Title

Using Road Network Spatial Clustering to Assess the Timing and Duration of Dining, Shopping, and Entertainment Activities in California

Author
Davis, Adam Wilkinson

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Santa Barbara

Using Road Network Spatial Clustering to Assess the Timing and Duration of Dining, Shopping, and Entertainment Activities in California

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

by

Adam Wilkinson Davis

Committee in charge:

Professor Konstadinos Goulias, Chair

Professor Susan Cassels

Professor Krzysztof Janowicz

June 2019

The dissertation of Adam Wilkinson Davis is approved.

____

Susan Cassels

____

Krzysztof Janowicz

____

Konstadinos Goulias, Committee Chair

June 2019

Using Road Network Spatial Clustering to Assess the Timing and Duration of Dining,

Shopping, and Entertainment Activities in California

by

Adam Wilkinson Davis

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to my committee, whose feedback and advice helped me refine my dissertation project and whose questions challenged me to think more deeply about the processes I was trying to study and how I was studying them.

I gratefully acknowledge the financial support my research has received from the US Department of Transportation and CalTrans.

None of the work I have done at UCSB would have been possible without the guidance and support of my advisor, Kostas Goulias. Thank you for making sure I was always well funded, had interesting data and a good computer, and for encouraging all of your students to explore the world both through research and travel. Your support has made grad school so much more fun than it might have been!

My friends in the GeoTrans Lab have been the best group of colleagues I could ask for. In particular, I thank Jae Lee and Elissa McBride for being so fun to write papers with and Srinath Ravulaparthy and Kate Deutsch-Burgner for helping get me started.

In addition, I thank all of the other friends at UCSB and elsewhere who put up with my excited texts and surely excessive links about cities, statistics, science, and literature.

I would like to thank all the people who make up R's uniquely helpful, supportive, and inclusive community.

I am proud to be a child of two wonderful, brilliant geographers, whose support and love has been vital in developing my curiosity and love of learning.

Finally, I would like to express my profound love and thanks to Nina Bingham, my partner in graduate school and life. Writing a dissertation is horrible. Writing dissertations together was so much better.

VITA OF ADAM WILKINSON DAVIS

EDUCATION

University of California, Santa Barbara
Ph.D. in Geography **2019**
Dissertation: "Using Road Network Spatial Clustering to Assess the Timing and Duration of Dining, Shopping, and Entertainment Activities in California"

University of California, Santa Barbara
M.A. in Geography **2015**
Thesis: "Investigating Place Attitudes in Santa Barbara, CA"

University of California, Berkeley
B.A. in Geography **2011**
Areas of Concentration: GIS, Cultural Landscapes, Economic Geography

High Distinction in Undergraduate Study

RELATED EXPERIENCE

UCSB Geography GeoTrans Lab
Graduate Researcher **Fall 2013 –**
Transportation modeling and econometric analysis, Python and R scripting, GIS, research planning

US Geological Survey, Menlo Park CA
Student Contractor (Geography Research Intern) **2011 – 2013**
Performed GIS analysis, Python scripting, database management, manuscript editing, figure creation, and land cover modeling research in support of the USGS Land Cover Trends Project

Freelance Cartographer **Fall 2010**
Designed and drafted maps to client's specifications for History Walks Paris http://historywalksparis.com

Pacifica Land Trust, Pacifica CA
GIS Consultant and Environmental Analyst **Summer 2010**
Collected and analyzed field (including GPS and dendrochronology) and remotely sensed data for habitat monitoring and restoration.

REFEREED JOURNAL PUBLICATIONS

McBride, E., **Davis, A.W.**, & Goulias, K. G. (2019). Fragmentation in Daily Schedule of Activities using Activity Sequences. *Transportation Research Record: Journal of the Transportation Research Board*

Davis, A.W., McBride, E., & Goulias, K. G. (2018). A latent class pattern recognition and data quality assessment of non-commute long-distance travel in California. *Transportation Research Record: Journal of the Transportation Research Board*

Davis, A.W., McBride, E., Zhu, R., Janowicz, K., & Goulias, K. G. (2018). Tour-based path analysis of long distance non-commute travel behavior in California. *Transportation Research Record: Journal of the Transportation Research Board*

McBride, E., **Davis, A.W.**, & Goulias, K. G. (2018). A spatial latent profile analysis to classify land uses for population synthesis methods in travel demand forecasting. *Transportation Research Record: Journal of the Transportation Research Board*

McBride E., **A.W. Davis**, J.H. Lee, and K.G. Goulias (2017) Incorporating Land Use in Synthetic Population Generation Methods and Transfer of Behavioral Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2668), 11-20.

Lee J. H., **Davis, A.W.**, Yoon, S. Y., and Goulias, K.G. (2017) Exploring Daily Rhythms of Interpersonal Contacts: Time of day dynamics of human interactions using latent class cluster analysis. *Transportation Research Record: Journal of the Transportation Research Board*, (2666), 56-68.

Lee, J. H., **Davis, A. W.**, & Goulias, K. G. (2017). Triggers of behavioral change: Longitudinal analysis of travel behavior, household composition and spatial characteristics of the residence. *Journal of Choice Modelling*, 24C, 4-21.

Lee, J. H., **Davis, A. W.**, Yoon, S. Y., & Goulias, K. G. (2016). Activity Space Estimation with Longitudinal Observations of Social Media Data. *Transportation*, 43(6), 955-977.

McBride, E., Lee, J. H., Lundberg, A., **Davis, A. W.**, & Goulias, K. G. (2016). Behavioral micro-dynamics of car ownership and travel in the Seattle metropolitan region from 1989 to 2002. *EJTIR*, 16(4), 735-753.

Sherba, J. T., Sleeter, B. M., **Davis, A. W.**, & Parker, O. (2015). Downscaling global land-use/land-cover projections for use in region-level state-and-transition simulation modeling. *AIMS Environmental Science*, 2(3), 623–647.

Blecha, J., & **Davis, A.** (2014). Distance, proximity, and freedom: Identifying conflicting priorities regarding urban backyard livestock slaughter. *Geoforum*, 57, 67–77.

Wilson, T. S., Sleeter, B. M., & **Davis, A. W.** (2014). Potential future land use threats to California's protected areas. *Regional Environmental Change*, 1–14.

REFEREED BOOK CHAPTERS

Davis, A. W., Lee, J. H., McBride, E. C., Ravulaparthy, S., & Goulias, K. G. (2019). California Business Establishment Evolution and Transportation Provision. In H. Briassoulis, D. Kavroudakis, & N. Soulakellis (Eds.), The Practice of Spatial Analysis: Essays in memory of Professor Pavlos Kanaroglou (pp. 295–323). Cham: Springer International Publishing.

Lee, J. H., **Davis, A.**, McBride, E., & Goulias, K. G. (2019). Statewide Comparison of Origin-Destination Matrices Between California Travel Model and Twitter. In C. Antoniou, L. Dimitriou, & F. Pereira (Eds.), Mobility Patterns, Big Data and Transport Analytics (pp. 201–228). Elsevier.

REFEREED CONFERENCE PROCEEDINGS

McBride, E., **Davis, A.W.**, & Goulias, K. G. (2019). Fragmentation in Daily Schedule of Activities Using Activity Sequences. Presented at the 98th Annual meeting of the Transportation Research Board

Chaniotakis, E., **Davis, A.W.**, Aifadopoulou, G., Antoniou, C., & Goulias, K. G. (2019). A Latent Class Cluster Comparison of Travel Behavior between Thessaloniki, Greece and San Diego, California. Presented at the 98th Annual meeting of the Transportation Research Board

Davis, A.W., McBride, E., & Goulias, K. G. (2018). A latent class pattern recognition and data quality assessment of non-commute long-distance travel in California. Presented at the 97th Annual meeting of the Transportation Research Board

Davis, A.W., McBride, E., Zhu, R., Janowicz, K., & Goulias, K. G. (2018). A tour-based path analysis of long distance non-commute travel behavior in California. Presented at the 97th Annual meeting of the Transportation Research Board

McBride, E., **Davis, A.W.**, & Goulias, K. G. (2018). A spatial latent profile analysis to classify land uses for population synthesis methods in travel demand forecasting. Presented at the 97th Annual meeting of the Transportation Research Board

Lee, J.-H., **Davis, A.W.**, McBride, E., & Goulias, K. G. (2018) Statewide Comparison of Origin-Destination Matrices between California Travel Model and Twitter. Presented at the 97th Annual meeting of the Transportation Research Board

Davis, A. W., Lee, J.-H., & Goulias, K. G. (2016). Taking Place Perception into Travel Behavior Research. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C.

McBride, E., Lee, J.-H., Lundberg, A., **Davis, A. W.**, & Goulias, K. G. (2016). Behavioral Microdynamics of Car Ownership and Travel in the Seattle Metropolitan Region from 1989 to 2002. Presented at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C.

Davis, A. W., Lee, J.-H., & Goulias, K. G. (2015). Analyzing Bay Area Bikeshare Usage in Space and Time. Presented at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

Lee, J.-H., **Davis, A. W.**, & Goulias, K. G. (2015). Exploratory Analysis of Relationships Among Long-Distance Travel, Sense of Place, and Subjective Well-being of College Students. Presented at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

Goulias, K. G., Lee, J.-H., & **Davis, A. W.** (2015). Longitudinal Mixed Markov Latent Class Analysis of the 1989 to 2002 Puget Sound Transportation Panel. Presented at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

AWARDS

Pyke Johnson Award (outstanding paper in the field of transportation systems planning and administration)	January 2018
Transportation Fellowship, UCConnect	2014 – 2017
Dean’s Fellowship, UCSB College of Letters and Science	2013 – 2014

TEACHING EXPERIENCE

University of California, Santa Barbara

Teaching Assistant for Geography 101 Transportation Futures **Fall 2017**
Ran labs and graded assignments.

Teaching Assistant for Geography 2 World Regions **Summer 2017**
Ran labs and graded assignments.

Teaching Assistant for Geography 185B Environmental Decision Making **Spring 2017**
Ran labs and graded assignments.

Co-Instructor for Freshman Seminar in Sustainable Smart Cities **Winter 2017**
Developed course material, facilitated group discussions.

Teaching Assistant for Geography 111A Transportation Planning and Modeling **Fall 2016**
Developed lab syllabus and assignments, conducted lab sections, graded assignments and exams.

Co-Instructor for Freshman Seminar in Autonomous Vehicles **Fall 2016**
Developed course material, facilitated group discussions.

Teaching Assistant for Geography 111A Transportation Planning and Modeling **Fall 2014**
Developed lab syllabus and assignments, conducted lab sections, graded assignments and exams.

University of California, Berkeley
Cartography Course Intern **Fall 2010**
Assisted cartography students with mapping projects.

GUEST LECTURES

UCSB – Smart Green Cities Course
"Lessons Learned About Sense of Place, Place Attitudes, and Social Media Data for Smart Cities" May 2018

UCSB – Transportation Planning and Modeling Course
"Using R for travel behavior data analysis and visualization" October 2017

Technical University of Berlin
Department of Transportation System Planning and Telematics
"Data Sources and Fusion Methods for Place Attractiveness and Destination Choice" March 2017

Czech Technical University (Prague)
Department of Transportation System Planning and Telematics
"Lessons learned about place attitudes and social media information". May 2016

UCSB – Introduction to Geographic Research Graduate Course
"What you need to know about human geography outside UCSB" February 2015

PROFESSIONAL SERVICE

Organizing Committee International Association of Travel Behavior Research Conference – July 2018

Co-Chair, Organizing Committee 2015 UCConnect Student Conference – Winter 2015

Reviewer, *Journal of Transportation Letters*

Reviewer, Transportation Research Board / Transportation Research Record ADD30 (Committee on Transportation and Land Development)

Reviewer, IATBR Conference (Winter 2018)

CONFERENCESS AND WORKSHOPS ATTENDED

RStudio Conference, Applied Machine Learning Workshop, Austin	January 2019
15th International Conference on Travel Behavior Research, Santa Barbara *"Destination Attractiveness and Neighborhood Identification: Case Studies in California" – Podium*	July 2018
American Association of Geographers Annual Meeting, New Orleans *"Commercial Neighborhood Attractiveness and Exclusion" – Podium*	April 2018
97th annual meeting of the Transportation Research Board, Washington *"A latent class pattern recognition and data quality assessment of non-commute long-distance travel in California" – Poster* *"A tour-based path analysis of long distance non-commute travel behavior in California" – Poster*	January 2018
96th annual meeting of the Transportation Research Board, Washington	January 2017
Smart Cities Symposium, Prague	May 2016
American Association of Geographers Annual Meeting, San Francisco *"Modeling Quantifiable Place Attributes" – Podium*	March 2016
UCConnect Student Conference, Riverside CA *"Taking Place Attitudes into Travel Behavior Research" – Poster*	February 2016
95th annual meeting of the Transportation Research Board, Washington *"Taking Place Attitudes into Travel Behavior Research" – Poster*	January 2016
14th International Conference on Travel Behaviour Research, Windsor UK *"Perception and Reality: Linking Metrics of Place Perception to Measurable Attributes of Place Using Cross-Classified SEM Analysis" – Podium*	July 2015

International Choice Modeling Conference 2015, Austin *"Detection and Measurement of Latent and Manifest Heterogeneity of Familiarity, Perceptions, and Attractiveness of Places using Multilevel Analysis" – Podium*	May 2015
American Association of Geographers Annual Meeting, Chicago *"Comparing Bikeshare Programs in Space and Time" – Podium*	April 2015
UCConnect Student Conference, Santa Barbara Co-Chair, Conference Organizing Committee	February 2015
94th annual meeting of the Transportation Research Board, Washington *"Analyzing Bay Area Bikeshare Usage in Space and Time" – Poster*	January 2015
ESRI Annual User Conference, San Diego CA	July 2014
California Geographical Society Annual Meeting, Los Angeles CA	May 2014
University of California Transportation Center Student Conference, Pomona, CA *"Next-Generation Real Time Activity-Travel Behavior Data Collection Using Smart Phones and Available Big Data" – Poster*	April 2014
Association of Pacific Coast Geographers Annual Meetings, Olympia WA	October 2012
ESRI Annual User Conference, San Diego CA	July 2012
Wise Use of Floodplain Management Workshop, College of Environmental Design UC Berkeley	March 2012
American Geophysical Union Fall meeting, San Francisco	December 2011
Association of Pacific Coast Geographers Annual Meeting, San Francisco	October 2011
ESRI Annual User Conference, San Diego	July 2010

ABSTRACT

Using Road Network Spatial Clustering to Assess the Timing and Duration of Dining, Shopping, and Entertainment Activities in California

by

Adam Wilkinson Davis

Activity-based models for travel behavior are an important tool in urban planning and transportation analysis because they simulate the lives of individual people minute-by-minute and mile-by-mile throughout urban space. These models produce realistic schedules that take into account each person's work, school, and personal life while also abiding by constraints imposed by time, space, and the need to be in the same place at the same time as other people. While these models have made great strides in accurately representing the ways people and households schedule activities throughout the day, they do not do as good a job at understanding the interconnections between space / place, what activities people do, and when they do them. One factor that contributes to this shortcoming is a general mismatch between the spatial distribution of activities and opportunities that these models consider and the spatial units used in modeling. This dissertation seeks to improve this aspect of spatial choice models by using a network-distance variant of density-based spatial clustering methods to extract activity centers by clustering the locations of entertainment, food service, and retail businesses.

This sort of spatial clustering requires an accurate means of identifying neighboring points in addition to well-chosen clustering parameters. Various simplified methods for calculating network distance are compared for their accuracy at measuring distance and identifying neighbors, and the least simplified method is chosen. A range of clustering parameters are tested, their results compared, and a final clustering is chosen that balances the need to have small, discrete clusters while also capturing as many businesses as possible and matching most activity locations.

The types, timings, and durations of activities matched to different clusters are explored in order to assess the potential effectiveness of these clusters. This analysis identifies a set of center-level metrics that influence activity participation and timing. Finally, the spatial variability of activity durations and travel times is investigated using hierarchical models based on the clusters and spatially autoregressive models. These models indicate that much of the spatial autocorrelation of activity duration can be understood as primarily reflecting differences in the opportunities available at the level of individual centers.

TABLE OF CONTENTS

1. Introduction

People choose places to live, places to work, and places and times to shop, dine, and be entertained. The models used in travel behavior research to understand these spatial choices and how they connect to other aspects of people's behavior have considerable shortcomings both in terms of the choice of measurement unit and in terms of determining what makes specific destinations attractive for particular purposes. Aggregation of individual data points into larger spatial units is a necessary step for many sorts of geographic analysis and is particularly important for spatial choice models for home, work, and activity locations, which require limited choice sets containing meaningful and distinct spatial units.

In this dissertation I hope to begin to bridge that gap by developing a method to perform this aggregation that can be used as an alternative to spatial units derived from census or administrative boundaries, which are often poorly aligned with the actual spatial distribution of activities and opportunities. I use density-based spatial clustering to identify commercial centers and develop a method of correctly identifying neighboring points using network distances that can be scaled to work for a large state. After identifying the centers, I investigate center-level impacts of land use and the mix of opportunities available in an area on the timing and duration of people's activities in these centers.

Chapter 2 provides a background on ***Modeling Spatial Choices*** with particular focus on activity-based travel behavior models, defining spatial units for choice models, spatial heterogeneity, and the use of accessibility as an explanatory variable for behavior.

In Chapter 3 ***Data Overview and Spatial Matching***, I introduce the 2012-13 California Household Travel Survey (CHTS) and the National Establishments Time Series (NETS), the major sources of data for my dissertation, and analyze the degree of correspondence between the spatial data they provide. The CHTS provides a one-day record of the activities, and locations of 108,778 California residents in 42,431 households. NETS is a comprehensive record of all business establishments in the United States, with annual records from 1990 through 2013; I extract a subset corresponding to retail, accommodation/ food service, and arts/entertainment business establishments active in California in 2012 with accurately coded coordinates for their locations. Using a sample of eight major chain shopping and eating destinations common throughout the state and identifiable by name, I investigate how closely geocoded destination locations in the CHTS match business establishment locations in the 2012 extract from NETS and named places found in OpenStreetMap in 2019. While the CHTS and NETS locations do not match perfectly, named destinations in CHTS are generally within 200 meters of a matching NETS business and a similar distance of a matching OpenStreetMap location. While NETS is not a perfect dataset, it is beneficial to have business data for the correct year.

Density-based spatial clustering requires an accurate distance measure in order to identify neighbors. In Chapter 3 ***Measuring Network Distance to Identify Neighboring Business Establishments***, I demonstrate the feasibility of performing accurate network distance calculations to identify neighboring business locations throughout California up to a 1000-meter maximum distance threshold. I compare these network distances to the distance calculation and neighbor identification results produced by two shortcut network distance methods and Euclidean distances. This comparison finds a high degree of error by both metrics for all alternative distance measures, although the simplified road network distances generally stabilize beyond about 500 meters. I also evaluate the tradeoffs involved in computation time and accuracy between different forms of network distance computations for larger point datasets or larger neighbor distances.

Chapter 4 ***Using Network-Distance DBSCAN to Identify Commercial Centers in California*** introduces network-distance DBSCAN as a method for identifying commercial centers from NETS business locations and explores the effects of a range of values for the method's two parameters, neighbor distance threshold (ε) and minimum neighbors threshold (minPts). I investigate the sensitivity of clustering results to parameter values both in terms of directly measurable attributes of the resulting centers, and using secondary information produced by matching CHTS destinations to centers. Finally, I choose a clustering with ε = 250 meters and minPts = 4 and illustrate the results with center maps for the San Francisco, Los Angeles, Santa Barbara, and Sacramento areas.

In Chapter 5 ***Exploratory Analysis of Activity Timing*** and Chapter 6 ***Spatial Analysis of Activity Duration and Travel Time*** I use the commercial centers identified in Chapter 4 as a grouping element to analyze timing, duration, and travel time for dining, entertainment, and shopping activities reported in the CHTS. Chapter 5 explores the impact of the local density and diversity of opportunities on the frequency and timing of these activities as a way to identify key variables related to activity timing that should be incorporated into future choice models. I also apply bootstrap resampling methods to compare the scale of local center-level differences in activity timing with those related to personal characteristics, daily schedules, and day of the week. Chapter 6 addresses the spatial dependency of activity duration and travel time at the sub-center level in order to provide insights into how well these centers work as a unit of analysis. Using spatially autoregressive linear models and hierarchical linear models, I find that activity duration generally has higher degrees of spatial autocorrelation than travel time using neighbors either identified by distance or by center membership.

I conclude with an overall discussion of results and future areas of work for this analysis. I also discuss potential improvements to commercial center identification that might make them more useful as a level for spatial choice models and other limitations of my analysis.

2. Modeling Spatial Choices

Activity-based models for travel behavior have seen tremendous improvements in understanding temporal and human aspects of choices, but this progress has been much more limited in its consideration of space and place. The research presented in this dissertation is aimed mainly at improving the understanding of space and place in this field. This chapter opens with a brief overview of Activity-Based Modeling systems used in travel behavior and transportation planning and of the discrete choice models on which these systems are built. I then discuss the concept of accessibility as it relates to these models, challenges that arise in choosing spatial units for this sort of analysis, and the ways that travel behavior researchers have addressed heterogeneity in time and space.

a. Activity-Based Models

The primary goal of travel behavior researchers is to understand the flow of people and vehicles through cities over the course of the day and to predict the impacts that changes to policy, infrastructure, and economic conditions might have on these flows. Early travel models attempted to do this by considering the density of people, workers, and students in different zones, and assigning flows between these zones based on their relative sizes and proximity. Activity-based models (ABMs) represent an improvement in both the realism and usefulness of this research by directly incorporating an understanding of the ways that individual people and households arrange their activities during the day.

ABMs attempt to simulate the daily life of every individual in a study area using a cascade of statistical models that account first for various long-term choices people make (how long do they stay in school, where do they work, where do they live) and then to progressively shorter-term and more specific ones (do they have to drop their kids off at school, do they have to buy groceries today), eventually producing a minute-by-minute schedule of the activities each simulated person will do during the day, where they will do those things, and how they will travel between those locations (K. G. Goulias et al., 2011; Rasouli & Timmermans, 2014; Vovsha, Bradley, & Bowman, 2005). The activity-scheduling portion of these models is most relevant to my dissertation. In this step, models simulate the more fixed / restrictive components of each person's schedule and align it with that of other members of their household; non-mandatory travel (shopping, eating, entertainment) is set within the spatial, temporal, and personal linking constraints set by these earlier choices. Activities are assigned to locations at the level of traffic analysis zones, and the models for assignment take into account the presence of the right kinds of workers, transportation infrastructure and the proximity of the potential destination to the person's other activities (Y. Chen et al., 2011b; K. G. Goulias et al., 2011).

A general flaw of ABMs is that they handle many highly interconnected decisions in sequence rather than simultaneously. This is necessary for computational reasons, but it may be particularly damaging for understanding the participation, scheduling, and location for activities like shopping, eating, and entertainment, since they involve many highly specific options and come near the very end of the schedule generation process. Work on activity spaces and potential path areas has influenced the spatiotemporal constraints included in models of personal travel decisions both for individuals (Fan & Khattak, 2008; Patterson & Farber, 2015) and within households (Neutens, Schwanen, & Miller, 2010; Yoon & Goulias, 2010), to ensure that models understand what is possible, rather than treating activity location choice as a central part of activity simulation (Vovsha et al., 2005). While much of the effort in improving activity based models has focused on making activity sequences more realistic and representative, the minimal and highly aggregated handling of space in general and activity location choices specifically has been acknowledged as a major shortcoming of ABMs since early in their development (Garling, Gillholm, Romanus, & Selart, 1997; Rasouli & Timmermans, 2014).

b. Discrete Choice Models

Discrete choice models are central to travel behavior research. These models start with the somewhat unrealistic assumption that when trying to decide what to buy, where to go, or how to get there, people consider all the available options and choose the one that brings them the most benefit or utility (Ben-Akiva & Lerman, 1985; Train, 2009). The development of alternative structures for these models is probably the most significant way in which the field seeks to provide increasingly useful and meaningful analysis of the choices people make. The general format for discrete choice models is built around a set of potential choices, each with an associated latent utility variable (U_{ni}) that is split into a systematic component (V_{ni}) and a random error component (ε_{ni}) in Equation 2.1:

$$U_{ni} = V_{ni} + \varepsilon_{ni} \quad \text{Equation 2.1}$$

The systematic component of utility is modeled as a linear combination of important attributes of each option and individual (X_{ni}) based on a set of coefficients (β) that correspond to people's preferences:

$$V_{ni} = \beta X_{ni} \quad \text{Equation 2.2}$$

Decision-makers are assumed to choose whatever option provides the highest utility, but because the true value of ε_{ni} is unknown, the models instead provide the probability of each person choosing each option based on the assumed distribution of the error term, which assumes a normal distribution in probit models and a Gumble / extreme value distribution in logit models, which makes the probabilities much easier to work with mathematically. For a simple multinomial logit model, the probability of individual n choosing option i from a menu of choices subscripted j in the denominator is:

$$P_{ni} = \frac{\exp(V_{ni})}{\sum_j \exp(V_{nj})} \ or \ P_{ni} = \frac{\exp(\beta X_{ni})}{\sum_j \exp(\beta X_{nj})} \quad \text{Equation 2.3}$$

Discrete choice models are popular because they express complex multivariate decisions relatively simply, but the basic form of these models does not address a number of factors that are obviously important to understanding people's choices, particularly the variability of people's individual preferences and the impact of similar alternatives on the models' performance (Greene & Hensher, 2003; Train, 2009).

Destination choice is a potentially difficult application of discrete choice modeling because the choice is contingent on where people live, work, and travel; what they want to do; and how wide or narrow a choice set they face. Since basic choice models assume that all options are completely independent (the independence of irrelevant alternatives), they can be swamped when the choice set includes numerous very similar alternatives. Various modified forms of choice models relax this assumption somewhat, but nesting structures address it directly by grouping related choices together (Train, 2009). While this model structure was originally intended to solve issues presented by mode choices, it has also been used as a way to decrease the severity of spatial autocorrelation among the utilities of neighboring locations, although spatially correlated choice models address the issue more directly (C. R. Bhat, 2000; C. R. Bhat & Guo, 2004).

Spatial choice models pose particular issues for choice-set generation, since these models assume that all potential choices are included. Even with the fairly large zones used in travel modeling, including all of the zones with grocery stores accessible to each person would overwhelm a model for shopping location. As a result, many models use randomly subsampled choice sets both for modeling and simulation, a solution that is not ideal but generally does not bias results for simpler models (Nerella & Bhat, 2004). In contexts where the number of potential choices is quite large, it is also often necessary to use an "unlabeled" model specification (which leaves out the intercept term for each alternative's utility equation). The process of generating choice sets is often handled before the model is estimated, but models that incorporate choice set generation endogenously are available (Nerella & Bhat, 2004; Swait, 2001). Two remaining issues facing spatial choice models are the need to spatially aggregate possible destinations and how to account for the effects of variable traffic conditions in destination choice models (Dill et al., 2014).

Spatial choice models for activity location typically include travel time and cost (which are generally the major negative drivers on the estimated utility of various options), personal characteristics of the traveler and their household and information about their daily schedule, and information about the attractiveness of each destination (usually based on the density of workers in the applicable employment category). Habitual behavior is a profoundly important component of many people's destination choices, but very few travel surveys contain sufficiently long-term data to make this workable (Schlich & Axhausen, 2003). Destination attractiveness is closely related to the concept of accessibility, which is the most widely applied spatial predictor of behavior. In the rest of this chapter, I will discuss the ways accessibility is measured and issues with the measurement units typically used both for accessibility and activity location choice sets. I will then briefly highlight the importance of accurate distance measurement and understanding behavioral heterogeneity between people and locations.

c. Accessibility

Travel behavior researchers and urban geographers often identify accessibility as a major factor in people's movement through space and in the attractiveness of destinations. Originally defined as "the potential of opportunities for interaction," (Hansen, 1959), measures of accessibility are generally intended to capture the density and diversity of potential opportunities for people's activities, measurable as a continuous field variable over space. Most common measures of accessibility are intended to characterize the overall patterns of opportunity density and the built environment over a region, rather than providing information about the specific places in which people pursue their activities.

The majority of accessibility indicators either consider a single origin or multiple anchor points like home and work locations around which availability and reach of opportunities are measured (S. Handy, 1993; D. M. Levinson, 1998; D. Levinson, Marion, Owen, & Cui, 2017). The range of accessibility people experience throughout the day, either at specific destinations or within their activity space, is linked to social interaction, activity scheduling, and task allocation within the household (Lee, Davis, Yoon, & Goulias, 2016; Patterson & Farber, 2015; Shliselberg, 2015; Yoon & Goulias, 2010). Diversity-centered accessibility has been measured by variables corresponding to the density of broad land use categories (Cervero & Kockelman, 1997) or using information entropy measures (Davis, 2015; de Abreu e Silva, Golob, & Goulias, 2006). Both methods have shown significant relationships with travel behavior, but they depend heavily on the classification scheme used.

One well-established formulation of accessibility is opportunity-based accessibility, which counts the number of potential destinations reachable within a certain period of time or by traveling a certain distance (Páez, Scott, & Morency, 2012). These are further enhanced by accounting for congestion and the expected opening and closing hours of businesses (Y. Chen et al., 2011b). This sort of accessibility can be measured as continuous field that varies over space, which makes it a valuable measure of the sorts of opportunities that people can access within a certain distance of their home, work, or school location but is less useful as a measure of the specific opportunities available in one place. Proximity to opportunities for complementary activities (like dining and entertainment) seems likely to affect destination choices, but a person would not, for instance, choose to go shopping in a residential area just because it was located midway between two major shopping centers that counted both of them in its accessibility. Opportunity-based accessibility is not a good measure of destination-level attractiveness.

At the other extreme, modeling specific opportunity locations (such as stores) individually is also infeasible for several reasons including inability of surveys to capture many destinations (i.e., there are many more business establishments in a region than the locations visited by respondents), mismatched geocoding between surveys and available databases of business establishments, and semantic mismatching between activity type and business establishment type. In addition, business locations cluster in space, whether measured by straight line or on a network, and this is particularly true for service industries which benefit from agglomeration economies achieved by attracting a larger pool of customers to a dense location (Kolko, 2010; Okabe, Okunuki, & Shiode, 2006; Ravulaparthy & Goulias, 2014; Yamada & Thill, 2004).

Accessibility can be understood as varying continuously over space, but choice models use choice sets made up of discrete options; in the case of activity destinations, these options take the form of spatially aggregated measurement units. To perform destination-level analysis of shopping, dining, and entertainment activities, we require a method for identifying potential destinations that are aggregated enough to consider the effects of surrounding opportunities while also being small to be meaningfully considered a destination. Questions remain about how best to perform the spatial aggregation of opportunity locations.

d. Problems with Census Units

Census spatial units (such as block groups and tracts) are commonly used in social and health sciences to define neighborhoods both because they are predefined and standardized and because they match the spatial resolution of other available datasets, but their size, shape, and boundaries may not be well-suited to measuring phenomena of interest to researchers (Bates, 2006; Weiss, Ompad, Galea, & Vlahov, 2007). Although they are designed to measure where people live, census units are often even a poor choice for delineating neighborhoods when studying housing in part because the census splits units at major roads, which assigns houses on opposite sides of the same street to separate units (Clapp & Wang, 2006).

In travel demand modeling, traffic analysis zones (TAZs) built up from census blocks are the typical spatial unit of analysis because the US Census and American Community Survey provide TAZ-level demographic and employment data, but a variety of analytical issues arise from their use. Páez and Scott (2004) point out two major issues with TAZs that relate to the Modifiable Areal Unit Problem (MAUP): the scale effect, by which the same analysis can lead to different conclusions depending on the resolution of the spatial units; and the zoning effect, by which different spatial partitioning leads to a wide range of possible analytical outcomes. This casts doubts upon models that incorporate zone-level spatial relationships. Solutions to the problem of developing the "right" zoning system have been proposed that attempt to minimize some of the negative impacts of spatial aggregation, including some that focus on capturing travel within a single zone using a variety of scales (Moeckel & Donnelly, 2015; Viegas, Martinez, & Silva, 2009).

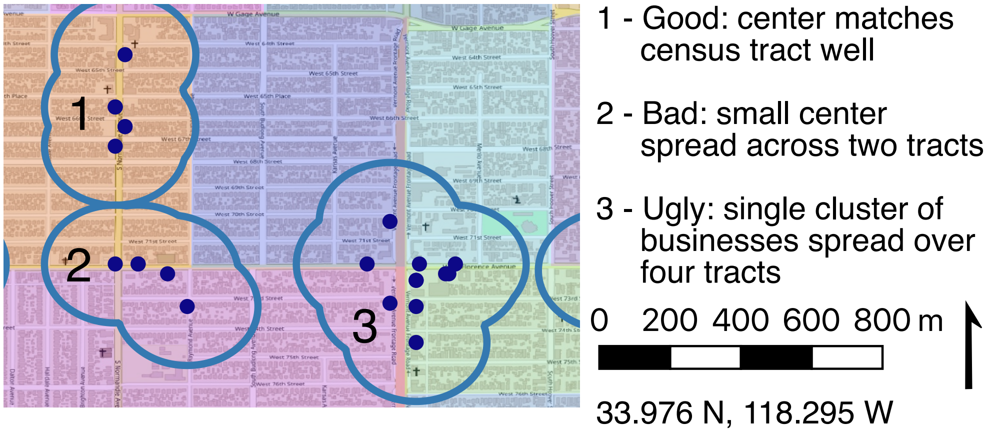

1
2
3
1 - Good: center matches census tract well
2 - Bad: small center spread across two tracts
3 - Ugly: single cluster of businesses spread over four tracts
0 200 400 600 800 m
33.976 N, 118.295 W

Figure 2.1 Three potential outcomes when businesses concentrated along roadways get mapped to census units. Census tracts shown in shaded colors. Euclidean-distance DBSCAN cluster search radius extents with minPts=5, eps=200 meters are shown in blue. Example 3 clusters along Vermont Ave, a major arterial in central Los Angeles.

Boundaries of these units are set based on "visible and identifiable features," like arterial roads (US Census Bureau, 2010), which are often the site of many businesses. As a result, commercial centers built around major roads and intersections are often split among multiple units. Figure 2.1 shows three potential cases of business aggregation to tracts from an area in central Los Angeles. The first example is the best case for small commercial centers, because all the businesses are located near the centroid of a single tract. In example 2, the businesses in an area cluster along a street dividing two tracts. In example 3, a set of businesses clusters along the intersection of two major roads that divide four tracts, which means that a tract-based spatial aggregation would effectively separate those businesses by as much as half a kilometer in an area with high population density, and much more in an area with larger tracts. The bubbles on the map correspond to the search radius for destinations in each of the clusters that result from our clustering result.

e. Other Ways of Defining Neighborhoods

While census units are very commonly used for delineation of urban space, they are not the only basis for such delineation. People often understand cities as being broken up into neighborhoods that are differentiated spatially and demographically and in terms of local culture, history, architecture, and/or the range of amenities available. Neighborhood definitions vary and depend heavily on the purpose for which boundaries are drawn; scholars looking for effects of neighborhood on health, wealth, and education have very different priorities than governments and businesses using neighborhoods to attract tourism or investment (Campbell, Henly, Elliott, & Irwin, 2009; Coulton, Korbin, Chan, & Su, 2001; LA Times Data Desk, 2018; San Francisco Association of Realtors, 2017). Attempts to extract neighborhood boundaries from surveys and crowdsourcing have found that people who live in an area have wildly varying ideas about neighborhood boundaries. An effort to crowdsource boundaries for Boston neighborhoods found that while people agreed on the general arrangement of named regions, they differed widely in their assessment of their relative sizes (Woodruff, 2013). Bae’s study of the cognitive boundaries of Los Angeles’s Koreatown neighborhood demonstrated that people who live in a place have varying understandings of its boundaries that may not align with boundaries defined by outsiders or government agencies (Bae & Montello, 2018).

An approach favored by travel behavior analysts is to dispense with neighborhood boundaries entirely and define “neighborhood” as a set of attributes drawn from a circular buffer around each individual household in the dataset. This approach seems adequate for travel behavior analysis, assuming a neighborhood affects behavior primarily through continuous attributes like open space area, road length, and number of nearby grocery stores, rather than through a softer concept such as the sense of being in a mixed use neighborhood (Cervero & Kockelman, 1997; Frank, Bradley, Kavage, Chapman, & Lawton, 2008; S. Handy, 1996). Because it doesn't attempt to draw boundaries, this method sidesteps the issue of vagueness, but may leave room for considerable uncertainty, depending on how the buffers intersect any polygonal data sources.

Travel behavior research that has considered neighborhoods as discrete entities has generally compared residential neighborhoods in separate parts of an urban area instead of attempting to divide an entire region (X. Cao, Mokhtarian, & Handy, 2007; Khattak & Rodriguez, 2005; Kitamura, Mokhtarian, & Laidet, 1997; Krizek, 2003). These studies have found significant effects of neighborhood-level urban form and land use on travel behavior, but these effects are often confounded by self-selection and the considerable impact of people's predispositions towards certain modes of transportation on both their travel and their choice of home neighborhood.

When precise neighborhood boundaries are not known or needed for analysis, geographers (particularly in this department) have employed hexagonal tessellations as a way to divide up space as a measurement framework for surveys about place attitudes (Davis, 2015; Deutsch, 2013; Deutsch-Burgner, Ravualaparthy, & Goulias, 2014; Montello, Friedman, & Phillips, 2014). This approach has also been used to identify employment sub-centers in Southern California and investigate the relationship between accessibility to jobs in these sub-centers and vehicle travel (Boarnet & Wang, 2019).

f. Urban Geography and the Arrangement of Urban Space

Since my goal is to identify centers of commercial activity rather than residential neighborhoods, it is important to consider how urban geographers and planners understand cities, including the arrangement of activities within them and the processes that shape this arrangement. Early urban geographers employed simplified models to understand the overall layout of the city, generally a densely-developed central region surrounded by concentric rings of decreasing rents, specializing in different activities (P. J. Smith, 1962). To the extent that generalized urban models are still employed, they generally understand activity as a spatially varied continuous urban field, with different priorities in different zones or realms (Couclelis, 1989; Friedmann & Miller, 1965; Godfrey, 1999).

