A New Era of Spatial Interaction:

Potential and Pitfalls

by

Taylor Matthew Oshan

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2017 by the
Graduate Supervisory Committee:

A. Stewart Fotheringham, Chair
Sergio S. J. Rey
Carson J. Q. Farmer
Trisalyn Nelson

ARIZONA STATE UNIVERSITY

December 2017

ABSTRACT

As urban populations become increasingly dense, massive amounts of new 'big' data that characterize human activity are being made available and may be characterized as having a large volume of observations, being produced in real-time or near real-time, and including a diverse variety of information. In particular, spatial interaction (SI) data – a collection of human interactions across a set of origins and destination locations – present unique challenges for distilling big data into insight. Therefore, this dissertation identifies some of the potential and pitfalls associated with new sources of big SI data. It also evaluates methods for modeling SI to investigate the relationships that drive SI processes in order to focus on human behavior rather than data description.

A critical review of the existing SI modeling paradigms is first presented, which also highlights features of big data that are particular to SI data. Next, a simulation experiment is carried out to evaluate three different statistical modeling frameworks for SI data that are supported by different underlying conceptual frameworks. Then, two approaches are taken to identify the potential and pitfalls associated with two newer sources of data from New York City – bike-share cycling trips and taxi trips. The first approach builds a model of commuting behavior using a traditional census data set and then compares the results for the same model when it is applied to these newer data sources. The second approach examines how the increased temporal resolution of big SI data may be incorporated into SI models.

Several important results are obtained through this research. First, it is demonstrated that different SI models account for different types of spatial effects and that the Competing Destination framework seems to be the most robust for capturing spatial structure effects. Second, newer sources of big SI data are shown to be very

useful for complimenting traditional sources of data, though they are not sufficient substitutions. Finally, it is demonstrated that the increased temporal resolution of new data sources may usher in a new era of SI modeling that allows us to better understand the dynamics of human behavior.

To my family: Mom, Dad, Kayley, Taryn, and Reid. Without their support I would not have taken the very unlikely path that has led me this work and the adventures and experiences required to complete it.

ACKNOWLEDGMENTS

Many thanks are in order. First and foremost, I would like to thank my entire committee for their support and feedback that helped to improve this work. I am particularly grateful to Dr. Stewart Fotheringham, my committee chair and doctoral supervisor, for his guidance and encouragement throughout the entire process of completing this research. I am also very grateful to Dr. Sergio Rey and Dr. Carson Farmer for their mentoring and support, especially in the pursuit of software development, without which this work would not be possible.

Next, I would like to thank my fellow doctoral students in the department. Without them, completing this research would have been exceptionally lonely! I am exceedingly appreciative of the many great conversations I had with Levi Wolf about work, life, and the more trivial minutiae of the graduate student experience.

A special thanks is in order to my girlfriend, Ramona Belfiore, for being indescribably supportive of me while I was writing this dissertation. She encouraged me until the very last word, when it counted most.

Finally, I am indebted to my family for supporting me throughout my graduate education. Indeed, it took some unexpected turns, but no matter where it took me, they were always waiting give me a warm welcome home. Spending time with them kept me centered and it was invaluable to know that they were always in my corner rooting for me.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 General overview

The majority of the global population now resides in cities (United Nations and Department of Economic and Social Affairs, 2014). It is not surprising then, that in the 'big data' era, massive amounts of data that characterize how people meet their economic needs, interact within social communities, and utilize shared resources such as transportation infrastructure are being made available through advances in technology and a sea-change in attitudes towards making data public. Understanding the drivers of urban dynamics can inform policies for increasing the resilience of cities and reduce the negative side-effects of urbanization (Batty, 2013). Therefore, it is crucial to harness the ever-increasing streams of data being produced by cities and their digitally equipped infrastructure. However, before these sources can be turned into actionable knowledge, there are challenges that must be overcome.

In the age of big data there has been an abundance of new datasets that may be characterized as having a large volume of observations, being produced in real-time or near real-time, and including a diverse variety of information (Kitchin, 2014b; Lovelace *et al.*, 2015). While big data have been hyped as the solution to many complex problems, many studies leveraging big data are often descriptive in nature and therefore fall short of answering interesting research questions that illuminate the processes that generate the data we observe. That is, big data do not necessarily imply an increase in knowledge or understanding. In order to ascribe meaning to the

torrents of data that are collected daily, they must first be tested against theories so that patterns within the data can be associated with social processes (Kitchin, 2014a). Adopting a model-driven approach to geospatial analysis solves an inherent problem of big data: their inability to speak for themselves. By building assumptions and expert opinions into the structure of a model, we can begin to solve this problem. The insights that result from applying data to an established model can confirm or reject hypotheses, rather than solely extracting patterns from data. Pure pattern detection is problematic because it is possible to observe spurious relationships amongst completely unrelated data. The process of developing and critically evaluating models that can accurately capture a given phenomenon is key to generating genuine insights (Farmer and Pozdnoukhov, 2012), and is a focus of this research agenda.

