of economic geography*, 1(1):27–50, 2001.

Joshua D. Gottlieb and Edward L. Glaeser. Urban resurgence and the consumer city. *Urban Studies*, 43(8):1275–1299, 2006.

M. J Greenwald and M. G Boarnet. Built environment as determinant of walking behavior: analyzing nonwork pedestrian travel in portland, oregon. *Transportation Research Record: Journal of the Transportation Research Board*, 1780(-1):33–41, 2001.

C. Angelo Guevara and Moshe E. Ben-Akiva. Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B: Methodological*, 48: 31–52, February 2013.

J. Y Guo and C. R Bhat. Operationalizing the concept of neighborhood: Application to residential location choice analysis. *Journal of Transport Geography*, 15(1):31–45, 2007.

Jessica Y. Guo, Chandra R. Bhat, and Rachel B. Copperman. Effect of the built environment on motorized and nonmotorized trip making: Substitutive, complementary, or synergistic? *Transportation Research Record: Journal of the Transportation Research Board*, 2010(-1):1–11, 2007.

S. Handy. Methodologies for exploring the link between urban form and travel behavior. *Transportation Research Part D: Transport and Environment*, 1(2):151–165, 1996a.

S. L Handy. Understanding the link between urban form and nonwork travel behavior. *Journal of Planning Education and Research*, 15(3):183–198, 1996b.

S. L Handy and D. A Niemeier. Measuring accessibility: an exploration of issues and alternatives. *Environment and planning A*, 29:1175–1194, 1997.

W. G Hansen. How accessibility shapes land use. *Journal of the American Institute of Planners*, 25(2):73–76, 1959.

Michael K. Hidrue. *The Demand for Conventional and Vehicle-to-grid Electric Vehicles: A Latent Class Random Utility Model*. University of Delaware, 2011.

B. Hillier and K. Tzortzi. Space syntax. *A Companion to Museum Studies*, page 282–301, 1976.

Robert Ishaq, Shlomo Bekhor, and Yoram Shiftan. A flexible model structure approach for discrete choice models. *Transportation*, page 1–16, 2013.

Tae Youn Jang. Count data models for trip generation. *Journal of Transportation Engineering*, 131(6):444–450, 2005.

Jerald Jariyasunant, Maya Abou-Zeid, Andre Carrel, Venkatesan Ekambaram, David Gaker, Raja Sengupta, and Joan L. Walker. Quantified traveler: Travel feedback meets the cloud to change behavior. *Journal of Intelligent Transportation Systems*, (just-accepted), 2013.

N. Jonnalagadda, J. Freedman, W. A Davidson, and J. D Hunt. Development of microsimulation activity-based model for san francisco: destination and mode choice models. *Transportation Research Record: Journal of the Transportation Research Board*, 1777 (-1):25–35, 2001.

John F. Kain and John M. Quigley. Measuring the value of housing quality. *Journal of the American Statistical Association*, 65(330):532–548, June 1970.

John F. Kain and John M. Quigley. The value of housing attributes. In *Housing Markets and Racial Discrimination: A Microeconomic Analysis*, page 190–230. NBER, 1975.

Wagner A. Kamakura and Gary J. Russell. A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, page 379–390, 1989.

Ryuichi Kitamura, Lidia P. Kostyniuk, and Kuo-Liang Ting. Aggregation in spatial choice modeling. *Transportation Science*, 13(4):325–342, 1979.

Kara Kockelman and Robert Cervero. Travel demand and the 3Ds: density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3):199–219, 1997.

K. J Krizek. Neighborhood services, trip purpose, and tour-based travel. *Transportation*, 30(4):387–410, 2003.

Paul F. Lazarsfeld and Robert K. Merton. Friendship as a social process: A substantive and methodological analysis. *Freedom and control in modern society*, 18(1):18–66, 1954.

B. H.Y Lee and P. Waddell. Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation*, 37(4):587–601, 2010.

B. H.Y Lee, P. Waddell, L. Wang, and R. M Pendyala. Reexamining the influence of work and nonwork accessibility on residential location choices with a microanalytic framework. *Environment and Planning A*, 42(4):913–930, 2010a.