Other fields are more interested in the processes that shape cities than the eventual form they take. Transportation-oriented scholars hold that existing transportation modes are the primary control on urban expansion, with increases in transportation speed leading over time to less centralized layouts (Muller, 2004). In addition to affecting the extent of cities, transportation infrastructure can also influence the shape of development. In Los Angeles and the areas surrounding San Francisco, urban expansion concentrated along streetcar lines radiating out from the city center, and these developments still serve as commercial cores for the inner suburbs (Bottles, 1987; Suisman, 2014; Wachs, 1984).

Marxist urban geographers acknowledge the historical importance of transportation in the expansion of cities but identify land rents as the primary driver, since these determine what people and purposes the city is built to serve and incentivize the periodic rebuilding of the inner city (Harvey, 2003; N. Smith, 1979). In addition to these economic and technological processes, physical geography plays a major role in shaping urban development. This is particularly notable in California. Much of San Diego and the San Francisco Bay Area are built on steep hilly terrain around large natural harbors, which limits buildable area and requires thoughtful design of transportation infrastructure (ConnectSF, 2018; Dailey, 2017), while inland cities like Sacramento have to contend with severe flooding and may avoid floodplain development. Earthquakes also occasionally force planners to substantially rethink urban design; for example, damaged and unsafe freeways may be removed to make way for infill development (King, 2014).

Most understandings of American cities include a dense, diverse Central Business District (CBD) or "downtown" as the focal point for the city's government, finance, and trade, with city government and the consumer (retail and entertainment) heart of the city located nearby (Murphy, 2017). While the outward appearance and experience of CBDs is fairly consistent – tall buildings, corporate offices, dense development, and high rents – the spatial layout is somewhat more variable, largely reflecting the infrastructural connections (harbors, rivers, and railroads) around which the cities were built (Hartman, 1950; Murphy, 2017). While some other parts of the US (particularly in the southwest) were developed largely after freeways became the primary mode of interurban transportation, most of California's major urban centers predate the expansion of freeways, and often pushed back against their expansion (Carlsson, 2009; Perez, 2017). While CBDs are generally the densest parts of cities, accessibility by car is often much higher in the inner-ring suburbs, where people have relatively easy access both to downtowns and to sub-centers spread throughout the region (Giuliano & Small, 1991; D. Levinson et al., 2017). Easy access to employment subcenters has been tied to lower Vehicle Miles Traveled (Boarnet & Wang, 2019). The historical development of subcenters poses something of a dilemma concerning whether suburbanization of housing or employment centers came first, and whether public infrastructure planning or private investment is the more important driver (Anas, Arnott, & Small, 1998; Gordon, Richardson, & Wong, 1986; Helsley & Sullivan, 1991; White, 1976).

In practice, cities reflect the combined influences of a range of processes, as well as the historical accumulation of investment and planning practices that produced each new development. Some unique attributes of California cities may have a bearing on the sorts of commercial centers I should identify. While Los Angeles is sometimes considered the archetypal 20th Century American city because of its expansive car-oriented suburban development (Krim, 1992), in some ways this title is a poor fit. While its downtown is arguably less dense than that of older cities, the Los Angeles region as a whole is much more densely populated and continuously developed than other Sun Belt cities (Singley, 2013; Wilson et al., 2012). Many American cities underwent “urban renewal” programs in the 1950s-70s to redevelop “blighted” (often predominantly black) neighborhoods near the city center into new commercial and residential centers, and California’s state government strongly emphasized these programs (Lai, 2012; Teaford, 2000; Thomas, Ritzdorf, & Hodne, 1997). The Western Addition redevelopment project in San Francisco is notable because it specifically justified the bulldozing of a middle class black neighborhood as an effort to re-build the Japanese-American downtown that was destroyed by the Japanese Internment (Lai, 2012). On the more positive side, California has also begun to push for more mixed-use infill development as a way to create “Sustainable Communities” that disincentivize car travel to decrease greenhouse gas emissions (OneBayArea, 2013; Steinberg, 2008).

Commercial centers come in a wide range of shapes and sizes. Dense downtowns offer a wide and diverse range of opportunities but should be expected to make up a smaller share of the commercial development of the region than sub-centers in inner- and middle-ring suburbs (Giuliano & Small, 1991; Helsley & Sullivan, 1991). These, in turn, should be expected to cluster along major roads (Bottles, 1987; Wachs, 1984). Suburban malls contain a great deal of the retail opportunities in California, and their uniquely car-oriented design, and immense parking requirements, and separation from existing centers (Ersoy, Hasker, & Inci, 2016) should make them fairly distinct from other sorts of centers, particularly as they decline (Parlette & Cowen, 2011; Schwartz, 2015). In residential neighborhoods that predate exclusionary zoning practices and in newer mixed-use areas, corner stores and small neighborhood centers meet some of residents' needs, while newer planned medium-density mixed-use developments often struggle to draw customers from larger centers (Bartlett, 2003; Grant & Perrott, 2011). Los Angeles's mini-malls (arguably the classic Los Angeles cultural landscape) thrive both because they serve local communities more directly than other businesses (Loukaitou-Sideris, 2002) and because they often incorporate a wider mix of services, including medical clinics (Sloane, 2003). Considerably less scholarship has been devoted to the retail geography of rural areas, but a study of older and developing rural commercial centers found that while the concentration of the market into fewer, larger stores (notably Wal-Marts) was a common theme in established areas, rural areas experiencing population growth tended to have a wider range of stores (Vias, 2004).

g. Measuring Distance on a Road Network

Distance measures are a central concern for both travel behavior analysis and spatial clustering, and choosing the correct method is vital to performing valid analysis (Boscoe, Henry, & Zdeb, 2012; X. (Jason) Cao, Mokhtarian, & Handy, 2009; S. L. Handy & Niemeier, 1997; Yamada & Thill, 2004). Euclidean distance is satisfactory for matching points that have slightly mismatched geocodes (Chapter 3), but measurement of distances between separate points should consider the limitations on travel between these points. People cannot travel through urban areas in straight lines from place to place, so distances between separate places in cities should be measured along a road network rather than by straight line. Transportation researchers generally acknowledge the importance of using road network distances in analysis, but shortcuts are often taken in the interests of computation time and complexity. These methods can entail assigning all locations to the nearest road intersection, road segment centroid, or census unit centroid. Simplified distance calculations are particularly attractive when large numbers of distance computations are required, such as for computing network centrality and accessibility (Boeing, 2018a; Y. Chen et al., 2011a; S. L. Handy & Niemeier, 1997; Ravulaparthy, Goulias, Sweeney, & Kyriakidis, 2013). Unfortunately, inaccuracies introduced by these methods can bias analysis, and may be particularly problematic for identification of neighbors and density calculations (Yamada & Thill, 2004).

h. Spatial Heterogeneity and Activity Timing

Spatial heterogeneity presents another set of issues, since the relationship between a location and the behavior of a person depends inherently on the location for reasons that are not recorded in the data, and this difference can change over the course of a day. Bhat and Zhao (C. Bhat & Zhao, 2002) show the impact of neglecting spatial heterogeneity in the context of stop-making decisions by households, but their proposed solution uses TAZs as the spatial units of analysis. The degree to which the attractiveness of different places varies over the course of the day has been less thoroughly addressed, but it is clearly an issue. Google Maps now displays plots showing the relative popularity at different days and times of individual destinations, and ratings sites like Yelp also indicate what days and times individual bars and restaurants are best visited. Some work has attempted to capture time of day signatures of major facilities and events (McKenzie & Janowicz, 2015; McKenzie, Janowicz, Gao, & Gong, 2015; Paul, Vovsha, Hicks, Livshits, & Pendyala, 2014), but work has been limited in travel behavior. Neglecting the variability of place attractiveness is a major issue for activity-based travel models that simulate activity participation by time of day and day of the week.

Although place-based analysis has been limited, travel behavior research has found a variety of ways to understand the mix of travel and activities people do in a day. Measures have included the amount of time spent sleeping, eating, working, and socializing; the number, modes, and lengths of trips made; and the number of people interacted with. Lee et al. investigated the relationships among these measures by developing a three-way latent class clustering model of daily schedules (Lee, Davis, Yoon, & Goulias, 2017) and other travel behavior researchers have noted the importance of understanding daily sequences of activities (C. R. Bhat et al., 2013). Geographers have focused more on the relationship between time and place; notably McKenzie et al. (2015) identified interesting differences between the temporal signatures of activity locations using social media data. These sorts of differences between places are not at all accounted for in existing activity-based models, which generally assume that all places that provide a certain amenity operate on the same schedule.

3. Data Overview and Spatial Matching

In this chapter, I introduce the major data sources used throughout this dissertation and discuss the overall quality and interoperability of the spatial data contained in these datasets. The bulk of the data used in this dissertation is drawn from two sources: the National Establishment Time Series (NETS) and the 2012-13 California Household Travel Survey (CHTS). NETS data is used as a catalog of opportunities for shopping, dining, socializing, and other activities that can be performed outside the house, and the place-based travel and activity diary in the CHTS provides a record of activities performed by California residents. I investigate the overall accuracy of the spatial match between these datasets using a set of major chain shopping and eating destinations that are common throughout the state and are identifiable by name; while the CHTS and NETS locations do not match perfectly, named destinations in CHTS are generally within 200 meters of a matching NETS business and a similar distance of a matching OpenStreetMap location.

a. Data Sources

The California subset of NETS contains a record of all business establishments (e.g. individual stores or offices) in California from 1990 to 2013 (Walls and Associates, 2017). This dataset is produced from Dun & Bradstreet business establishment data and licensed by Walls and Associates. For this dissertation, I use customer-facing businesses in the retail, food service, and entertainment categories (NAICS 2-digit codes 44, 45, 71, and 72) with at least three employees that were located in California in 2012. Businesses with 1 or 2 employees were excluded because many of them appeared to be home addresses rather than storefronts; for instance, a home-based business that sells products online is retail, but it does not provide a place for people to physically shop.

NETS data contains some degree of imprecision in business classification and geographic location, and it is necessary to remove business locations that are not coded with sufficient precision to match precisely with the road network. Establishments in NETS businesses are classified according to the North American Industrial Classification System (NAICS). At the lowest level, NAICS includes over a thousand different distinct business types (United States Office of Management and Budget, 2017). Because it was designed to describe the economic relationships between businesses, NAICS is not ideally suited for the sorts of questions I ask in my dissertation; the system recognizes 360 types of manufacturers for example, but only 4 types of restaurants.

Most businesses in NETS are geocoded down to their position on the road segment and side of the road (“Block Face” level), but some locations are coded only down to the centroid of their census tract, block group, or zip code (Walls & Associates, 2013). This inconsistency is more common for non-storefront locations than it is for the types of businesses I use here: 94.8% of the retail/entertainment/food service businesses with at least three employees are geocoded precisely, as are 93.9% of the other businesses active in California. I exclude locations with imprecise geocodes from my analysis, decreasing the total number of businesses from 193,820 establishments in the target categories with at least 3 employees in 2012 to 183,772 establishments that have precise geocodes. These businesses break down by category as follows: Retail 110,117 (NAICS 2-digit 44 and 45); Arts, Entertainment, and Recreation 17,301 (NAICS 71); and Accommodation and Food Service 56,354 (NAICS 72).

OpenStreetMap (OSM) provides an attractive free alternative to proprietary NETS data as a source of business locations. OSM is a free-to-use community-provided source of spatial data for shops, restaurants, and other points of interest in a fairly consistent data format available worldwide for the present day ("OpenStreetMap," 2019). The place classification scheme used by OSM conveys a great deal more detail than NETS does in the types of places I used for this analysis which makes it an especially appealing option as a source of business location data. My primary reason for using NETS data instead of OSM is that whereas NETS contains data for 2012, it is not easy to access OSM data from previous years (though a 40GB global data file is available online for 2013 https://planet.openstreetmap.org/planet/full-history/2013/). Free-to-use data sources like OSM would be a preferable source for similar studies in the future, as long as care is taken to extract the data concurrently with the travel survey data collection process.

While I had access to NETS data paid for under a previous project, I did not have any contemporary road network data as high quality as what is available from OpenStreetMap (from which I extracted the statewide road network in Spring 2018). This road network is likely somewhat different from what existed in 2012-13, but there have been no major new freeways or arterial roads built in California since then, apart from those in new residential developments that this dissertation does not concern.

The California Household Travel Survey (CHTS) contains demographic and travel-related information for 108,778 people in 42,431 California households, with people surveyed on every day from February 2012 until February 2013 (NUSTATS, 2013). Each person was asked to fill out an activity travel diary containing a complete record of every place they visited between 3AM on their assigned survey day and 3AM the next day. The activity-travel diary provides locations, descriptions, travel modes, and timings for all activities performed in a single assigned day in the life of all the people in the survey. The households selected for the survey were spatially stratified by county in order to ensure the collection of sufficient data to model the behavior of rural travelers, but sampling probabilities for the Southern California residents used in our analysis are fairly consistent across counties. The CHTS provides estimated sample weights for each household that could be used with most of the analysis methods I present in my dissertation, but previous analysis has suggested that they are likely unreliable (McBride, Davis, Lee, & Goulias, 2017). For schedule comparison analysis, only out-of-home activities were considered. Table 3.1 contains frequencies for all of the out of home activity types reported in the CHTS.

Locations visited by people in the CHTS were geocoded during the initial data collection using a program that accessed a Google Maps API (NUSTATS, 2013). This process stored locations for each household, so that if multiple people in a single household visited the same location, it was coded identically in all of their place logs, but this program does not appear to have provided fully consistent results between households. I found a considerable amount of spatial variability in the locations recorded for places that should have been coded the same, possibly because different people described the places slightly differently. For instance, multiple people who appeared to visit the same Costco had geocodes that varied by over a hundred meters in one direction or the other.

Table 3.1 Out-of-home Activities Reported in the CHTS

Activity Purpose	Higher-level Category	Acts
Non-work related activities (social clubs, etc)	Civic / Religious	339
Volunteer work/activities	Civic / Religious	563
Civic/religious activities	Civic / Religious	7,366
Eat meal at restaurant/diner	Dining	18,149
Drive through meals	Drive through	4,717
Drive through other	Drive through	1,270
Entertainment	Entertainment	7,209
Exercise/sports	Exercise	329
Indoor exercise	Exercise	4,875
Outdoor exercise	Exercise	10,225
Service private vehicle	Maintenance	5,058
Household errands	Maintenance	9,138
Personal business	Maintenance	6,958
Health care (doctor, dentist, eye care, chiropractor, veterinarian)	Medical	6,007
Study / schoolwork	School	62
After school or non-class-related sports/physical activity	School	788
All other after school or non-class related activities	School	1,227
Meals at school/college	School	2,337
In school/classroom/laboratory	School	11,254
Shopping for major purchases or specialty items	Shopping	2,843
Routine shopping	Shopping	34,326
Social/visit friends/relatives	Social	21,563
Loop trip	Transportation	7,260
Change type of transportation/transfer	Transportation	23,637
Pickup/drop off passenger(s)	Transportation	24,187
Work-sponsored social activities	Work	106
Training	Work	264
All other work-related activities at my work	Work	1,069
Meals at work	Work	5,085
Work-related	Work	9,878
Work/job duties	Work	30,432
Other	Other	4,498

b. Point Matching

This dissertation targets local clusters of opportunities rather than individual locations to explore the effects of density and diversity on the attractiveness of destinations, but it is worth considering whether it would be practical to use business and activity location datasets together at the level of individual points. To do this, I match the locations of eight stores and restaurants with multiple locations in California that appear frequently in the CHTS.

I identified the unique place names in the CHTS Places dataset and counted the occurrences of each. CHTS respondents visited a total of 492,321 places on their respective survey day, with 64,033 unique place names. Many of the most frequent names referred to private residences or generic transportation infrastructure sites (e.g., “Bus Stop”, “Transit Stop”). Costco was the most common business name and appeared in the travel diaries of 2,250 people as “COSTCO.” Manual name matching and word searching identified another 58 forms (e.g., “COSTCO GOLETA”) for a total of 2,450 reported visits. All places with at least 1,000 combined appearances in the CHTS are shown in Table 3.2.

For eight most commonly visited businesses in the CHTS (as shown Table 3.2), I searched for all locations in the 2012 NETS dataset with a business name containing some variant of the business name in question (e.g., “WALMART” or “WAL-MART”). For comparison, I queried OSM on January 7, 2019 for places with matching names. The spatial attribute of NETS data corresponds to single points, but OSM provides spatial data in a range of different structures, and relevant businesses are stored as a mix of points, some as polygons (“buildings”). For the comparison analysis, I converted polygons to point data using their centroids. Figures 3.1 and 3.2 show the cumulative share of destinations for each business name within a given distance of a matching NETS and OSM point, respectively.

Table 3.2 Commonly appearing place names in CHTS Travel Diary Places Dataset

Place (Grouped)	Unique Spellings in CHTS Place file	Times Visited in CHTS Place file	Used in Place Match figures
Private Residence	27	239,612	
Transportation Infrastructure	34	13,253	
Costco	60	2,450	Yes
Walmart	15	2,246	Yes
Government Office	2	2,167	
Safeway	14	1,979	Yes
Target	10	1,805	Yes
McDonalds	9	1,483	Yes
Starbucks	5	1,418	Yes
Trader Joes	6	1,237	Yes
Church	8	1,201	
Home Depot	38	1,056	
CVS	14	1,053	
Vons	21	1,053	Yes
Kaiser (Hospital)	141	1,017	

For seven of the eight chain businesses considered here, NETS contains a matching location to 75% of CHTS destinations within about 250 meters, as shown in Figure 3.1. This match is slightly better for Starbucks and grocery stores (Safeway, Trader Joe's, and Von's) than for other big box stores (Costco, Target, and Walmart), but all of the stores stabilize between 80-90% once the search radius is expanded to 350 meters. Less than 25% of CHTS visits to McDonald's restaurants are even within a kilometer of a business called McDonald's in NETS, which likely reflects the company's franchised structure, in which individuals or small companies own and operate most locations that carry the company's brand.

OSM data from 2019 matches CHTS locations comparably well to NETS, as shown in Figure 3.2, but different place types stabilize at somewhat more varied match rates, possibly reflecting varying rates of turnover between different companies. McDonalds matches substantially better in this dataset, whereas Starbucks matches much worse than it did for NETS. NETS and OSM have generally similar match rates (Figure 3.3) despite the intervening years in which locations people visited in 2012 could have closed or changed ownership. This suggests that OSM might be a more reliable source for this type of analysis. However, since OSM also contains businesses that opened after 2012 that should not be considered in analysis of travel behavior in that year, NETS is probably a preferable data source for that year. Destination analysis would be made easier if future travel diary surveys chose a specific place dataset to use both in data collection and analysis.

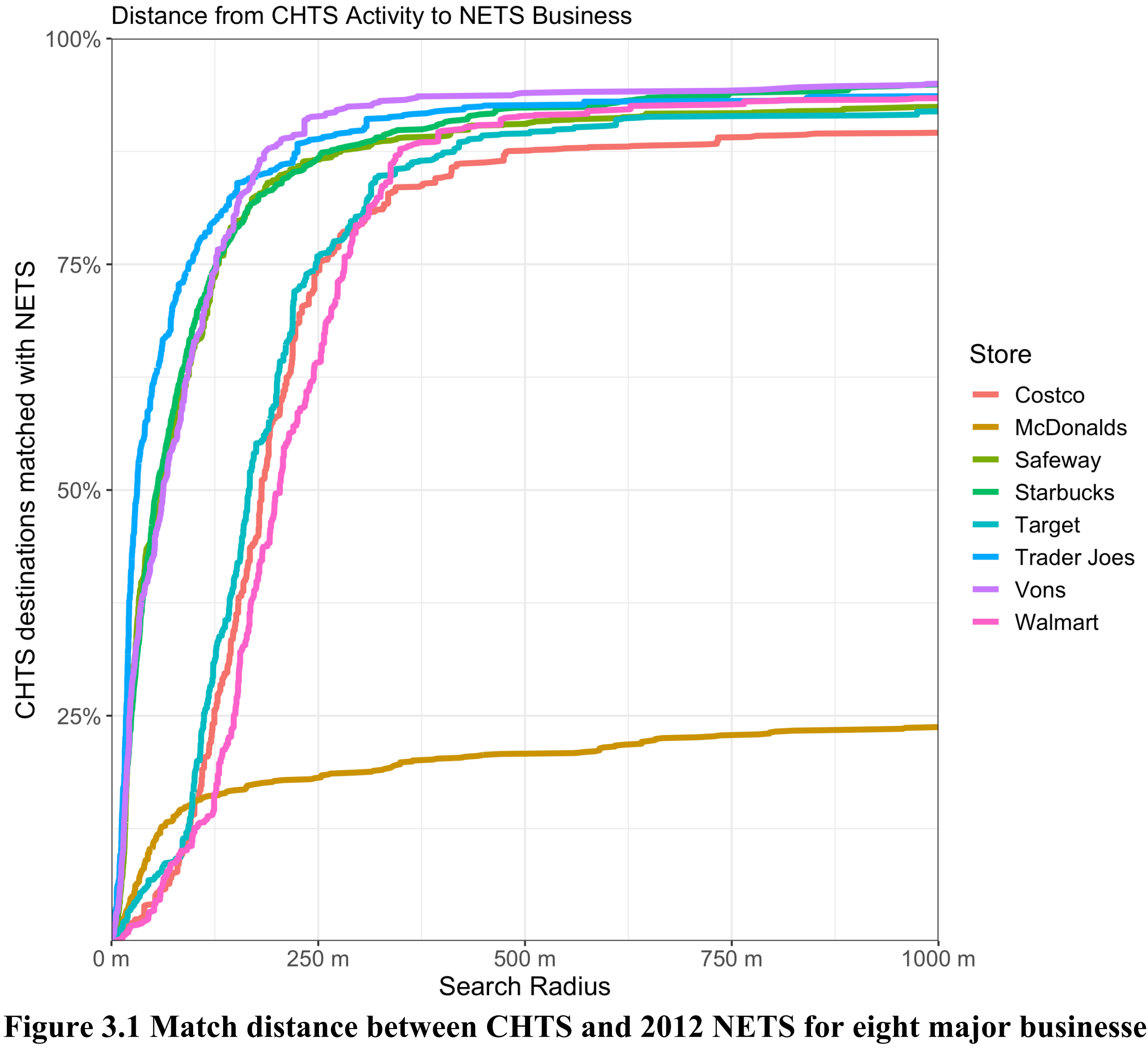


Figure 3.1 Match distance between CHTS and 2012 NETS for eight major businesses in California. NETS contains a matching location for most destinations within 250 meters but matches much less poorly for McDonald's.

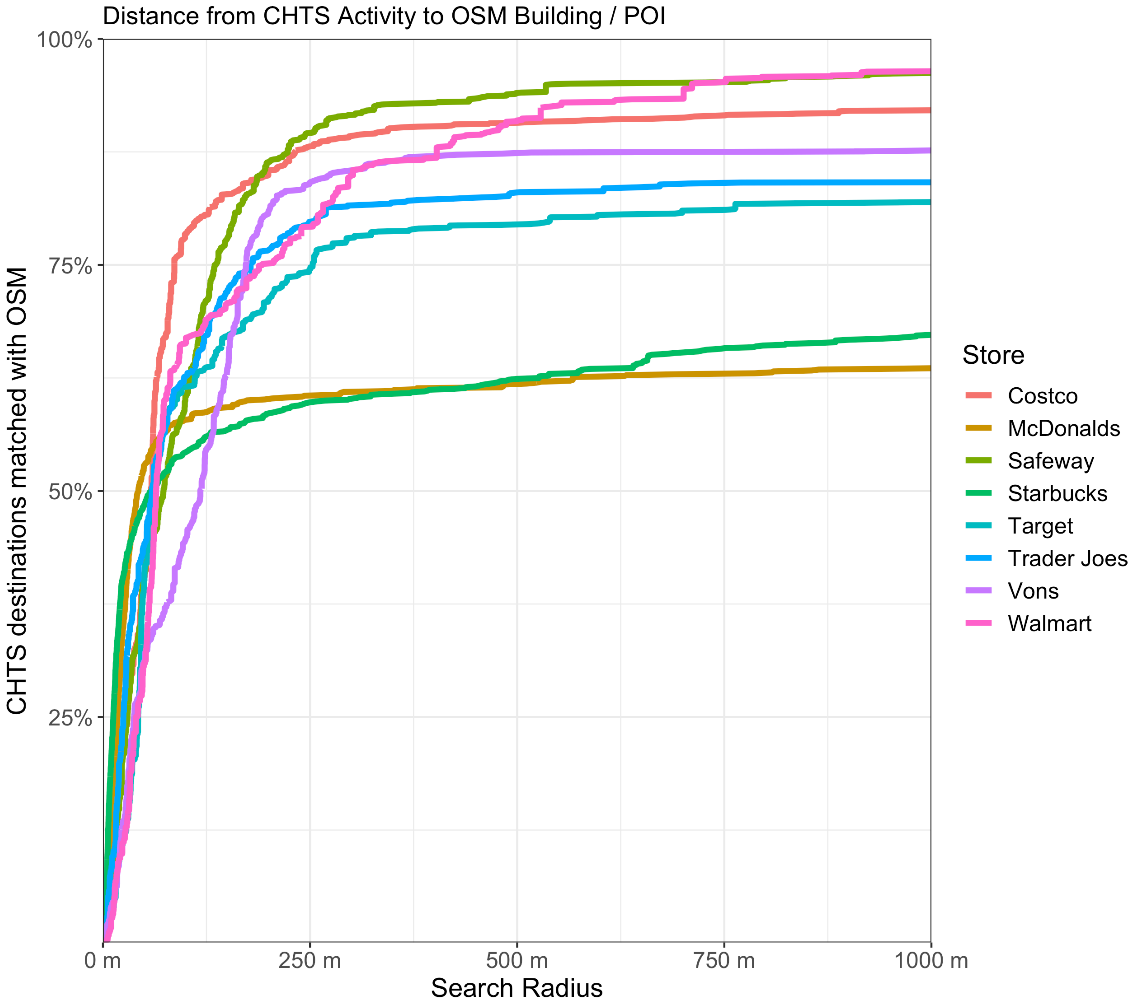


Figure 3.2 Match distance between CHTS and 2019 OSM for eight major businesses in California. OSM reaches a matching location for most destinations within 250 meters but matches worse for McDonald's and Starbucks.

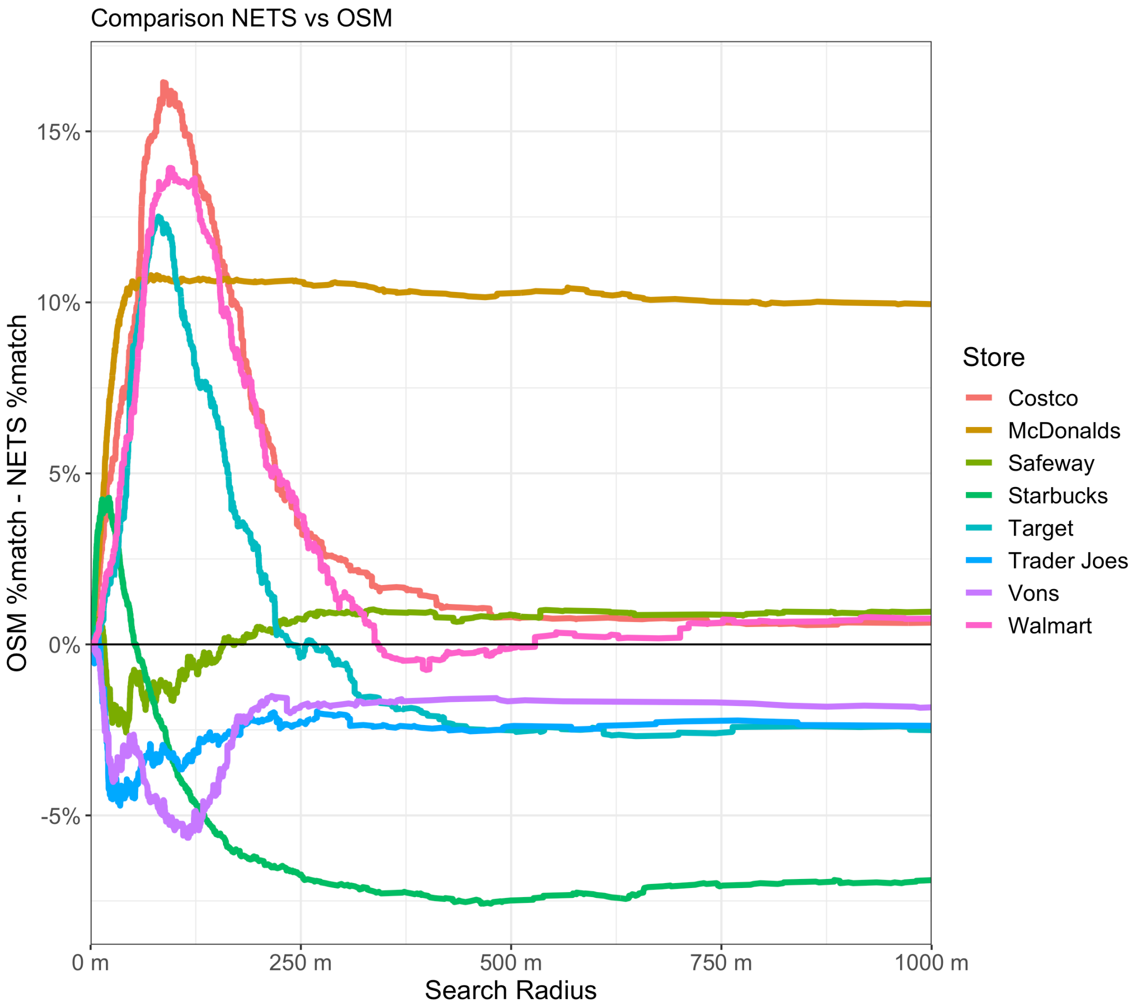


Figure 3.3 Difference in match rate over distance between OSM and NETS. OSM is generally better for McDonalds and generally worse for Starbucks. For other stores, OSM matches slightly sooner but match rates are very close for all search radii over about 300 meters.

4. Identifying Neighboring Business Establishments on the Road Network

This chapter discusses network distance computation and neighbor identification using National Establishments Time Series business establishment location data joined to an OpenStreetMap road network. Accurate identification of neighboring points is a central requirement for performing density-based spatial clustering (discussed in Chapter 5). The correct method for calculating distances on a road network is to add each location as its own node in the network by splitting the nearest road segment at the point's location, which allows shortest path distances to be calculated directly to or from that location. This process can be complicated and time consuming for large datasets, and shortcuts are sometimes used.

I assess the error introduced in distance calculation and neighbor identification from three simplified methods of distance computation: Euclidean / straight-line distance between points, network distance based on snapping all points to the nearest intersection, and network distance based on snapping all points to the centroid of the nearest road segment (see Figure 4.1). While the two other network-based methods become increasingly accurate at distances over about 500 meters, they introduce substantial error to the neighbor identification process at closer distances. My main findings are as follows:

- By dividing the state up into one-kilometer square chunks and processing the network of each one separately, it is possible to join a set of POIs to a large statewide transportation network and identify all network neighbors within a reasonable amount of time on a personal computer.

- Node-to-node and segment-to-segment distances are not much faster to calculate than accurate network distances and these approximations are only slightly more accurate than Euclidean distance is, at least for distances up to 1 kilometer.
- For distances of over 750 meters, average error from nearest-intersection and segment-based distance calculations stabilizes at 60 meters. Since the error from individual pairs of points is limited by the length of their respective road segments, the relative magnitude of this error decreases over distance.
- Simplified network distance measures are particularly bad for identifying neighbors within the distance range I consider most seriously for the density-based clustering exercise in Chapter 5.

Examples of these of the four distance measures compared here are shown for three business locations in Isla Vista, CA in Figure 4.1.

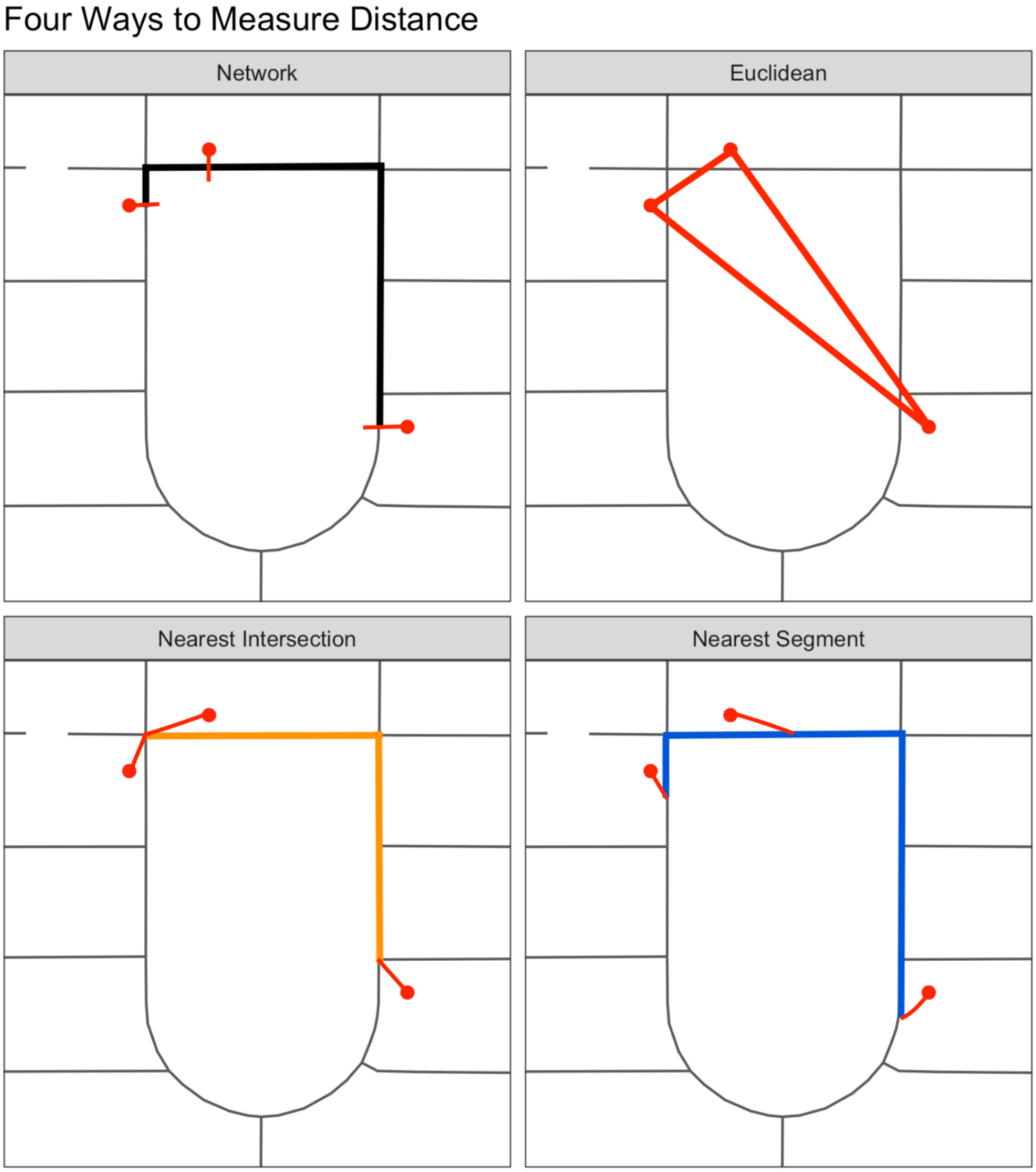


Figure 4.1 Point-to-point network distance along with three simplified ways of measuring distance. Euclidean distance (top-right) is very quick to compute but does not reflect the reality of movement along road networks. Nearest Intersection distance (bottom-left) and nearest segment distance (bottom-right) simplify network distance calculation by moving points to existing locations on the network either at intersections / nodes or segments / links.

a. Subsampling the Network to Attach POIs

For this dissertation, OpenStreetMap road network data was preprocessed using OSMnx, a Python library that downloads OSM road data, checks and cleans network topology, and exports it in multiple formats (Boeing, 2017). This preprocessing step provides an edge list that can easily be used for routing and spatial data that can be used for accurate point-to-polyline-matching, since it maintains the full spatial detail of the original Open Street Map data. For the analysis presented in this chapter, the statewide network contains 1,443,374 intersections and 1,883,110 road segments with a total length of 466,110 kilometers. Of these road segments, 86,542 have at least one business establishment and 32,222 have at least two.

The first step for calculating the network distance between business establishments is to position each location on the road network and determine its distance from the endpoints of its respective road segment. To distinguish them from network nodes and sample points, in this section I refer to business establishment locations as points of interest (POIs). It is possible to identify the individual closest points on a polyline dataset to a set of POIs analytically, but joining POIs to regularly spaced points sampled along their respective nearest road segments is a simpler process and can be made as precise as desired. It is much faster to identify nearest neighbors between two sets of points than to identify the exact nearest points on a polyline dataset. Even more importantly, regular subsampling of the road segment makes it straightforward to directly calculate the distance between each POI and the two endpoints of its road segment.

This process entails sacrificing some degree of precision in positioning POIs in exchange for processing speed but can be made arbitrarily precise by sampling points at smaller intervals. Decreasing the sampling interval increases the processing time of this step but does not affect the processing time of the network distance step, since that only requires the distance between each point and two nodes. In order to avoid visual clutter, the example figures used in this section use a 5-meter point spacing, but for the analysis in the rest of this chapter and this dissertation, distances are calculated using a 1-meter spacing between sample points. By using 1-meter spacing of sample points, I calculate POI position on the road network with error distributed uniformly between ±0.5 meters and road network distances with error distributed triangular between ±1 meter. The point position matching analysis presented in Chapter 3 demonstrates that this level of uncertainty is far below the level inherently present in the activity and business location datasets used in this analysis.

I position business establishment locations on the road network with an automated process using the following steps in R. Figures 4.2 and 4.3 demonstrate this process on a subset of business establishments and road segments in Isla Vista, CA, which is adjacent to UCSB's campus.