Human activities that require traversing physical space, such as commuting, shopping, dining, or socializing, are drivers of urban dynamics. Accurate descriptions of these dynamics require precise measures of spatial contexts that propel and attract potential movers, as well as an understanding of the costs of interacting within complex social and physical environments. Spatial interaction models, which seek to explain and predict aggregate movement patterns, are a set of tools that have a long history of being employed by academics, and professionals (Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989; Sen and Smith, 1995; Roy, 2004). Traditionally, these models analyze data that represent movements between a set of origins and destinations, which are represented by large areal units (i,e,. states, counties, cities), and have been collected over extended periods of time (decades, years and months). The core assumption of spatial interaction models is that the magnitude of movements between two locations will increase as a function of their attractiveness, but that it will decrease as the physical separation (i.e., distance or time) between locations

increases. As such, spatial interaction models, have been used to study migration, transportation, residential mobility, retailing, attendance at events and universities, patronage of medical facilities, and economic interactions around the world, utilizing many different variables as proxies for size and separation.

While spatial interaction has a long history, it was not until the mid-20th century that the processes underlying spatial interaction became of widespread interest to regional scientists and geographers. Following a relative 'trough' in spatial interaction research over the past few decades, there has been a renewed interest in human movement under the banner of 'human mobility'. This is primarily due to the widespread availability of spatially and temporally disaggregate mobility datasets from sources such as automated transportation systems, mobile phone records, GPS trajectories, and social media, often described under the umbrella of big data (Arribas-Bel, 2014), as well as the availability of increased computational power to handle these new data. However, this new thrust of research has moved away from trying to understand processes and tends to focus on predicting the movement of individuals (Song *et al.*, 2010b; Lin *et al.*, 2013; Pirozmand *et al.*, 2014; Do and Gatica-Perez, 2014) or establishing regularities (Brockmann *et al.*, 2006; González *et al.*, 2008; Han *et al.*, 2009; Bazzani *et al.*, 2010; Song *et al.*, 2010a; Liang *et al.*, 2012; Wang *et al.*, 2014). A result of this trend has been the promulgation of so-called 'universal' spatial interaction models (Lenormand *et al.*, 2012; Simini *et al.*, 2012; Yan *et al.*, 2013) that predict movement using non-parametric specifications that do not allow for model building or hypothesis testing. In contrast, parametric spatial interaction models produce parameters that may be interpreted as the strength and nature of the attributes of a place that generate its attractiveness against the costs that must be

overcome to travel to that place. Therefore, they are a key tool for *understanding* the underlying decision-making processes that generate spatial interaction data.

Despite the widespread applicability of parametric spatial interaction models, there does not yet exist a methodology that exploits the real-time qualities of emerging datasets; one where the temporal dimension of a movement phenomenon, such as commuting, is considered for increasingly finer time periods. Furthermore, existing models generally seek to explain and predict the number of movements that occur based on generalized locational attributes (i,e,. average population), rather than specific indicators, such as points-of-interest or subway station usage that can also describe destinations. Parametric spatial interaction models may thus be improved by leveraging the richness of big data to provide new insights into the dynamic mechanisms that facilitate human movement within cities. However, it remains largely unknown whether or not these new data sources have limitations compared to more traditional spatial interaction data (i.e., the decennial census). Therefore, exploring big spatial interaction data and incorporating them into the destination choice modeling framework is a crucial task that is necessary for modernizing the geographical sciences toolkit, especially for studying and governing urban areas, which are increasing in number, density, and importance (United Nations and Department of Economic and Social Affairs, 2014). Furthermore, the destination choice framework that is popular within the spatial choice literature is generalized to consider the case of origin choice and is therefore referred to as location choice. Knowledge of location choice processes is important for policy development, which can be useful on its own, as well as a factor within other regional models, such as land-use/land-change, market analysis, or location-allocation.

A key inquiry in parametric spatial interaction modeling is the nature of distance-

decay – how space limits human activity – and how this effect may change for different types of flows and for different regions. In particular, spatial structure, which refers to the organization of locations in space, has been an on-going interest in spatial interaction models for 50 years. Debate about the effects of spatial structure in spatial interaction models crescendoed in the late 70's and early 80's with the goal of determining whether or not distance-decay measurements were strictly behavioral or were influenced by spatial structure. A proposed solution was the competing destination model (Fotheringham, 1983a), which considers several aspects of individual behavior in relation to spatial structure and spatial information processing. Over the last few decades, spatial structure has remained an important topic, and several additional techniques from the econometrics (LeSage and Pace, 2008) and spatial statistics (Griffith, 2007) paradigms have been proposed that harness increased computational power to use more complex statistical methods to account for spatial structure within spatial interaction models. While all these parametric spatial interaction models draw on a common foundation, they are distinct in terms of methodology and theory, with newer techniques shifting the focus away from distance-decay in favor of the concept of spatial autocorrelation.

It is clear then that the spatial interaction landscape has become increasingly diverse in terms of available model specifications, data, and underlying theories for investigating location choice. Therefore, our understanding of spatial interaction and spatial decision-making processes may be enhanced by a) making connections between similarities of different spatial interaction model specifications; b) clarifying the differences between spatial interaction specifications; and c) investigating the potentials and pitfalls of new sources of big spatial interaction data in the context of the breadth of available modeling specifications. Following these themes, several

important research questions will be pursued in this manuscript. First, "can we detect useful spatial-temporal dynamics in spatial interaction data that can be used to enhance spatial interaction models? If so, are these dynamics sensitive to certain spatial and temporal resolutions?" Second, "what are the advantages and disadvantages to different spatial interaction model specifications?" and "does each specification adequately define and capture spatial structure?" Finally, "what are some of the drivers of urban location choice?"