Brian HY Lee, Paul Waddell, Liming Wang, and Ram M. Pendyala. Reexamining the influence of work and nonwork accessibility on residential location choices with a microanalytic framework. *Environment and planning. A*, 42(4):913, 2010b.

C. B Leinberger and M. Alfonzo. Walk this way. *National Post. Accessed online at www. cleinberger. com/AdminHome. asp*, 2005.

Steven R. Lerman. Location, housing, automobile ownership, and mode to work: a joint choice model. *Transportation Research Record*, (610), 1976.

K. Lynch. *The image of the city*. Mit press, 1960.

June Ma and Konstadinos G. Goulias. Application of poisson regression models to activity frequency analysis and prediction. *Transportation Research Record: Journal of the Transportation Research Board*, 1676(1):86–94, 1999.

K. Manaugh and A. El-Geneidy. Validating walkability indices: How do different households respond to the walkability of their neighborhood? *Transportation Research Part D: Transport and Environment*, 2011.

John W. Matthews and Geoffrey K. Turnbull. Neighborhood street layout and property value: the interaction of accessibility and land use mix. *The journal of real estate finance and economics*, 35(2):111–141, 2007.

D. McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, page 13–29, 1980.

Daniel McFadden. *Modelling the choice of residential location.* Institute of Transportation Studies, University of California, 1978.

Gabriel Metcalf. The san francisco exodus, October 2013. URL `http://www.theatlanticcities.com/housing/2013/10/san-francisco-exodus/7205/`.

H. J Miller. Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures. *Geographical Analysis*, 31(1):1–26, 1999.

A. V Moudon, C. Lee, A. D Cheadle, C. Garvin, D. Johnson, T. L Schmid, R. D Weathers, and L. Lin. Operational definitions of walkable neighborhood: theoretical and empirical insights. *Journal of Physical Activity & Health*, 3:99, 2006.

R. F Muth. *Cities and housing: the spatial pattern of urban residential land use.* University of Chicago Press, 1969.

A. C. Nelson. The new urbanity: The rise of a new america. *The ANNALS of the American Academy of Political and Social Science*, 626(1):192–208, 2009.

Jon P. Nelson. Accessibility and the value of time in commuting. *Southern Economic Journal*, page 1321–1329, 1977.

Jeffrey P. Newman and Vincent L. Bernardin Jr. Hierarchical ordering of nests in a joint mode and destination choice model. *Transportation*, 37(4):677–688, 2010.

Doina Olaru, Brett Smith, and John HE Taplin. Residential location and transit-oriented development in a new rail corridor. *Transportation Research Part A: Policy and Practice*, 45(3):219–237, 2011.

James Parks and Joseph Schofer. Characterizing neighborhood pedestrian environments with secondary data, 2006.

R. M. Pendyala. Phased implementation of a multimodal activity-based travel demand modeling system in florida. volume II: FAMOS users guide. Technical report, 2004.

R. M Pendyala, R. Kitamura, A. Kikuchi, T. Yamamoto, and S. Fujii. Florida activity mobility simulator: Overview and preliminary validation results. *Transportation Research Record: Journal of the Transportation Research Board*, 1921(-1):123–130, 2005.

A. R. Pinjari, R. M. Pendyala, C. R. Bhat, and P. A. Waddell. Modeling residential sorting effects to understand the impact of the built environment on commute mode choice. *Transportation*, 34(5):557–573, 2007.

J. Pucher and J. L Renne. Socioeconomics of urban travel: evidence from the 2001 NHTS. *Transportation Quarterly*, 57(3):49–77, 2003.

John M. Quigley. Housing demand in the short run: An analysis of polytomous choice. In *Explorations in Economic Research, Volume 3, number 1*, page 76–102. NBER, 1976.

John M. Quigley. Consumer choice of dwelling, neighborhood and public services. *Regional Science and Urban Economics*, 15(1):41–63, 1985.

José Ramírez, Caroline Schaerer, and Philippe Thalmann. *Hedonic methods in housing markets [electronic resource]: pricing environmental amenities and segregation.* Springer, 2008.