1. Find closest road segment to each POI. Subsampling polylines is taxing, so to save processing time, only run this process on road segments that are joined to at least one POI. Separate the POIs into groups based on which road segment they are nearest to.
2. Generate sample points on each road segment of interest (circles on road in Figure 4.2). The corresponding function in *sf* generates regularly spaced points starting at half the sample spacing from the from the first vertex of the line segment. E.g., for a sample interval of 5 meters, the first sample point will be 2.5 meters from the *from* node of the given road segment.

3. For all the POIs attached to a specific road segment, determine which sample point is nearest. Arrange the POIs by sample point index.
4. Calculate distance along the road segment from each POI to the previous point and the next point. For the first and last POI on a road, these will be the one of two endpoint nodes. Add links to the network corresponding with lengths equal to the respective distances between the POI and the two adjacent points (Figure 4.3). The distance from each point to the next is equal to the difference between their sample point indices times the sample point spacing. The distance from the first POI to the *from* node is $(i - 0.5) \times s$, where i corresponds to the index of the sample point and s corresponds to the sample spacing. The distance from the last POI to the *to* node is $d - (i - 0.5) \times s$, where d corresponds to the full length of the road segment.

Once the original network has been updated with a new node and a new link for each POI, distances can be calculated along the network from any POI to any other POI. In rare cases, the process of attaching POIs to the road network moves results in distances that are smaller than the Euclidean distance between the POIs. In these cases, I use the larger of the two distance measures.

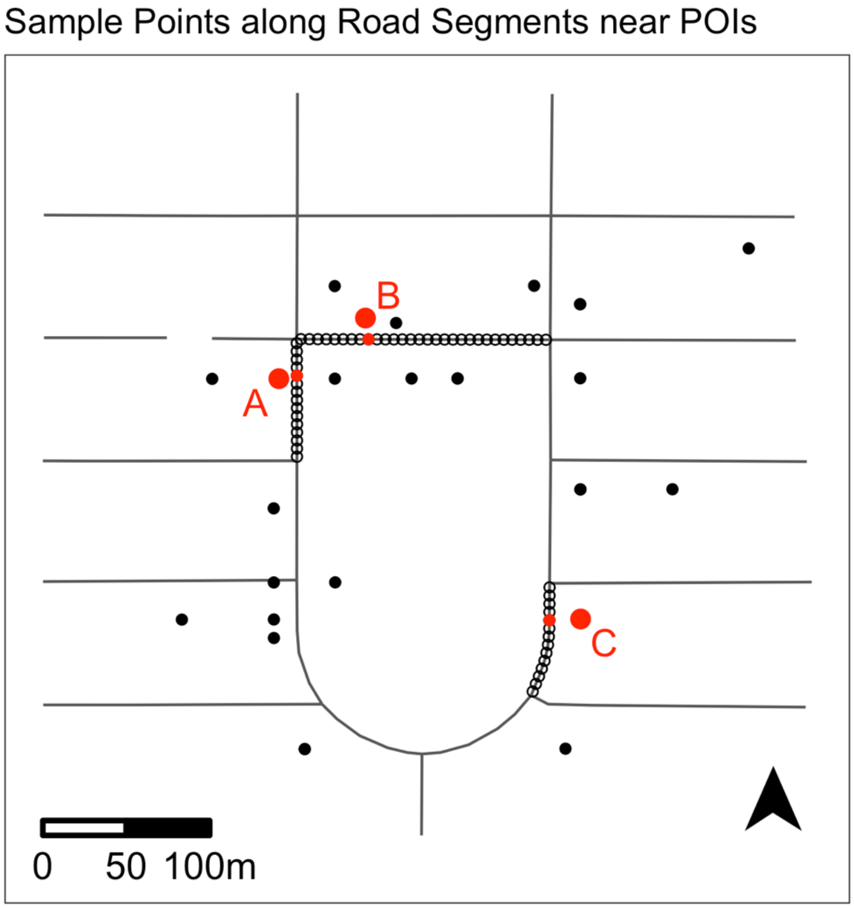


Figure 4.2 Identify the road segment nearest to each POI and sample it at a regular interval. The nearest sample point to each POI is an approximation of the nearest point on the line to that point.

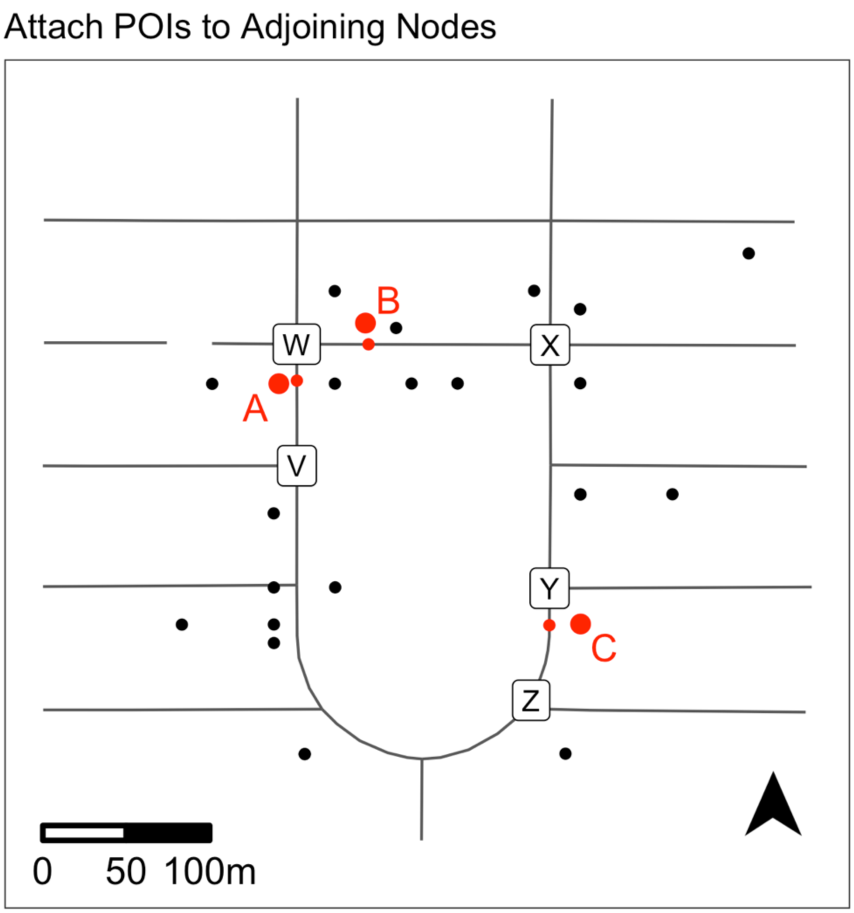


Figure 4.3 Add POIs to the network as new nodes and link them to the two nodes on the road segment they are attached to. In this example, point A will have links to nodes V and W; point B will have links to W and X, and point C will have links to Z and Y. When there are multiple POIs on a single segment, links will be created between them in sequence.

b. Network Distance and Neighbor Identification

Once the joining process is complete, the network includes all of the business establishments added as nodes, with links connecting them to the endpoints of their road segment. The next step is to compute distances.

The Dijkstra algorithm (Dijkstra, 1959) is a method to compute the shortest path distance between pairs of nodes on a network. This method spans the entire network, visiting nodes at successively increasing distance from an origin node. Because it performs a blind search out from the origin and visits all nodes in every direction that are nearer to the origin before reaching the destination, the basic form of the Dijkstra algorithm is a fairly inefficient means of calculating distances between single pairs of points (Zeng & Church, 2009). Navigation services can provide distances and travel times between pairs of points with relative ease using improved algorithms like A* and other methods that precompute the lengths of frequently used routes (Delling, Sanders, Schultes, & Wagner, 2009; Zeng & Church, 2009) or span the network simultaneously from the origin and destination to save time, but full distance matrices are often more costly to provide. The inefficiency of this method for calculating single network distances is not problematic for identifying network neighbors, since the process of identifying all neighbors of a node within a given distance is identical to identifying the route between that node and a single other node at that distance. A method similar to this was described but not fully explored by Yiu and Mamoulis (2004), and a similar method was developed for identifying particularly accident-prone stretches of road (Zhang, Han, & Kim, 2018).

It is possible to modify the Dijkstra algorithm to identify all neighbors within a set distance and stop scanning the network once that distance is reached. I achieve a similar effect by chunking both the network and list of nodes I'm interested in, and passing spatial subsets of each to the distance matrix calculator in the R package *igraph* (Csardi & Nepusz, 2006), which implements the algorithm more efficiently in C than I could in R or Python. After initial exploration of the spatial distribution of business establishment locations in California, I determined that 1 kilometer was a reasonable maximum threshold for neighbor distance, and I would likely choose as smaller value of ε, the neighbor distance in DBSCAN. I divided up California business locations into 1-kilometer cells; larger or smaller cells might improve efficiency further, but this size was acceptable for my purposes. This yielded a list of *from* nodes for my distance computation. I then buffered each region by the maximum neighbor distance (1 kilometer) and selected all other business locations and road segments that fell within this larger box. Since road distance is always at least as long as straight-line distance, the combination of the *from* points and the additional points inside the larger buffer contains all the possible network neighbors of the *from* nodes.

c. Accuracy Tradeoffs for Simplified Distance Computation Methods

In order to demonstrate the importance of using correct distance calculation methods, I also calculated straight-line (Euclidean), nearest-intersection, and nearest-segment network distances between all pairs of business establishments located within 1 kilometer of each other by actual network distance. Since Euclidean distance represents the shortest possible distance between two points using projected coordinates, in every case where network distance is smaller than the corresponding Euclidean distance, Euclidean distance is used instead. Statewide, there are 13.6 million pairs of retail, entertainment, and food service businesses within 1 kilometer of each other by road. Table 4.1 shows the upper triangle of the correlation matrix for these four distance measures statewide. The three simplified distance computation methods provide results that are generally close to accurate, and their accuracy increases at greater distances, but the distance errors are substantial, particularly for neighbor identification.

Table 4.1 Correlation of four distance measures for California retail, entertainment, and food service establishments located within 1 kilometer on the road network of each other.

	Nearest-Intersection	Nearest-Segment	Euclidean
Correct Network	0.964	0.955	0.944
Nearest-Intersection		0.895	0.926
Nearest-Segment			0.913

Euclidean distances always underestimate road network distances, but the difference varies spatially both with the overall structure of an area's road network and with the relative orientation of the straight-line distance between a pair of points and the rest of the network (Boeing, 2018; Ravulaparthy, 2013; Ravulaparthy et al., 2013). Since road network distance is always at least as long as Euclidean distance, I include a distance computation produced by scaling Euclidean distance to the average ratio between the two metrics. For California customer-serving business locations within 1 kilometer of each other, road network distance is 28% longer than Euclidean distance, on average; I include this adjusted distance measure in the figures below.

Nearest-intersection and nearest-segment network distances are calculated in the same way as point-based network distance, but instead of adding new links from each business to the two nearest intersections, businesses are assigned to their nearest intersection or to the centroid of their nearest road segment. This location reassignment is the only source of error for these methods relative to network distance calculated by adding new network nodes for each business location, and the largest potential error from either method is equal to half the combined length of the two points' respective road segments. Errors for each method are greatest when points are displaced the most: nearest-intersection distance is most accurate between pairs of establishments located close to road intersections, whereas nearest-segment distance is most accurate when both establishments are located at the centers of their respective road segments. As a result, these two methods tend to have errors of opposite magnitudes.

To compare the overall error of these measures over distance, I divide all the business pairs into 5-meter-wide bins and calculate the average of the mean absolute error for each of the approximate distance measures using Equation 4.1:

$$MAE_d = mean\left(\left|altdist_{i,j} - truedist_{i,j}\right|\right); \quad netdist_{i,j} \in d \qquad \text{Equation 4.1}$$

Given a distance range d, identify all pairs of points that have a “true” network distance within that range. The mean absolute error for the range is equal to the mean of the absolute values of the difference between alternative distances for those pairs of points and the “true” network distance measurement. For consistency, both other network distance measures are corrected to never be less than the corresponding Euclidean distance. The mean average error in these bins is shown in Figures 4.4 in meters and 4.5 by percent of actual distance. At very short distances, Euclidean distance performs slightly better than either of the simplified measures based on the road network. Both approximate network distance measures are more accurate for measuring distances over 300 meters. At distances above 500 meters, the relationship among the four measures stabilizes. Because the error introduced by displacing points from their actual position on the road segment is fixed for each point, the overall average error of both network methods remains constant in absolute terms (around 60 meters, on average) and dwindles as a percent of the measurement. Because errors in Euclidean distance relate to the overall straightness of the road network, these errors remain constant as a percent of the distance measure (about 18% in California on average) and increase linearly in meters.

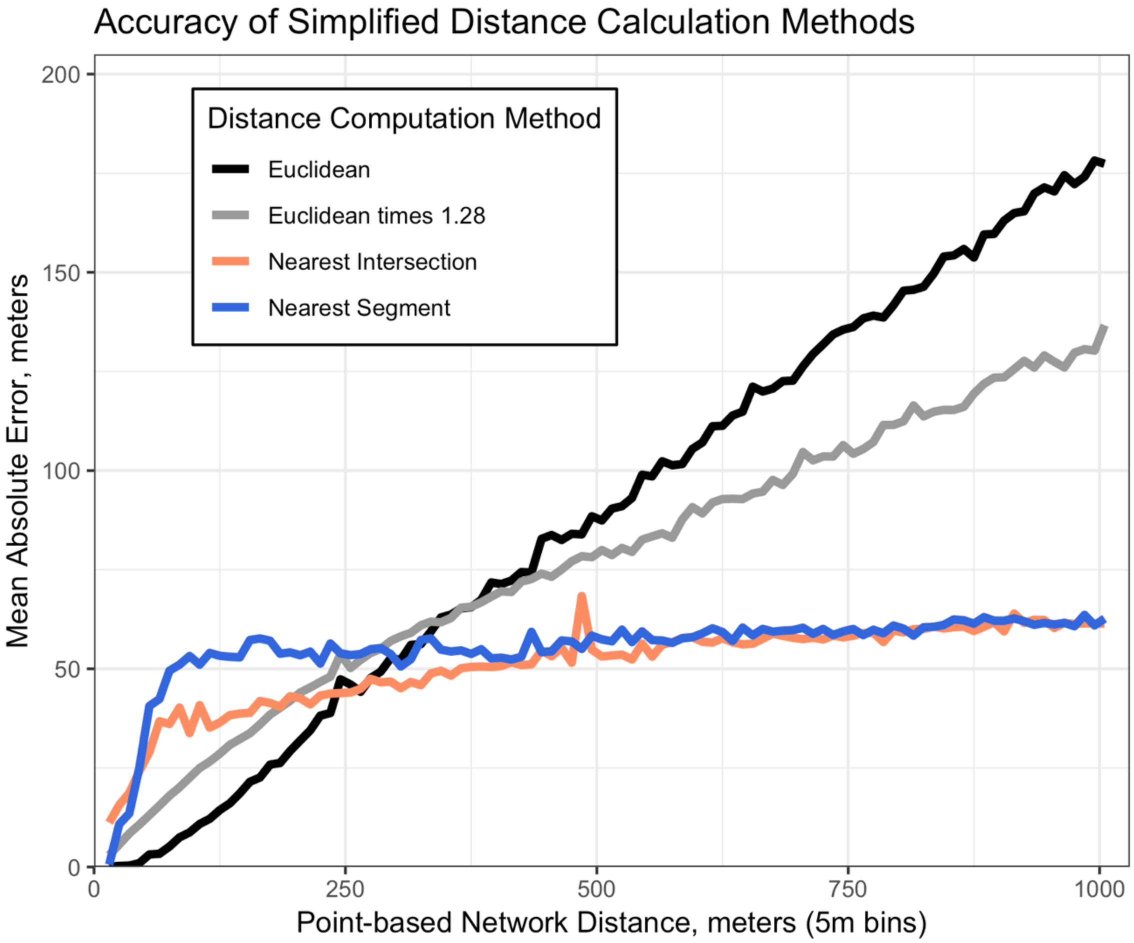


Figure 4.4 Mean absolute error of four approximate distance measures. At smaller distances, Euclidean distance performs better. Euclidean distance errors appear to increase in a linear fashion beyond 1 kilometer, whereas network errors stabilize to about 65 meters from about 750 meters on.

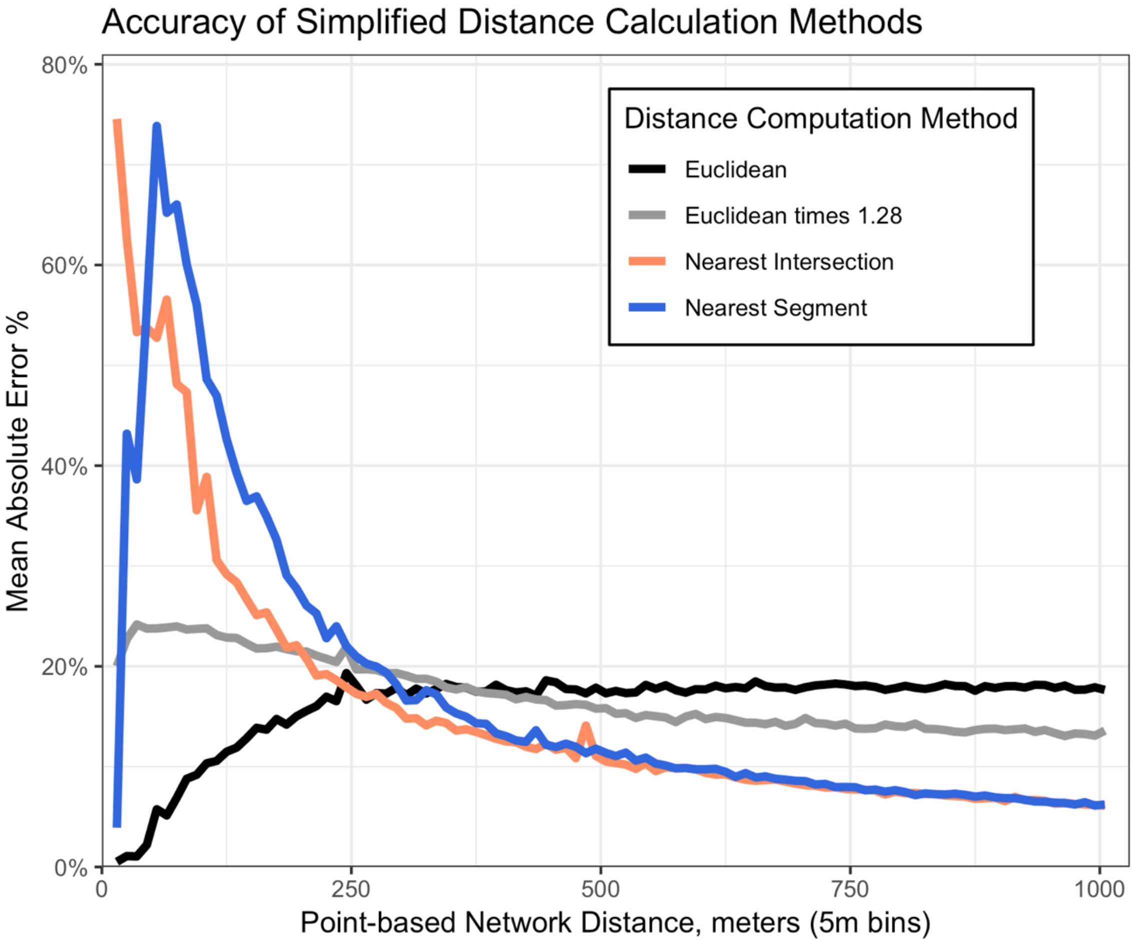


Figure 4.5 Percent absolute error for four approximate distance measures. At smaller distances, unmodified Euclidean distance performs much better, but stabilizes to around 18% for distances over 500 meters. Network distances improve at larger distances since their error can never be more than half the combined length of the road segments of the points in question.

These errors of distance measurement impact that accuracy of identifying fixed-radius neighbors between pairs of points, a major input to density-based clustering methods like DBSCAN. To test the impact of using incorrect distance measures to identify neighbors for my dataset, I used the adjusted Euclidean distance and intersection- and segment-based network distance to perform neighbor identification at a range of neighbor thresholds spaced every 10 meters between 50 and 600 meters, a range of neighbor distance thresholds I felt likeliest to use as ε, DBSCAN's neighbor distance threshold, for my final commercial center clustering in the next chapter. Two points are classified as neighbors using a specific distance computation method if the distance between them is less than or equal to ε, and non-neighbors if the distance is greater than ε. Figures 4.6-8 show the result of these tests, with the range from 150 to 400 meters highlighted. In each case, the neighbor detection results from the three simpler distance measures is compared against results from actual network distances. Since these measures require enumeration of the negative results of a test as well as the positives, I used Euclidean distance to identify all pairs of points within ε by straight line, since only these points could possibly be within ε by other measures of distance. I include Euclidean distance on the overall accuracy plot to show the overall share of points within ε by straight-line distance that also are by network distance.

As Figure 4.6 shows, adjusted Euclidean distance and the two simplified network measures produce neighbor identification results with an accuracy of between 82% and 96%. Nearest-segment network distance is the most-accurate alternative at values of ε below 100 meters, but nearest-intersection distance is slightly more accurate for values of ε within the central target range of 150-400 meters. Separating the overall accuracy measure into sensitivity (Figure 4.7) and specificity (Figure 4.8), which correspond respectively to a test's accuracy when encountering values that should be classified as positive and negative, highlights the main difference between how the two network-based distance approximations perform for neighbor identification. Using nearest-intersection distance produces more false positives because it pulls points on multiple road segments to the same intersection network node, assigning them a distance of 0 meters, which I then correct to the Euclidean distance between the points. Nearest-segment distance produces more false negatives because it pulls businesses clustered around major intersections away from each other to the centroids of their corresponding roads. Since more than 75% of businesses that are within a given value of ε by Euclidean distance are by network distance as well (black curve on Figure 4.6), the overall accuracy of nearest-intersection distance is slightly higher despite its much higher rate of false positives.

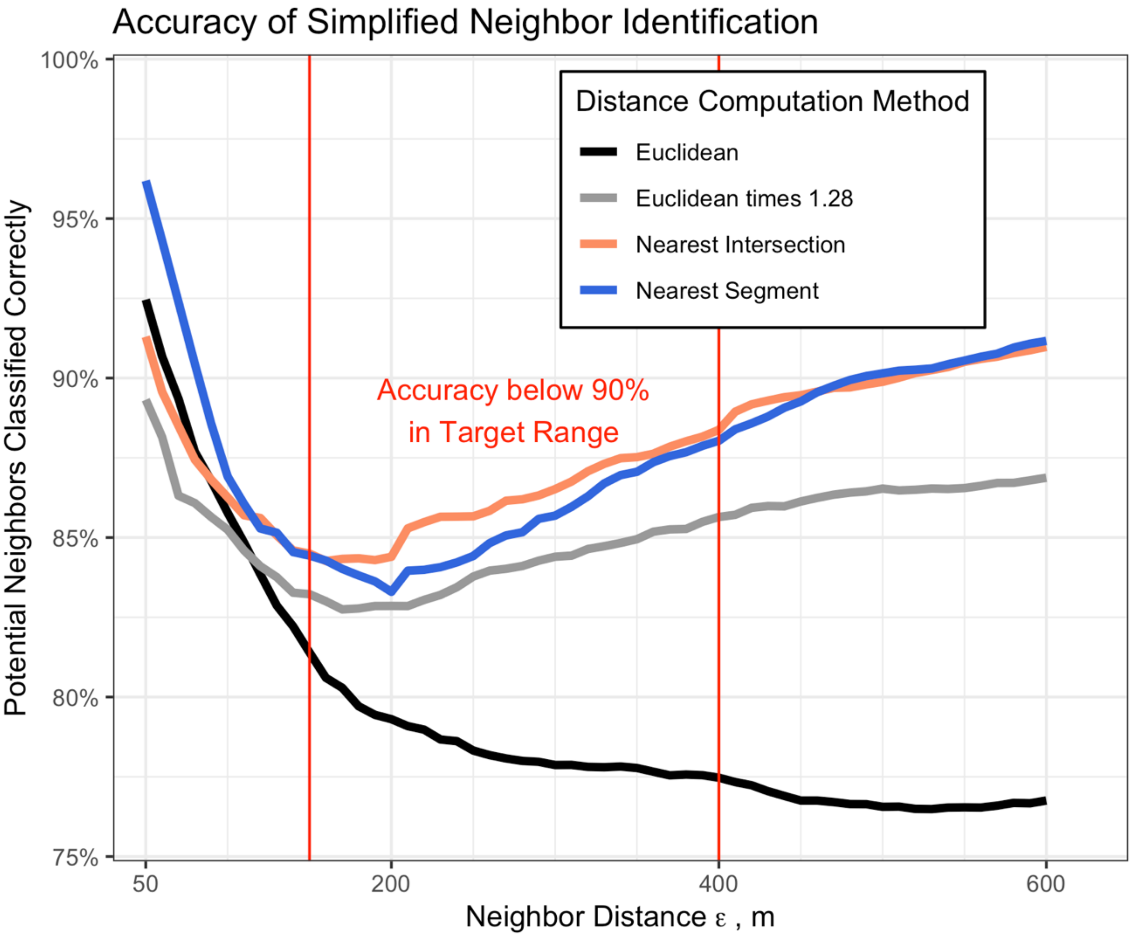


Figure 4.6 Accuracy of neighbor identification using Euclidean distance and three simplified distance measures at a range of thresholds for neighbor distance ε, measured in meters. Accuracy is percent of all potential neighbors (pairs of business establishments with a Euclidean distance less than or equal to ε) classified correctly as either neighbors (network distance less than or equal to ε) or non-neighbors (network distance greater than ε).

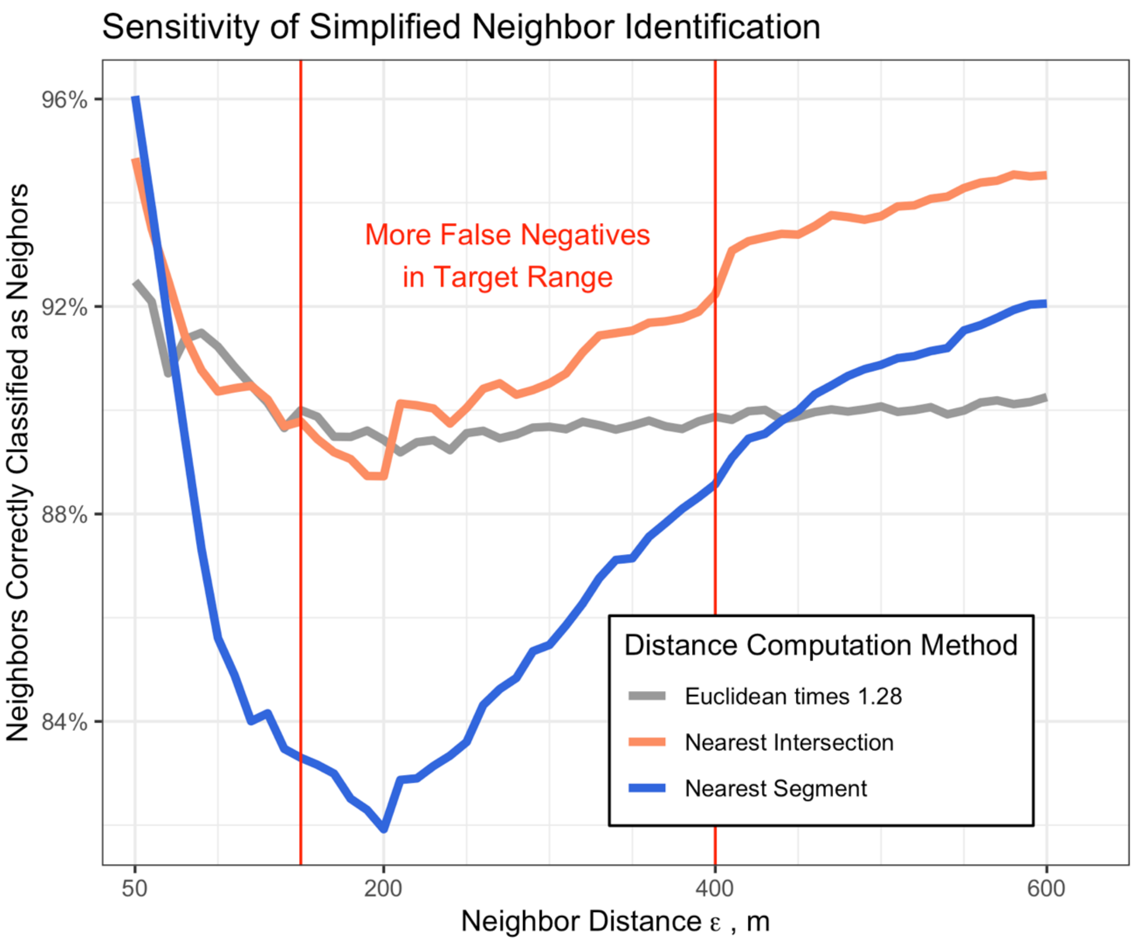


Figure 4.7 Sensitivity (true positive rate) of neighbor identification using three simplified distance measures at a range of thresholds for neighbor distance ε, measured in meters. Sensitivity is percent of all actual neighbors (network distance less than or equal to ε) that are correctly identified as neighbors using simplified distance measures.

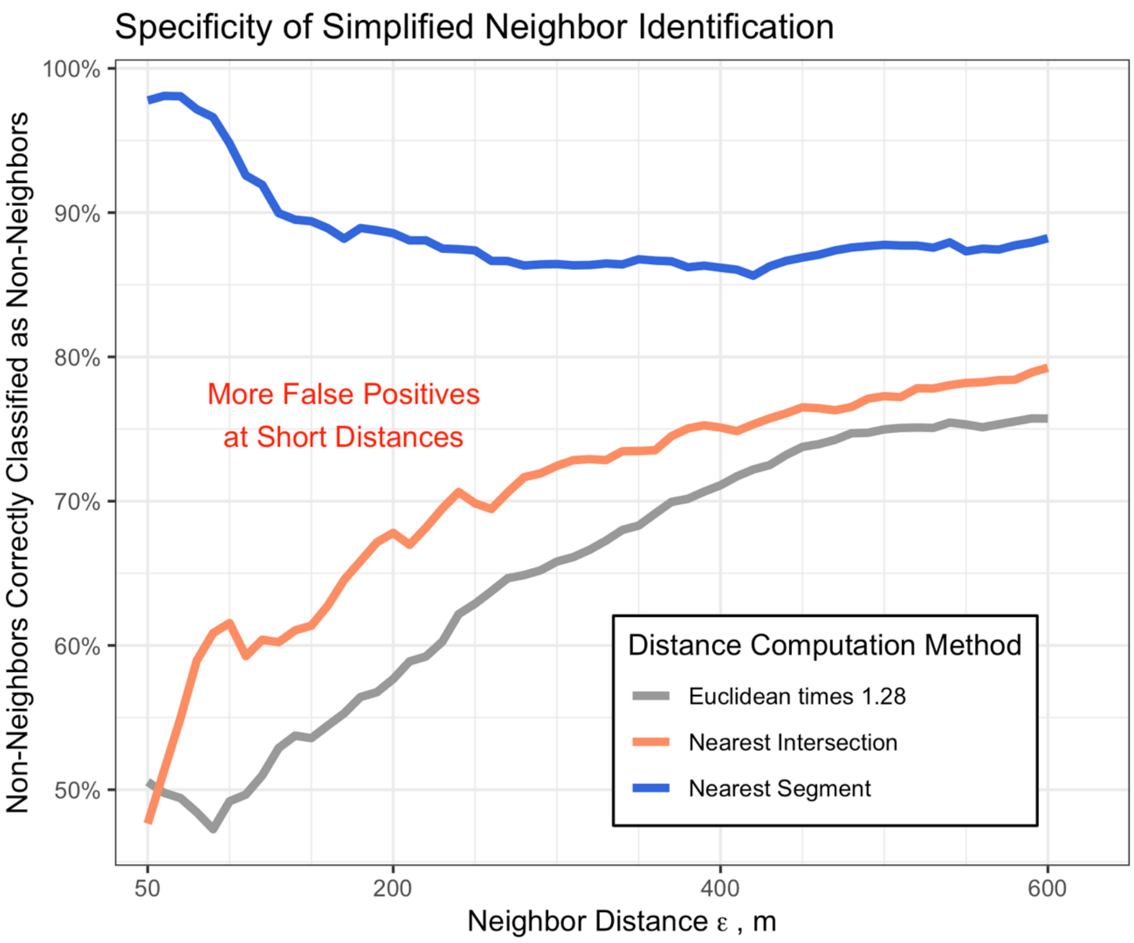


Figure 4.8 Sensitivity (true negative rate) of neighbor identification using three simplified distance measures at a range of thresholds for neighbor distance ε, measured in meters. Specificity is percent of all non-neighbors (network distance greater than ε) that are correctly identified as non-neighbors using simplified distance measures.

d. Interpolated Network Distance

For the problem of identifying neighboring retail, food service, and entertainment business establishments in California, it was feasible to add all the locations to the network directly as new nodes, but this might not be the case for a considerably denser point dataset. Igraph's method for generating a shortest path distance matrix for a network has a time complexity of O(s*|E|log|E|+|V|), where |V| is the number of vertices, |E| the number of edges and s the number of sources, or points from which distances need to be computed (igraph core team, 2015). Adding a new point to the road network will increase |V| by 1, and |E| by either 1 (if existing links are split at each point) or 2 (if new links are created between each point and the original nodes, as I did for simplicity of code). Since there are nearly ten times as many road segments and intersections in California than customer-serving business locations, adding all the business locations only slightly increases the network size and runtime for each source. While changes to |V| and |E| are somewhat limited by the size of the original network, the sources (s) term directly reflects the number of added points, since network distances must be computed from each point of interest. In this case, the total runtime of the process was still acceptable (about 40 minutes for the whole state using 1 kilometer square chunks and a maximum distance threshold of 1 kilometer when run on a 2017 MacBook Pro), but if the density of POIs were much higher (and were especially concentrated on a few road segments), it would make sense to limit the network-spanning process to existing nodes that connect links containing POIs and account separately for the distance between on-segment positions and road intersections, since these distances are fixed for a given point.

Tests indicated that the added matrix operations added negligible processing time compared to the process of spanning the road network. The number of POIs is ***P***, and the number of nodes is ***N***:

1. Subset the network to an appropriate size for the desired maximum neighbor distance (the chunking process described earlier in this chapter).
2. All POIs lie on road segments that have a *from* node and a *to* node. Identify all nodes that are either a *from* or *to* node for at least one POI. Compute the ***N***-by-***N*** network distance matrix among these nodes.

3. Pull distances from these node-distance matrices to create four new ***P***-by-***P*** distance matrices. The four matrices will contain distances between the POIs' respective *from* nodes (rows) and *from* nodes (columns); *to* and *from*; *from* and *to*; *to* and *to*.
4. Add the distance from each POI to its *from* and *to* nodes in the corresponding row and column in each matrix. When pairs of points are on the same segment, subtract the distances instead of adding them (this can bet done with *ifelse* in R). The resulting four matrices now correspond to a set of four candidate shortest-path distances between all pairs of POIs, with each one essentially representing a distance based on departing the first POI and arriving at the POI in two specific directions.
5. Use *pmin* or an equivalent function to get the cell-by-cell minimum value from these four matrices, providing the overall shortest-path distances.

e. Conclusions

This chapter demonstrated that it is feasible to use the road network to identify fixed-distance neighbors up to 1000 meters for all of California. These results are used as an input to the density-based clustering process described in Chapter 5. Comparisons suggested that simplified network distance measures are generally very accurate for calculating large numbers of distances if the target distance is over a kilometer, but nearest-intersection distance is only slightly faster to run and nearest-segment distance is much slower.

This analysis has a few notable limitations. It ignored one-way roads and turning restrictions because I am most interested in identifying neighbors at spatial scales that are reasonably walkable, but these features are vitally important for calculating road network distances and directions for driving purposes. Additionally, the methods discussed here would be considerably slower with a denser distribution of points or if longer neighbor distances were required. I did not test different chunk sizes since 1000x1000 meters worked well enough for my purposes, but there are likely more efficiencies to be found there (or by directly coding a stop function into the shortest path algorithm).

5. Using Network-Distance DBSCAN to Identify Commercial Centers in California

There are a number of reasons why the sorts of spatial units used in travel behavior analysis are a particularly poor fit for modeling centers of commercial activity, as well as the form of commercial centers, given the urban economic and historical geography of California. To extract commercial centers for this analysis, I propose using spatial clustering to identify concentrated centers of retail, food service, and entertainment locations that are likely to attract high densities of shopping, dining, and entertainment-related activities. This clustering should identify discrete areas with a dense and diverse range of opportunities for designated activities, and the resulting centers should "make sense" in familiar areas and capture various spatial arrangements of business locations. In practical terms, it must be possible to unambiguously join individual activity locations to a specific center; preliminary tests suggested that enclaves (highly dense clusters largely surrounded by another cluster with minimal gaps in between) were particularly problematic for this goal. While clusters of relatively consistent size would be better for future applications with destination choice models, not all activity and travel analyses require that, and it may be acceptable to capture all of a "downtown" or "main street" area in one cluster.

This chapter presents a network-distance implementation of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) as a method to identify centers of concentrated commercial centers retail, food service, and entertainment business locations in California. It discusses the methods by which parameters of this clustering algorithm were tuned to produce a final clustering with the desired characteristics, presents local maps of my final clustering, and discusses how the new method relates to other attempts by geographers and planners to divide and characterize urban areas.

DBSCAN is a well-established deterministic spatial clustering algorithm that identifies sets of densely packed points but does not assign points in low-density areas to any cluster. I chose DBSCAN for this application for a number of key considerations:

- Unlike many clustering methods, it does not require a set number of clusters as an input and can identify clusters of various shapes and sizes.
- It separates areas of relatively high density into discrete clusters to enable comparative analysis.
- It distinguishes between areas of high expected activity density and areas of relatively low expected activity density. This is particularly beneficial here because the activity locations reported in the CHTS represent a relatively small share of the businesses in California, whereas NETS (the data source used in clustering) is ostensibly exhaustive.
- Since it only requires distances between points (and not other information about point locations), DBSCAN can easily be adapted to use any distance measure.