1.2 Research objectives

This research evaluates a variety of spatial interaction model specifications in light of newly available data sources in the context of urban location choice. Considering the avalanche of new spatial interaction data sources with increasingly finer spatial and temporal resolutions, the main goal of this dissertation will be to investigate how the added detail can be used to better predict and understand spatial interaction processes. To that end, this research will seek to fulfill three primary goals: 1) compare and contrast parametric spatial interaction model specifications that account for spatial structure; 2) build parametric spatial interaction models using two newer sources of urban transportation data – bike-share cycling trips and taxi trips in New York City (NYC) – and new measures of urban structure that may be useful indicators of location attractiveness; and 3) explore the spatial and temporal dynamics in spatial interaction data. These three primary goals may each be further broken down into several objectives that will be carried out to in order achieve the larger goals. The individual objectives are as follows:

1. Compare and contrast parametric spatial interaction model specifications

- Review parametric spatial interaction model specifications
- Evaluate the similarity and robustness of specifications using simulations

2. Build parametric spatial interaction models of bike trips and taxi trips in New York City

    - Calibrate production-constrained models of location choice using different specifications
    - Calibrate origin-specific production-constrained models to investigate spatial non-stationarity

3. Explore the spatial and temporal dynamics in spatial interaction data

    - Review the nature of spatial interaction data and characterize existing applications
    - Calibrate temporal subset production-constrained models to investigate temporal non-stationarity

1.3 Thesis structure

These goals and objectives will be addressed in the subsequent eight chapters of this dissertation. A summary of each chapter is as follows:

Chapter 2: Spatial interaction models An in depth review of spatial interaction models is provided. This includes both parametric and non-parametric spatial interaction models. A large focus throughout this chapter is on the development of the idea of spatial structure in spatial interaction, methods to account for it, and how it effects distance-decay. Finally, several metrics for assessing spatial interaction models are highlighted.

Chapter 3: Spatial interaction in the era of big data In this chapter, the concept of big data is defined in the context of spatial interaction and some problems associated with it are highlighted. Next, it summarizes exploratory analysis methods and applications for spatial interaction data. Lastly, it reviews research related to bike-share cycling trips and taxi trips, especially in the context of spatial interaction and within New York City.

Chapter 4: Spatial interaction in New York City Here, the data sources that will be used for analysis are presented. This includes basic spatial and temporal patterns of bike-share trips and taxi trips in New York City, as well as location attributes. These attributes include traditional census variables and non-traditional variables made available by the city or one of its governing agencies.

Chapter 5: A simulation-based investigation of spatial structure Several simulation experiments are carried out in this chapter to compare and contrast parametric model specifications that account for spatial structure. First, data will be simulated using the data-generating processes for each model (including a null model that is free from spatial structure effects) and then each model will be calibrated on all of the synthetic datasets. The purpose is to explore the extent that the models may capture similar effects. Second, the data will be aggregated to larger spatial units to assess how robust each specification is to measurement error induced from the aggregation process.

Chapter 6: Local models of location choice in New York City In this chapter several attraction-constrained models of location choice will be calibrated on the bike and taxi data. Then these models will be extended to localized origin-specific models so that the parameters can be visually analyzed for spatial non-stationarity.

Chapter 7: Temporal subset models of location choice in New York City Production-constrained models of location choice are first established for the bike and taxi data. The models will then be calibrated on increasingly finer temporal subsets to explore if the parameter estimates are still reliable and how urban behavior changes over time.

Chapter 8: Discussion and Conclusions Finally, the previous seven chapters will be discussed in terms of the major findings and implications for the different themes within this dissertation. Limitations and future work will also be addressed.

1.4 Moving forward

The previous sections introduced the work that will be carried out in this dissertation, including an overall motivation, the specific research goals, and a sketch of the structure of the chapters ahead. In the next chapter, the spatial analysis technique of interest – spatial interaction models – will be introduced and reviewed.

Chapter 2

SPATIAL INTERACTION MODELS

2.1 Introduction

Spatial interaction models are a class of models used to understand and predict aggregate flows of people, information, or goods over space. The underlying hypothesis for spatial interaction models is that the volume of flows between an origin and destination is a function of the potential at an origin, the attractiveness of a destination, and the cost of overcoming the separation between the origin and destination. It is the generalizability of this hypothesis that has allowed spatial interaction models to be applied in diverse settings and has been of consistent interest to geographers and regional scientists, as well as allied disciplines (Haynes and Fotheringham, 1984; Fotheringham and O'Kelly, 1989; Sen and Smith, 1995; Roy, 2004; Oshan *et al.*, 2014; Farmer and Oshan, 2017). In particular, there is often special attention paid to the role of physical separation within these models, which is usually captured by inter-location distances.