Stephanie Yates Rauterkus and Norman G. Miller. Residential land values and walkability. *The Journal of Sustainable Real Estate*, 3(1):23–43, 2011.

Martin G. Richards and Moshe Ben-Akiva. A simultaneous destination and mode choice model for shopping trips. *Transportation*, 3(4):343–356, December 1974. ISSN 0049-4488, 1572-9435.

Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *The journal of political economy*, 82(1):34–55, 1974.

AnnaLee Saxenian. *Regional advantage: Culture and competition in Silicon Valley and Route 128.* Harvard University Press, 1996.

A. Sevtsuk. *Path and Place: A Study of Urban Geometry and Retail Activity in Cambridge and Somerville, MA*. PhD thesis, Massachusetts Institute of Technology, Dept. of Urban Studies and Planning, 2010.

Claude Shannon. *The mathematical theory of communication.* The University of Illinois Press, Urbana, paperback ed. edition, 1948.

Daniel Silver, Terry Nichols Clark, and Clemente Jesus Navarro Yanez. Scenes: Social context in an age of contingency. *Social Forces*, 88(5):2293–2324, 2010.

Daniel Silver, Terry Nichols Clark, and Christopher Graziul. 12 scenes, innovation, and urban development. *Handbook of creative cities*, page 229, 2011.

Donald Lee Snyder. Random point processes. 1975.

Kenneth L. Sobel. Travel demand forecasting by using the nested multinomial logit model. *Transportation Research Record*, 775:48–55, 1979.

Yan Song and Gerrit-Jan Knaap. Measuring the effects of mixed land uses on housing values. *Regional Science and Urban Economics*, 34(6):663–680, 2004.

M. Southworth. Designing the walkable city. *Journal of urban planning and development*, 131:246, 2005.

M. Southworth and P. M Owens. The evolving metropolis: studies of community, neighborhood, and street form at the urban edge. *Journal of the American Planning Association*, 59(3):271–287, 1993.

SPUR. In san francisco, the boom is back, December 2012. URL `http://www.spur.org/publications/article/2012-12-18/san-francisco-boom-back`.

Brian Deane Taylor and Camille NY Fink. *The factors influencing transit ridership: a review and analysis of the ridership literature*. University of California Transportation Center, 2003.

C. M Tiebout. A pure theory of local expenditures. *The journal of political economy*, 64 (5):416–424, 1956.

Kenneth Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

Kenneth E. Train. EM algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.

TRB. *Metropolitan travel forecasting: current practice and future direction*, volume 288. Natl Academy Pr, 2007.

A. Vij, A. Carrel, and J. L Walker. Capturing modality styles using behavioral mixture models and longitudinal data. In *2nd International Choice Modelling Conference, Leeds*, 2011.

Akshay Vij. Incorporating the influence of latent modal preferences in travel demand models. *Dissertation*, 2013. URL `http://uctc.net/research/UCTC-DISS-2013-04.pdf`.

P. Waddell. Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297–314, 2002a.

P. Waddell. UrbanSim: modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297–314, 2002b.

P. Waddell, H. Ševcıková, D. Socha, E. Miller, and K. Nagel. Opus: An open platform for urban simulation. In *Computers in Urban Planning and Urban Management Conference, London*, 2005.

Paul Waddell. *Accessibility and Residential Location: The Interaction of Workplace, Residential Mobility, Tenure and Location of Choices.* 1996.

Paul Waddell. A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. *Environment and Planning B*, 27(2):247–264, 2000.

Paul Waddell. Parcel-level microsimulation of land use and transportation: The walking scale of urban sustainability. In *Proceedings of the 2009 IATBR Workshop on Computational Algorithms and Procedures for Integrated Microsimulation Models*, 2009.

Paul Waddell. Draft technical documentation: San francisco bay area UrbanSim application. 2013.

Paul Waddell and Firouzeh Nourzad. Incorporating nonmotorized mode and neighborhood accessibility in an integrated land use and transportation model system. *Transportation Research Record: Journal of the Transportation Research Board*, 1805(1): 119–127, 2002.

Paul Waddell, Brian J. L. Berry, and Irving Hoch. Residential property values in a multinodal urban area: New evidence on the implicit price of location. *The Journal of Real Estate Finance and Economics*, 7(2):117–141, September 1993.