DBSCAN and similar methods are used for a wide range of data clustering purposes, particularly when it is necessary to identify particularly high-density regions from a set of point observations dispersed over a larger area. Many applications draw from geotagged social media data, and specific applications have attempted to identify distinct points of interest from sets of posts recorded with random error in their geolocation (Maddimsetty, 2018, Orenstein et al., 2014) and extract regions of interest from large collections of geotagged photos (Hu et al., 2015). Spatial clustering is relatively new in transportation research, although it has recently been used to identify major tourist regions in Florida using geotagged tweets (Hasnat & Hasan, 2018) and in traffic accident hotspot analysis, where a method using road-network DBSCAN performs much better than existing methods at identifying dangerous stretches of road (Zhang et al., 2018). The most similar application I found came in the form of a blog post by the author of the package I used to extract OSM road network data: Boeing uses the NETS dataset and OSM roads to demonstrate the importance of using road network (rather than Euclidean) distances in clustering, particularly when clustering is done for an irregular road network broken up by streams and freeways (Boeing, 2018b). That application used node-based distances rather than true network distances, which Chapter 3 suggests is only slightly better-suited for this purpose than straight-line distance.

Although it is straightforward to implement and generally provides useful results, DBSCAN has a number of notable drawbacks that arise when points are not distributed consistently over the study area, either through directional dependence (anisotropy) or varied density. If points demonstrate consistent directional trends in space – either locally or in the whole study area – using Euclidean distance may not be appropriate; ADCN presents one potential solution by favoring the major axis of an ellipse constructed from each point's potential neighbors (Mai, Janowicz, Hu, & Gao, 2018). This dissertation presents another approach using road network distance, which is applicable when the points being clustered are constrained to a network both in terms of location and interaction.

Conventional DBSCAN can struggle to distinguish clusters when density varies so widely throughout the study area that a single set of parameters will fail to detect actual clusters in low-density areas and/or fail to identify breaks within large clusters in high-density areas. OPTICS addresses this by providing a hierarchical clustering that resembles DBSCAN but with a neighbor distance threshold that varies over space (Ankerst, Breunig, Kriegel, & Sander, 1999). Testing OPTICS with NETS data and Euclidean distances revealed numerous enclave clusters at every value of the cluster separation parameter Xi. OPTICS could be run with road network distances as well, but the implementation in the R package *dbscan* does not accept a fixed radius nearest neighbors list or a sparse distance matrix for OPTICS (Hahsler, Piekenbrock, Arya, & Mount, 2018).

The rest of this chapter presents an overview of the DBSCAN algorithm and results from testing a range of parameters with network-DBSCAN and concludes with a justification for the final clustering model used in this analysis.

a. DBSCAN Method Overview

To extract clusters, DBSCAN requires a set of data points, a distance function, and two parameters: the fewest total points a point must neighbor to count as a core point (minPts), and the maximum distance between two points for them to count as neighbors (ε). For this analysis, I create a list of fixed-radius (ε) network-distance neighbors for all business locations, and use this as an input to an efficient implementation of the DBSCAN algorithm in the R package *dbscan* (Hashler et al., 2018).

DBSCAN identifies clusters using a three-step process:

1. For each observation p, count the total points with a distance of less than ***ε*** from p (including p). Flag all p that have at least minPts neighbors as core points.
2. Identify all core points within ε of each other as neighbors and extend transitive neighbor-status to all the core points neighboring each of these, so that all core

points are neighbors if they are within ε or can be connected by a string of other neighboring core points.

3. Each group of neighboring core points is a single cluster. All points that do not meet the minPts threshold but are neighbors of one or more core points count as edge points for all clusters containing core points that they neighbor. All points that neither meet the minPts threshold nor are adjacent to any points that do are classified as noise. Edge and noise points are not used when attaching activity locations to clusters.

In order to cluster points using DBSCAN, we must choose values for its two parameters: minPts and ε (maximum distance between neighbors). Attempts have been made to automate the selection of parameters for density-based clustering methods, but these parameters still be chosen with consideration for the phenomenon being clustered (Karami and Johansson, 2014). The minimum points parameter controls the minimum acceptable size for a cluster; various rules of thumb have been proposed for selecting this value, with many recommendations placing the minimum usable value at around two times the number of dimensions in the data unless there are numerous duplicate points in the data, which in this case suggests a minimum acceptable cluster size of around 4 or 5 (Sander, Ester, Kriegel, & Xu, 1998; Schubert, Sander, Ester, Kriegel, & Xu, 2017). The main downside from using too small a value for minPts is that multiple relatively distant clusters will be joined if there are lines of points connecting them. Neighbor distance threshold ***ε*** should be set at a scale that is meaningful for the data clustered (Schubert et al., 2017), often using the subjective "elbow method," which looks for the distance at which the plot of distance to kth-nearest-neighbors flattens. Nearest-neighbors curves for this dataset are shown in Figure 5.1

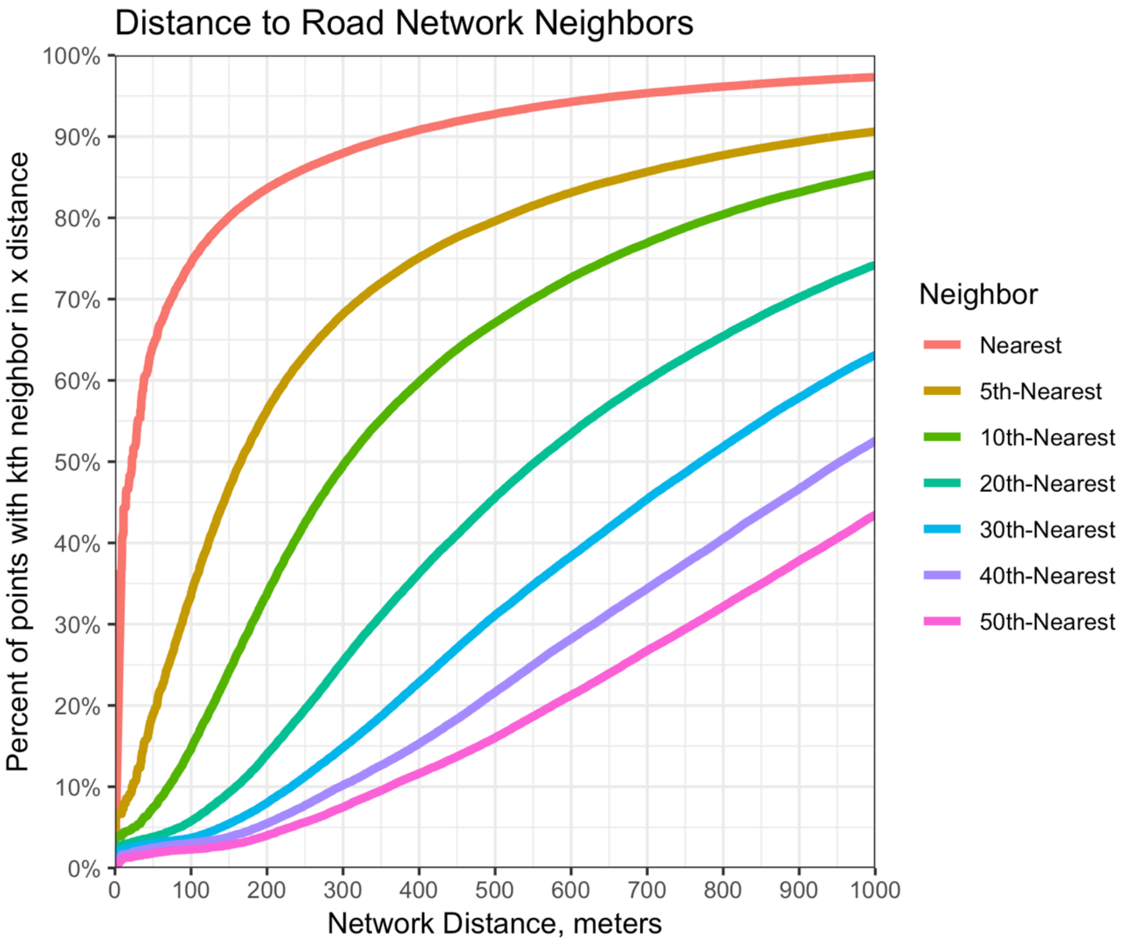


Figure 5.1 Network-distance kth-neighbor plot for California customer-serving businesses. "Elbows" are clearly visible in the curves for k=1, 5, and 10.

For this analysis, I consider all distances under 1 kilometer (roughly 15 minutes walking for young and middle aged people, per Knoblauch, Pietrucha, & Nitzburg, 1996) to be potentially acceptable distances. I test clustering results for a range of distances up to 1 kilometer, but the network-distance nearest-neighbors plot provides some useful information (Figure 5.1). Each curve is built by determining the distance to the kth nearest neighbor of each point, arranging by distance from smallest to largest, and converting to a cumulative percentage. The height of the kth-neighbor curve at a specific distance is the share of locations that have their kth-nearest neighbor within that distance. The single nearest-neighbor curve shows that most business locations are very close to other businesses – 75% reach their nearest neighbor within 100 meters, and the 5th and 10th nearest neighbor curves show that half of all businesses reach those neighbors within about 160 meters and 300 meters, respectively. The fifth-nearest neighbor curve appears to "elbow" in the range of 200-400 meters, which suggests this may be the appropriate value of ε to use for a clustering with minPts=5, the lowest value considered in the initial search.

b. Tuning DBSCAN Parameters

It is useful to explore potential options within a range that makes sense for the data. Since the goal is to identify major destinations for non-mandatory activities, we use locations at which people pursued these activities as a piece of secondary information to compare the clustering results. For business location data, a low value minPoints threshold seems to be ideal (especially since a development with 5 stores and restaurants would count as a local center). Lower neighbor distance limits make it possible to distinguish between clusters, but very low distances capture a smaller share of relevant activity locations, in part due to a mismatch between geocoding results (Chapter 3).

Given a network neighbors list, DBSCAN runs very quickly, so I test parameters over a range of ε, every 50 meters between 100 and 1000 meters, and a range of minPts spaced every 5 between 5 and 100 points, producing a grid of results. Figure 5.2 shows the percent of business locations in a cluster for each of these 380 sets of parameters, and Figure 5.3 shows median cluster size. Previous cluster experiments with Euclidean distance worked well with a minPts of 5 and ε of 200 meters. Unsurprisingly, the largest share of businesses in clusters occurs when the smallest clusters are permitted (minPts=5) and the threshold for neighbor status is at its largest (ε=1000 meters), but at relatively low values of minPts, the share of businesses in a cluster is not particularly responsive to changes in ε. This suggests that testing varied values for ε might be the better approach to reach the desired properties once a set of potentially usable clusters has been identified. Median cluster size appears to be strongly related to the value of the minPts parameter, with higher values of minPts only detecting larger clusters. There is a slight tendency towards larger clusters when ε is larger because nearby clusters are linked, which is not desirable given the goal of identifying small separate centers when present. This and the approximate elbow locations for the smaller values of minPts suggest that a value of ε below 500 meters would be a good choice.

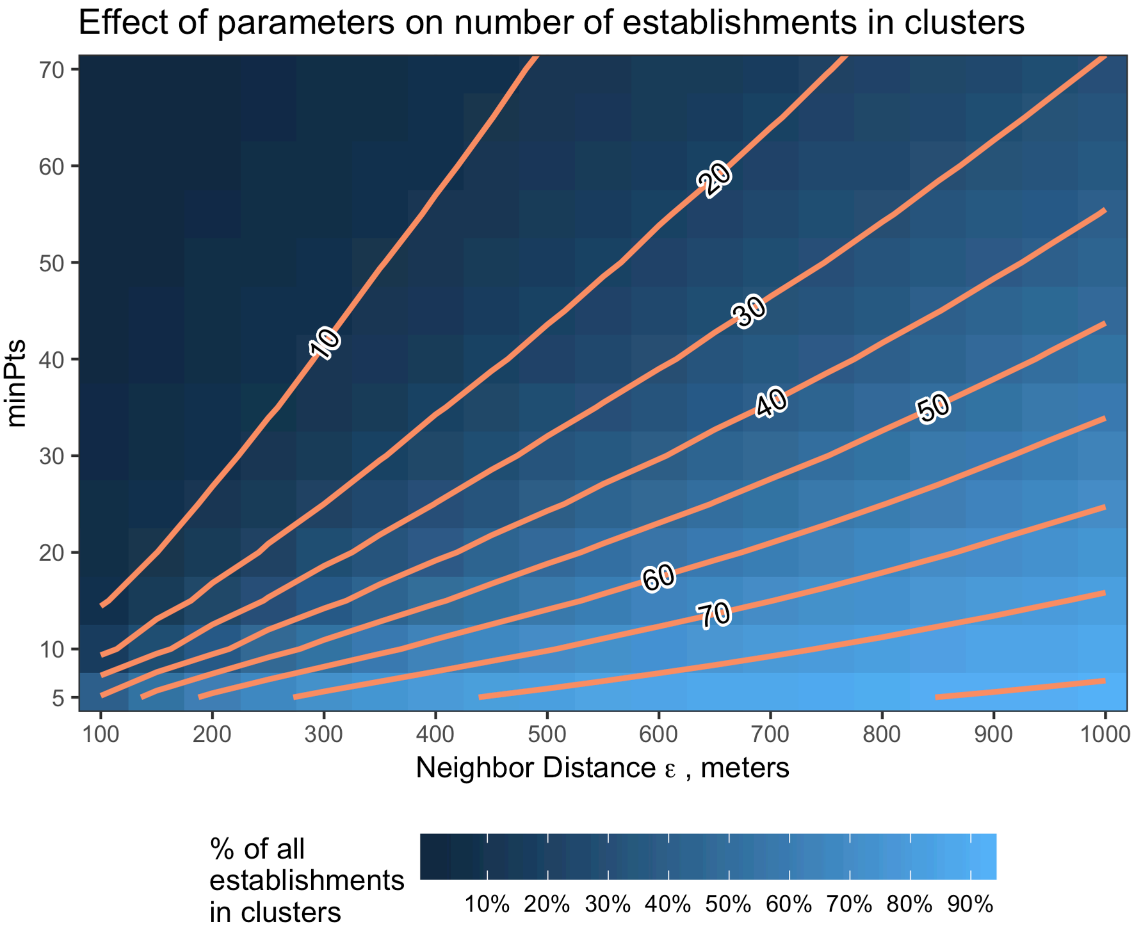


Figure 5.2 Percent of businesses classified using network DBSCAN with 380 sets of parameters. Small values of minPts paired with high values of epsilon include most relevant businesses in centers but also combine many centers across large areas.

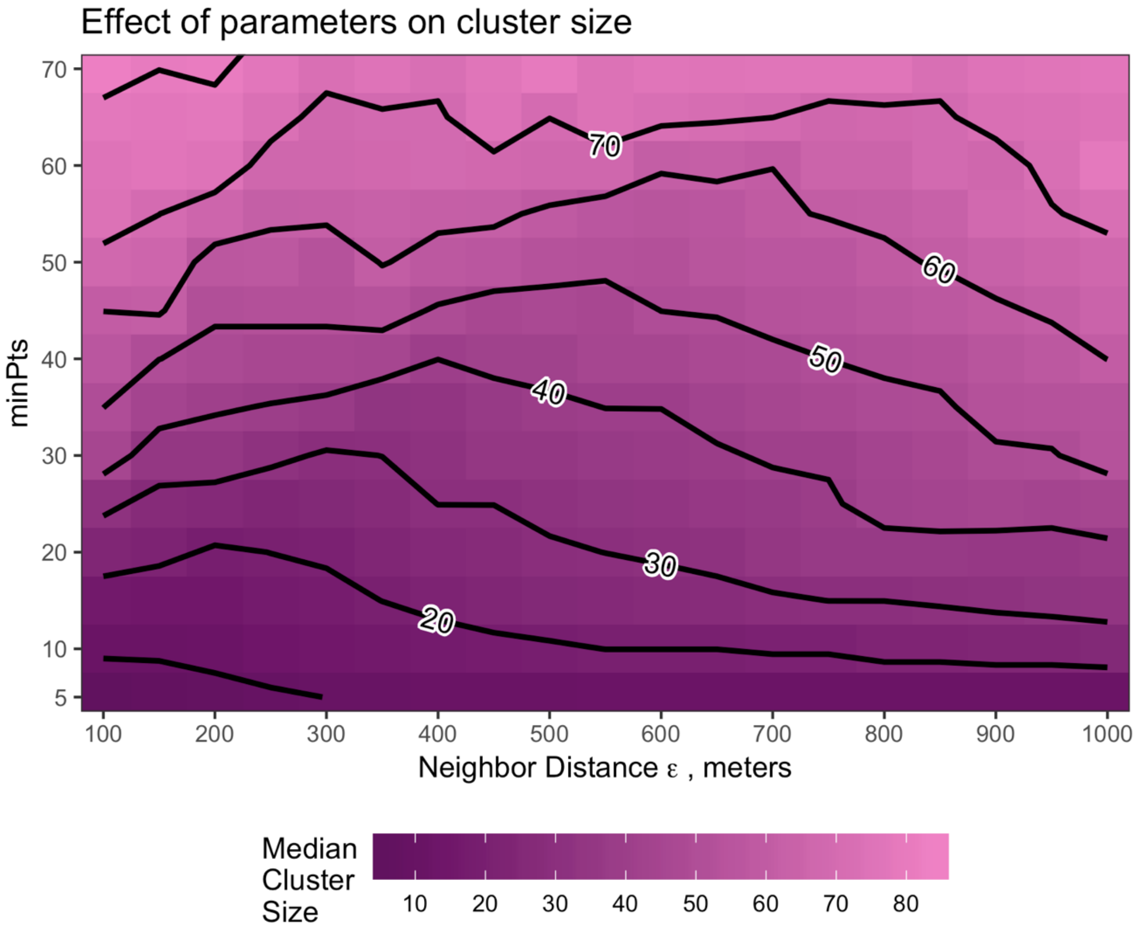


Figure 5.3 Median cluster size from network DBSCAN with 380 sets of parameters. Clusters tend to be smaller with lower values of minPts, but size has an irregular relationship with neighbor distance.

Mapping the clusters produced by a few of these tests indicates that pairing a low ε with a high minPts (upper left in Figures 5.2 and 5.3) identifies only the largest urban cores. Increasing the ε while maintaining a high value of minPts (upper right) expands these major downtown clusters laterally and adds a few other business areas. Lower values of minPts allow for the detection of far more clusters in a much larger share of the state. High ε parameters paired with the lowest values of minPts (lower right) classify almost all developed parts of the state into clusters and join all nearby clusters so that, for instance, most of Los Angeles County is covered by a single cluster.

Since the goal of this research is to identify areas of densely packed opportunities for shopping, dining, entertainment, and socializing, I use classified CHTS activity locations as secondary information to aid in selecting a final clustering. For each set of parameters, I assign every activity location to the cluster of the nearest business in a cluster as long as the distance was no greater than 200 meters. Figure 5.4 compares the various clustering results in terms of the share of shopping activities and home locations within a cluster, and the results are much the same for entertainment and dining. My goal for this analysis is to maximize the inclusion of observed shopping destinations while minimizing the share of home locations in clusters, since although some people live downtown, commercial centers should not cover areas that are primarily residential. The frontier at which home locations are minimized and shopping locations are maximized corresponds to the lowest values of minPts, with lower values of ε capturing fewer home and shopping locations. The tradeoff between excluding home locations and including shopping locations appears to shift with ε between about 150 and 300 meters. Below this range, increasing $\boldsymbol{\varepsilon}$ greatly increases the number of shopping destinations assigned to a center without adding more home locations; above this range small improvements in matching shopping destinations come at the expense of a considerable increase in the residential areas assigned to a center. These results also suggest that it may be worthwhile to test a lower value of minPts with correspondingly small values of ε.

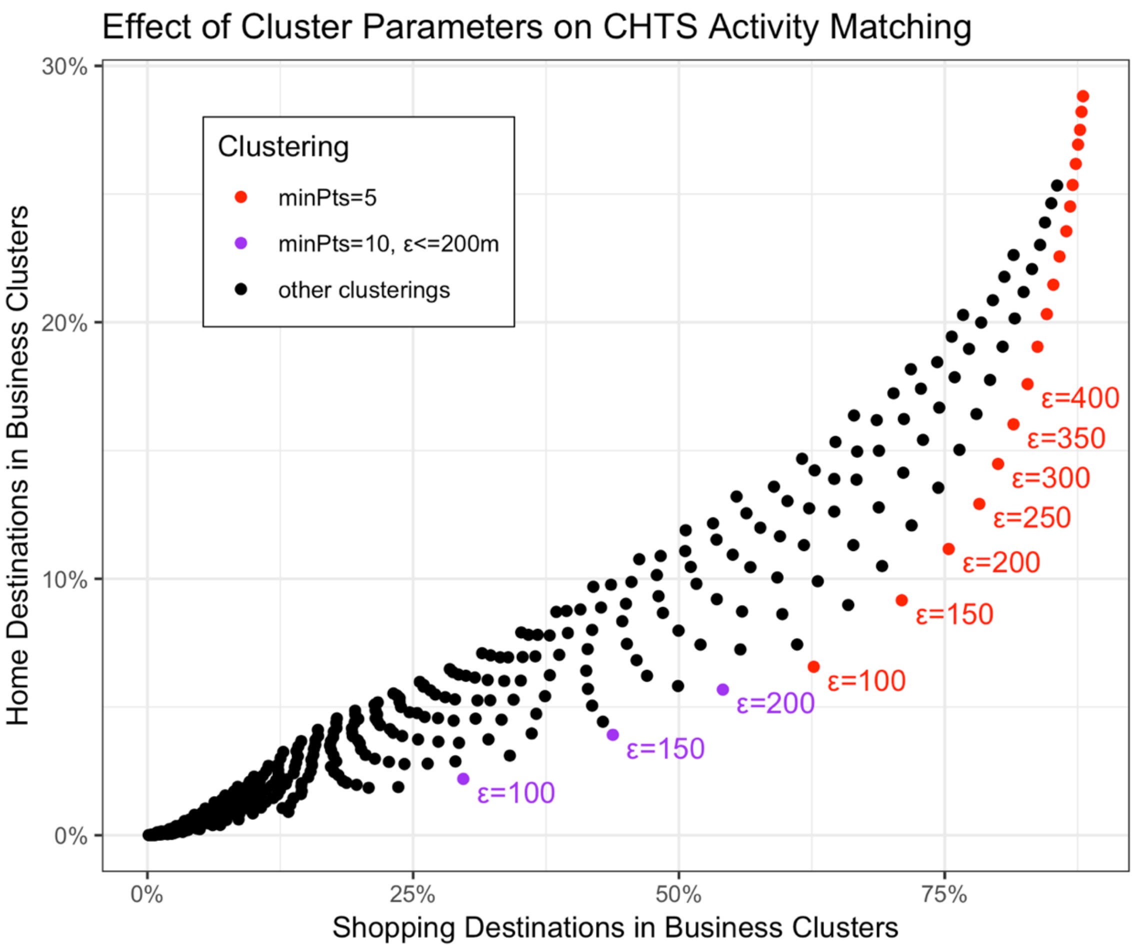


Figure 5.4 Relative shares of CHTS shopping and residential locations falling in clusters of NETS businesses at various sets of parameters. The lowest value of minPts forms the frontier plot.

Since small values of minPts seemed to perform best by all my initial tests, I decided to run an additional round of clusterings for minPts=4 and 5, with ***ε*** spaced every 25 meters up to 300 meters and every 50 meters up to half a kilometer. As Figure 5.5 shows, the clusters with minPts=4 represent only a slight improvement over minPts=5 in terms of match rate for shopping destinations but might provide substantially different cluster shapes for the same overall match rate. In particular, the minPts=4 clusters may achieve higher overall clustering rates without combining as many neighboring clusters. Mapping the results from a number of these clusterings indicated that the main tradeoff in cluster form was between the expansion of large clusters at higher values of ε, which prevents me from distinguishing between different parts of a city, and the appearance of small enclave clusters within and overlapping clusters around the major centers at lower values of ε, which makes it difficult to definitively match activities to a specific cluster.

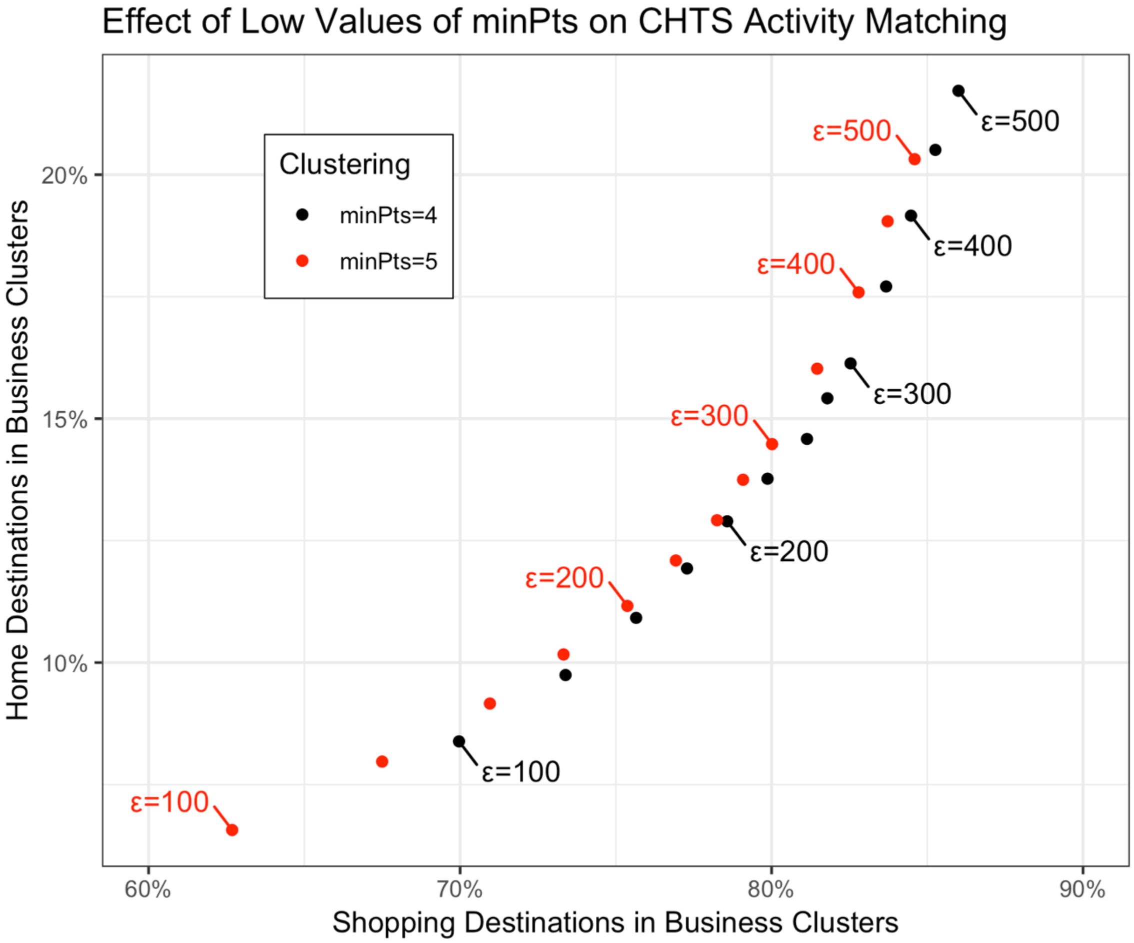


Figure 5.5 Relative shares of CHTS shopping and residential locations falling in clusters of NETS businesses for clusters with minPts set to 4 or 5.

c. Clustering Issues

In choosing my final clustering, two major cluster form issues hampered my analysis: sprawling multi-center clusters were detected when the neighbor distance threshold (ε) was set too high, and small overlapping clusters cropped up around larger centers when the threshold was set too low. These results reflect reality to some degree and are not wrong *per se* – agglomerations capture areas of generally continuous high commercial density, and the separation between small adjacent clusters does indicate a localized decrease in density – but they do not form a useful basis for analysis. I explore these results in a pair of maps of central LA (Figures 5.6 and 5.7) produced from clusterings with ε values that bracket my final selection. I represent clusters in these maps using polygons created by combining the 200 meters Euclidean distance circular buffer around all the within cluster business establishment points, which corresponds to the area in which CHTS activity locations could potentially be matched to these centers.

In areas where density is consistently high, even moderate values of ε erase the distinctions between urban subcenters by linking strings of large clusters together across whole regions. In Figure 5.6, a clustering with minPts set to 4 and a neighbor distance threshold (ε) of 300 meters identifies two very large centers that span most of central Los Angeles. In addition, it identifies a single continuous cluster along Ventura Blvd, which links the downtown areas of several cities on the southern edge of the San Fernando Valley (Suisman, 2014), and it combines multiple distinct shopping areas in Glendale into a roughly star-shaped cluster. Since I am interested in studying the effect of locally available opportunities on people's time use, I would prefer to identify centers that a person could reasonably walk across rather than allowing "local" to span tens of miles while ignoring smaller centers nearby.

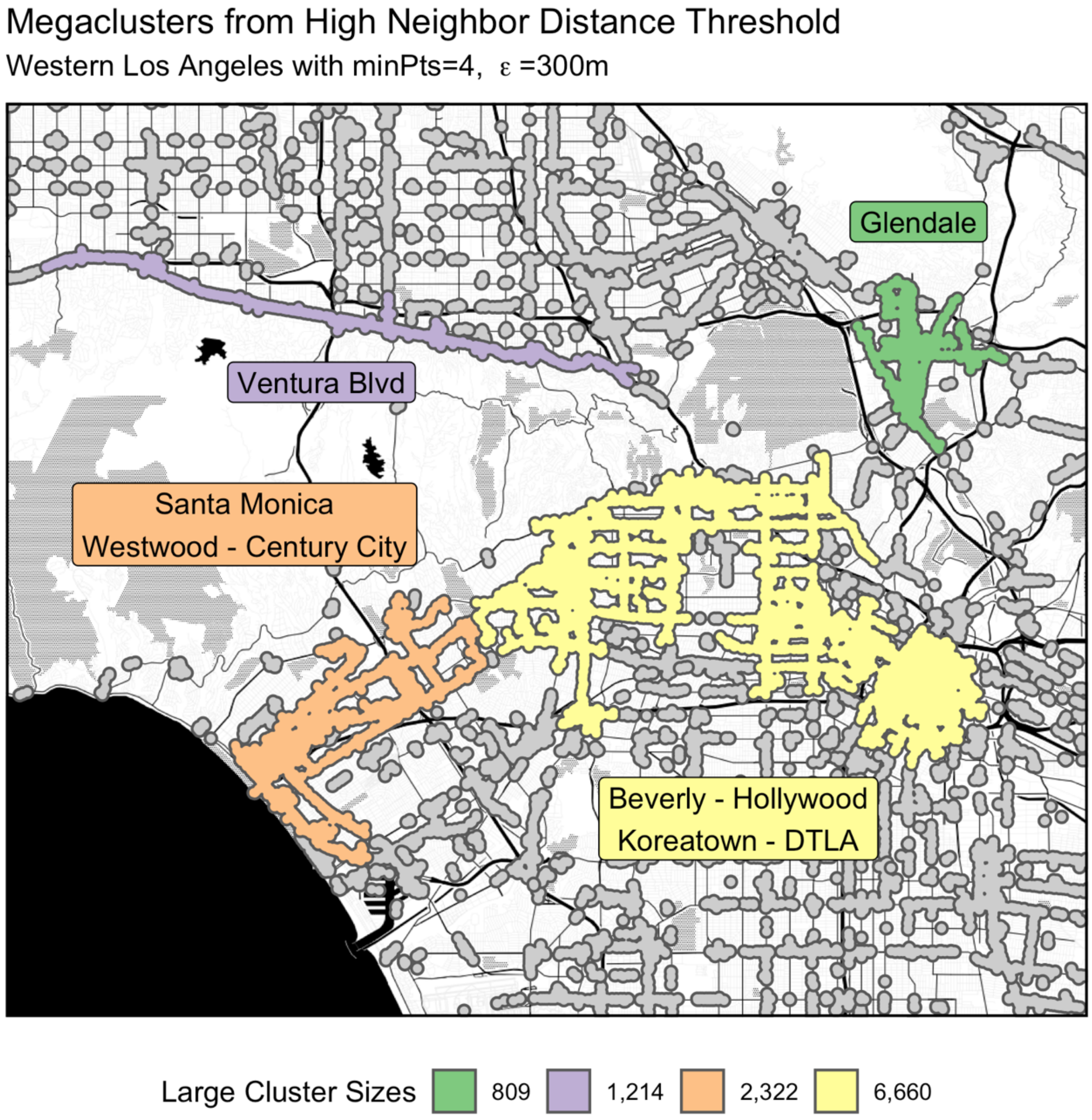
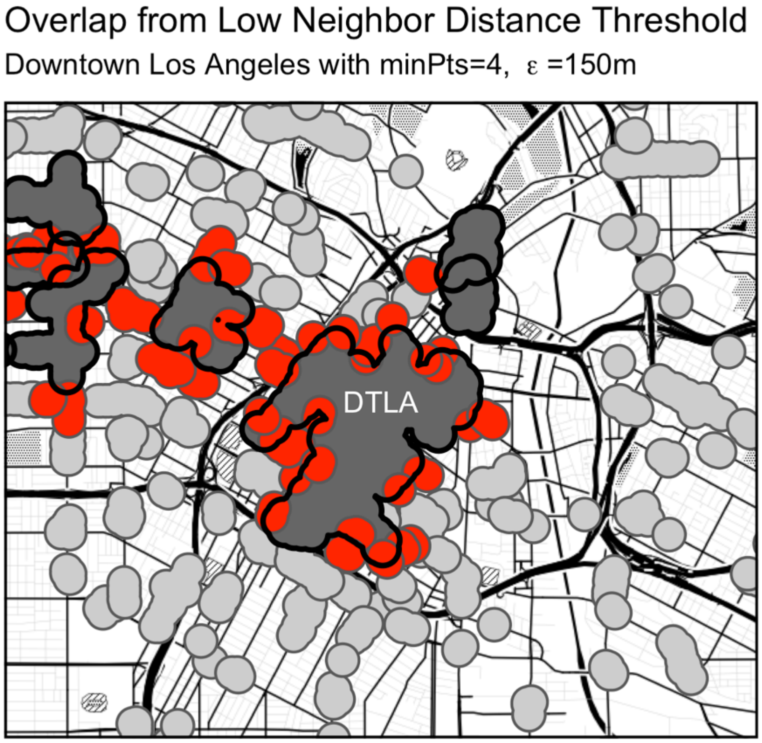


Figure 5.6 Reasonably large neighbor distances group multiple centers together across a large region. Using these clusters as an analysis unit would make it impossible to distinguish between activities taking place in Downtown Los Angeles (east side of the yellow cluster) from those happening in Beverly Hills (west side). Larger neighborhood threshold values produce more extreme results.

Using too-small values of ε paired with low minPts can cause other problems in dense areas, as these clusterings create intricate patterns of nearly-adjacent clusters separated by much smaller distances than the overall accuracy of the CHTS data suggests would be appropriate. The map of overlapping clusters around downtown LA in Figure 5.5

demonstrates how this effect can occur when distance thresholds are too far below the level at which superclusters begin to form. Centers highlighted in red are small (15 or fewer business locations) and have buffer boundaries that overlap those of a cluster with at least 50 business locations. While some of these overlaps are relatively small, some of the small clusters appear to be nearly surrounded by areas in which CHTS locations would be joined to a larger cluster. This added degree of uncertainty when linking CHTS points to commercial centers would introduce unwanted ambiguity into my analysis. Interestingly, this effect emerged only when I was using network distance or OPTICS-based classification.


Overlap from Low Neighbor Distance Threshold

Downtown Los Angeles with minPts=4, ε =150m

Figure 5.7 Small neighbor distance thresholds identify numerous small centers adjacent to and sometimes within larger dense centers. Small clusters that overlap much larger ones are shown in red.

One way to measure uncertainty in cluster assignment is to consider the number of activities that are within the match distance (200 meters) of business establishments in more than one cluster. Table 5.1 shows the share of CHTS shopping, dining, and entertainment activities located within range of a business in a cluster for several sets of cluster parameters as well as the total number of clusters detected and the size of the largest cluster. Activity match rates in parentheses indicate the share of activity locations within 200 meters of business establishments in more than one cluster. Clusterings with minPts=4 detect more clusters and match a higher share of business establishments at a given ε, but they also appear to detect more edge clusters, leading to higher rates of ambiguous matches. Match rates for shopping and dining activities are comparably high, but entertainment activities match center locations at a much lower rate, largely because the CHTS does not distinguish entertainment activities at commercial establishments or public areas from those in private residences (e.g., a Super Bowl party could either be described as "Entertainment" or "Social/Visit friends and relatives").

Increasing ε lowers the total number of clusters detected by connecting nearby clusters and also expands the largest clusters. When ε is set to 250 meters or lower, the largest cluster detected is in downtown San Francisco, but at 275 meters or higher, the downtown Los Angeles cluster merges with the corridor of development stretching west to Santa Monica. These large clusters may be appropriate for analyses that seek to identify all areas of high opportunity density in a region, but because I want to be able to compare the activities of people in nearby areas, I would prefer to keep them separate. Unfortunately, there are no clustering pairs that substantially subdivide the central San Francisco cluster or separate the clusters between Santa Monica and Downtown LA into more than four large pieces without also producing myriad small clusters.