Parametric models, which involve deriving unknown parameters by calibrating the model on observed data, may be used both for explaining spatial interaction, as well as making predictions. In contrast, non-parametric models do not have any parameters that need to be calibrated and therefore do not provide any interpretative or explanatory power. Consequently, non-parametric models are limited to predicting spatial interaction. In the review of spatial interaction models that follows, parametric models will first be presented. A key inquiry in parametric spatial interaction modeling

is how space limits human activity and how this effect may change for different types of flows, for different groups of individuals, and for different spatial contexts. In particular, spatial structure in spatial interaction will be a core theme within this review of parametric spatial interaction models. Subsequently, non-parametric models will be discussed. Lastly, some metrics for assessing model fit will be reviewed for use in empirical work.

2.2 Parametric spatial interaction models

2.2.1 The gravity model

The earliest spatial interaction models originated from a physical analogy based on Newton's law of gravitational attraction between two bodies (see for example Zipf, 1946), where the number of flows between two locations is given by the product of the populations of the origin and destination, divided by the distance between them. This relationship can be generalized in the following manner,

$$T_{ij} = k\frac{V_i^{\mu}W_j^{\alpha}}{d_{ij}^{\beta}} \tag{2.1}$$

where T represents an $n \times m$ matrix of flows between n origins (subscripted by i) to m destinations (subscripted by j), V and W are $n \times 1$ and $m \times 1$ vectors of origin and destination attributes, respectively, d is an $n \times m$ matrix of the costs to overcome the physical separation between i and j (usually distance or time), k is a scaling factor, and μ, α, and β are exponential parameters. This model is often simplified further by assuming that some or all of the exponential parameters are unity and therefore are not included in equation (2.1). When data for T, V, W, and d are available we can estimate the exponential parameters (also called calibration), which summarize the

effect that each model component contributes towards explaining the system of known flows (T). In addition, known parameters can be used to predict unknown flows when there are deviations in model components (V, W, and d) or the set of locations in the system is altered (Fotheringham and O'Kelly, 1989). Despite its usefulness, this simplistic model lacks an analytical derivation and a theoretical behavioral framework.

2.2.2 Entropy maximization and the family of gravity-type spatial interaction models

Wilson (1967, 1969, 1970, 1971, 1973) provided a formal framework by applying the statistical theory of entropy-maximization to analytically derive a 'family' of gravity-type spatial interaction models, henceforth referred to as just spatial interaction (SI) models. This framework seeks to assign flows between a set of origins and destinations by finding the most probable configuration of flows out of all possible configurations, without making any additional assumptions. These models can also be obtained using an information minimization framework (Fotheringham and O'Kelly, 1989). By including information about the total inflows and outflows at each location (also called constraints), the following family of models can be obtained,

Unconstrained

$$T_{ij} = V_i^{\mu} W_j^{\alpha} f(d_{ij}) \tag{2.2}$$

Production-constrained

$$T_{ij} = A_i O_i W_j^{\alpha} f(d_{ij}) \tag{2.3a}$$

$$A_i = \Big(\sum_j W_j^{\alpha} f(d_{ij}) \Big)^{-1} \tag{2.3b}$$

Attraction-constrained

$$T_{ij} = B_j V_i^{\mu} D_j f(d_{ij}) \tag{2.4a}$$

$$B_j = \Big(\sum_i V_i^{\mu} f(d_{ij}) \Big)^{-1} \tag{2.4b}$$

Doubly-constrained

$$T_{ij} = A_i B_j O_i D_j f(d_{ij}) \tag{2.5a}$$

$$A_i = \Big(\sum_j B_j D_j f(d_{ij}) \Big)^{-1} \tag{2.5b}$$

$$B_j = \Big(\sum_i A_i O_i f(d_{ij}) \Big)^{-1} \tag{2.5c}$$

where O_i and D_j are the total number of flows emanating or terminating at an origin or destination, A_i and B_j are balancing factors that ensure these totals are preserved in the predicted flows, and d_{ij} takes on a functional form, referred to as the distance-decay function. This is most commonly either a power function,

$$f(d_{ij}) = d_{ij}^{\beta} \tag{2.6}$$

or an exponential power function,

$$f(d_{ij}) = exp(\beta * d_{ij}) = e^{\beta * d_{ij}} \tag{2.7}$$

and β is expected to take on negative values that indicate the decaying nature on the effects of increasing physical separation on the propensity for flows to occur. The former distance-decay function is justified through Wilson's max-entropy derivation

where it is the natural result of the analytic framework while the latter distance-decay function is justified when the analyst believes there is a logarithmic evaluation of transport costs and is also the specification that arises from a physical analogy to Newton's law of gravity. Vries *et al.* (2009) and Martínez and Viegas (2013) provide some examples of alternative specifications, thought these are less popular in the literature and often harder to interpret. Finally, μ, α, and β are parameters to be estimated through model calibration.