Paul Waddell, Liming Wang, Billy Charlton, and Aksel Olsen. Microsimulating parcel-level land use and activity-based travel: Development of a prototype application in san francisco. *Journal of Transport and Land Use*, 3(2), 2010.

J.L. Walker and J. Li. Latent lifestyle preferences and household location decisions. *Journal of Geographical Systems*, 9(1):77–101, 2007.

WalkScore. Walk score methodology, July 2011. URL `http://www2.walkscore.com/pdf/WalkScoreMethodology.pdf`.

Brett Wallace, Fred Mannering, and G. Scott Rutherford. Evaluating effects of transportation demand management strategies on trip generation by using poisson and negative binomial regression. *Transportation Research Record: Journal of the Transportation Research Board*, 1682(1):70–77, 1999.

J. W Weibull. An axiomatic approach to the measurement of accessibility. *Regional Science and Urban Economics*, 6(4):357–379, 1976.

Rachel Weinberger and Matthias N. Sweet. Integrating walkability into planning practice. In *Transportation Research Board 91st Annual Meeting*, 2012.

Huw CWL Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and Planning A*, 9(3):285–344, 1977.

H. J. Wootton and G. W. Pick. A model for trips generated by households. *Journal of Transport Economics and Policy*, page 137–153, 1967.

Sadayuki Yagi and Abolfazl Mohammadian. Joint models of home-based tour mode and destination choices: Applications to a developing country. *Transportation Research Record: Journal of the Transportation Research Board*, 2076(-1):29–40, December 2008. doi: 10.3141/2076-04.

Dennis Zielstra and Hartwig H. Hochmair. Comparison of shortest path lengths for pedestrian routing in street networks using free and proprietary data. In *Proceedings of Transportation Research Board-91st Annual Meeting*, page 22–26, 2012.

Appendix A

Variable Listing

The table below lists variables that can hypothetically be used in this framework. Many variables are not available due to data limitations, and additional variables may be added as the data becomes available; an X marks each variable that was computed and tested in this model. Thanks to Nicola Szibbo for help with this table.

Built Environment Variables

Avail	Explanation	Measured
	Sidewalk completeness	% street frontage with sidewalks
X	Pedestrian route directness	route ft./direct ft. ratio
X	Ped/Bike Factor	Street network density + sidewalk completeness + bike land completeness/3
X	Street network extent	Miles/1000 residents
X	Street network density	Street centerline miles/square mile or intersection (sum of valences) per square mile/1300
X	Street connectivity	Ratio of intersections to total intersections plus cul-de-sacs
	Pedestrian infrastructure	% of streets with sidewalks
	Street Type	% of streets that are collectors, highways, freeways, major roads, arterials and local roads
	Street Width	# of lanes
	Pedestrian Network Coverage	% of total centerline distance
	Pedestrian Crossing Distance	curb to curb ft
	Bike route completeness	% of street routes (arterial + collector) with a bike lanes on one side or if possible parallel lanes
	Sidewalk completeness	% of streets with sidewalks on both sides + .5% streets with sidewalks on one side
	Bicycle Network Coverage	% of total centerline distance
	Street lighting	% of street network with street lights
	Sidewalk width	pavement width of sidewalk
X	Building Age	# of years old
X	Population density	Persons/square mile
X	Residential Density	DU/acre
X	FAR	ratio of a parcel's commercial floor area to the parcel's land area dedicated to commercial uses
X	Mix of uses	# of homes within 1/2 mile of site
X	Height	# of stories