Table 5.1 Activity location and cluster size results for a range of cluster parameters with low values of minPts and ε. In general, higher values of *ε* correspond to higher overall match rates and lower rates of ambiguous matching, but fewer and much larger clusters in major downtowns

minPts	*ε* (m)	Activities Located in a Business Cluster (ambiguous between 2 or more)			Total Clusters	Largest Cluster (city)
		Dining	Entertainment	Shopping		
4	150	78.0% (14.8%)	39.7% (7.9%)	75.0% (13.6%)	9,348	2,889 (SF)
4	175	79.7% (10.3%)	41.0% (5.3%)	76.7% (10.3%)	8,714	3,397 (SF)
4	200	81.0% (7.2%)	42.5% (3.8%)	78.1% (7.4%)	8,129	4,028 (SF)
4	225	82.1% (4.8%)	43.5% (2.3%)	79.4% (5.2%)	7,549	4,382 (SF)
4	250	82.9% (3.3%)	44.4% (1.5%)	80.7% (3.5%)	7,030	4,996 (SF)
4	275	83.7% (2.4%)	45.3% (1.0%)	81.3% (2.5%)	6,636	6,514 (LA)
4	300	84.4% (1.7%)	46.1% (0.7%)	82.1% (1.8%)	6,233	6,660 (LA)
5	150	73.1% (10.9%)	36.2% (5.8%)	70.4% (10.1%)	7,377	2,846 (SF)
5	175	75.6% (7.8%)	38.0% (4.1%)	72.8% (8.1%)	6,971	3,201 (SF)
5	200	77.3% (5.7%)	39.6% (3.2%)	74.8% (6.0%)	6,609	3,547 (SF)
5	225	78.8% (3.6%)	41.0% (1.7%)	76.4% (4.1%)	6,232	4,263 (SF)
5	250	80.2% (2.5%)	42.0% (1.0%)	77.7% (2.8%)	5,859	4,935 (SF)
5	275	81.1% (1.9%)	42.7% (0.6%)	78.6% (1.9%)	5,566	6,354 (LA)
5	300	82.0% (1.4%)	43.6% (0.5%)	79.6% (1.3%)	5,255	6,577 (LA)

For the remainder of this analysis, I use the network-DBSCAN clustering with minPts=4 and ε=250 meters (bolded in Table 5.1). As with any model selection problem, this choice of clustering is somewhat arbitrary, but \it meets the requirements I set out better than any alternative. This result catches large shares of the activities I am interested in and limits cluster overlap (both in terms of linking actual CHTS locations and in visual inspection of mapped cluster results) while maintaining separations between large clusters in major downtowns. Clusterings that use smaller distances to link neighbors match substantially lower shares of CHTS locations and have much higher rates of cluster overlap.

d. Final Clusters

This section investigates the general makeup of the commercial centers identified by network-DBSCAN with minPts=4 and ε=250 meters and mapped results to determine how well this method identified different types of centers statewide. Table 5.2 shows the general size distribution of commercial centers, as well as the general division of business types in clusters of various sizes. While smaller centers are much more numerous, centers of at least 26 business locations contain 64% of all businesses in a cluster. Retail accounts for roughly 60% of businesses across all sizes of centers and retail establishments are present in all centers containing at least 14 businesses. Accommodation and food service businesses are somewhat less common, making up about 1/3 of the businesses across a range of center sizes. Arts and entertainment businesses are much less common overall but concentrate in larger centers. The median Accommodation/ Food Service and Retail businesses are located in centers with 45 and 46 establishments, respectively; in contrast, the median arts and entertainment business is located in a center with 58 establishments. Nearly 40% of all entertainment establishments are located in centers with over 100 business establishments.

Table 5.2 Distribution of business establishments by type and center size. While small centers are more numerous, larger ones contain a larger share of all businesses. Arts and entertainment businesses are particularly concentrated in large centers.

		Accommodation / Food		**Arts / Entertainment**		**Retail**	
Center Size	Centers	Estabs	Percent	Estabs	Percent	Estabs	Percent
≤ 5 estabs	2,496	2,686	28.9%	810	8.7%	5,803	62.4%
6 to 10	1,875	4,774	33.6%	991	7.0%	8,442	59.4%
11 to 25	1,587	9,303	36.5%	1,545	6.1%	14,668	57.5%
26 to 50	600	7,411	35.1%	1,136	5.4%	12,596	59.6%
51 to 100	300	7,251	34.1%	1,262	5.9%	12,734	59.9%
101 to 500	164	9,084	32.9%	1,705	6.2%	16,799	60.9%
≥ 501	8	5,309	33.9%	1,996	12.7%	8,377	53.4%

Different centers are made up of different proportions of these business types. This variability is particularly pronounced for smaller centers in densely developed areas, which are more likely to specialize in specific industries than similar-sized centers in less developed areas, which tend to serve a wider range of purposes (Giuliano & Small, 1991; Helsley & Sullivan, 1991). A simple classification helps differentiate centers in the maps that follow. Retail businesses make up a substantial component of nearly every center (Table 5.2), so I emphasize areas with particularly high concentrations of accommodation and food service or arts and entertainment businesses. To do this, I identify the upper tercile for percent food and percent entertainment weighted by the number of businesses in each center. The upper tercile for %food is calculated by sorting the centers in ascending order by %food and computing the cumulative sum of the number of establishments in each center; the upper tercile is the lowest value of %food that corresponds to a cumulative sum greater than two thirds of the total set of business establishments. Roughly two thirds of businesses are in centers with a lower %food, and roughly one third are in centers with a higher %food. The same calculation is done for %entertainment.

One third of clustered business locations are in centers with at least 39.4% food service businesses; centers above this concentration are shown in magenta on the maps. One third of clustered business locations are in centers with at least 8.3% arts and entertainment businesses, and centers with at least this high a concentration are shown in light blue. Centers above both thresholds are shown in dark blue, and centers below both thresholds are shown in grey. The relatively low cutoff for entertainment-oriented centers flagged many very small centers with only one or two relevant businesses, so centers with fewer than two businesses in the corresponding class were excluded even if they are over the threshold percentage. The maps show the results of this clustering in the San Francisco Bay Area, Greater Los Angeles, Santa Barbara, and Sacramento. As in the previous section, these maps represent clusters using polygons produced by buffering all clustered business establishment locations by 200 meters.

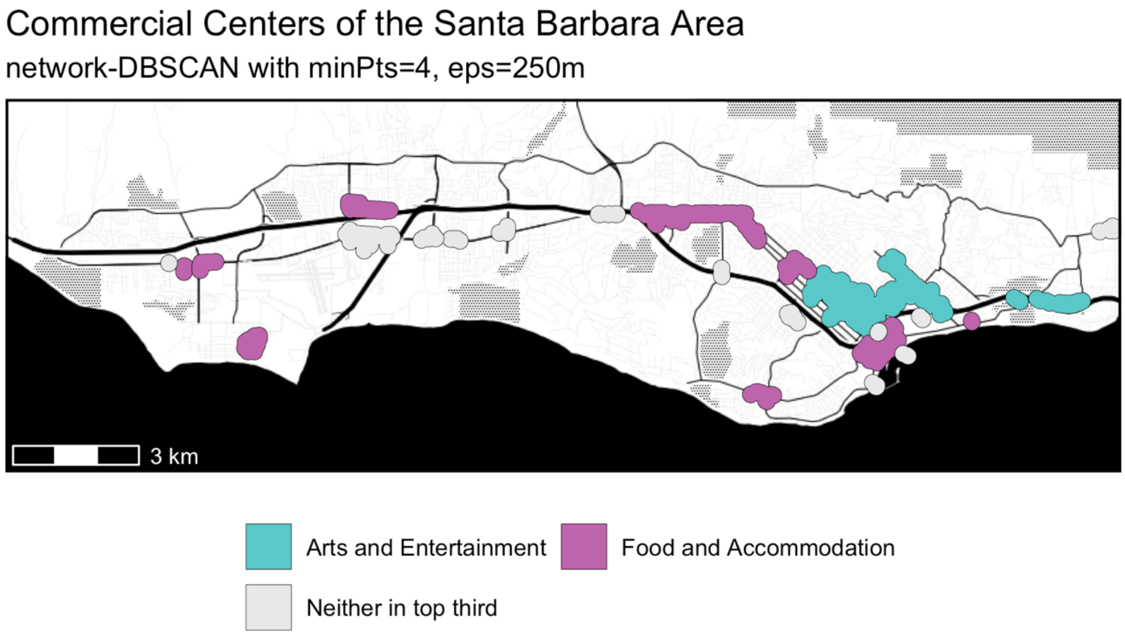

Commercial Centers of the Santa Barbara Area
network-DBSCAN with minPts=4, eps=250m
3 km
Arts and Entertainment
Food and Accommodation
Neither in top third

Figure 5.8 Commercial centers in the Santa Barbara Area. These centers mainly cluster

Santa Barbara’s centers (Figure 5.8) mostly stretch along State Street and Hollister Avenue (visible on the map as the thinner road south of the highway in the western half of the map and north of the highway connecting the two sets of centers in the eastern half). This surface street runs down the center of the area and serves as the main street for both Goleta and Santa Barbara. The large downtown center has a relatively large concentration of arts and entertainment businesses (particularly theaters and music venues) but is below the food service cutoff because of the high concentration of stores. Centers along the waterfront appear to specialize in food service, as do many of the smaller centers spread throughout Goleta. The small center in Isla Vista has a particularly high concentration of food service businesses to serve the adjacent college campus.

Centers for the San Francisco Bay Area and Greater Los Angeles

In order to identify as many smaller centers as possible, I selected a final clustering with several very large “downtown” clusters that would need to be broken up for choice modeling or many other applications. This result is acceptable for the activity timing and duration analyses presented in the next two chapters, since downtown San Francisco and Downtown Los Angeles are fairly well-understood as commercial neighborhoods. However, these large clusters are also much bigger than transportation analysis zones, which likely makes them an even worse fit as part of a choice set for a destination choice model. Future work will explore either hierarchical clustering or varying clustering parameters over space in order to make the final clusters more consistent throughout the study area.

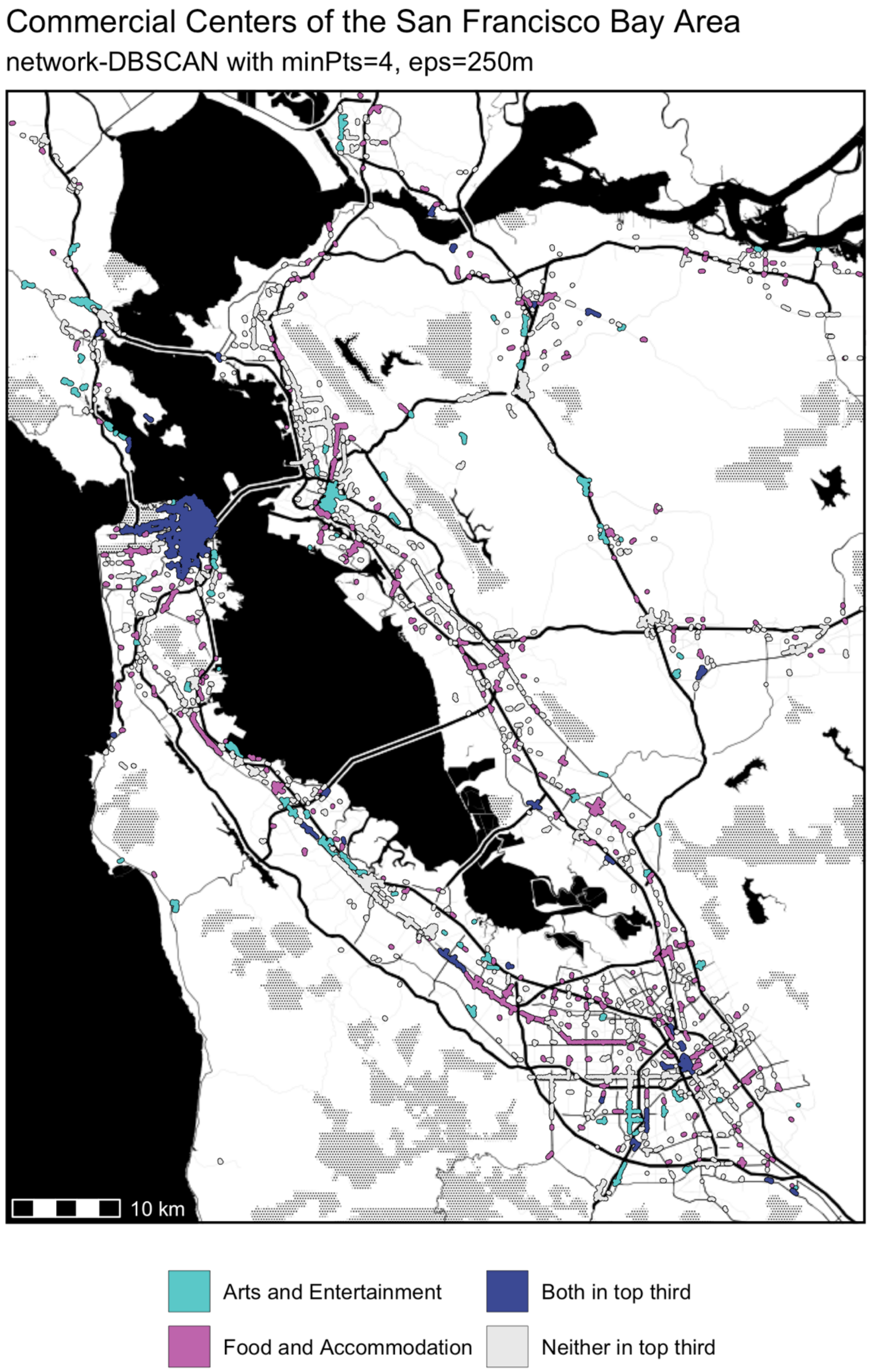

Commercial Centers of the San Francisco Bay Area
network-DBSCAN with minPts=4, eps=250m
10 km
Arts and Entertainment
Both in top third
Food and Accommodation
Neither in top third

Figure 5.9 Centers in the San Francisco Bay Area. Downtown San Francisco is by far the largest center in the area, and protected open spaces limit development of the hilly areas surrounding the bay.

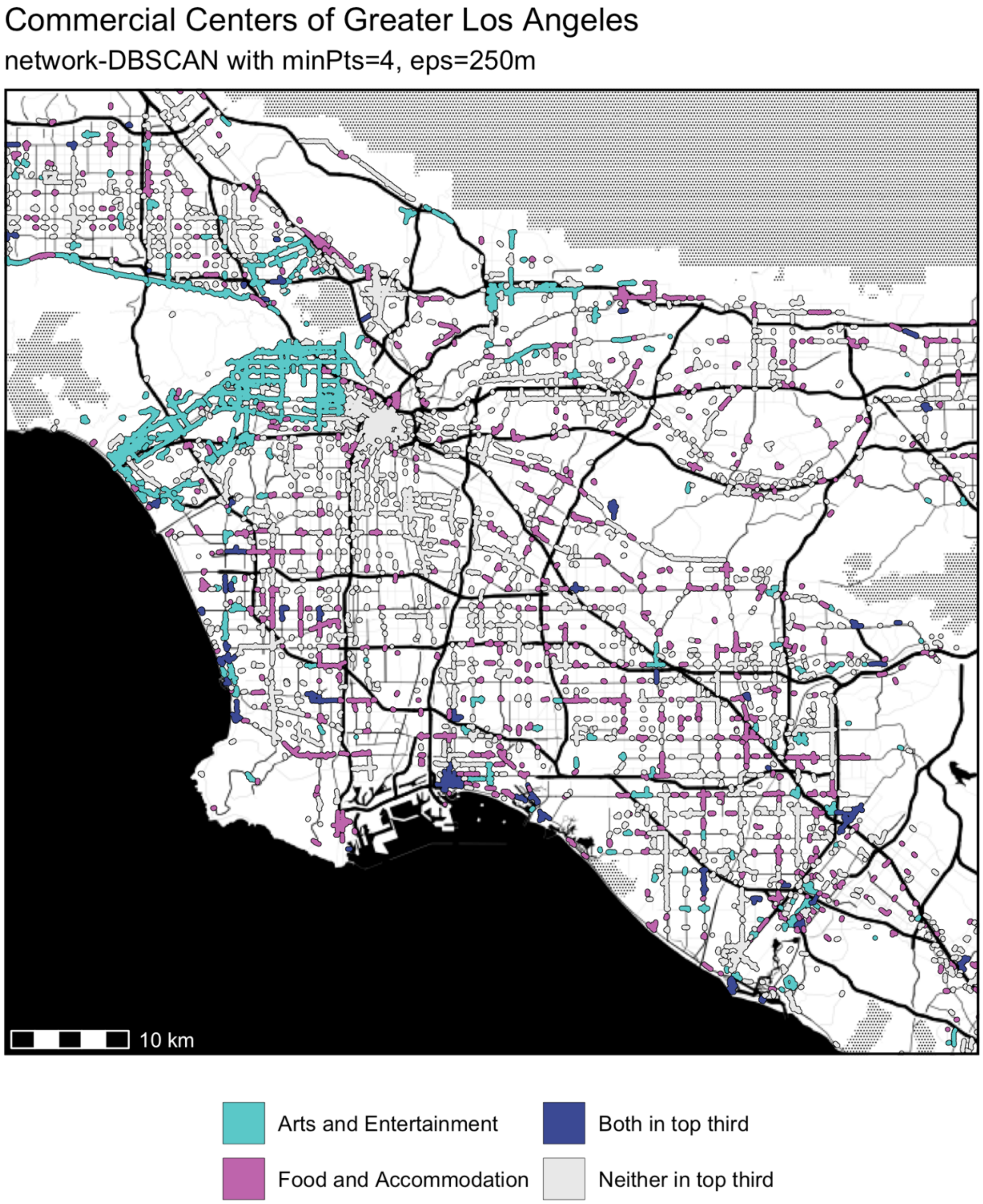


Figure 5.10 Centers of the greater Los Angeles area. Note the dense development, and regularly spaced food-service subcenters spread throughout the region.

Medium-sized commercial centers in both regions appear to be made up of a mix of roughly circular centers (malls and smaller city centers) and longer axial developments (mostly called boulevards in Los Angeles and avenues in the Bay Area). Both regions are dotted with much smaller centers, which exploration in person and in Google Street View suggested are mostly individual strip malls and small commercial intersections in suburban areas. I was not satisfied with the method's ability to subdivide the large cluster in downtown San Francisco, which with nearly every set of parameters expanded into nearby but culturally distinct neighborhoods like the Mission and Richmond Districts. Otherwise, development in the San Francisco Bay area appears fairly hemmed in by the bay and by protected open spaces that cover roughly a third of the region's land area (Bay Area Open Space Council, 2017). Apart from the major dense corridor in the north, greater Los Angeles has a distribution of small-to-medium-sized centers spread over the developed area; this difference in spatial distribution may in part reflect the overall difference in densities between the two regions. While San Francisco is one of the densest cities in the US, the surrounding area is substantially less densely developed, and the Los Angeles – Long Beach – Anaheim metropolitan area is the most densely developed in the country (Wilson et al., 2012).

The area north and east of downtown Sacramento (Figure 5.11) appears to be a microcosm of the results from greater Los Angeles, with a set of linear centers following a gridded pattern out from the center along major arterial roads (visible in the faint black lines connecting many of the smaller centers). Unlike the other two regions, the areas of highest food density appear to be concentrated around the dense downtown cluster rather than spread throughout the region.

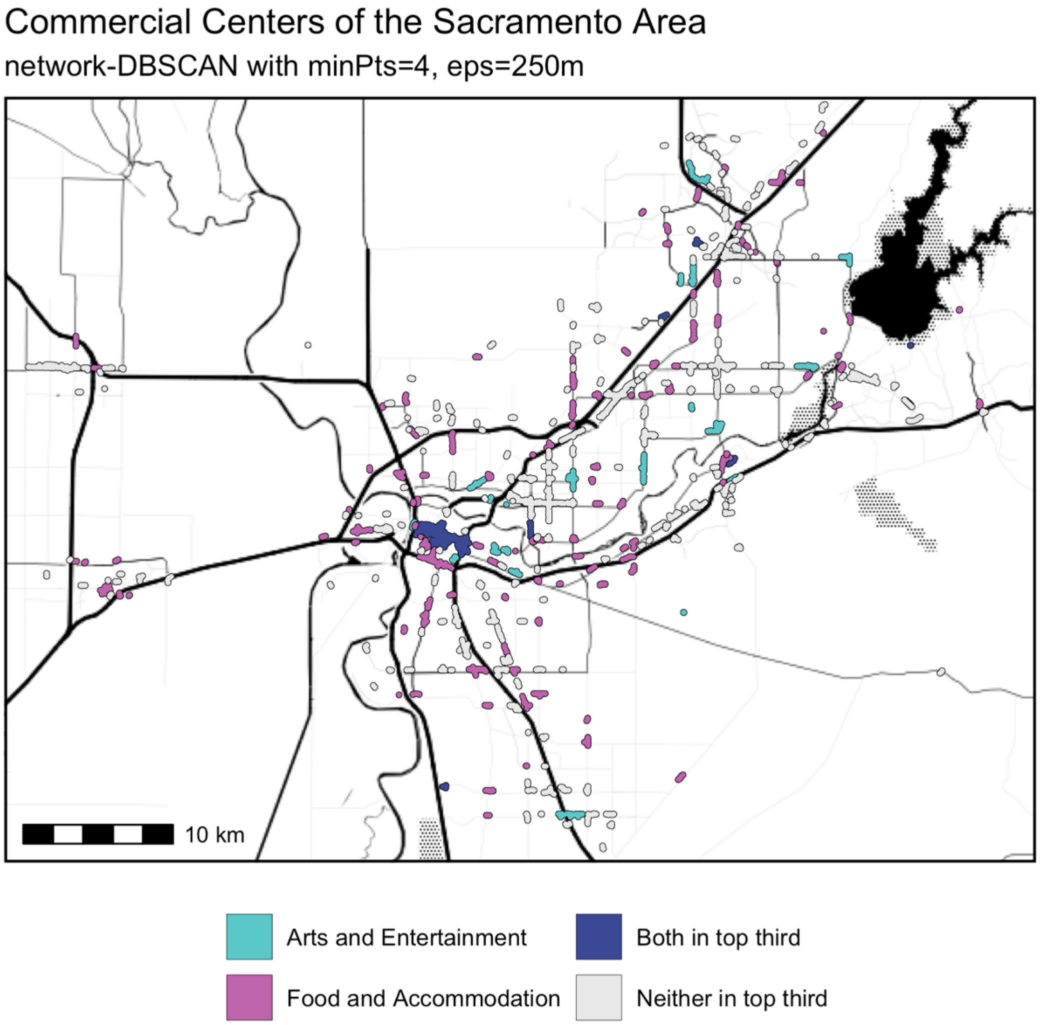


Figure 5.11 Sacramento is gridded like Los Angeles, but centers are much sparser

e. Conclusions

Given the right set of parameters and accurately identified neighbors, network-distance DBSCAN produces usable clusters from locations of customer-serving businesses. A clustering with minPts=4 and ε=250 meters identifies centers of varying shapes and sizes throughout California but does not differentiate between sub-regions of the most densely developed urban cores. I use this clustering in the analysis for the remaining chapters because it catches almost all activity locations, and because it does not require centers to be completely distinct.

Identifying commercial centers directly from business location data avoids the issue of boundaries dividing densely developed areas that is present when census-based spatial units are used, but these centers are subject to many of the same measurement uncertainty issues that apply to any other method of grouping multiple observations into a single unit. The center classifications used in the maps or any other center-level variable cannot describe the full range of places within the center and leaves analysis subject to biases caused by the ecological fallacy. One case where this is an issue is the identification of opportunity-dense zones for various types of activities: a large center with a fairly small overall share of entertainment businesses might well provide more overall opportunities for entertainment than a small center with nine businesses, one of which is a small movie theater, but the latter would be classified as "high density" for Arts and Entertainment using the criteria I set for the maps. Still, some form of spatial aggregation is necessary for many sorts of analysis, particularly given the spatial mismatches identified in Chapter 3.

6. Exploratory Analysis of Activity Timing

Time geography generally understands people's ability to move through space as being limited by a set of constraints first outlined by Hägerstrand in 1970 (Hägerstrand, 1970). According to this model, people are hemmed in by *capability* constraints imposed by the physical limitations of human life (particularly the needs for sleep and food) and the modes of transportation available to a person; *coupling* constraints imposed by the need to be in the same place at the same time as other people and material goods to work, shop, eat, and socialize; and *authority* constraints that limit where a person is allowed to be and where they are excluded (e.g., you can only go to a store when it's open and can only drive on roads). These constraints define the areas that a person might visit during the day but do not say much about where and when they are likely to visit. Golledge's concept of anchor points adds to this model by identifying places that are particularly important, meaningful, and fixed in space-time within a person's life (Golledge & Stimson, 1997). Home, work, and school are obvious anchor points, but each person's preferred destinations for daily activities like shopping, exercising, and socializing are anchor points as well. While people are free to travel wherever they'd like (subject to Hägerstrand's constraints), human behavior is not randomly scattered within each person's potential space-time prism. It is difficult to get a sense of people's regular destinations apart from home, work, and school from a single-day survey, but it is possible to identify major centers of activity using the methods in Chapter 5 and to investigate overall patterns of activity scheduling that are tied to personal characteristics and daily schedules. To do this, we first need to understand patterns of time use in each derived center.

The duration and frequency of various types of out-of-home activities vary substantially across the week, as well as by the structure of the household and individual characteristics. Discrete-continuous time-use budgeting models (i.e., models that jointly examine the types of activities and duration in each activity during a day) indicate that employment, age, income, access to vehicles, and general location within an urban area are important predictors of the total amount of time people spend on specific activities (C. R. Bhat, 2005; Calastri, Hess, Daly, & Carrasco, 2017). These models rarely consider activity timing within the day. In contrast, the work done by Lee, McBride, and Goulias in identifying daily activity patterns using latent class analysis and fragmentation / sequence analysis methods and tying these schedules to land use and accessibility (Lee et al., 2017; McBride, Davis, & Goulias, 2019) have provided insights into the sequential scheduling of various activities within the day without directly tying these activities to people’s needs. By linking people’s activity choices to the locations and fixed schedules of their anchor points and those of other members of their households and limiting destination choices to locations with sufficient opportunities, activity-based models acknowledge that activities are not evenly distributed in space or time. Activity choices and timings do not necessarily follow simple rules that are consistent across time and space, and many forms of spatial and temporal heterogeneity are important to understanding people’s activity scheduling.

Restaurants and bars demonstrate particular schedule variability. The graphs of “Popular Times” Google Maps for various stores, restaurants, cafes, and bars show clear variations over the course of the week and between ostensibly similar businesses, and this variability is also clearly observable in person. For example, restaurants in Isla Vista serve people who work or go to school at UCSB around lunch time and local residents in the evening, and appear equally busy at both times of day, whereas many fancier restaurants elsewhere are only open for dinner. Santa Barbara's Funk Zone is busy with wine-tasting on weekend afternoons, but bars elsewhere are full most evenings but empty during daytime. Some places operate by such specific calendars that an aggregate model based on survey data would be highly unlikely to represent them accurately – for instance the Tap Room on Ortega St in Downtown Santa Barbara opens early many mornings to provide a viewing venue for fans of English soccer matches taking place several time zones away.

These temporal variations reflect real qualitative differences between the types of activities pursued by different people at different places, but these differences are captured well by neither the California household travel survey nor NETS business datasets. The CHTS distinguishes between "drive through meals" and "eat a meal at restaurant," and between "routine shopping" (for groceries, clothing, and household maintenance) and "shopping for major purchases or specialty items," but provides no further distinctions within "Entertainment" and "Social" activities. While these latter two categories presumably do reflect different primary purposes and only one inherently requires other people, either could take place at a special purpose venue (a music club), a restaurant or bar, a public park, or at someone's house. Compounding the difficulty of classifying activities is uncertainty created by using NAICS classifications, which identify only four types of restaurants (United States Office of Management and Budget, 2017) and also provide no clear way of distinguishing between a music club that serves alcohol and any other bar, or between a bar that serves some food and a restaurant (the distinction in that case depends on what "primarily" serving either food or alcohol means in the specific state).

The issues that arise when trying to classify activities and locations represents a mix of vagueness and uncertainty, as defined by Lukasiewicz and Straccia (2008). Vagueness refers to descriptive characteristics that exist on a continuum between complete presence and absence, which makes classification difficult or fuzzy – getting dinner with friends represents both an eating activity and a socializing one, and bars that serve food are somewhere in between a bar and a restaurant. Uncertainty refers to potential inaccuracy caused by the absence of information necessary to classify something, which is the case when activity and place datasets use overly broad or mismatched classifications. The CHTS provides a mechanism through which people can address vagueness in activity classification by reporting multiple (potentially simultaneous) activities in a specific place, but individuals appear to have reported their activities very differently from each other, and only 4.6% of CHTS locations outside of work, school, and home have more than one reported activity.

a. Timing Variation Day-to-Day

This chapter takes an exploratory approach to investigate heterogeneity of activity scheduling and timing within the day for three specific activity types: dining at restaurants, entertainment, and shopping. Personal characteristics, pre-existing obligations, and weekly plans can affect activity scheduling both by making certain activities more or less likely and by influencing how these events are arranged over the course of the day. A well-designed spatial choice model would provide a more complete approach to understanding spatial choice and timing (particularly the combined effects and interactions of multiple variables), and my aim here is to identify key variables that should be included in models in Chapter 7 as well as choice models in the future. Additionally, the findings presented here provide insight into the relative impact of personal characteristics and place attributes on activity scheduling. This chapter focuses on activity scheduling and timing heterogeneity and explores their relationships.

I consider a range of personal, schedule-related, and land use / centers variables for this analysis. Table 6.1 contains the overall counts and relative frequencies of activities by a number of different personal characteristics, and Table 6.2 contains the corresponding information for variables about centers. Categorical variables used to group different activities are as follows: day of the week (and holiday), life-cycle stage, tour type (origin/destination and number of intermediate stops), size of the destination center, the mix of chain and independent businesses in the destination center, and the relative proportion of food service and entertainment (and by extension retail) businesses in the center. The term *destination centers* is used here to mean centers for which we have CHTS records of people use as one of their destinations in their daily travel pattern.

The CHTS was designed as a 365-day survey in order to get roughly round-the-calendar representation statewide. I separate federal holidays (plus Christmas Eve, New Year's Eve, and the day after Thanksgiving) from the other days of the week, since closings of schools, stores, and offices on these days substantially impact people's work and nonwork activity schedules. The lower representation of Mondays in the results reflects the fact that in addition to the four holidays that always fall on a Monday (Martin Luther King Day, President's Day, Memorial Day, and Labor Day), three other date-specific holidays were on Mondays in 2012, the year of the survey (Veteran's Day Observed, Christmas Eve, and New Year's Eve).

Life cycle stage is a commonly used variable in travel behavior research, which is designed to track the combined impact of age, household makeup, and employment. This chapter uses a 12-level classification, slightly simplified from the version used by Goulias and Lee (K. Goulias, 2009; Lee & Goulias, 2015). Categories include: 1) All people under 18; 2) College/University student; 3) Home-duties with no children; 4) Home-duties with children; 5) Part time worker (<40 hours per week) with no children; 6) Part time worker with children; 7); Full time worker (>=40 hours per week) with no children; 8) Full time worker with children; 9) Looking for a job; 10) Disabled; 11) Retired; and 12) All other people.

Tour types were classified based on how a sequence of trips relates to anchor points. A tour (also called a trip chain in the literature) is any series of trips starting at one location and returning to the same location; trip chaining is the act of sequencing destinations so that one tour can provide transportation to multiple activities. Each separate location where one stops to participate in one or more activites is called a stop. For this analysis, I identified all the reported visits to the anchor points home, work, and school using the spatial coordinates and place names people provided; I defined trip chains as all sequences of stops that led from one anchor point to another. Home-based tours (trip chains that start and end at home without including a stop at either a school or work location) are the most common and presumably are made in order to pursue activities at one of their destinations. In contrast, commute trip chains (sequences of trips that start at home and end at work/school or vice versa) presumably have “getting to work/school” or “returning home” as their primary purpose, and other stops are included when it is convenient. For this analysis, I grouped all non-home-based tours and chains together and subdivided the two categories further by the number of intermediate stops included between the start and end anchor points.

Center size is initially broken down in the same way it was during the cluster tuning steps in Chapter 5. NETS contains information about whether each business is independent or part of a larger company or business chain; I break these up into terciles using the same frequency-based method used when classifying centers in Chapter 5. The "high chain" category contains centers with at least 39% chain businesses; the "mid chain" category contains centers with less than 39% but more than 23% chain businesses, and the "low chain" category contains centers with less than 23% chain businesses. While there are different numbers of centers in each cluster, one third of all businesses are in each of the three groupings. I break up the relative shares of food service and entertainment businesses similarly, with the class breaks placed at 29% and 39% for food service and at 3.5% and 8.3% for entertainment. The terciles are calculated separately for the individual measures, so the resulting nine cross-groupings do not each contain $1/9^{th}$ of all the businesses in centers.

Table 6.1 shows the distribution of activities as a function of the person- and tour-level variables discussed above, as well as the total number of people in each life cycle stage and surveyed on each day of the week. For life-cycle stage and day of the week, the total number of people is fixed, so I calculate Per-Person Per-Day (PPPD) activity participation rates as the ratio between number of activity occurrences and number of people. For the breakdown by tour type, PPPD would not make sense, since people make different numbers and types of tours on different days, so I provide percent of relevant activities on those tours. Life cycle stage does not appear to have a very strong impact on activity participation. People who reported disabilities and did not work participated in substantially fewer activities of all three types (and notably have the lowest frequency for entertainment activities). Day of the week is strongly related to activity participation: the three activity types I consider are pursued more often from Friday-Sunday than during the rest of the week, and all three are more common on Saturdays than on any other day. Holidays have lower rates than the weekend days, but this probably depends a lot on the specific holiday. Multi-purpose home-based tours contain a much larger share of shopping activities, whereas single-purpose home-based tours are somewhat more likely to be made for the purpose of dining and much more likely to be made to an entertainment destination (dinner *or* a movie?). Trip chains that have work or school as a start or end point contain a higher share of dining and entertainment activities than home-based tours do, possibly suggesting people do routine household replenishment shopping on days when they don’t have to commute.

Table 6.1 Relative frequencies of dining, entertainment, and shopping activities across different groups of people in CHTS.

			Dining		Entertainment		Shopping	
	Grouping	People	Events	PPPD	Events	PPPD	Events	PPPD
Life-Cycle Stage	Child	22,312	3,026	0.14	1,658	0.07	5,005	0.22
	College Student	7,797	1,505	0.19	537	0.07	2,394	0.31
	Home-duties no kids	2,040	345	0.17	135	0.07	981	0.48
	Home-duties with kids	3,307	586	0.18	284	0.09	1,731	0.52
	Part time worker no kids	9,667	2,161	0.22	753	0.08	4,265	0.44
	Part time worker with kids	4,569	878	0.19	351	0.08	2,016	0.44
	Full time worker no kids	20,844	4,596	0.22	1,415	0.07	7,546	0.36
	Full time worker with kids	12,551	2,577	0.21	972	0.08	4,022	0.32
	Looking for Work	3,772	566	0.15	249	0.07	1,519	0.40
	Disabled	3,808	464	0.12	156	0.04	1,394	0.37
	Retired	16,651	3,446	0.21	1,295	0.08	7,202	0.43
	All other	1,795	343	0.19	126	0.07	716	0.40
Day of the Week	Monday	12,585	1,626	0.13	541	0.04	3,750	0.30
	Tuesday	15,224	2,537	0.17	835	0.05	5,093	0.33
	Wednesday	15,113	2,602	0.17	781	0.05	4,955	0.33
	Thursday	15,777	2,756	0.17	916	0.06	4,988	0.32
	Friday	15,027	3,153	0.21	1,050	0.07	5,126	0.34
	Saturday	15,385	3,779	0.25	1,972	0.13	7,343	0.48
	Sunday	15,975	3,383	0.21	1,487	0.09	6,202	0.39
	Holiday	4,027	657	0.16	349	0.09	1,334	0.33
		Tours	Events	%	Events	%	Events	%
Tour / Trip Chain Type	Home-based Tour (1 stop)	52,022	5,224	30.6	2,650	15.5	9,194	53.9
	Home-based Tour (2 stops)	15,238	3,722	27.5	1,213	8.9	8,619	63.6
	Home-based Tour (3 stops)	7,523	2,441	26.5	788	8.6	5,982	64.9
	Home-based Tour (4+ stops)	8,490	3,796	25.7	1,127	7.6	9,852	66.7
	Other Tour (1 stop)	14,827	1,775	46.5	438	11.5	1,603	42.0
	Other Tour (2 stops)	5,828	759	36.9	254	12.3	1,046	50.8
	Other Tour (3 stops)	2,696	739	40.7	329	18.1	747	41.2
	Other Tour (4+ stops)	2,913	1,855	44.4	660	15.8	1,662	39.8

Table 6.2 shows the distribution of activities as a function of the center characteristics discussed above. As shown in Chapter 5, larger centers have a notably higher share of entertainment businesses, and they appear to attract a much greater share of entertainment

and dining activities. Somewhat smaller centers appear somewhat more likely to be retail-focused. Centers with high proportions of chain businesses (like malls) have much higher shares of shopping activities, whereas centers with a larger share of independent businesses tend to have much larger shares of dining and entertainment activities. This effect is nearly as substantial as the difference between the high-entertainment / high-food category and the other centers. Given that high, mid, and low-chain centers each have the same total number of businesses, the larger number of centers in the Low Chain category indicate that these centers are typically smaller than the centers in the High Chain and particularly Mid Chain categories. Unsurprisingly, the relative mix of businesses in a center is strongly related to the mix of activities that people pursue in that center. The mix of activities outside of centers is strongly influenced by the uncertainty in the Entertainment activity category, discussed above.

Table 6.2 Relative frequencies of destinations for dining, entertainment, and shopping activities across different center types.