The so-called unconstrained model (or total-trip constrained model) is given in (2.2), which does not conserve the total inflows or outflows during parameter estimation. The production-constrained and attraction-constrained models are given in (2.3a) and (2.4a). These models conserve either the number of total inflows or outflows at each location and are therefore useful for building models that allocate individuals either to a set of origins or to a set of destinations. Finally, the doubly-constrained model is given in (2.5a), which conserves both the inflows and the outflows at each location during model calibration. The quantity of explanatory information provided by each model is given by the number of parameters it provides. As such, the unconstrained model provides the most information, followed by the two singly-constrained models, with the doubly-constrained model providing the least information. Conversely, the model's predictive power increases with higher quantities of built-in information (i.e. total in or out-flows) so that the doubly-constrained model usually provides the most accurate predictions, followed by the two singly-constrained models, and the unconstrained model supplying the weakest predictions (Fotheringham and O'Kelly, 1989). An exposition of these spatial interaction models, their applications, and various extensions is provided by Wilson (2010a). These models were later given a behavior-based theoretical foundation through the economic framework of utility

theory (McFadden, 1974; McFadden, 1977) and have been shown to be equivalent to discrete choice models derived within a utility-maximization framework plus any aggregation bias (Anas, 1983).

While these spatial interaction models have been calibrated using linear programming, and non-linear optimization, regression is perhaps most frequently used. They can be linearized and included within an ordinary least squares regression specification by taking the logarithm of both sides of a given model (Fotheringham and O'Kelly, 1989). For the unconstrained model, this yields the so-called log-linear or log-normal gravity model,

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij} \tag{2.8}$$

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j - \beta d_{ij} \tag{2.9}$$

which can be expressed more generally in terms of a regression specification as

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j + \beta \ln d_{ij} + \epsilon \tag{2.10}$$

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j + \beta d_{ij} + \epsilon \tag{2.11}$$

where ϵ is a a normally distributed error term with a mean of 0 and β is still expected to take on a negative value. These two specifications differ in that the distance variable is logged in equation 2.10 whereas in equation 2.11 it is not. This difference arises depending on whether a power function (equation 2.6) or an exponential function (equation 2.7) is plugged into equation (2) before linearizing it. However, there are several limitations of the log-normal gravity model, which include (I) flows are often counts of people or objects and should be modeled as discrete entities; (II) flows are often not normally distributed; (III) downward biased flow predictions due to producing estimates for the logarithm of flows instead of actual flows; (IV) zero flows

are problematic since the logarithm of zero is undefined. Therefore, the Poisson log-linear regression specification for the family of spatial interaction models was proposed over ordinary least squares regression (Flowerdew and Aitkin, 1982; Flowerdew and Lovett, 1988). This specification assumes that the number of flows between i and j is drawn from a Poisson distribution with mean, $\lambda_{ij} = T_{ij}$, where λ_{ij} is assumed to be logarithmically linked to the linear combination of variables,

$$\ln \lambda_{ij} = k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij} \tag{2.12}$$

and exponentiating both sides of the equation yields the unconstrained Poisson log-linear gravity model,

$$T_{ij} = \exp(k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij}) \tag{2.13}$$

where equations 2.12 and 2.13 refer to the unconstrained model with a power function distance-decay. Using fixed effects for the balancing factors in equations (3-5), the constrained variants of the family of spatial interaction models can be specified for Poisson regression as,

Production-constrained

$$T_{ij} = \exp(k + \mu_i + \alpha \ln W_j - \beta \ln d_{ij}) \tag{2.14}$$

Attraction-constrained

$$T_{ij} = \exp(k + \mu \ln V_i + \alpha_j - \beta \ln d_{ij}) \tag{2.15}$$

Doubly-constrained

$$T_{ij} = \exp(k + \mu_i + \alpha_j - \beta \ln d_{ij}) \tag{2.16}$$

where μ_i are origin fixed effects and α_i are destination fixed effects that achieve the same results as the balancing factors A_i and/or B_j in equations (2.2-2.5c) (Tiefelsdorf

and Boots, 1995). Notice that k is the estimated intercept and must be included in these log-linear models to ensure the total number of flows is conserved, despite not being included in the maximum entropy models where such conservation is typically implied during model calibration in this non-linear form. Using Poisson regression is more representative of flows and satisfies limitations (I-II) and it also alleviates limitations (III-IV) since we no longer need to take the logarithm of T_{ij} (Fotheringham and O'Kelly, 1989). The specifications in equations 2.13 - 2.16 are typically estimated within a generalized linear modeling (GLM) framework using iteratively weighted least squares, which is known to converge to the Poisson maximum likelihood estimates

A popular extension of these models is to calibrate a separate model for all flows from each origin to all of the destinations, which is often called an origin-specific local model. Then a set of parameter estimates for each model term is obtained for each origin and can be mapped in order to explore any spatial variation. One way to achieve this is to filter the overall dataset into subsets that only include flows from a single origin and then to calibrate an entirely separate model for each subset. Another method is to specify a single regression model that appropriately segments the data based on each origin. This can be done by introducing interaction terms into the regression where every variable interacts with a categorical variable that indicates which flow observations start at which origin. The main difference between these methods is that the latter assumes a common variance amongst the individual subsets of the data, which can result in slightly different parameter estimates. Recent literature has also proposed geographically-weighted techniques for estimating local parameter for spatial interaction models (Nakaya, 2001; Nissi and Sarra, 2011; Kalogirou, 2015; Kordi and Fotheringham, 2016). In the following sections, the central

role of origin-specific spatial interaction models in diagnosing spatial structure effects will be highlighted.