X	Residential Type	Attached or detached
X	Single family parcel size	sq ft.
X	Single-family housing share	# of single-family units/total dwellings (%)
	Unit type	# of Studio, 1, 2 or 3+ bedrooms
X	Single-family Dwelling Density	DU/acre
X	Multi-family Dwelling Density	Du/acre
X	Multi-family housing share	# of multi-family units/total dwellings (%)
	Housing affordability	% of below market rate units
X	Growth compactness	Persons/square mile
X	Setback requirements	lot line distance
	Parking ratios	cars/unit
X	Amenities Proximity	walking ft. to closest grocery
X	Transit Proximity	walking ft. to closest stop
X	Amount of development within walking distance of transit	sq. ft
X	Housing proximity to recreation	% of dwellings within 1/4 mile of park
X	Property value	Land and building value in dollars per acre
X	High residential density	Residential acres with a density higher than 12 dwelling units/acre
X	Low residential density	Residential acres with a density lower than 3 dwelling units/acre
X	Medium residential density	Residential acres with a density between 3 and 12 units/acre
X	Service and retail	% of land for Big Box retail, vs. strip retail, vs.neighborhood retail
X	Highway retail	Acres of highway retail (gas stations, mini-marts, fast food) land uses and regional high-density shopping mall
X	Retail counts	Number of retail stores
X	Office	Acres of office land uses
X	Connected ratio	Percentage of intersections that are not dead ends
X	Bus stop counts	Number of bus stops

	Sidewalk coverage	Miles of sidewalks
	Park space available	Park acres/1000 people
X	Open space	Percent of total sketch area in open space land-use classes
X	Park proximity	% of walk ft. to closest park
	Open Space Connectivity	0-1 Index
	Shade Tree Density	# of shade trees/block length (shade trees/m)
	Imperviousness	Acres/capita
	Environmental resources	acres per capita vernal pools, wetlands etc.
	Orientation of buildings	% of south facing orientation of buildings or street grid
X	Transit Proximity to BART	# of ft.
X	Transit Proximity to Caltrain	# of ft.
X	Third places	# of coffee shops and/or bars
	Average Slope Gradient	Average slope gradient within the area
X	Block Size	Length (ft), width (ft)
X	Employee density	Number of employees/area of industrial and commercial land, in the residence area
	Neighborhood Unsafety	Neighborhood monthly criminal cases within area
	Sociodemographic Variables	
Avail	**Explanation**	
X	Residential population	Total number of people living in the area
X	Rent own ratio	Renters to total residents
X	Median age household	Median age of residents
X	Female-headed	Number of female-headed households
X	Median income	Median income of households in the area
X	Black ratio	Ratio of African-American population to total population
X	Latino ratio	Ratio of Latino-American population to total population
	Daytime population density	Population density during daytime hours
X	Children ratio	Ratio of children to total population - proxy for percent of families
X	Old ratio	Ratio of population over 70 to total population
X	College ratio	Ratio of college students to total population

X	Low-income ratio	Ratio of low-income households to total households
X	High-income ratio	Ratio of high-income households to total households
X	Average Household Size	Average number of members per household
X	Families with own children	Ratio of households with children present to total households
X	Average Family Size	Average number of members per household for households with children present
	Vacant Housing Units	Housing units without occupants
	Median Gross Rent	Median rent per unit
X	Percent Employed	Ratio of employed population to total population
X	Percent Unemployed	Ratio of unemployed population to total poulation
X	Area in square miles	Total area in miles
X	Total population	Total population in area
X	Total housing units	Total number of residential units in area
X	Total male population	Ratio of male population to total population
X	Total female population	Ratio of female population to total population
X	Population under school age, under 5 years	Ratio of children under 5 years to total population
X	School age population, 5-17 years	Ratio of children between 5 and 17 years to total population
	English spoken at home	Ratio of households that speak English at home
	Spanish spoken at home	Ratio of households that speak Spanish at home
	Chinese spoken at home	Ratio of households that speak Chinese at home
	Percent native born	Percent born in the United States
	Occupied housing units	Ratio of household units that are occupied by the owner
	Renter occupied housing units	Ratio of household units that are occupied by renters
	Median year householder moved into housing unit	Median year the household moved into the structure

Median gross rent as percent of household income in dollars	Median ratio of rent to household income
One person households	Percent of one person households
Two or more person non-family households	Percent of two or more non-family member households
Percent below poverty level	Percent of households below poverty level
Percent of workers driving to work	Percent of workers driving to work
Percent of occupied housing units with vehicle available	Percent of households with vehicle
Percent enrolled in public school (grades Pre-K to 12)	Percent of children age 5-18 enrolled in public school
Percent high school graduates, 25 years and over	Precent high school graduates among non-children