			Dining		Entertainment		Shopping	
	Grouping	Centers	Events	%	Events	%	Events	%
Center Size	5 or fewer	2,496	961	31.7%	283	9.3%	1,786	58.9%
	6 to 10	1,875	1,604	29.2%	376	6.9%	3,506	63.9%
	11 to 25	1,587	3,394	30.4%	489	4.4%	7,264	65.2%
	26 to 50	600	2,540	31.6%	413	5.1%	5,080	63.2%
	51 to 100	300	2,986	33.7%	631	7.1%	5,251	59.2%
	101 to 500	164	3,959	35.1%	998	8.8%	6,332	56.1%
	501 or more	8	1,630	38.7%	650	15.4%	1,928	45.8%
Chains	High Chain	2,139	6,555	27.8%	1,135	4.8%	15,908	67.4%
	Mid Chain	1,669	5,417	34.5%	1,142	7.3%	9,145	58.2%
	Low Chain	3,222	5,102	40.0%	1,563	12.3%	6,094	47.8%
Business Type Mix	High Ent & High Food	576	2,387	41.3%	761	13.2%	2,628	45.5%
	High Ent & Mid Food	320	1,482	35.6%	408	9.8%	2,274	54.6%
	High Ent & Low Food	1,035	1,406	32.8%	510	11.9%	2,377	55.4%
	Mid Ent & High Food	314	1,925	36.5%	424	8.1%	2,918	55.4%
	Mid Ent & Mid Food	260	2,411	34.9%	538	7.8%	3,956	57.3%
	Mid Ent & Low Food	218	1,026	27.1%	258	6.8%	2,507	66.1%
	Low Ent & High Food	1,761	2,984	36.2%	378	4.6%	4,874	59.2%
	Low Ent & Mid Food	577	1,643	26.4%	245	3.9%	4,334	69.7%
	Low Ent & Low Food	1,969	1,810	24.4%	318	4.3%	5,279	71.3%
	Destinations not in Centers		3,419	22.6%	4,091	27.0%	7,644	50.4%

b. Timing Variation Within the Day

In order to investigate the heterogeneity of activity timing over the course of the day, I produced activity participation and flux (activity starts – ends) sequences for groupings of activities shown in the tables above. These curves were generated with a temporal resolution of 5 minutes for the tables and 15 minutes for visual clarity in the charts; very few reported events in the activity diaries are recorded as having started and ended on minutes not divisible by 5 (and particularly quarter-hours), which suggests increased precision may be inappropriate.

To calculate activity participation and flux, I created a comprehensive list of activities in a certain category (such as *shopping* or *dining*), and note their start and end times, as well as the grouping characteristics listed above. I classified each activity by the life-cycle stage, weekday, tour type, or center type it falls into, and converted the start and end times to the temporal resolution for the analysis. For this paper, start times not on an even quarter hour are rounded *down* to the nearest time evenly divisible by 5 minutes, and uneven end times are rounded *up*. This ensures that every activity will be counted for at least one time period.

Two statistics are used for this analysis: *flux* (the difference between activity starts and stops in that grouping category at a given time point) and *percent active* (the share of ongoing activities of a given type normalized by the total number of activities of that type that take place in that grouping category over the course of the day). The *percent active* of relevant activities (subscripted i) happening in group c at time t can be calculated either as the cumulative sum of the activity flux for group c at time t or as as follows: first identify how many total activities of the type take place at any time during the day in t; then for each time point, identify how many of those activities started on or before that time and end after it. The ratio between these quantities is the share of relevant activities active at a given time point in a given center type.

$$Percent\ Active_{ct} = 100 \times \frac{\sum_i (inclust(i,c) \times active(i,t))}{\sum_i inclust(i,c)} \qquad \text{Equation.6.1}$$

$$inclust(i,c) = \begin{cases} 1 \mid location_i \in center_c \\ 0 \mid location_i \notin center_c \end{cases}$$

$$active(i,t) = \begin{cases} 1 \mid start_i \geq t\ \cap end_i < t \\ 0 \mid start_i < t\ \cup end_i \geq t \end{cases}$$

To get a sense of the measurement accuracy of both the flux and percent present calculations, I use a bootstrapping process to generate 100 new sets of activities by resampling from the original set of activities in the CHTS. The calculation for flux and percent active was then performed for every grouping variable and time point for each of these bootstrap activity schedules.

Variables affecting activity timing during the day

While the variability of activity participation over the day can best be seen graphically, I first wanted a measure of the overall impact of the various grouping variables used on daily schedules. To determine how different each group's schedule was from all the others, I calculated the correlation between the sequence of *flux* values produced by a specific grouping (e.g., *Centers with a high proportion of chain businesses*) and the sequence produced by all activities that took place outside that grouping (in that case, all activities in centers with a medium or low proportion of chain businesses as well as all activities that took place outside of centers). Since very few people participate in these activities at night, I extracted only values between 8 AM and 11 PM, spaced every 5 minutes for a total of 181 time points.

Tables 6.3 and 6.4 show the middle 90 bootstrap runs for each estimated correlation. Higher correlation coefficients indicate that a group's schedule is very similar to the overall schedule of activities, and lower coefficients indicate that a variable is substantially different from the "normal" schedule. The spreads in values produced by the bootstrapping indicate that a relationship is particularly uncertain.

Table 6.3 shows the correlation ranges for the three activities from schedules cross-classified by life-cycle stage, day of the week, and tour type.   For example, the within group activity flux correlation of children for dining is between 0.710 and 0.796 (right?).

People in different life cycle stages have much more varied schedules than do people surveyed on different days of the week. Parents with young children have somewhat different schedules from people without children, making them slightly farther from the overall typical schedules across all three activity types, but the difference between people in each job classification with kids and those without kids is smaller than the differences between job hours. While Table 6.1 showed that people who do not work for pay engage in many of the same activities as people who work outside the home, the two groups appear to do those activities on very different schedules. Full-time workers, particularly those without kids, have the most “normal” dining schedules (the bootstrap correlation interval is tighter than for other groups), possibly because regular work schedules are somewhat structured around typical meal times. Different days of the week appear to have their own schedules, but somewhat surprisingly, none of the days stands out as particularly different from “normal,” although holidays are unique. While different tour types all keep meal times in roughly the same place, retail and entertainment timings vary substantially, particularly for activities attached to commute trips.

Table 6.3 Activity Flux correlations for person-level and tour-level variables. Cells show the bootstrapped middle-90% ranges for correlation between the activity flux schedule produced for each group and that produced from all activities not in that group (8AM-11PM only).

	Grouping	Dining	Entertainment	Shopping
Life-Cycle Stage	Child	0.710 - 0.796	0.570 - 0.692	0.664 - 0.752
	College Student	0.508 - 0.691	0.290 - 0.492	0.437 - 0.581
	Home-duties no kids	0.326 - 0.555	0.204 - 0.434	0.392 - 0.571
	Home-duties with kids	0.354 - 0.529	0.265 - 0.442	0.583 - 0.699
	Part time worker no kids	0.680 - 0.795	0.452 - 0.616	0.489 - 0.631
	Part time worker with kids	0.562 - 0.705	0.263 - 0.461	0.378 - 0.531
	Full time worker no kids	0.774 - 0.830	0.541 - 0.682	0.604 - 0.722
	Full time worker with kids	0.731 - 0.823	0.521 - 0.652	0.581 - 0.685
	Looking for Work	0.321 - 0.488	0.239 - 0.458	0.425 - 0.567
	Disabled	0.264 - 0.479	0.228 - 0.448	0.351 - 0.507
	Retired	0.720 - 0.805	0.569 - 0.701	0.608 - 0.700
	All other	0.313 - 0.518	0.181 - 0.381	0.217 - 0.390
Weekday	Monday	0.587 - 0.722	0.348 - 0.493	0.466 - 0.579
	Tuesday	0.654 - 0.736	0.372 - 0.529	0.470 - 0.577
	Wednesday	0.722 - 0.799	0.423 - 0.554	0.414 - 0.580
	Thursday	0.661 - 0.755	0.298 - 0.465	0.513 - 0.629
	Friday	0.681 - 0.771	0.433 - 0.586	0.561 - 0.680
	Saturday	0.548 - 0.668	0.573 - 0.690	0.642 - 0.748
	Sunday	0.637 - 0.733	0.441 - 0.570	0.629 - 0.723
	Holiday	0.161 - 0.339	0.301 - 0.482	0.303 - 0.453
Tour / Trip Chain Type	Home-based Tour (1 stop)	0.629 - 0.700	0.601 - 0.712	0.660 - 0.736
	Home-based Tour (2 stops)	0.639 - 0.746	0.520 - 0.628	0.652 - 0.755
	Home-based Tour (3 stops)	0.592 - 0.702	0.367 - 0.531	0.498 - 0.628
	Home-based Tour (4+ stops)	0.605 - 0.704	0.345 - 0.489	0.551 - 0.669
	Other Tour/Chain (1 stop)	0.550 - 0.636	0.076 - 0.279	0.185 - 0.355
	Other Tour/Chain (2 stops)	0.482 - 0.606	-0.034 - 0.178	0.064 - 0.258
	Other Tour/Chain (3 stops)	0.425 - 0.579	0.322 - 0.474	0.026 - 0.216
	Other Tour/Chain (4+ stops)	0.520 - 0.660	0.250 - 0.415	0.192 - 0.389

Activity timing varies nearly as much based on opportunities available in an area as it does from person to person and generally more than it does on different days of the week. The timing of entertainment activities is the most variable. Business type mix appears to have a greater impact on dining and shopping than center size does. Surprisingly, while the mix of independent and chain businesses in a center has a profound impact on the relative mix of activities people pursue there, center size and business type mix are much more strongly related to activity timing.

Table 6.4 Center Types and Timing variability

	Grouping	Dining	Entertainment	Shopping
Center Size	5 or fewer	**0.469 - 0.610**	0.263 - 0.432	0.358 - 0.491
	6 to 10	0.554 - 0.671	0.322 - 0.482	0.388 - 0.526
	11 to 25	0.653 - 0.765	0.121 - 0.342	0.572 - 0.684
	26 to 50	0.667 - 0.776	0.277 - 0.492	0.501 - 0.627
	51 to 100	0.599 - 0.704	0.188 - 0.385	0.493 - 0.613
	101 to 500	0.710 - 0.801	0.321 - 0.522	0.500 - 0.633
	501 or more	**0.486 - 0.641**	0.387 - 0.531	**0.267 - 0.434**
Chains	High chain	0.738 - 0.803	0.364 - 0.521	0.607 - 0.721
	Mid chain	0.665 - 0.776	0.363 - 0.518	0.536 - 0.654
	Low chain	0.771 - 0.838	0.424 - 0.564	0.457 - 0.600
Business Type Mix	High Ent & High Food	0.669 - 0.762	0.360 - 0.533	0.369 - 0.507
	High Ent & Mid Food	0.503 - 0.669	0.283 - 0.471	0.260 - 0.430
	High Ent & Low Food	0.576 - 0.678	0.335 - 0.508	0.370 - 0.516
	Mid Ent & High Food	0.582 - 0.696	0.159 - 0.358	0.339 - 0.472
	Mid Ent & Mid Food	0.569 - 0.676	0.300 - 0.490	0.369 - 0.521
	Mid Ent & Low Food	0.482 - 0.624	0.074 - 0.245	0.356 - 0.525
	Low Ent & High Food	0.655 - 0.753	0.216 - 0.442	0.492 - 0.622
	Low Ent & Mid Food	0.518 - 0.649	0.155 - 0.419	0.492 - 0.618
	Low Ent & Low Food	0.556 - 0.678	0.153 - 0.319	0.541 - 0.673
	Destinations not in Centers	0.719 - 0.790	0.561 - 0.657	0.684 - 0.776

Business Center Classification and Activity Scheduling Results

For visualization, it can be helpful to group similar schedules together. It is possible to extract groups of schedules through clustering either people's activities at different times using latent class analysis (using each person's schedule as an observation and each time period as a variable, with categorical activity identifiers in each cell), as done by Lee et al (2017). Transposing that dataset to use different sequences grouped as above (with activity counts or fluxes in the cells) for factor analysis or latent profile analysis would be another approach. For a previous version of this chapter, I clustered centers based on size and the relative share of retail activities, dividing the centers groups of large and small centers, and then splitting the smaller ones based on the relative density of retail businesses. The following figures show schedules for three activity types produced from three groups: 1) activities taking place in centers of at least 501 business establishments (red); 2) activities taking place in all other centers (green); and 3) activities taking place outside of centers (blue). These plots indicate subtle but significant differences in activity scheduling between areas with different densities of opportunities; the most consistent pattern is that major downtowns appear to operate on a clock that is shifted about an hour later.

For Figures 6.1-4, the observed pattern is shown as solid line, and the results of 100 bootstrap runs of schedule regeneration are shown in semitransparent lines to provide a sense of the uncertainty of these measures. The vertical axis in these plots shows what share of relevant activities that take place in a particular type of center during the entire day are ongoing at a given time, as shown in Equation 6.1, and the slope is a linear function of the *flux*. If one curve is consistently higher than another, the activities it contains must last longer, on average, since they get counted at multiple time points. When differences are reported as significant in the text, but a specific number of runs is not provided, assume the comparison held in at least 95 of the runs (equivalent to a p<0.05), unless it is clear that none of the paths overlap, in which it is safe to assume that the comparison is significant at a level equivalent to p<0.01.

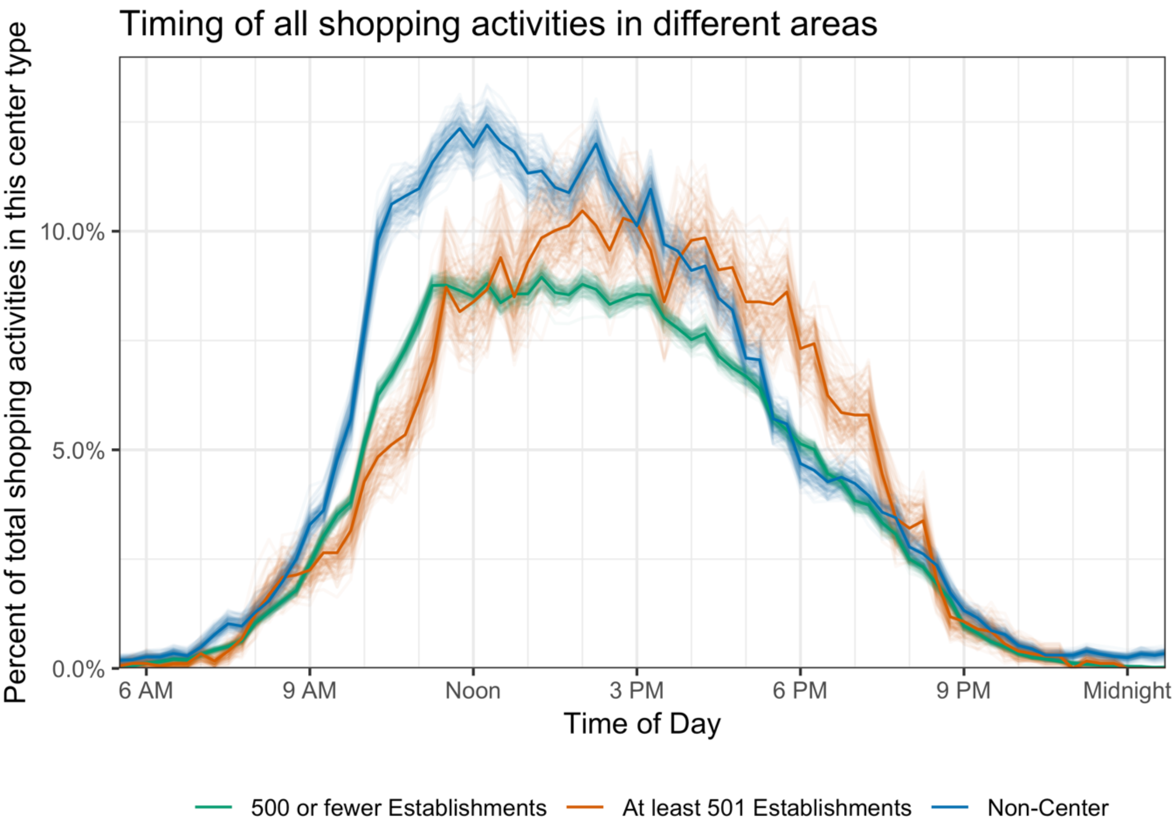


Figure 6.1 Timing of Shopping Activities in large centers, other centers, and non-centers. Dark curves show results from original sample; fainter lines are from bootstrapped samples.

People go shopping at roughly the same times of day everywhere, but the gap between activity participation rates in the two center types shown in Figure 6.1 suggests that there is real variation in the duration of these activities from place to place. Non-center shopping destinations tend to get busy between 9 and 10 AM and maintain a consistent level of business throughout the day. Smaller centers receive shopping activities at roughly the same time as non-centers, but individual activities tend to be shorter. People start visiting stores in large downtown areas in the late morning and continue to shop into the evening in larger numbers in these centers than elsewhere.

Retail opportunities in low-density areas may correspond to visits to small corner stores in residential neighborhoods or big box stores that take up enough space to have few neighbors within 200 meters of their geocoded location, leading the clustering method to identify them as noise points. The three curves have substantially different variances, as implied by range of bootstrap curves shown in the figures. Only a small share of shopping activities took place in the largest centers, so this curve has the widest variance; other centers have the least variance. It is also worth noting that the shopping trips used in this analysis are drawn from two different activity purposes listed in CHTS, which distinguishes between shopping for a major purchase and everyday/routine shopping. These activities differ somewhat in average duration, but their spatial distribution is relatively even between different center types.

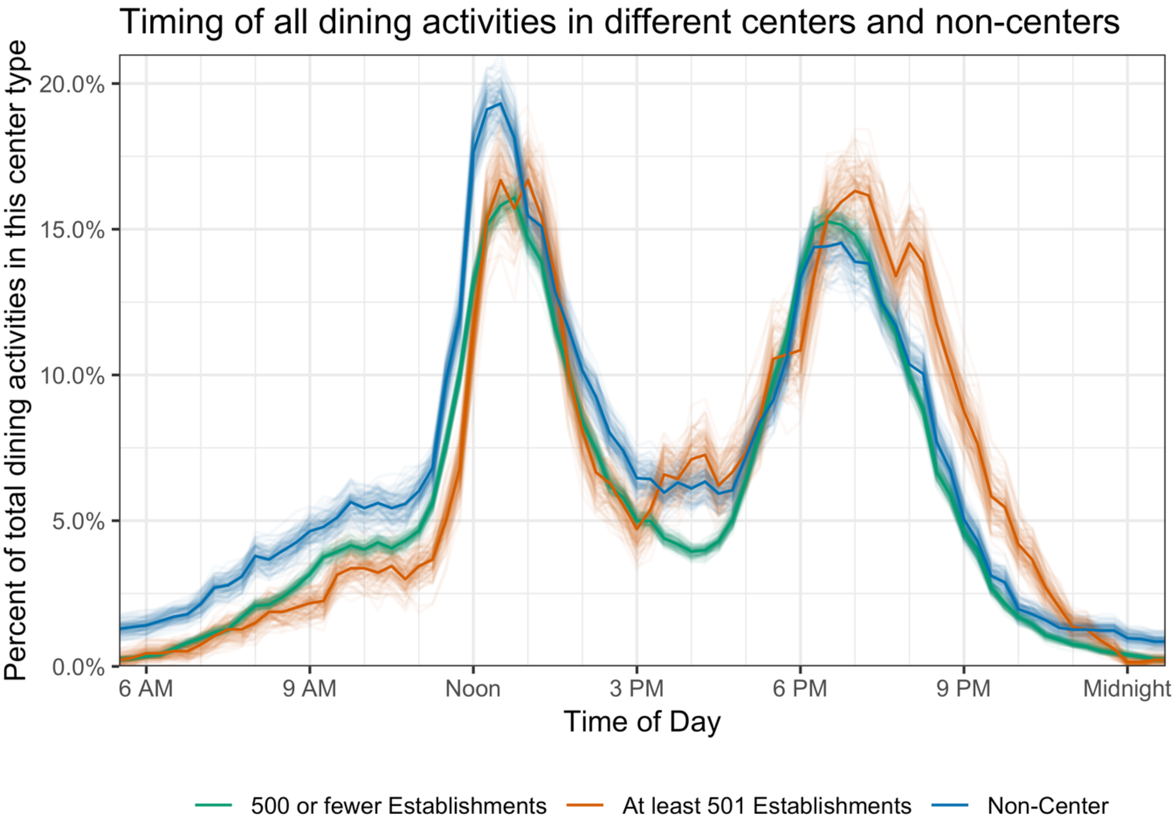


Figure 6.2 Timing of Dining Activities in major centers, smaller centers, and non-center areas. Dark curves show results from original sample; fainter lines are from bootstrapped samples.

As expected, dining activities are most common around lunch and dinner times across all business center types as well as less dense areas, but Figure 6.2 shows some differences between location types. Early morning and midday dining trips make up a larger share of such activities in non-center areas than in centers. Dinner is the main meal for which people go to both center types shown here, unlike non-center areas, which draw slightly more people at lunch. The patterns for major centers and smaller centers are generally similar until midafternoon, although lunch appears to start about 30 minutes later in large centers. Dinner starts at similar times in all three types of places, but the dinner period lasts about 45-60 minutes longer in large centers.

The timing of dinner activities is shown in more detail in Figure 6.3. Both lower-density categories reach their maximum concentration of diners between 6:15 and 6:30 PM, but large centers do not until after 7:00 PM. In addition, the number of people getting dinner in major centers begins to decline roughly an hour later than it does in the smaller centers, an effect that is borne out by the bootstrapping: major centers are significantly higher than both other center types for most times from 7:15-11:15, and at almost all time points after 8 PM. In addition to starting later, diners in major centers also spend slightly longer at meals starting between 17:00 and 22:00 (77.7 minutes, on average) than in smaller centers (71.2 minutes), or non-centers (66.5 minutes). Interestingly, despite the later starts and somewhat longer meals, 13.7% (77/561) of people who get a meal at any time after 16:00 in major centers list an entertainment activity after dinner, whereas the rate is 8.8% (475/5391) for smaller centers. This potential for activity pairing may represent a substantial pull to major downtowns, which present a wider range of entertainment opportunities.

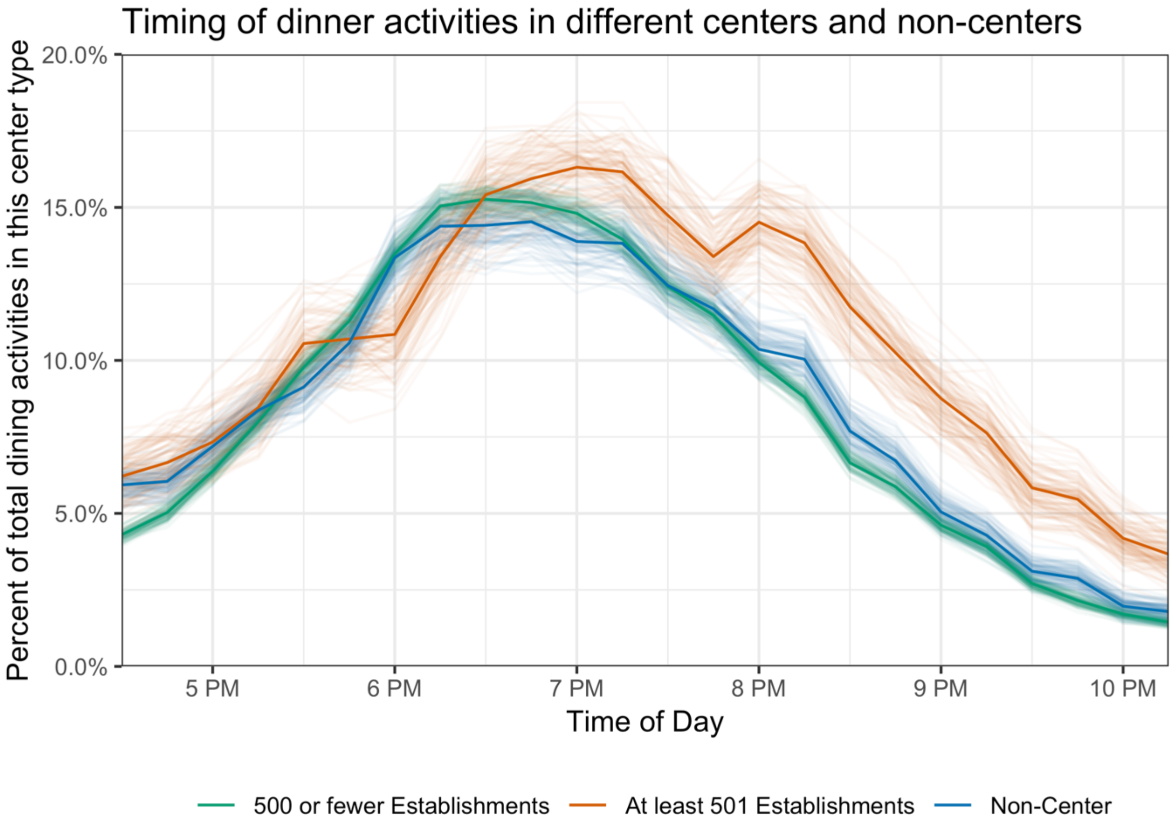


Figure 6.3 Timing of PM Dining activities. Dark curves show results from original sample; fainter lines are from bootstrapped samples.

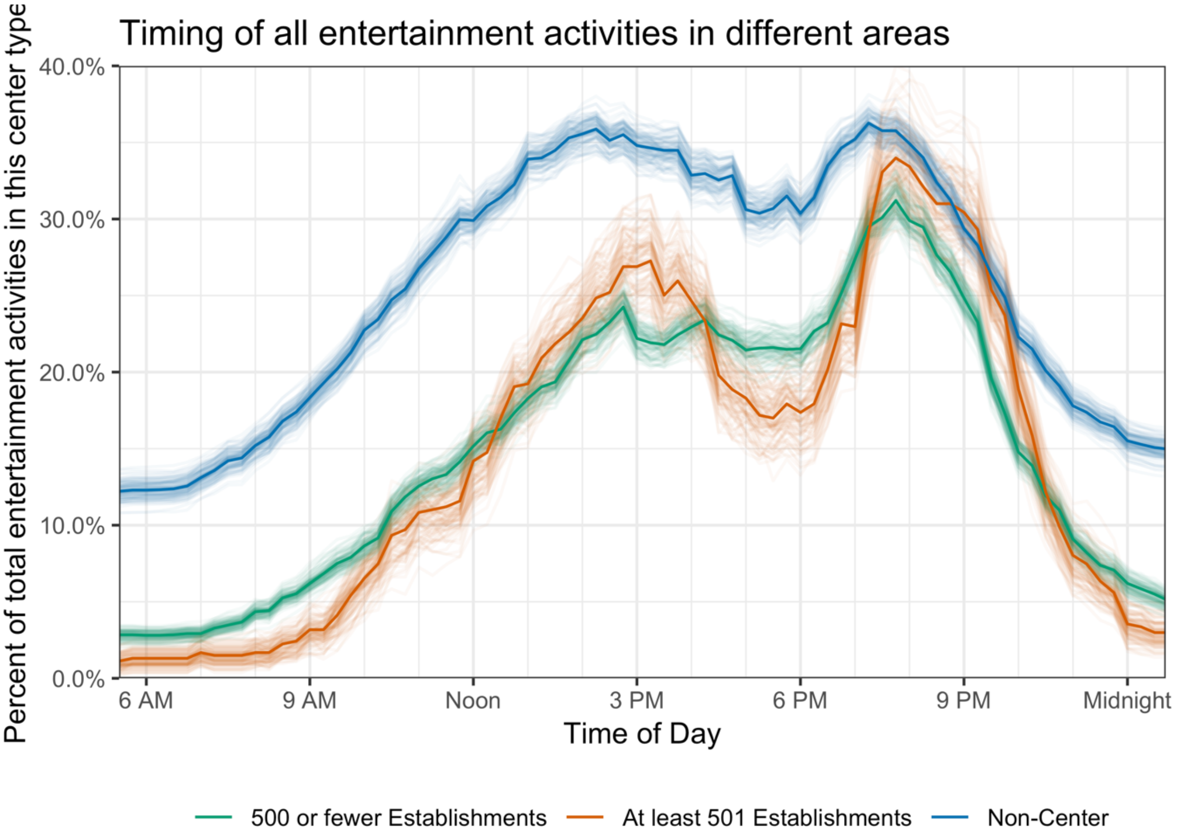


Figure 6.4 Timing of Entertainment Activities. Dark curves show results from original sample; fainter lines are from bootstrapped samples.

Figure 6.4 shows entertainment activities (e.g., movies, watching sports, listening to live music). This is one of the less common activities in the sample, which results in a somewhat noisier plot than the other activities. Entertainment is primarily an afternoon and evening activity in the centers, but the mix of afternoon and evening activities may differ be somewhat from place to place. Entertainment activities last similar lengths (about 2.5 hours, on average) in both center types. Interestingly, activities listed as “entertainment” outside of centers appear to be qualitatively very different in that a significant number of these activities last overnight.

c. Conclusions

Spatial variation calculated using commercial centers has a strong impact on activity scheduling and contributes to timing heterogeneity roughly as much as life cycle stage and day of the week. Day of the week and the mix of chain and independent businesses in a center appear to have a stronger impact on the overall rate of participation in various activities than on the timing of those activities, whereas life cycle stage, center size, and the mix of businesses in a center may have a relatively large impact on timing. This analysis of activity timing also suggests that the activity types listed in the CHTS are too narrow to adequately describe leisure activities.

7. Spatial Analysis of Activity Duration and Travel Time

This dissertation focuses on the development and application of spatial clustering methods in order to identify areas that provide lots of opportunities for people to shop, eat food, and be entertained. Given this focus, it is important to consider other avenues through which the spatial relationships between activity locations and attributes can be understood. The **duration** of activities people pursue in places and the **travel time** people are willing to accept in order to reach them are two attributes of activities that depend on the opportunities, context, and transportation infrastructure of the places in which people do them.

Broad categories like "shopping" and "dining" provide limited insight into what an activity entails, and duration can be a useful additional attribute to consider, since it makes it possible to distinguish between quick fast food meal and a longer sit-down meal with family at a restaurant, for example. Models for activity duration are also an important component of activity-based travel modeling systems because knowing how long someone is likely to spend at a place is vital to understanding how different destinations are linked through trip chaining. While the duration of a specific activity depends most on the details of what the person needs to accomplish (the duration of a trip to a grocery store varies greatly depending on how much needs to be bought, but time spent at a movie theater is typically right around two hours), the duration of individual shopping activities has been shown to relate to urban geography and the temporal constraints imposed by the rest of the traveler's schedule, particularly from other destinations on the same tour (C. R. Bhat, 1996; Schwanen, 2004). Travel time is strongly related to activity duration and frequency, but investigations of this relationship have focused mainly on the public health impacts of limited access to grocery stores close to home (Cannuscio et al., 2013; X. Chen & Clark, 2016; Grebitus, Lusk, &

Nayga, 2013) or the ability of unique and “vivid” destinations to attract people from much farther away than would be expected for ostensibly similar activities (Alter & Balcetis, 2011; Darley & Lim, 1999).

I anticipate finding spatial dependence in the activity duration and corresponding travel times from a number of causes. Nearby activity locations might capture different people engaging in very similar activities at the same specific business; somewhat more widely-spaced activities may reflect the mix of features and specialties at the level of a small center or a section of a larger one; activities spaced over wider areas (such as between nearby centers) may reflect local transportation infrastructure and conditions.

This chapter aims to investigate the spatial dependency of activity duration and travel time both before and after controlling for other variables that likely affect them using simple linear models, multi-level models, and spatially autoregressive models. Greater Los Angeles is used as the study area in order to simplify the calculation of spatial statistics. I am particularly interested in differentiating between the spatial dependency of these variables within a center and the relationship between nearby centers. Overall, I find that there is mild but significant spatial autocorrelation in duration and (to a lesser degree) travel time between nearby activity locations, but it is difficult to measure the relationships between nearby centers because of the widely varying activity densities in these centers. Using simple linear models to account for known information about the people who perform these activities, and other information about the tours they take place on decreases the overall degree of spatial autocorrelation somewhat, but multilevel models (using center membership as the grouping variable) and spatially-lagged models have a considerably stronger impact.

a. Data Preparation

For this analysis, I investigate 6,756 activity and travel durations for shopping (3,940), dining (2,236), and entertainment (580) locations in greater Los Angeles drawn from the California Household Travel Survey (CHTS). These activities represent time expenditures of 5,717 distinct persons (belonging to 4,713 households) that visited one or more of 1,016 centers in Greater Los Angeles. To extract activity durations and travel times, I first identified all tours and trip chains between home, work, and school locations in the dataset, as discussed in Chapter 6. Because there are strong relationships between activities on the same tour, I decided to avoid repeating observations in the models by selecting the single longest-duration activity on a tour as the "primary" purpose and calculating the total duration spent at other destinations on the same tour for use as an explanatory variable in the model. Total travel time was calculated by adding the durations of all the trips on the tour from the departure from an anchor point until the return to another (or the same) anchor point. To identify the primary mode type used on the tour, I first aggregated the two dozen options that CHTS provided into seven general categories: bike, walking, bus, personal vehicle (car / motorcycle), rail, other mass transit (mostly shuttles, paratransit, and taxis), and other modes. While people in the Los Angeles area utilize a wider range of modes when travelling for other purposes, these trips were overwhelmingly made by personal vehicle (85.6%), with walking (9.7%) making up the largest share of the rest.

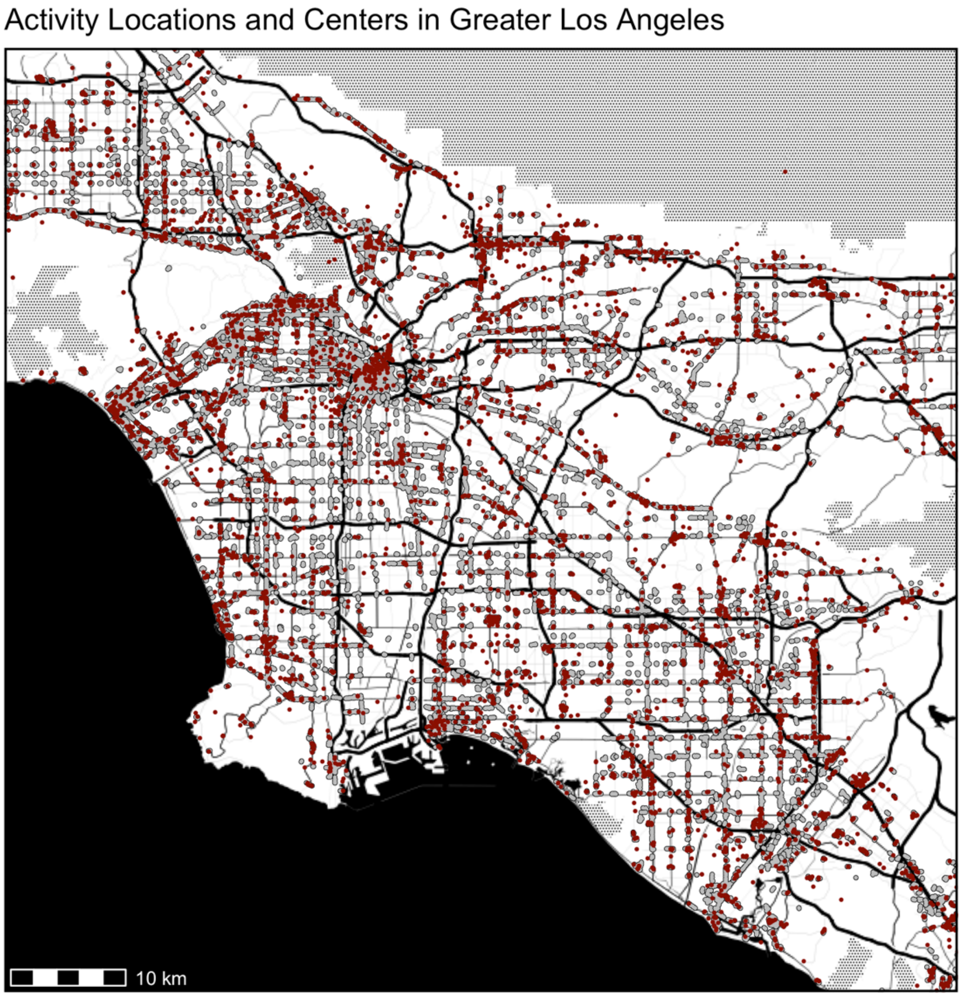


Figure 7.1 Activity locations (dark red) and centers used in this analysis. The study area covers the Los Angeles basin and some of the surrounding areas.

In order to investigate higher-level spatial relationships in activity duration and travel time, activity locations were attached to the commercial centers identified in Chapter 5. Many smaller centers do not contain any activity locations from the survey, as shown in Figure 7.1, which provides a map of activity and center locations. Table 7.1 shows the number of activities recorded in centers of various sizes. Unsurprisingly, larger centers generally had a much larger number of activities, and a relatively small share of the smallest centers contained any activity locations at all. This supports the use of the aggregative approach to center classification used in Chapter 6, but it makes center-level analysis that relies on specific locations problematic, since center-level residuals and means cannot be reliably computed from centers with very few activity locations.

Table 7.1 Number of activity locations recorded in centers of different sizes. Most centers with at least 11 business establishments have at least one activity, and almost all centers of over 50 have at least give activities recorded.

Center Size	Centers	Any Activity	At least 5 Activities	% of All Businesses	% of All Activities
5 or fewer	558	105	8	5.1%	3.0%
6 to 10	574	237	35	9.0%	8.3%
11 to 25	485	334	94	16.1%	17.1%
26 to 50	191	174	87	13.9%	15.1%
51 to 100	102	101	86	14.9%	18.5%
101 to 500	59	59	57	20.2%	20.7%
501 or more	6	6	6	20.7%	17.3%

Durations vary considerably across the three activities, as shown in Figure 7.2: shopping has a mean duration of 45.7 minutes, dining takes 67.8 minutes on average, and entertainment activities average 179 minutes. Entertainment activities are typically the longest and are also the most varied. The three activity categories are much more similar than activity durations, but entertainment activities still draw generally longer trips, possibly because these locations are rarer. Many of the activities come from multi-destination tours: 41.8% represent the only stop, and 25.1% have exactly one other stop. Of the tours with any time spent at other locations, the average is 50.1 minutes.