2.2.3 Intervening opportunities

Perhaps the second oldest hypothesis pertaining to human movement is that of intervening opportunities, which posits that 'the number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities' (Stouffer, 1940). It was later refined such that intervening opportunities were defined as those locations within a circle with a radius given by the distance between the origin and destination under consideration (Stouffer, 1960). This is in contrast to the gravity model, which focuses directly on costs associated with physical distance between locations. Indeed there is much research comparing the two frameworks (Kaltenbach, 1972; Haynes *et al.*, 1973; Dison and Hale, 1977; Smith, 1980; Elffers *et al.*, 2008). In particular, Okabe (1976) shows that under certain conditions the two models act very similar. Furthermore, several frameworks have been set forth which seek to include both gravity and intervening opportunities type effects (Wills, 1986; Ulyssea-Neto, 1993; Cascetta *et al.*, 2007). Wilson provides a derivation of the intervening opportunities models using an entropy-maximizing framework (Wilson, 1967) and additional model forms and procedures have been proposed (Schmitt and Greene, 1978; Rogerson, 1986; Akwawua and Pooler, 2001; Afandizadeh and Hamedani, 2012; Nazem *et al.*, 2015). For brevity, technical specifications for the intervening opportunities model will not be provided. While the intervening opportunities model has its own rich literature, it has been the subject of much less recent research when compared to gravity models.

As such, only its basic concept and history is introduced so that it may serve as a reference within subsequent sections of this review.

2.2.4 Spatial structure and spatial dependence in spatial interaction

2.2.4.1 Background

Throughout the 1970's and 1980's there was much debate about the role of the structure of a spatial system in spatial interaction models. The intense interest in the matter can be best captured in a series of publications, associated comments and replies, and subsequent reviews (Curry, 1972; Johnston, 1973; Cliff *et al.*, 1974; Curry *et al.*, 1975;Cliff *et al.*, 1975; Johnston, 1975; Cliff *et al.*, 1976; Sheppard *et al.*, 1976; Fotheringham and Webber, 1980; Griffith and Jones, 1980, Fotheringham, 1981, Sheppard, 1984). At the root of the problem was the fact that *unintuitive* spatial patterns could be observed for a set of local distance-decay parameter estimate values that resulted from an origin-specific gravity-type spatial interaction model, the cause of which was at first uncertain and contested. As further evidence towards a spatial structure effect accrued, a consensus formed that whatever was causing the unlikely patterns in the parameter estimates could be skewing the behavioral interpretability of them. This meant that it was unclear whether spatial variability associated with distance-decay was caused by behavioral and perceptual differences of individuals over space or by other factors. Over the next few decades, several theories and technical specifications to account for potential spatial structure effects were proposed. However, the contemporary spatial interaction corpus is far from a consensus in terms of what causes spatial structure effects and how to best account for them.

2.2.4.2 Speculation

Even before there was active debate, several scholars theorized that the structure of locations were important in determining the magnitude of flows between origins and destinations. A review of some of these theories is provided by Griffith (1976) who portends that the potential relationship between location hierarchy and spatial interaction was first described by Heide (1963), Olsson (1967), and Claeson, (1968; 1969), and was subsequently explored analytically by Glejser (1969), Pedersen (1970) and Long and Uris (1971). However, one of the earliest attempts at isolating the spatial structure effect more generally was put forth by Curry (1972). Many potential issues in spatial interaction models, including the shape of the study area, aggregation and representation bias, *spatial autocorrelation* in the locational attributes, model estimation techniques, and interdependence between spatial interaction and locational distributions are discussed by Curry, though none are definitively linked to any measurable problem in spatial interaction models. One important idea from Curry's work was that even if there is no explicit relationship between spatial interaction and distance, if spatial interaction is a function of the location's populations, which are themselves a function of their neighbor's populations, then spatial interaction has an implicit relationship with distance. Curry's (1972) development of this idea included one of the earliest uses of the term *map pattern* in relation to spatial interaction to describe the clustering of locations, a pattern which was being used to infer spatial dependence between attributes of locations (i.e., spatial autocorrelation). Therefore, an association was made between map pattern and spatial autocorrelation. Though there is no use of the term *spatial structure*, there are several mentions of *locational structure*, which is in reference to origin and destination attributes.

Soon after, Johnston (1973, 1976) put forth the hypothesis that the spatial pattern observed in local distance-decay parameters was caused by the variation in the distributions of distances between each origin and the set of destinations, which would later be described as the conditional distance distribution (Tiefelsdorf, 2003). Johnston tests his hypothesis by calibrating a spatial interaction model, which assumes a continuous measure of distance on data generated by an intervening opportunities model, which uses a ranked measure of distance. A weakness in Johnston's theory, however, is that it assumes that all of the flows between each origin and destination are constant. Such a hypothetical scenario implies that there is no relationship between the spatial interaction flows and distance. Sheppard (1979b) and Fotheringham (1981) demonstrate that this an unrealistic assumption in gravity-type spatial interaction models where there is, in fact, an explicit relationship between flows and distance and therefore, Johnston's theory is not applicable. In addition, Johnston's use of the term *map pattern* refers to the distribution of locations in space (i.e., spatial clustering), though it is in reference to how this causes inter-location distance variation (1973), rather than in reference to locational attributes, and spatial structure is described as both 'the distribution of origins and destinations' and the 'influence of system geometry' (Johnston, 1976).