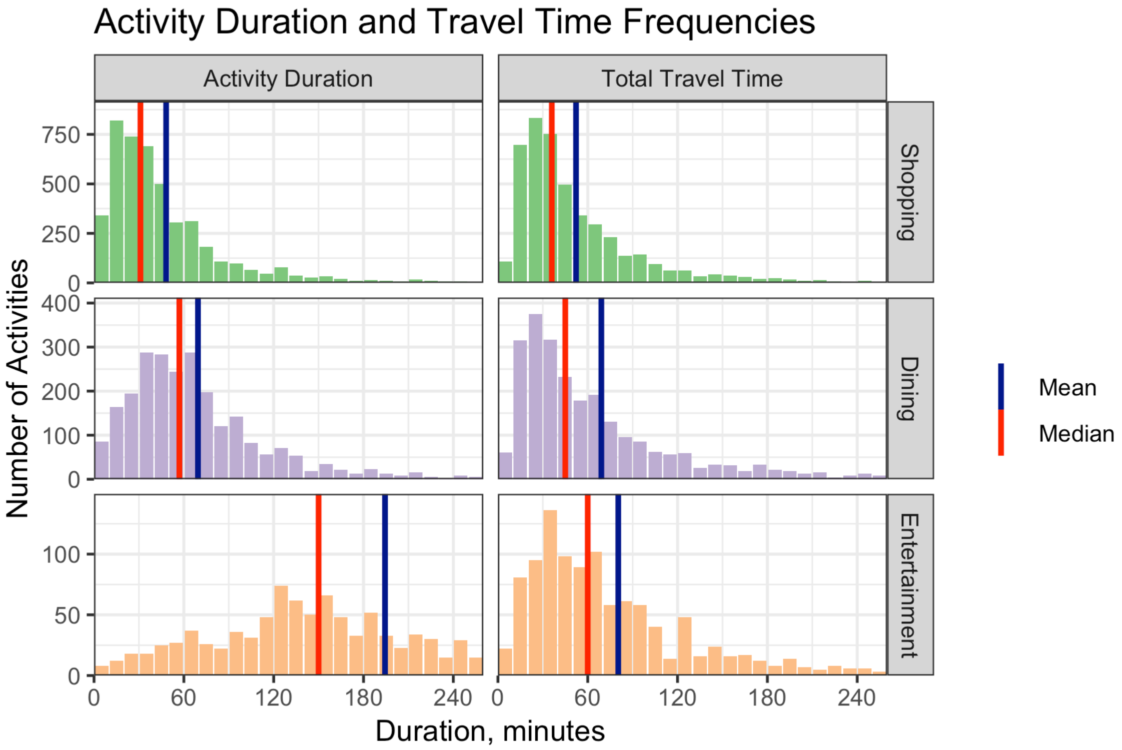


Figure 7.2 Distribution of durations and travel times by activity type.

b. Methods

Moran's I Correlogram

Moran's I is a commonly used measure for identifying spatial autocorrelation that identifies the degree of spatial autocorrelation present in a dataset using in generally the same terms as Pearson's correlation coefficient (Anselin, 2003; Bivand, Pebesma, & Gómez-Rubio, 2013; Cliff & Ord, 1981; Moran, 1950). Unlike multivariate correlation methods, which use paired values of different variables to determine whether high/low values of one variable align with high/low values of the other, Moran's I compares multiple observations of the same variable measured in different places, with "neighboring" observations paired for comparison (Equation 7.1). In the equation, N is the total number of points, and W is the sum of all values in the weight matrix; w_{ij} is the value of the spatial weights matrix for points i

and j. Significance tests can be performed on Moran's I by converting its value to a z-score and testing it against the normal distribution (steps for calculating the variance of Moran's I can be found in Cliff & Ord, 1981).

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \qquad \text{Equation 7.1}$$

The results of this calculation are controlled using the spatial weights matrix, which has one row and one column per observation and contains zeros for non-neighboring point pairs and positive values for point pairs that are considered neighbors. The values in the non-zero cells are generally set either to a function of distance or a single consistent value (if neighbor status is considered a binary value). The choice of weights matrix strongly influences the result and should be chosen to reflect the expected underlying structure of the correlation. By adjusting the spatial weights matrix to connect a continuously expanding range of points, it can be used to produce a correlogram, a plot that shows the overall spatial autocorrelation of observations at different distance ranges.

Spatially Autoregressive Models using Lagged Dependent Variables

In this section I describe various forms of regression models used here to explain the variation of duration or travel time to a center as a function of characteristics of the person, location visited, and tour-level travel behavior, as well as to account for spatial dependency. Each model form represents a different data generating process. These models are often compared against a baseline simple linear regression model (Equation 7.2) that does not address spatial dependency between the observed value of the dependent variable (Y) at one location and the values of the same variable elsewhere. The models presented here contain spatial explanatory variables (X) of the location where each activity episodes happens.

$Y = X\beta + \epsilon;\ \epsilon \sim N(0, \sigma^2)$ Equation 7.2

The random error term is assumed to be homoscedastic with mean zero. Y is a vector of 6,756 dependent variable values, X is a matrix with 6,756 rows and one column for each explanatory variable and an additional column of ones. The vector β contains one regression coefficient for each column of X, including the additional that corresponds to the intercept, and ε is the random error term also a vector of 6,756 values.

The spatial lag model (SLM) is a linear regression model that treats the Y values for each unit of analysis (in this case each distinct activity episode) as a function of the Y values for activity episodes in the neighborhood (Equation 7.3). The symbols used here are the same as Equation 7.2 with the addition of W, which is a spatial weights matrix (like the one in Equation 7.1) that identifies the spatial neighborhood for each observed Y, and the spatial correlation parameter ρ. This model implies that the duration of an episode at a center is a function of the duration of other episodes in the same center or other close by centers. This form of spatial dependency may be due to the existence of places that are designed specifically for short activities (e.g., fast food places at lunch time) or longer ones (e.g., concert halls, movie theaters). In the models for travel time, this form of spatial dependency may address locations where people are likely to experience long travel times due to congestion. The random error term is assumed to be independent between observations, as in Equation 7.2.

$Y = X\beta + \rho\, WY + \epsilon;\ \epsilon \sim N(0, \sigma^2)$ Equation 7.3

The spatial error model provides an alternative to SLM. In spatial error models, the dependent variable is not directly dependent on other values in the neighborhood, but the random error term (u) has a spatially correlated structure, shown in Equation 7.4. In this equation, the spatial weights matrix W is moved from the main equation into the equation for the error term and is multiplied by the spatial correlation parameter λ. This implies that spatial dependency is due to unobserved factors not included in the explanatory variables of the regression.

$Y=X\beta+u; \quad u=\lambda Wu+\epsilon; \epsilon\sim N(0, \sigma^2)$ Equation 7.4

Another variant of spatial dependency model includes both a spatial lag for the dependent variable and a spatial lag for the random error term (Equation 7.5), which combines components from Equations 7.3. and 7.4.

$Y=X\beta+\rho WY+u; u=\lambda Wu+\epsilon; \epsilon\sim N(0,\sigma^2)$ Equation 7.5

If we think spatial dependency is due to spatial similarity among the explanatory variables, a model with spatial lags of the X variables is another option (Equation 7.6). The symbols used here are the same as Equation 7.2 with the addition of W, which is the matric of the spatial neighborhood for each observed X and the spatial correlation parameter γ.

$Y=X\beta+\gamma WX+\epsilon; \epsilon\sim N(0, \sigma 2)$ Equation 7.6

Finally, a model that encompasses all of the forms of spatial dependency included in the above models is shown in Equation 7.7. This is a combination of the specification in Equations 7.3, 7.4, and 7.6.

$Y=X\beta+\rho WY+\gamma WX+u; u=\lambda Wu+\epsilon; \epsilon\sim N(0, \sigma^2)$ Equation 7.7

Spatial regression literature suggests to select the most appropriate model based on theoretical consideration, past evidence of other studies, and a series of LaGrange multiplier tests that compare different formulations against the simple liner regression model. Preliminary LaGrange multiplier tests of the results of linear models indicated that a model with spatially lagged Y values was likely to be more appropriate for this dataset than models using spatially correlated errors (Anselin, 2010).

Multilevel Models

Multilevel models for spatial dependency are based on the suspicion that two different levels of spatial dependency might explain the variation of the dependent variable Y. In the examples presented here it is possible that unobserved characteristics of the centers are not included in the regression specification and they are a spatial trend at the centers level that is separate from any spatial correlation at the episode level. To test the possibility of this dependency we formulate a regression model as a linear mixture model and allow for the intercept of the regression to vary randomly from a center to another. Equation 7.8 shows the form for a multilevel model with random intercepts. This equation modifies the simple linear regression equation by adding a term corresponding to the effect of each center, A_{cent}. A_{cent} is in turn explained by center-level characteristics ($X_{cent}\beta_{cent}$) and a random term, a_{cent}, that accounts for the remaining portion of the variance of Y that is consistent at the level of each individual center. The center-level random term (a_{cent}) provides a unique intercept for each center, but instead of estimating each of these values, the model estimates their standard deviation σ_{cent} as a parameter of the model and provides a rough guess for the true value of a_{cent} for each center.

$Y=X_{act}\beta_{act}+A_{cent}+\epsilon;\ \epsilon\sim N(0,\sigma_{act}^2)$

$A_{cent} = X_{cent}\beta_{cent}+ a_{cent};\ a_{cent}\sim N(0, \sigma_{cent}^2)$. Equation 7.8

In cases where there is thought to be spatial dependency both between observations within each unit and between units, it is possible to estimate a that simultaneously incorporates both forms of spatial dependency in a hierarchical structure. This more complex model is difficult to estimate due to sparseness of the weight matrices and spatial correlations involved and is discussed more at the end of the chapter.

c. Results

Spatial Autocorrelation of Variables of Interest

To assess the overall extent of spatial autocorrelation in the data, I made a Moran's I correlogram for activity durations and total travel time. I tested multiple schemes for producing the weights matrix and determined that ten nearest neighbors with row standardized weights gave the clearest results. Because the overall spatial distribution of activity locations varies over the study area, neighbors identified using a fixed radius may be less useful. Each lagged step in the correlogram represents links formed by adding the neighbors identified in the previous lag, such that where lag 1 matches all points to their 10 nearest neighbors, lag 2 matches points to any of the 10 nearest neighbors of their existing neighbors that they were not already matched. The correlogram (Figure 7.3) indicates that both variables have significant spatial autocorrelation out to three lags, and that Moran's I is substantially higher for duration than for travel time within the two nearest sets of neighbors.

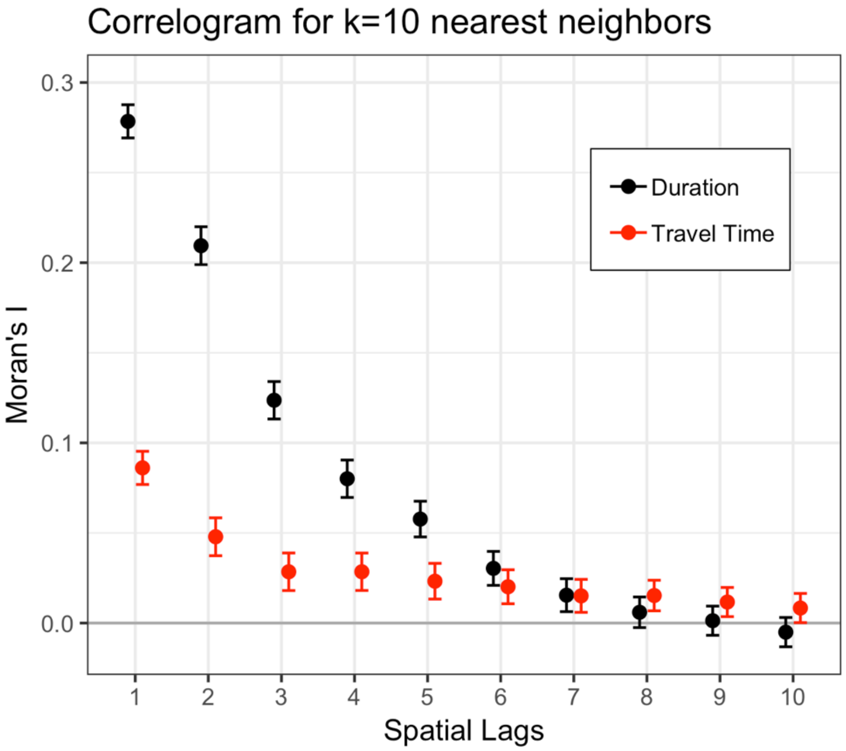


Figure 7.3 Moran's I Correlogram of activity durations and travel times using k=10 nearest neighbors to populate the W matrix. Error bars correspond to ± 2 standard deviations, roughly corresponding to a z-test with a threshold of p=0.05. Expected value of I in the case of randomness is shown with the grey line slightly below 0.

Model Results

I ran four models for each of the two response variables: one linear model with a set of explanatory values relating to the activity, traveler, travel mode, tour, and the center in which the activity took place; one linear model with those variables plus the other response variable and the total duration of other activities on the tour; one multilevel model grouped by center with all the previously included variables and an intercept for each center; and one spatially autoregressive model with spatial lags. Results for the models for activity duration and travel time are shown in tables 7.2 and 7.3, respectively.

Apart from the variables shown in these tables, I tested specifications including variables such as gender, number of kids in the household, income, and various other attributes of the commercial centers in which the activities occur but found them to be insignificant predictors of both duration and travel time.

Lagrange multiplier tests performed on the final linear models indicated that there was likely a significant degree of spatial autocorrelation in the residuals and suggested that a model with spatial lags was an appropriate way to address this issue. The spatial coefficient rho is significantly different from zero in spatial lags models for both activity duration and travel time, which indicates that including spatial linkage improves the models. The residual spatial autocorrelation Lagrange multiplier test on the residuals of the spatial model indicates that there is unlikely to be any residual spatial autocorrelation.

The models presented here explain a relatively small share of activity duration and travel time (R^2=0.325 for duration and 0.330 for travel time from the more complex linear models), but the coefficients capture differences that are highly significant and often quite substantial, and the results provide useful information about the spatial dependency of the variables in question. The variables modeled generally contain a high degree of randomness, especially given the number of specifics about these activities that were unavailable (notably more specific details about the activities), and hazard-duration models often provide a more appropriate fit for activity duration (C. R. Bhat, 1996), but past research did not account for spatial dependency. For both variables, individual coefficients are fairly stable between the different models, which suggests that substantial new information is added by including space in the model, whether in the form of group intercepts or spatially lagged dependent variables.

Models for activity duration (Figure 7.2) consistently show that dining and particularly entertainment activities take substantially longer than shopping ones, and the difference between entertainment and shopping activities is nearly two hours in every model. People appear to be more willing to take active modes of transportation (walking and biking) to shorter-duration activities, whereas activities reached by transit generally lasted longer than those reached by car. This may reflect a difference in the sorts of activities people pursue close to home and those they have to travel farther for. The reference category for tour type is home-based tours, and activities on these types of tours (and on overnight tours) tend to be substantially longer than activities on tours that start or end at locations with a fixed time. Work-based tours and commutes to work are associated with particularly short activity durations (unsurprisingly), suggesting that these are particularly likely to be quick stops for coffee or minor errands. The only variables that change by more than their standard error between the linear and multilevel models and the spatial model are average business size in the center (the only variable explicitly attached to space that was significant in this set of models) and entertainment, which perhaps indicates that different places specialize in entertainment activities. Although the two models are not nested and may not be directly comparable using conventional fit metrics, the nearest-neighbors spatially-lagged model appears to be a generally similar fit for this data than the multilevel model.

Table 7.2 Results from models for activity duration, values are coefficients and (standard errors). Groups of dummy variables are indicated with numerals. Reference categories are as follows: 1: shopping activities, 2: all other life cycle stages, 3: personal vehicles, 4: home-based and all other tours.

Terms for Activity Duration Models	Regression without Time X (Eq. 7.2)	Regression with Time X (Eq. 7.2)	Regression with Center Intercepts (Eq. 7.8)	Spatially Lagged Model (Eq. 7.3)
(Intercept)	51.50 (2.15)	50.08 (2.13)	50.79 (2.49)	28.72 (2.15)
Dining[1]	22.18 (1.91)	19.32 (1.90)	18.97 (1.85)	17.37 (1.80)
Entertainment[1]	137.16 (2.56)	130.11 (2.56)	119.86 (2.55)	112.47 (2.51)
Disabled[2]	7.67 (4.89)	5.56 (4.82)	6.02 (4.62)	6.00 (4.57)
Home Duty[2]	7.08 (3.48)	7.34 (3.42)	6.19 (3.26)	5.24 (3.25)
Part Time[2]	-4.38 (2.41)	-4.48 (2.37)	-4.53 (2.26)	-4.63 (2.25)
Full Time[2]	-1.73 (2.02)	-2.37 (1.99)	-3.01 (1.90)	-3.33 (1.88)
Walk[3]	-15.15 (2.94)	-11.08 (2.91)	-10.19 (2.83)	-11.15 (2.75)
Bike[3]	-11.64 (7.68)	-11.90 (7.56)	-11.93 (7.23)	-11.65 (7.17)
Bus[3]	35.37 (7.02)	49.23 (6.97)	42.11 (6.68)	43.87 (6.61)
Rail[3]	3.18 (15.32)	14.12 (15.11)	10.98 (14.60)	7.93 (14.33)
Other Transit[3]	31.42 (8.26)	22.86 (8.15)	27.18 (7.78)	23.01 (7.73)
Other Mode[3]	-25.24 (13.18)	-22.38 (12.98)	-31.04 (12.40)	-26.46 (12.30)
Mode Changes	13.38 (2.62)	12.39 (2.59)	9.20 (2.47)	8.41 (2.45)
Total Stops	-3.06 (0.54)	-7.38 (0.60)	-6.02 (0.58)	-5.98 (0.570)
Work-Based Tour[4]	-24.16 (4.30)	-20.13 (4.25)	-23.23 (4.06)	-22.46 (4.02)
Late-Night Tour[4]	75.20 (5.68)	59.40 (5.70)	48.78 (5.49)	52.63 (5.40)
Commute to Home[4]	-18.92 (3.38)	-18.06 (3.33)	-19.40 (3.18)	-18.10 (3.16)
Commute to Work[4]	-28.35 (6.34)	-26.34 (6.25)	-24.18 (5.95)	-25.07 (5.92)
Mean Employees per Establishment	0.25 (0.04)	0.22 (0.04)	0.21 (0.06)	0.08 (0.04)
Travel Time		0.17 (0.02)	0.13 (0.02)	0.12 (0.02)
Duration at other Destinations		0.21 (0.02)	0.14 (0.02)	0.16 (0.02)
other			Intercept SD=28.06	Rho=0.39 (0.01)
R^2	0.325	0.346		
logLikelihood	-46423	-46294	-45967	-45928
AIC	92890	92636	91983	91906
BIC	93044	92804	92158	92081

Models for travel time (Table 7.3) also show that dining and entertainment activities are linked to longer-duration trips than shopping ones, but the differences are much less substantial than in the models for activity duration. The entertainment part of this effect decreases substantially in models that include activity duration (suggesting that it's not so much that people are willing to travel farther for entertainment, but that they're generally willing to travel farther for longer-duration activities). People with disabilities tend to have longer tours, possibly reflecting inequities in the transportation system. Cars are much more flexible than fixed-route transit and are particularly advantageous for multiple-destination trips, which may explain why the other modes are generally used on shorter trips. As is with the case for activity duration, tours with strict temporal constraints, such as those taken on the way to work or in the middle of the work day, generally include much shorter trips. Substantially more center attributes are significantly linked effect in models for travel time than in models for duration, and people appear to be willing to travel farther when going to larger centers and those with a higher share of independent business establishments. The spatial coefficient rho is much smaller for this model than it is for activity duration, but it is still clearly significant. Additionally, the multilevel model appears to do a slightly better job than the spatial model for travel time.

Table 7.3 Results from models for total travel time, values are coefficients and (standard errors). Groups of dummy variables are indicated with numerals. Reference categories are the same as in Table 7.2

Terms for Travel Time Models	**Regression without Duration X (Eq. 7.2)**	**Regression with Duration X (Eq. 7.2)**	**Regression with Center Intercepts (Eq. 7.8)**	**Spatially Lagged Model (Eq. 7.3)**
(Intercept)	1.27 (2.71)	-4.51 (2.70)	-2.79 (2.95)	-11.37 (2.80)
Dining^1	14.30 (1.36)	12.09 (1.35)	12.36 (1.36)	12.05 (1.34)
Entertainment^1	21.04 (1.83)	6.45 (2.09)	6.26 (2.10)	5.71 (2.08)
Disabled^2	11.34 (3.46)	10.58 (3.41)	9.90 (3.41)	10.46 (3.38)
Home Duty^2	0.13 (2.46)	-0.31 (2.42)	-0.60 (2.42)	-0.44 (2.40)
Part Time^2	-0.65 (1.71)	-0.36 (1.68)	-0.46 (1.68)	-0.60 (1.67)
Full Time^2	2.43 (1.43)	2.49 (1.41)	2.35 (1.41)	2.29 (1.40)
Walk^3	-16.72 (2.11)	-14.38 (2.08)	-14.92 (2.08)	-14.40 (2.06)
Bike^3	6.71 (5.44)	8.49 (5.35)	7.99 (5.34)	8.58 (5.31)
Bus^3	-36.95 (4.98)	-34.68 (4.94)	-36.35 (4.94)	-35.43 (4.91)
Rail^3	-13.26 (10.86)	-7.37 (10.70)	-12.75 (10.77)	-11.16 (10.62)
Other Transit^3	35.52 (5.85)	31.20 (5.76)	31.09 (5.75)	30.52 (5.72)
Other Mode^3	-4.46 (9.34)	-0.81 (9.18)	-0.51 (9.17)	-1.39 (9.11)
Mode Changes	7.18 (1.86)	6.20 (1.83)	5.68 (1.83)	5.81 (1.82)
Total Stops	15.55 (0.38)	14.69 (0.40)	14.72 (0.40)	14.69 (0.39)
Work-Based Tour^4	-18.30 (3.06)	-15.59 (3.00)	-16.29 (3.00)	-16.56 (2.98)
Late-Night Tour^4	69.18 (4.02)	59.73 (4.01)	59.17 (4.00)	59.14 (3.96)
Commute to Home^4	3.66 (2.39)	6.43 (2.36)	6.26 (2.36)	6.61 (2.34)
Commute to Work^4	-2.17 (4.49)	1.61 (4.42)	1.63 (4.41)	2.14 (4.39)
Mean Employees per Establishment	0.19 (0.03)	0.17 (0.03)	0.17 (0.03)	0.14 (0.03)
Businesses (/100)	0.19 (0.06)	0.17 (0.06)	0.24 (0.16)	0.12 (0.06)
Independent Business Fraction	16.91 (3.35)	19.72 (3.30)	18.03 (3.64)	17.32 (3.28)
Duration at this Destination		0.09 (0.01)	0.08 (0.01)	0.08 (0.01)
Duration at other Destinations		0.15 (0.02)	0.14 (0.02)	0.14 (0.02)
other			Intercept SD=7.48	Rho=0.17 (0.02)
R^2	0.319	0.341		
logLikelihood	-43616	-43483	-43439	-43443
AIC	87279	87017	86932	86939
BIC	87447	87199	87121	87128

Residual Correlograms

Lagrange multiplier tests indicated that spatial autocorrelation is present in the linear models but is eliminated by the spatially lagged model, but it is important to examine the extent of this, particularly since the spatially lagged model only accounts for spatial autocorrelation at the first lag. Figure 7.4 shows the correlograms for the residuals from all the models; the correlation of the original data is shown in black. All of the models reduce the overall spatial autocorrelation somewhat, which indicates that some of the apparent spatial autocorrelation was explainable with other variables. The spatially-lagged model essentially eliminates spatial autocorrelation, which is exactly what its specification is designed to do. For both dependent variables, the multilevel model (purple) substantially outperforms both linear models at reducing spatial autocorrelation, with residual values of Moran's I below 0.02 in each case, and for activity duration the multilevel model effectively eliminates spatial autocorrelation of residuals. For the second and third lags, the multilevel model and spatially autoregressive model are both generally better than the linear model results but are essentially indistinguishable from each other. Overall, this indicates that grouping activity locations at the level of commercial clusters and using multilevel models is not only appropriate for this sort of analysis, but it accounts for spatial autocorrelation approximately as well as spatial autoregressive models while also providing a more useful interpretation and simpler model structure.

In order to investigate spatial autocorrelation at the center level, I calculated the average location of all the activities in each center and treated this as a center point location for identifying neighboring centers. I calculated correlograms for center-level mean residuals and fixed effects (group intercepts in the multilevel model) using values from centers with at least five activity locations, since group-level means are likely to be unreliable for small samples. These plots show generally very low values of Moran's I and vary wildly from lag to lag, as shown in Figure 7.5.

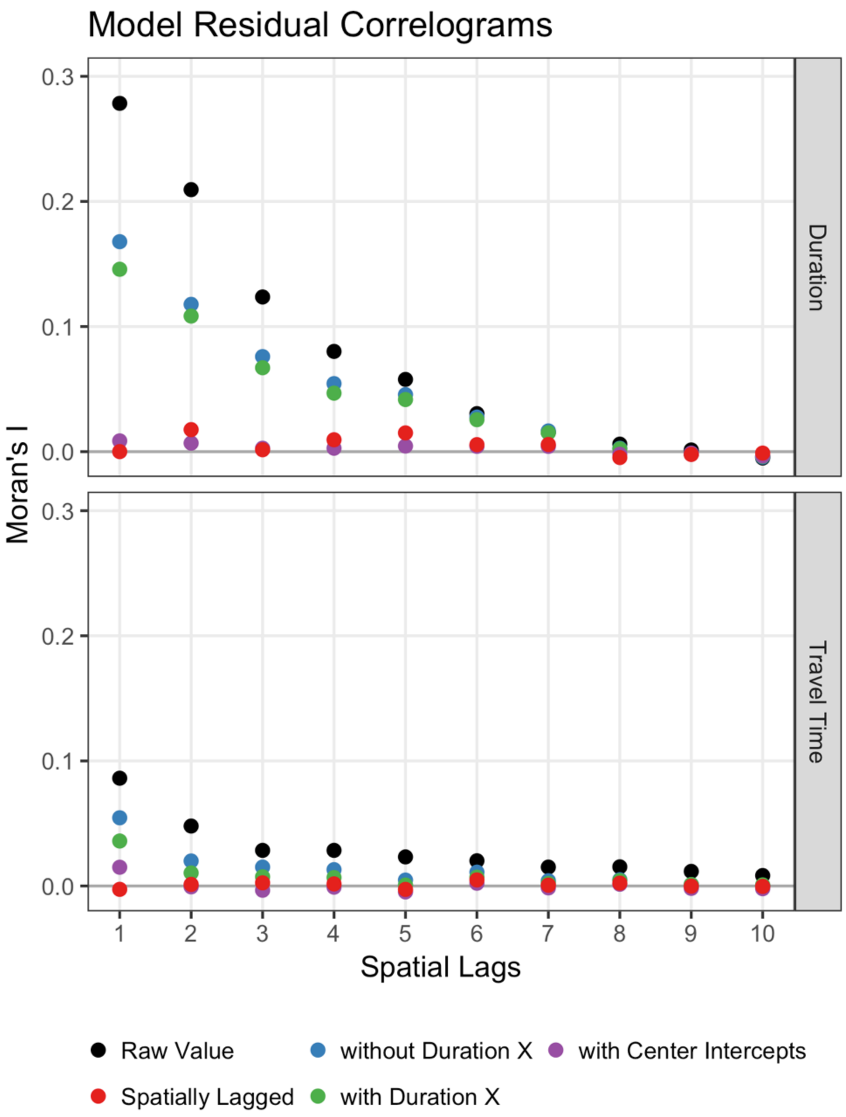


Figure 7.4 Moran's I Correlogram of residuals from models for activity durations and travel times using k=10 nearest neighbors. Using a multilevel model to identify a unique intercept for every center eliminates as much spatial autocorrelation as the spatially lagged model. This plot does not show error bars, but values above 0.01 generally have a significant p-value for this dataset.

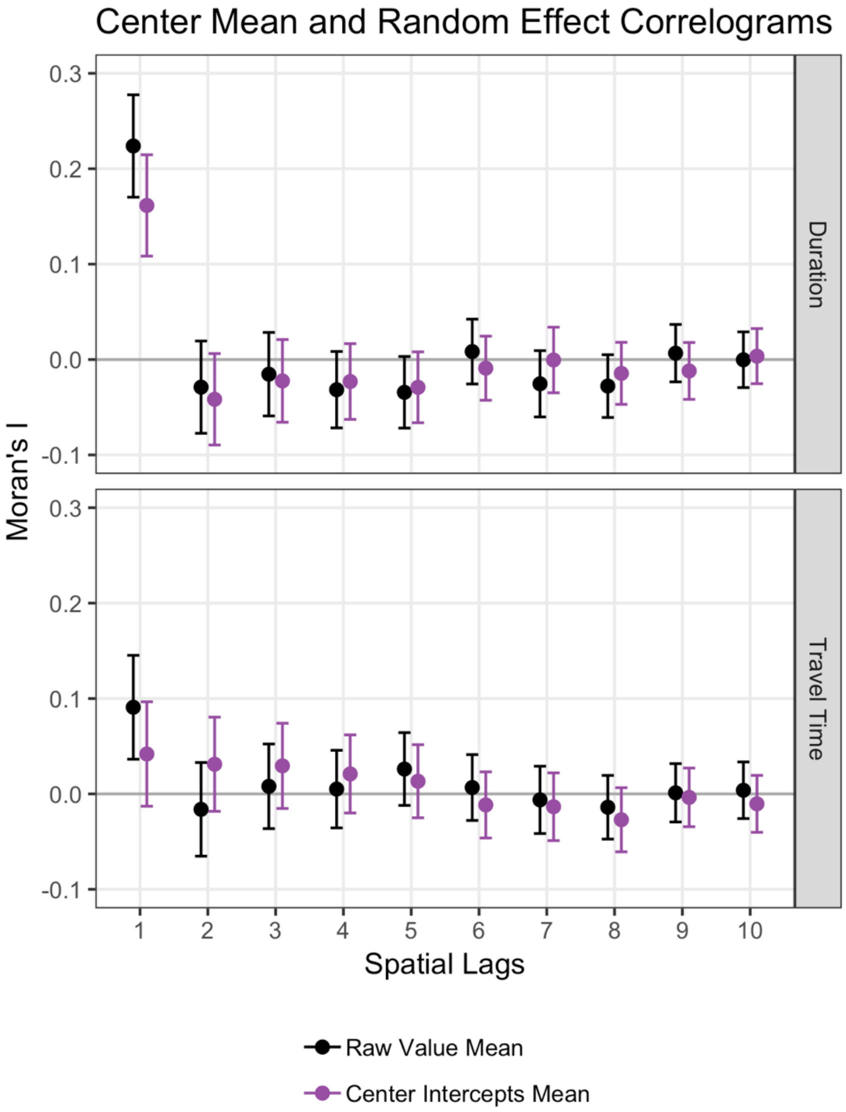


Figure 7.5 Moran's I Correlogram of center-level mean values and multilevel model conditional means for each center using k=10 nearest neighbors.

d. Discussion and Conclusions

Model Results

Activity duration and travel time have a moderate degree of spatial autocorrelation that can be partially reduced by accounting for the effects of various contextual variables through ordinary linear regression. The multilevel and spatially autocorrelative lagged models presented here both do a much more complete job of explaining the data while also nearly eliminating spatial autocorrelation from model residuals. While both models represent substantial improvements over the linear formulations, the results suggest that activity duration has more substantial spatial autocorrelation than travel time does, whether measured by the effect of the spatial coefficient rho in the spatially lagged model or the variance of center-level effects estimated by the multilevel model. This may reflect the fact that activity duration is tied to a single place, whereas travel time is presumably influenced by the traveler's point of origin, route, and the range of transportation links between them. For the variables in question, regression coefficients change minimally between the simpler models and the models that incorporate spatial effects get nearly identical coefficients to the final ones, but accounting for spatial dependency still adds a significant improvement to the model. This suggests that the spatial model in large part captures relationships that were otherwise not accounted for.

While these models provided a great deal of insight into the relationships between nearby activities (whether measured as neighbors or grouped into centers), model results about center effects are a mixed bag. Relatively few center-level variables were significant in the models, and the categorical and continuous variables that measure the mix of businesses in each center showed no relationship at all to the variables in question. The need of removing small centers from the cluster level analysis made it very difficult to determine whether spatial autocorrelation exists at that level. I skirted this issue in Chapter 6 by looking at the overall influence of a number of center-level variables, but that is not an option when modeling activities in specific places.

Hierarchical Spatially Autoregressive Models

While the models I present in this chapter incorporate either multilevel or spatially autoregressive components, hierarchical spatially autoregressive models address effects at both of these levels simultaneously, allowing for spatial dependency both at the level of individual observations and between higher-level entities, as well as providing for mixed effects and other features of multilevel modeling (Dong & Harris, 2015). A study run with simulated data found that when spatial variability is present both within and between groups, using models that account only for one of these effects tend to overestimate significance (Xu, 2014). These models have been implemented in R (Dong & Harris, 2016), and I tried to apply them to this dataset but ran into a number of issues. Points at the lower level appear to only be able to have other points in their same group as neighbors, and the wide mix of center sizes in my data made it impossible to estimate models on the full dataset. Additionally, when I tested the models on a subset of the centers with a large number of activities, the model function usually crashed without explanation presumably due to the sparseness of the weights matrices involved. A test model run with a very limited set of explanatory variables and a spatial subset of the data returned significant spatial coefficients both at the lower level and between centers, so I will explore this area in the future.

Mixed Effect Multilevel Models and Model Interpretation

Spatial autocorrelation is an effect rather than a process. The similarity between nearby observations does not reflect some inherent distance-based similarity relationship, but rather other processes that operate over space at various scales. While the clearest examples of this can be found in physical geography – nearby places don’t have similar climates just because they’re close to each other but rather because they receive similar amounts of solar radiation and are similarly situated with respect to global circulation patterns – similar explanations can be made for this sort of data. Activity duration does not exhibit spatial autocorrelation because people know how long other people spend in a place and adjust their schedules accordingly, but rather nearby places attract activities of similar durations because they offer similar amenities to people. Spatially autoregressive models are a useful way to handle the problems spatial autocorrelation causes in data analysis, and they are particularly useful for making predictions (much more than multilevel models that estimate effects from small local samples can be), but they are fundamentally a way to capture unmeasured variables and effects at the local level without necessarily explaining them.

Given that the multilevel models seem to fully capture the spatial autocorrelation present in this dataset, this spatial autocorrelation can be interpreted reflecting the similarity of activities within a center. This similarity likely reflects both the mix of opportunities available in that center as well harder-to-measure characteristics like sense of place and attractiveness, both of which are strongly related to people's willingness to spend time in commercial areas (Deutsch, 2013). I also tested models that allowed the effect of activity type to vary from center to center, essentially calculating a separate intercept for every combination of center and activity type. These models returned singular results, which indicates that they could not separate center-level direct effects from the variability of activity types between centers. This provides a means to handle spatial autocorrelation in a model while potentially *explaining* more of it, not in that we necessarily would know how to predict a center's effect, but rather in terms of ascribing the similarities between nearby places to specialization rather than spatial dependency alone.

8. Conclusions

As of 2017, passenger transportation accounted for roughly 17% of US greenhouse gas emissions, and emissions have increased considerably over the last few decades (despite substantial improvements in vehicle efficiency) because of increases travel demand, much of which is tied to urban sprawl (US EPA, 2019, p. 2.29). California's SB375 and similar regulations in other states and countries require regional and local governments to develop "sustainable community strategies," which generally encourage mixed use and infill developments (Steinberg, 2008). Travel demand is understood as primarily deriving from the need to leave the house to work, shop, socialize, and pursue other activities, rather than an innate desire to move around. Sustainable community plans aim to decrease the demand for personal vehicle travel and increase the use of walking and biking by allowing people to meet these needs closer to home or in areas with high access to public transit.

While the overall, region-scale relationship between urban density and personal vehicle use / vehicle miles traveled is fairly well-established (Cervero & Kockelman, 1997; Gim, 2012), numerous questions remain about the relationship between the density and diversity of opportunities provided by the built environment and the activities people pursue in it at finer spatial scales. Understanding this relationship is particularly important for planning infill developments, since these developments are relatively small and operate within the context of existing cities. Activity-based models have allowed travel behavior research to greatly improved the understanding and modeling of activity scheduling and interactions at the level of individual people and households, but the field has given much less focus to improving the way these models account for the interactions between space and place and the timing and types of people's activities. The ability of existing models to incorporate spatial heterogeneity and the shortcomings of the measurement units currently used to group potential destinations have both been acknowledged as areas in which the field needs to improve.

My goal for this dissertation is to develop new spatial units that can be used to examine travel behavior from the perspective of the destination, the opposite viewpoint of conventional travel behavior analysis. Census polygons and Transportation Analysis Zones are poorly aligned with the spatial distribution of activities and opportunities in urban areas, and spatial clustering can be used to identify more useful groupings. In order to get a sense of the usefulness of the spatial opportunity clusters I have extracted, I investigate the variability of the types, timings, and durations of activities that people pursue in them.

The core finding from the more methods-oriented chapters of this dissertation (Chapters 4 and 5) is that road-network-distance DBSCAN clustering of business location can be used to identify regional sub-centers in a way that is useful for spatiotemporal analysis of shopping, dining, and entertainment activities. I highlight the importance of using accurate road distance measures and demonstrate the feasibility of performing this sort of analysis statewide. Using known activity locations as secondary information in choosing clustering parameters allowed me to tune the results of this clustering to better suit the needs of this analysis. The resulting centers are likely too variable in size to be directly employed in a spatial choice model, but they investigation provided a useful exercise in balancing the risks of aggregation and disaggregation in spatial clustering.

While the centers I identified in Chapter 5 are not completely ready for their planned applications, they still provide a useful framework for investigating the spatial variability in activity participation, timing, and duration (Chapters 6 and 7). In Chapter 6, I find that activity timing varies roughly as much as a function of center attributes as it does by day of the week and life cycle stage. In Chapter 7, I investigate the spatial dependency of activity duration and find that using centers in a hierarchical linear model accounts for the spatial autocorrelation present in these variables nearly as well as using a spatially autoregressive model. The hierarchical model provides an estimate of the scale of unaccounted-for variability between centers.

The spatial dependency results shown in Chapter 7 are particularly important because they provide a first look at the usefulness of these centers for studying the spatial heterogeneity of the attractiveness or uniqueness of destinations. These model results show that by grouping activities spatially based on spatial clusters identified from opportunity locations, it is possible to account for all of the spatial dependency in activity duration. Thus, spatial autocorrelation present in the raw data can be understood as reflecting differences between centers that specialize in or are particularly attractive for certain types of activities, and the centers I identified can be understood as representing meaningfully distinct places.