2.2.4.3 The debate

In response to both Curry and Johnston, Cliff *et al.* (1974) provide a simulation study to explore the potential effects of spatially autocorrelated locational variables on the parameter estimates of spatial interaction models. Importantly, in this paper we begin to see a deeper association of the terms map pattern and spatial autocorrelation.

It is also this work that sparked the most intense debate about spatial structure, which resulted in a series of comments, replies, and follow-up papers (Johnston, 1975; Cliff *et al.*, 1975; Curry *et al.*, 1975; Cliff *et al.*, 1976; Sheppard *et al.*, 1976; Johnston, 1976) that led to several useful outcomes.

First, Johnston (1975) clarifies his argument so as to distinguish the difference between the effects of conditional distance distributions from those arising from spatially autocorrelated location attributes (i.e., population mass terms), although admitting that they are potentially interdependent and using map pattern to describe their combined influence, which convoluted the two concepts along with the map pattern terminology. Second, Curry *et al.* (1975) demonstrate that the simulation design of Cliff *et al.* (1974) was flawed in that it does not truly include significant levels of spatial autocorrelation in the location attributes. By correctly specifying a simulation to include accurate levels of spatially autocorrelated location attributes, they are able to observe that the multicollinearity between location attributes and distance, which is implied by spatial autocorrelation, results in a misspecified model where the distance-decay parameter estimates includes the effects of distance from two sources (i.e., explicit and implicit) leading to biased and inconsistent estimates. Thirdly, we can begin to see the importance of the use of simulation as a tool for investigating spatial structure. For example, a portion of the previously described debate focuses squarely on how to specify the spatial associations underlying the spatial autoregressive specifications (Curry *et al.*, 1975; Cliff *et al.*, 1976; Sheppard *et al.*, 1976). Simulation remains an under-appreciated topic in contemporary spatial interaction research and will be further discussed in subsequent sections. In the wake of this debate, it started to become clear that the problem of spatial structure was really one of misspecification related to how locations were clustered in space, however,

the nature of the relationship had yet to be quantified and the language to describe it remained ambiguous.

2.2.4.4 Criticisms

The tenacity of the debate on spatial structure even grabbed the attention of those from outside the discipline of spatial analysis. In particular, Sayer (1977) took a critical stance towards the inadequacies of spatial interaction models that had been highlighted by the debate. He argues that spatial interaction models, and quantitative urban modeling more generally, do not include enough of the diverse and complex spatial processes involved in generating SI. Specifically, Sayer highlights how the assumption of temporal stationarity and generic populations are untenable in urban settings where dynamics and diversity are more likely. While he believes that these limitations can be alleviated by describing more complex real-world spatial-temporal processes and relationships, he is clear in his skepticism of empirical modeling as the means for doing so. This is perhaps most evident based on his conclusion of the spatial structure debate that, 'mathematization beyond the call of duty' and lack of consideration of conceptual roots of the problem combined to mystify the issues and led the debate into an inconclusive technical impasse" (Sayer, 1977). However, in hindsight we know that this is false, as scholarship on spatial structure has continued to the present with several proposed solutions that will be outlined later. Furthermore, though he considers the use of mathematics to be exaggerated, his concerns of complexity have consistently been engaged without the need to abandon the entire framework of spatial interaction models. For instance, work on dynamic spatial interaction models (Harris and Wilson, 1978; Fotheringham and Knudsen, 1986; Nijkamp and Reggiani, 1988;

Wilson, 2008; Birkin and Heppenstall, 2011), and disaggregate spatial interaction models (Birkin *et al.*, 2010; Newing *et al.*, 2015). Nevertheless, Sayer's call for a wider breadth of complexity and diversity may be employed as a point of evaluation for solutions proposed to remedy the spatial structure effects. A particularly poignant observation is that the analytical approach of Curry (1972) essentially couches the problems of one spatial concept, that of distance-decay, in a new spatial concept, that of spatial autocorrelation, which can be equally distracting from more pertinent issues (Sayer, 1977).

Interestingly, Sheppard (1979a, 1979b, 1979c) begins to similarly cast doubt on some aspects of spatial interaction models, though his concerns are stated more in terms of specific research questions and within ongoing research avenues rather than a sweeping and general critique. Most notably, he calls for further work on spatial behavior, such as how individuals search and learn within their spatial environment, and how to further reconcile disaggregate theory and aggregate modeling techniques (Sheppard, 1979c).

2.2.4.5 Further defining the nature of spatial structure: the debate diverges

In the wake of the spatial structure debate, two distinct approaches were taken in order to explain the nature of the effects of spatial structure in spatial interaction models. The first approach, which was taken by Fotheringham and Webber (1980) was to account for spatial structure by explicitly modeling the interdependence between spatial interaction and locational attributes. Here spatial structure is taken as the differences in spatial opportunities over space (i.e., spatial clustering of locations and associated attributes), while map pattern refers to the bias in distance-decay parameter

estimates when there is a feedback from spatial interaction to spatial structure that is not explicitly modeled. For example, in a study of migration between urban centers, to accurately obtain parameters for the effect of urban centers' size on SI, there also needs to be a sufficient model for urban growth as well. Since growth is interdependent upon SI, these two things should actually be modeled with a simultaneous equation system (SES) that is calibrated with either two-stage least squares or iteratively weighted least squares, rather than ordinary least squares. Generally, parameter estimates will be biased and inconsistent if any of the independent variables are a function of the dependent variable (i.e., independent variables are not independent of the error term), regardless of the scale that spatial interaction occurs. This type of misspecification is also called endogeneity in the econometrics literature. Fotheringham and Webber (1980) derived the exact nature of this bias, which demonstrates that if the SES is not utilized, then spatial interaction can be systematically over- or under- estimated by the model. Their proposed technique is also flexible in that it is expected that a unique SES will be specified for each modeling task at hand, since there are likely different relationships between spatial interaction and locational attributes and the underlying spatial structure for different types of spatial interaction phenomena. In fact, Fotheringham and Webber provided a second example within the context of retail shopping trips and show that there can be more than one spatial structure effect present in a system.