This also highlights the potential usefulness of some future method of center extraction for identifying discrete options for a spatial choice model. Spatial choice models that incorporate the dependency between the (latent) utility values of nearby locations have been developed, but they are not used as commonly as they should be. Because these centers appear to capture most of the spatial dependency between activities directly, they simplify this process although it still may be necessary to account for the similarity of nearby centers.

Limitations

The shopping, eating, and entertainment activities that people pursue are not as consistent and predictable from day-to-day as their work and school activities. While the activities studied in this dissertation do not happen on regular and externally enforced schedules, many of them do have a degree of regularity both temporally, in the form of household weekly plans for grocery shopping and other household maintenance activities, and spatially in the form of habitual destinations. These patterns cannot be captured with single-day (or even two-day) activity diaries, but the few longer activity-travel surveys that have been conducted indicate that habit is a major component of people's travel behavior.

One aspects of the way that commercial centers were defined and the ways activity locations were attached to them are worth noting: both of these processes were done without consideration for the specific types of businesses present or the match between the activity type and business type. It would be possible to identify separate clusters for different business types, although this would lessen the ability to understand the relationships between paired activities (like shopping and eating). Alternatively, business type could be combined with road network distance to create a combined dissimilarity metric that could be directly employed in cluster detection. This modification might make more sense for place data that divides these types of businesses into a larger number of categories than NETS does. While changing the way business category is considered in clustering has some tradeoffs, using the relationship between activity type and business type to match activities to clusters when there are multiple candidates in a specific area seems like a generally good idea and will be employed in future iterations of this analysis.

Final clustering parameters were chosen because they identified centers that seemed correct when visually inspected on a map and matched appropriate activity locations better than other clusterings tested, as discussed in Chapter 5. Since these criteria are both fairly subjective, it is worth considering whether the cluster results could be optimized based on their internal properties alone. Substantial research has gone into optimizing cluster parameter selection without the use of secondary data (Schubert, Sander, Ester, Kriegel, & Xu, 2017), and future research will consider how to apply these methods to commercial center clustering.

One well-established limitation of the DBSCAN algorithm is that it is most effective when the underlying data has fairly consistent density. This is clearly not the case for the business establishment location data used in this dissertation, and the variability of business density and cluster size across California results in the detection of overly large clusters in major city centers and the presence of ambiguous and overlapping clusters in less dense areas and on the fringes of larger clusters. OPTICS was specifically designed to address this issue by taking a hierarchical approach to cluster detection, but as noted in Chapter 5, this method identifies highly ambiguous clusters when used for business location data. An alternative approach would be to allow clustering parameters to vary over space, with lower values of ε used in city centers and higher values used in less dense areas. Since this analysis aims to employ commercial centers for the analysis of activity participation rather than to describe the spatial clustering of business locations, it may be unnecessary to assign a single set of clustering parameters for the whole state.

Next Steps

The natural next step for the analysis presented in this dissertation is to use these centers as choice options in a discrete choice model for destinations for eating, shopping, and entertainment activities. While the current version of the clustering performs relatively well at identifying distinct centers in areas of mostly residential development, it produces centers that are much too big in areas with a consistently high density of commercial businesses. The boundaries of large centers likely reflect the overall distribution of opportunities better than TAZes, but since they are much larger than TAZes, they are less useful for studying where people travel within downtowns. One approach to this would be to subsample large centers using K-Means or a divisive hierarchical clustering method like DIANA. For these methods, care would need to be taken to determine an appropriate number of centers into which to divide each large center, likely a function of the total number of businesses or the area spanned by a center. Subdividing existing centers would likely increase center-level spatial autocorrelation, since to the degree that the large center had any characteristics that were consistent throughout it, these characteristics would then be shared by all the sub-centers created from it. A cleverly-designed partial choice set might account for this somewhat, and nested choice models would be another viable option, since these essentially turn the choice problem into a hierarchical one.

Using spatial clustering to identify centers of activity that are more realistic than TAZes, but the precision and accuracy of spatial data in both surveys and business locations (explored in Chapter 3) set a ceiling on our ability to match reported activity locations with known business locations. The uncertainty involved when performing this matching is one reason that destination choice modeling requires the use of aggregate spatial units. A project that could greatly improve the immediate usability of activity travel diaries for this sort of analysis would be to incorporate a standard place dataset (derived from OpenStreetMap or another source) directly into the data collection process and distributing this dataset to researchers. Surveys already ask respondents to report place names, but geocoding mainly uses reported addresses; instead, respondents could match their destinations by name and approximate location to potential destinations in this dataset. While this process might slightly increase the burden for survey respondents, it would allow better understanding of activities for which destination choice is made daily while also potentially allaying privacy concerns by diminishing the need to provide precise geocodes for more sensitive destinations like home, work, and school.

9. Bibliography

Alter, A. L., & Balcetis, E. (2011). Fondness makes the distance grow shorter: Desired locations seem closer because they seem more vivid. *Journal of Experimental Social Psychology*, *47*(1), 16–21. https://doi.org/10.1016/j.jesp.2010.07.018

Anas, A., Arnott, R., & Small, K. A. (1998). Urban Spatial Structure. *Journal of Economic Literature*, *36*(3), 1426–1464. Retrieved from JSTOR.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 49–60. https://doi.org/10.1145/304182.304187

Anselin, L. (2003). Spatial Econometrics. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics* (pp. 310–330). https://doi.org/10.1002/9780470996249.ch15

Anselin, L. (2010). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, *20*(1), 1–17. https://doi.org/10.1111/j.1538-4632.1988.tb00159.x

Bae, C. J.-H., & Montello, D. R. (2018). Representations of an Urban Ethnic Neighbourhood: Residents' Cognitive Boundaries of Koreatown, Los Angeles. *Built Environment*, *44*(2), 218–240. https://doi.org/10.2148/benv.44.2.218

Bartlett, R. (2003). Testing the "Popsicle Test": Realities of Retail Shopping in New "Traditional Neighbourhood Developments." *Urban Studies*, *40*(8), 1471–1485. https://doi.org/10.1080/0042098032000094397

Bates, L. K. (2006). Does neighborhood really matter? Comparing historically defined neighborhood boundaries with housing submarkets. *Journal of Planning Education and Research*, *26*(1), 5–17.

Bay Area Open Space Council. (2017, July 13). Protected land in the Bay Area increased by 36,265 acres since 2013. Retrieved April 17, 2019, from Bay Area Open Space Council website: https://openspacecouncil.org/protected-land-bay-area-increased-36265-acres-since-2013/

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. In *MIT Press Series in Transportation Studies*: *Vol. 9*. Cambridge, MA: The MIT Press.

Bhat, C. R. (1996). A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity. *Transportation Research Part B: Methodological*, *30*(3), 189–207. https://doi.org/10.1016/0191-2615(95)00029-1

Bhat, C. R. (2000). A multi-level cross-classified model for discrete response variables. *Transportation Research Part B: Methodological*, *34*(7), 567–582. https://doi.org/10.1016/S0191-2615(99)00038-7

Bhat, C. R. (2005). A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological*, *39*(8), 679–707. https://doi.org/10.1016/j.trb.2004.08.003

Bhat, C. R., Goulias, K. G., Pendyala, R. M., Paleti, R., Sidharthan, R., Schmitt, L., & Hu, H.-H. (2013). A household-level activity pattern generation model with an application for Southern California. *Transportation*, *40*(5), 1063–1086. https://doi.org/10.1007/s11116-013-9452-y

Bhat, C. R., & Guo, J. (2004). A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transportation Research Part B: Methodological*, *38*(2), 147–168. https://doi.org/10.1016/S0191-2615(03)00005-5

Bhat, C., & Zhao, H. (2002). The spatial analysis of activity stop generation. *Transportation Research Part B: Methodological*, *36*(6), 557–575. https://doi.org/10.1016/S0191-2615(01)00019-4

Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R* (2nd ed.). In *Use R!* (2nd ed.). Retrieved from https://asdar-book.org/

Boarnet, M. G., & Wang, X. (2019). Urban spatial structure and the potential for vehicle miles traveled reduction: the effects of accessibility to jobs within and beyond employment sub-centers. *The Annals of Regional Science*, *62*(2), 381–404. https://doi.org/10.1007/s00168-019-00900-7

Boeing, G. (2017). *OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks*. *65*, 126–139. https://doi.org/10.1016/j.compenvurbsys.2017.05.004

Boeing, G. (2018a). A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood. *Environment and Planning B: Urban Analytics and City Science*, 2399808318784595. https://doi.org/10.1177/2399808318784595

Boeing, G. (2018b, April 5). Network-Based Spatial Clustering. Retrieved April 5, 2019, from Geoff Boeing website: https://geoffboeing.com/2018/04/network-based-spatial-clustering/

Boscoe, F. P., Henry, K. A., & Zdeb, M. S. (2012). A Nationwide Comparison of Driving Distance Versus Straight-Line Distance to Hospitals. *The Professional Geographer*, *64*(2), 188–196. https://doi.org/10.1080/00330124.2011.583586

Bottles, S. L. (1987). *Los Angeles and the Automobile: The Making of the Modern City*. University of California Press.

Calastri, C., Hess, S., Daly, A., & Carrasco, J. A. (2017). Does the social context help with understanding and predicting the choice of activity type and duration? An application of the Multiple Discrete-Continuous Nested Extreme Value model to activity diary data. *Transportation Research Part A: Policy and Practice*, *104*, 1–20. https://doi.org/10.1016/j.tra.2017.07.003

Campbell, E., Henly, J. R., Elliott, D. S., & Irwin, K. (2009). Subjective Constructions of Neighborhood Boundaries: Lessons from a Qualitative Study of Four Neighborhoods. *Journal of Urban Affairs*, *31*(4), 461–490. https://doi.org/10.1111/j.1467-9906.2009.00450.x

Cannuscio, C. C., Tappe, K., Hillier, A., Buttenheim, A., Karpyn, A., & Glanz, K. (2013). Urban Food Environments and Residents' Shopping Behaviors. *American Journal of Preventive Medicine*, *45*(5), 606–614. https://doi.org/10.1016/j.amepre.2013.06.021

Cao, X. (Jason), Mokhtarian, P. L., & Handy, S. L. (2009). Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings. *Transport Reviews*, *29*(3), 359–395. https://doi.org/10.1080/01441640802539195

Cao, X., Mokhtarian, P. L., & Handy, S. L. (2007). Do changes in neighborhood characteristics lead to changes in travel behavior? A structural equations modeling approach. *Transportation*, *34*(5), 535–556. https://doi.org/10.1007/s11116-007-9132-x

Carlsson, C. (2009, June 11). Revisiting the San Francisco Freeway Revolt. Retrieved April 18, 2019, from Streetsblog San Francisco website: https://sf.streetsblog.org/2009/06/11/revisiting-the-san-francisco-freeway-revolt/

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

Chen, X., & Clark, J. (2016). Measuring Space–Time Access to Food Retailers: A Case of Temporal Access Disparity in Franklin County, Ohio: The Professional Geographer: Vol 68, No 2. *The Professional Geographer*, *68*(2), 175–188.

Chen, Y., Ravulaparthy, S., Deutsch, K., Dalal, P., Yoon, S. Y., Lei, T., … Hu, H.-H. (2011a). Development of Indicators of Opportunity-Based Accessibility. *Transportation Research Record*, *2255*(1), 58–68. https://doi.org/10.3141/2255-07

Chen, Y., Ravulaparthy, S. K., Deutsch, K., Dalal, P., Yoon, S. Y., Lei, T., … Hu, H.-H. (2011b). Development of Indicators of Opportunity-Based Accessibility. *Transportation Research Record: Journal of the Transportation Research Board*, *2255*, 58–68. https://doi.org/10.3141/2255-07

Clapp, J. M., & Wang, Y. (2006). Defining neighborhood boundaries: Are census tracts obsolete? *Journal of Urban Economics*, *59*(2), 259–284. https://doi.org/10.1016/j.jue.2005.10.003

Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: models & applications*. London: Pion.

ConnectSF. (2018). *Appendix A: The History of Transportation in San Francisco*. Retrieved from https://connectsf.org/wp-content/uploads/ConnectSF-Vision-Report_Appendix-A_The-History-of-Transportation-in-SF.pdf

Couclelis, H. (1989). Macrostructure and Microbehavior in a Metropolitan Area. *Environment and Planning B: Planning and Design*, *16*(2), 141–154. https://doi.org/10.1068/b160141

Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping Residents' Perceptions of Neighborhood Boundaries: A Methodological Note. *American Journal of Community Psychology*, *29*(2), 371–383. https://doi.org/10.1023/A:1010303419034

Csardi, G., & Nepusz, T. (2006). *The igraph software package for complex network research* [C, R]. Retrieved from https://igraph.org/

Dailey, K. (2017, October 3). Cable Car History [Text]. Retrieved April 18, 2019, from San Francisco Municipal Transportation Agency website: https://www.sfmta.com/getting-around/muni/cable-cars/cable-car-history

Darley, W. K., & Lim, J. (1999). Effects of store image and attitude toward secondhand stores on shopping frequency and distance traveled. *International Journal of Retail & Distribution Management*, *27*(8), 311–318. https://doi.org/10.1108/09590559910288596

Davis, A. W. (2015). *Investigating Place Attitudes in Santa Barbara, CA*. University of California, Santa Barbara, Santa Barbara, CA.

de Abreu e Silva, J., Golob, T. F., & Goulias, K. G. (2006). Effects of Land Use Characteristics on Residence and Employment Location and Travel Behavior of Urban Adult Workers. *Transportation Research Record: Journal of the Transportation Research Board*, *1977*, 121–131.

Delling, D., Sanders, P., Schultes, D., & Wagner, D. (2009). Engineering Route Planning Algorithms. *Algorithmics of Large and Complex Networks. Lecture Notes in Computer Science*. Springer.

Deutsch, K. (2013). *An Investigation in Decision Making and Destination Choice Incorporating Place Meaning and Social Network Influences* (Doctoral Dissertation). University of California, Santa Barbara, Santa Barbara, CA.

Deutsch-Burgner, K., Ravualaparthy, S., & Goulias, K. (2014). Place happiness: its constituents and the influence of emotions and subjective importance on activity type and destination choice. *Transportation*, *41*(6), 1323–1340. https://doi.org/10.1007/s11116-014-9553-2

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, *1*(1), 269–271. https://doi.org/10.1007/BF01386390

Dill, J., Broach, J., Deutsch-Burgne, K., Xu, Y., Guensler, R., Levinson, D., & Tang, W. (2014). Multiday GPS Travel Behavior Data for Travel Analysis: The Effect of Day-to-Day Travel Time Variability on Auto Travel Choices. *Urban Studies and Planning Faculty Publications and Presentations*. Retrieved from https://pdxscholar.library.pdx.edu/usp_fac/137

Dong, G., & Harris, R. (2015). Spatial Autoregressive Models for Geographically Hierarchical Data Structures. *Geographical Analysis*, *47*(2), 173–191. https://doi.org/10.1111/gean.12049

Dong, G., & Harris, R. (2016). *HSAR: An R Package for Integrated Spatial Econometric and Multilevel Modelling*. Presented at the GIS Research UK 2016, Greenwich, UK.

Ersoy, F. Y., Hasker, K., & Inci, E. (2016). Parking as a loss leader at shopping malls. *Transportation Research Part B: Methodological*, *91*, 98–112. https://doi.org/10.1016/j.trb.2016.04.016

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. 226–231. AAAI Press.

Fan, Y., & Khattak, A. (2008). Urban Form, Individual Spatial Footprints, and Travel: Examination of Space-Use Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *2082*, 98–106. https://doi.org/10.3141/2082-12

Frank, L., Bradley, M., Kavage, S., Chapman, J., & Lawton, T. K. (2008). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation*, *35*(1), 37–54. https://doi.org/10.1007/s11116-007-9136-6

Friedmann, J., & Miller, J. (1965). The Urban Field. *Journal of the American Institute of Planners*, *31*(4), 312–320. https://doi.org/10.1080/01944366508978185

Garling, T., Gillholm, R., Romanus, J., & Selart, M. (1997). Interdependent Activity and Travel Choices: Behavioural Principles of Integration of Choice Outomes. In D. F. Ettema & H. J. P. Timmermans (Eds.), *Activity-Based Approaches to Travel Analysis*. Retrieved from https://papers.ssrn.com/abstract=2649173

Gim, T.-H. T. (2012). A meta-analysis of the relationship between density and travel behavior. Transportation, 39(3), 491–519. https://doi.org/10.1007/s11116-011-9373-6

Giuliano, G., & Small, K. A. (1991). Subcenters in the Los Angeles region. *Regional Science and Urban Economics*, *21*(2), 163–182. https://doi.org/10.1016/0166-0462(91)90032-I

Godfrey, B. J. (1999). The Geography of James E. Vance Jr. (1925-1999). *Geographical Review*, *89*(4), 580–589. Retrieved from JSTOR.

Golledge, R. G., & Stimson, R. J. (1997). *Spatial Behavior: A Geographic Perspective*. New York: The Guilford Press.

Gordon, P., Richardson, H. W., & Wong, H. L. (1986). The Distribution of Population and Employment in a Polycentric City: The Case of Los Angeles. *Environment and Planning A: Economy and Space*, *18*(2), 161–173. https://doi.org/10.1068/a180161

Goulias, K. (2009, December). *Travel Behavior Dynamics from a Lifespan Development Perspective*. 13–18. Jaipur, India.

Goulias, K. G., Bhat, C. R., Pendyala, R. M., Chen, Y., Paleti, R., Konduri, K. C., … Hu, H. (2011). Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California: Design, implementation, preliminary findings, and integration plans. *2011 IEEE Forum on Integrated and Sustainable Transportation Systems*, 164–169. https://doi.org/10.1109/FISTS.2011.5973624

Grant, J., & Perrott, K. (2011). Where Is the Café? The Challenge of Making Retail Uses Viable in Mixed-use Suburban Developments. *Urban Studies*, *48*(1), 177–195. https://doi.org/10.1177/0042098009360232

Grebitus, C., Lusk, J. L., & Nayga, R. M. (2013). Effect of distance of transportation on willingness to pay for food. *Ecological Economics*, *88*, 67–75. https://doi.org/10.1016/j.ecolecon.2013.01.006

Greene, W. H., & Hensher, D. A. (2003). A latent class model for discrete choice analysis: contrasts with mixed logit. *Transportation Research Part B: Methodological*, *37*(8), 681–698. https://doi.org/10.1016/S0191-2615(02)00046-2

Hägerstrand, T. (1970). What About People in Regional Science? *Papers in Regional Science*, *24*(1), 7–24. https://doi.org/10.1111/j.1435-5597.1970.tb01464.x

Hahsler, M., Piekenbrock, M., Arya, S., & Mount, D. (2018). dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms (Version 1.1-2). Retrieved from https://CRAN.R-project.org/package=dbscan

Handy, S. (1993). *Regional Versus Local Accessibility: Implications for Nonwork Travel*. Retrieved from https://escholarship.org/uc/item/2z79q67d

Handy, S. (1996). Methodologies for exploring the link between urban form and travel behavior. *Transportation Research Part D: Transport and Environment*, *1*(2), 151–165. https://doi.org/10.1016/S1361-9209(96)00010-7

Handy, S. L., & Niemeier, D. A. (1997). Measuring Accessibility: An Exploration of Issues and Alternatives. *Environment and Planning A: Economy and Space*, *29*(7), 1175–1194. https://doi.org/10.1068/a291175

Hansen, W. G. (1959). How Accessibility Shapes Land Use. *Journal of the American Institute of Planners*, *25*(2), 73–76. https://doi.org/10.1080/01944365908978307

Hartman, G. W. (1950). The Central Business District--A Study in Urban Geography. *Economic Geography*, *26*(4), 237–244. https://doi.org/10.2307/141260

Harvey, D. (2003). The right to the city. *International Journal of Urban and Regional Research*, *27*(4), 939–941. https://doi.org/10.1111/j.0309-1317.2003.00492.x

Hasnat, M. M., & Hasan, S. (2018). Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies*, *96*, 38–54. https://doi.org/10.1016/j.trc.2018.09.006

Helsley, R. W., & Sullivan, A. M. (1991). Urban subcenter formation. *Regional Science and Urban Economics*, *21*(2), 255–275. https://doi.org/10.1016/0166-0462(91)90036-M

Hu, Y., Gao, S., Janowicz, K., Yu, B., Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, *54*, 240–254. https://doi.org/10.1016/j.compenvurbsys.2015.09.001

igraph core team. (2015). igraph Reference Manual. Retrieved April 9, 2019, from Igraph.org website: https://igraph.org/c/doc/igraph-Structural.html

Karami, A., & Johansson, R. (2014). Choosing DBSCAN Parameters Automatically using Differential Evolution. *International Journal of Computer Applications*, *91*(7).

Khattak, A. J., & Rodriguez, D. (2005). Travel behavior in neo-traditional neighborhood developments: A case study in USA. *Transportation Research Part A: Policy and Practice*, *39*(6), 481–500. https://doi.org/10.1016/j.tra.2005.02.009

King, J. (2014). Loma Prieta quake left legacy of repair, renewal. *San Francisco Chronicle*. Retrieved from https://www.sfgate.com/bayarea/article/Loma-Prieta-quake-left-legacy-of-repair-renewal-5816995.php

Kitamura, R., Mokhtarian, P. L., & Laidet, L. (1997). A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation*, *24*(2), 125–158. https://doi.org/10.1023/A:1017959825565

Knoblauch, R. L., Pietrucha, M. T., & Nitzburg, M. (1996). Field Studies of Pedestrian Walking Speed and Start-Up Time. *Transportation Research Record*, *1538*(1), 27–38. https://doi.org/10.1177/0361198196153800104

Kolko, J. (2010). Urbanization, Agglomeration, and Coagglomeration of Service Industries. In E. L. Glaeser (Ed.), *Agglomeration Economics* (pp. 151–180). Retrieved from http://www.nber.org/chapters/c7983.pdf

Krim, A. (1992). Los Angeles and the anti-tradition of the suburban city. *Journal of Historical Geography*, *18*(1), 121–138. https://doi.org/10.1016/0305-7488(92)90280-M

Krizek, K. J. (2003). Residential Relocation and Changes in Urban Travel: Does Neighborhood-Scale Urban Form Matter? *Journal of the American Planning Association*, *69*(3), 265–281. https://doi.org/10.1080/01944360308978019

LA Times Data Desk. (2018). Mapping L.A. - Neighborhoods. Retrieved April 15, 2019, from Los Angeles Times website: http://maps.latimes.com/neighborhoods/

Lai, C. (2012). The Racial Triangulation of Space: The Case of Urban Renewal in San Francisco's Fillmore District. *Annals of the Association of American Geographers*, *102*(1), 151–170. https://doi.org/10.1080/00045608.2011.583572

Lee, J. H., Davis, A. W., Yoon, S. Y., & Goulias, K. G. (2016). Activity space estimation with longitudinal observations of social media data. *Transportation*, *43*(6), 955–977. https://doi.org/10.1007/s11116-016-9719-1

Lee, J. H., Davis, A.W., Yoon, S. Y., & Goulias, K. G. (2017). Exploring Daily Rhythms of Interpersonal Contacts: Time of Day Dynamics of Human Interactions Using Latent Class Cluster Analysis. *Transportation Research Record: Journal of the Transportation Research Board*, *2666*, 58–68. https://doi.org/10.3141/2666-07

Lee, J. H., & Goulias, K. G. (2015, January). *Examining Differences and Commonalities of Life Cycle Stages in Daily Contacts and Activity-Travel Time Allocation*. Presented at the 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

Levinson, D. M. (1998). Accessibility and the journey to work. *Journal of Transport Geography*, *6*(1), 11–21. https://doi.org/10.1016/S0966-6923(97)00036-7

Levinson, D., Marion, B., Owen, A., & Cui, M. (2017). The City is flatter: Changing patterns of job and labor access. *Cities*, *60*, 124–138. https://doi.org/10.1016/j.cities.2016.08.002

Loukaitou-Sideris, A. (2002). Regeneration of Urban Commercial Strips: Ethnicity and Space in Three Los Angeles Neighborhoods. *Journal of Architectural and Planning Research*, *19*(4), 334–350. Retrieved from JSTOR.

Lukasiewicz, T., & Straccia, U. (2008). Managing uncertainty and vagueness in description logics for the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 291–308. https://doi.org/10.1016/j.websem.2008.04.001

Maddimsetty, R. P. (2018, March 19). You Are (Probably) Here: Better Map Pins with DBSCAN & Random Forests. Retrieved July 27, 2018, from foursquare-eng website: https://engineering.foursquare.com/you-are-probably-here-better-map-pins-with-dbscan-random-forests-9d51e8c1964d

Mai, G., Janowicz, K., Hu, Y., & Gao, S. (2018). ADCN: An anisotropic density-based clustering algorithm for discovering spatial point patterns with noise. *Transactions in GIS*, *22*(1), 348–369. https://doi.org/10.1111/tgis.12313

McBride, Elizabeth C., Davis, A. W., Lee, J. H., & Goulias, K. G. (2017). Incorporating Land Use into Methods of Synthetic Population Generation and of Transfer of Behavioral Data. *Transportation Research Record*, *2668*(1), 11–20. https://doi.org/10.3141/2668-02

McBride, Elizabeth C., Davis, A. W., & Goulias, K. G. (2019). Fragmentation in Daily Schedule of Activities using Activity Sequences. *Transportation Research Record*, 0361198119837501. https://doi.org/10.1177/0361198119837501

McKenzie, G., & Janowicz, K. (2015). Where is also about time: A location-distortion model to improve reverse geocoding using behavior-driven temporal semantic signatures. *Computers, Environment and Urban Systems*, *54*, 1–13. https://doi.org/10.1016/j.compenvurbsys.2015.05.003

McKenzie, G., Janowicz, K., Gao, S., & Gong, L. (2015). How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, *54*, 336–346. https://doi.org/10.1016/j.compenvurbsys.2015.10.002

Moeckel, R., & Donnelly, R. (2015). Gradual rasterization: redefining spatial resolution in transport modelling, Gradual rasterization: redefining spatial resolution in transport modelling. *Environment and Planning B: Planning and Design*, *42*(5), 888–903. https://doi.org/10.1068/b130199p

Montello, D. R., Friedman, A., & Phillips, D. W. (2014). Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, *28*(9), 1802–1820. https://doi.org/10.1080/13658816.2014.900178

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, *37*(1/2), 17–23. https://doi.org/10.2307/2332142

Muller, P. O. (2004). Transportation and Urban Form: Stages in the Spatial Evolution of the American Metropolis. In S. Hanson & G. Giuliano (Eds.), *The Geography of Urban Transportation* (3rd ed., pp. 59–85). New York: Guilford Press.

Murphy, R. E. (2017). *The Central Business District: A Study in Urban Geography*. https://doi.org/10.4324/9781315131153

Nerella, S., & Bhat, C. R. (2004). Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models. *Transportation Research Record*, *1894*(1), 11–19. https://doi.org/10.3141/1894-02

Neutens, T., Schwanen, T., & Miller, H. J. (2010). Dealing with Timing and Synchronization in Opportunities for Joint Activity Participation. *Geographical Analysis*, *42*(3), 245–266. https://doi.org/10.1111/j.1538-4632.2010.00792.x

NUSTATS. (2013). *2010-2012 California Household Travel Survey Final Report* (p. 137). Retrieved from California Department of Transportation website: http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/Files/CHTS_Final_Report_June_2013.pdf

Okabe, A., Okunuki, K., & Shiode, S. (2006). SANET: A Toolbox for Spatial Analysis on a Network. *Geographical Analysis*, *38*(1), 57–66. https://doi.org/10.1111/j.0016-7363.2005.00674.x

OneBayArea. (2013, July 18). *Plan Bay Area: Strategy for a Sustainable Region*. Retrieved from http://files.mtc.ca.gov.s3.amazonaws.com/pdf/Plan_Bay_Area_FINAL/pbafinal/index.html

OpenStreetMap. (2019). Retrieved May 7, 2019, from OpenStreetMap website: https://www.openstreetmap.org/about

Orenstein, D. E., Frenkel, A., & Jahshan, F. (2014). Methodology Matters: Measuring Urban Spatial Development Using Alternative Methods , Methodology Matters: Measuring Urban Spatial Development Using Alternative Methods. *Environment and Planning B: Planning and Design*, *41*(1), 3–23. https://doi.org/10.1068/b38017

Páez, A., & Scott, D. M. (2004). Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal*, *61*(1), 53–67. https://doi.org/10.1007/s10708-005-0877-5

Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, *25*, 141–153. https://doi.org/10.1016/j.jtrangeo.2012.03.016

Parlette, V., & Cowen, D. (2011). Dead Malls: Suburban Activism, Local Spaces, Global Logistics. *International Journal of Urban and Regional Research*, *35*(4), 794–811. https://doi.org/10.1111/j.1468-2427.2010.00992.x

Patterson, Z., & Farber, S. (2015). Potential Path Areas and Activity Spaces in Application: A Review. *Transport Reviews*, *35*(6), 679–700. https://doi.org/10.1080/01441647.2015.1042944

Paul, B. M., Vovsha, P. S., Hicks, J. E., Livshits, V., & Pendyala, R. M. (2014). Extension of Activity-Based Modeling Approach to Incorporate Supply Side of Activities, Extension of Activity-Based Modeling Approach to Incorporate Supply Side of Activities: Examples for Major Universities and Special Events, Examples for Major Universities and Special Events. *Transportation Research Record*, *2429*(1), 138–147. https://doi.org/10.3141/2429-15

Perez, J. (2017, September 29). The Los Angeles Freeway and the History of Community Displacement. Retrieved April 18, 2019, from The Toro Historical Review website: https://thetorohistoricalreview.org/2017/09/29/the-los-angeles-freeway-and-the-history-of-community-displacement/

Rasouli, S., & Timmermans, H. (2014). Activity-based models of travel demand: promises, progress and prospects. *International Journal of Urban Sciences*, *18*(1), 31–60. https://doi.org/10.1080/12265934.2013.835118

Ravualaparthy, S. (2013). *Spatial Perspectives in Business Establishment Behavioral Modeling: A Case-Study Analysis in Santa Barbara County* (Doctoral Dissertation). University of California, Santa Barbara, Santa Barbara, CA.

Ravulaparthy, S. K., & Goulias, K. (2014). Characterizing the Composition of Economic Activities in Central Locations. *Transportation Research Record: Journal of the Transportation Research Board*, *2430*, 95–104. https://doi.org/10.3141/2430-10

Ravulaparthy, S. K., Goulias, K. G., Sweeney, S., & Kyriakidis, P. C. (2013, February 24). *Exploring the Spatial and Temporal Patterns of Business Concentration and Dispersion: A Case-study Analysis for County of Santa Barbara*. Presented at the 52nd Annual Meeting of Western Regional Science Association, Santa Barbara, CA.

San Francisco Association of Realtors. (2017, December). San Francisco Neighborhood Map. Retrieved April 16, 2019, from Bay Area Market Reports website: https://www.bayareamarketreports.com/trend/san-francisco-neighborhood-map

Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194. https://doi.org/10.1023/A:1009745219419

Schlich, R., & Axhausen, K. W. (2003). Habitual travel behaviour: Evidence from a six-week travel diary. *Transportation*, *30*(1), 13–36. https://doi.org/10.1023/A:1021230507071

Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems (TODS)*, *42*(3). Retrieved from http://delivery.acm.org/10.1145/3070000/3068335/a19-schubert.pdf?ip=169.231.189.174&id=3068335&acc=AUTHOR%2DIZED&key=CA367851C7E3CE77%2E022A0CC51A76093F%2E4D4702B0C3E38B35%2E78C63DFD53B41BF7&__acm__=1532824342_0ff95e6b63ba31bfaad2b748d010f852

Schwanen, T. (2004). The determinants of shopping duration on workdays in The Netherlands. *Journal of Transport Geography*, *12*(1), 35–48. https://doi.org/10.1016/S0966-6923(03)00023-1

Schwartz, N. D. (2015, January 3). The Economics (and Nostalgia) of Dead Malls. *The New York Times*. Retrieved from https://www.nytimes.com/2015/01/04/business/the-economics-and-nostalgia-of-dead-malls.html

Shliselberg, R. (2015). Accessibility and Spatial Interaction. *Transport Reviews*, *35*(6), 814–816. https://doi.org/10.1080/01441647.2015.1058302

Singley, P. (2013). Los Angeles: Between Cognitive Mapping and Dirty Realism. In R. El-Khoury & E. Robbins (Eds.), *Shaping the City: Studies in History, Theory & Urban Design* (2nd ed., pp. 98–134). London: Routledge.

Sloane, D. C. (2003). Medicine in the (Mini) mall: An American health care landscape. In C. Wilson & P. Groth (Eds.), *Everyday America. Cultural landscape studies after JB Jackson* (pp. 293–308). Berkeley: University of California Press.

Smith, N. (1979). Toward a Theory of Gentrification A Back to the City Movement by Capital, not People. *Journal of the American Planning Association*, *45*(4), 538–548. https://doi.org/10.1080/01944367908977002

Smith, P. J. (1962). Calgary: A Study in Urban Pattern. *Economic Geography*, *38*(4), 315. https://doi.org/10.2307/142261

Steinberg, D. Sustainable Communities and Climate Protection Act of 2008. , Pub. L. No. Senate Bill No. 375, California Health and Safety Code (2008).

Suisman, D. R. (2014). *Los Angeles Boulevard: Eight X-Rays of the Body Public* (25h Anniversary). ORO Editions.

Swait, J. (2001). Choice set generation within the generalized extreme value family of discrete choice models. *Transportation Research Part B: Methodological*, *35*(7), 643–666. https://doi.org/10.1016/S0191-2615(00)00029-1

Teaford, J. C. (2000). Urban Renewal and Its Aftermath. *Housing Policy Debate*, *11*(2), 443–465. https://doi.org/10.1080/10511482.2000.9521373

Thomas, J. M., Ritzdorf, M., & Hodne, C. (1997). Urban Planning and the African American Community. *Psyccritiques*, *42*(12). https://doi.org/10.1037/000678

Train, K. E. (2009). *Discrete Choice Methods with Simulation* (Second Edition). New York: Cambridge University Press.

United States Office of Management and Budget. (2017). *North American Industrial Classification System 2017 Manual*. Retrieved from https://www.census.gov/eos/www/naics/2017NAICS/2017_NAICS_Manual.pdf

US Census Bureau. (2010). 2010 Geographic Terms and Concepts - Census Tract. Retrieved July 26, 2018, from https://www.census.gov/geo/reference/gtc/gtc_ct.html

US EPA. (2019). Inventory of U.S. Greenhouse Gas Emissions and Sinks 1990-2017 (No. EPA 430-R-19-001). Retrieved from United States Environmental Protection Agency website: https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks

Vias, A. C. (2004). Bigger stores, more stores, or no stores: paths of retail restructuring in rural America. *Journal of Rural Studies*, *20*(3), 303–318. https://doi.org/10.1016/j.jrurstud.2003.10.003

Viegas, J. M., Martinez, L. M., & Silva, E. A. (2009). Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones, Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones. *Environment and Planning B: Planning and Design*, *36*(4), 625–643. https://doi.org/10.1068/b34033

Vovsha, P., Bradley, M., & Bowman, J. L. (2005). *Activity-Based Travel Forecasting Models in the United States: Progress since 1995 and Prospects for the Future*. Retrieved from https://trid.trb.org/view/759300

Wachs, M. (1984). Autos, Transit, and the Sprawl of Los Angeles: The 1920s. *Journal of the American Planning Association*, *50*(3), 297–310. https://doi.org/10.1080/01944368408976597

Walls & Associates. (2013). National Establishment Time-Series (NETS) Database: 2012 Database Description. Retrieved July 21, 2017, from http://exceptionalgrowth.org/downloads/NETSDatabaseDescription2013.pdf

Weiss, L., Ompad, D., Galea, S., & Vlahov, D. (2007). Defining Neighborhood Boundaries for Urban Health Research. *American Journal of Preventive Medicine*, *32*(6, Supplement), S154–S159. https://doi.org/10.1016/j.amepre.2007.02.034

White, M. J. (1976). Firm suburbanization and urban subcenters. *Journal of Urban Economics*, *3*(4), 323–343. https://doi.org/10.1016/0094-1190(76)90033-4

Wilson, S. G., Plane, D. A., Mackun, P. J., Fischetti, T. R., Goworowska, J., Perry, M. J., & Hatchard, G. W. (2012). *Patterns of Metropolitan and Micropolitan Population Change: 2000 to 2010* (No. C2010SR-01; p. 102). Retrieved from United States Census Bureau website: https://www.census.gov/prod/cen2010/reports/c2010sr-01.pdf

Woodruff, A. (2013, June 27). Neighborhoods as seen by the people [Bostonography]. Retrieved April 16, 2019, from https://bostonography.com/2013/neighborhoods-as-seen-by-the-people/

Xu, H. (2014). Comparing Spatial and Multilevel Regression Models for Binary Outcomes in Neighborhood Studies. *Sociological Methodology*, *44*(1), 229–272. https://doi.org/10.1177/0081175013490188

Yamada, I., & Thill, J.-C. (2004). Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography*, *12*(2), 149–158. https://doi.org/10.1016/j.jtrangeo.2003.10.006

Yiu, M. L., & Mamoulis, N. (2004). Clustering Objects on a Spatial Network. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 443–454. https://doi.org/10.1145/1007568.1007619

Yoon, S., & Goulias, K. (2010). Impact of time-space prism accessibility on time use behavior and its propagation through intra-household interaction. *Transportation Letters*, *2*(4), 245–260. https://doi.org/10.3328/TL.2010.02.04.245-260

Zeng, W., & Church, R. L. (2009). Finding shortest paths on real road networks: the case for A*. *International Journal of Geographical Information Science*, *23*(4), 531–543. https://doi.org/10.1080/13658810801949850

Zhang, Y., Han, L. D., & Kim, H. (2018). Dijkstra's-DBSCAN: Fast, Accurate, and Routable Density Based Clustering of Traffic Incidents on Large Road Network. *Transportation Research Record*, *2672*(45), 265–273. https://doi.org/10.1177/0361198118796071