The second approach of Griffith and Jones (1980) was more exploratory in nature in that it sought to examine the correlations between different components of doubly-constrained spatial interaction models for several cities where each model consisted of a separate set of census tracts that pertain only to a single city. Their main goal was to answer the question, 'is the rate of distance decay in spatial interaction models

independent of the spatial structure associated with the corresponding origins and destinations?' In this scenario, spatial structure is defined as the convolution of the spatial configuration of areal units and their linkages and the interdependence/spatial autocorrelation that can exist across these units. While they found that there is moderate-to-strong correlation in the accessibility for each cities' system of locations and in the origin and destination balancing factors for each cities' system of locations, little-to-no spatial autocorrelation was found in the locational attributes for each cities' system of locations. A principle components analysis was then used to make the conjecture that spatial interaction and geometry are inseparable and that 'there exists a fundamental geometric dimension relating to the geographic distribution of workers/jobs'. These two combined findings imply that there may be interdependence between spatial interaction and locational attributes, though spatial autocorrelation in the locational attributes is not the appropriate metric to measure such interdependence. Finally, using a regression analysis of which variables explain the most variation in the distance-decay parameters, they surmised that an 'indirect relationship exists between distance-decay and geometric pattern, with this relationship being reasonably sensitive to changes in the geometry of destinations'.

It is important to note that the results from Griffith and Jones (1980) are speculative in that they do not analytically tie their observations to any defined source of bias that might be occurring in the distance-decay parameter estimates in contrast to Fotheringham and Webber (1980) who explicitly defined an interdependence and the resulting bias that arises when the interdependence is not accounted for. They did, however, offer a rudimentary version of the simultaneous spatial autoregressive (SAR) model, though it is not estimated due to several technical complications. Furthermore,

it seems that Griffith and Jones (1980) confounded several sources of potential bias. One source is described in the following passage

> In small urban settings, distance-decay and spatial-structure effects blend together. This is caused by two factors. On the one hand, the scale of analysis will often employ areal units that because of their sizes mask most of the spatial structure effects. On the other hand, in small urban areas spatial propinquity will cause movers to be unable to discriminate clearly between the two effects.

which refers to bias that is caused by aggregation of the data to areal units. Aggregation may induce measurement error that can indeed obfuscate the nature of the true underlying spatial structure or any other variable, though this is an issue with the data and not a description of a spatial interaction process. Previously, the debate had been centered on distinct locations that were abstracted as points in space so that the effects of areal aggregation were not a major issue. However, the work of Griffith and Jones (1980) marked a major shift in how locations are abstracted in space within the debate. This important detail is not stressed, though it has implications for how spatial structure is approached. Specifically, this shifts the focus from inter-distance relationships to 'the geometric linkages between areal units' (Griffith and Jones, 1980). Where distance is measured between aggregate areal unit centroids, instead of specific locations, aggregation bias is likely to occur (Webber, 1980; Okabe and Tagashira, 1996; Tagashira and Okabe, 2002).

Another source of bias that is entirely separate from aggregation is later described by Griffith and Jones (1980)

> To summarize, one controversial issue of spatial interaction modelling is whether or not the rate of distance decay is independent of the geographic structure associated with origins and destinations. In other words, do the propensity of origins to emit interactees and the propensity of destinations to attract interactees vary as the geometry of origins and destinations changes? Furthermore, do these propensities change with variations in

> the nature and degree of spatial autocorrelation latent in the geographic distributions of origin and destination totals?

which describes the case in which flows at one location are a function of the flows in nearby locations. This is the contemporary theoretical basis of the SAR model, which will be explored in more detail later. However, this theory is not given any contextual basis in terms of migration processes, which is an important detail because the pertinence of the SAR model is evaluated on a theoretical basis, since spatial autocorrelation in the dependent variable is generally only troublesome if it cannot be accounted for by independent variables. Therefore, the exploratory study of Griffith and Jones does not isolate a particular cause or measurement of spatial structure effects.

After reviewing much of the existing literature, Fotheringham (1981) raised the bar by highlighting weaknesses in existing theories about spatial structure effects and proposes that an adequate theory should apply to the entire family of spatial interaction models, regardless of the model calibration technique that is employed. He defined spatial structure as the "size and configuration of origins and destinations in a spatial system", which contrasted Griffith and Jones's (1980) explicit focus on geometry and linkages. Fotheringham's review included a summary of several studies employing location-specific local spatial interaction models and suggested that the presences of spatial structure effects can be diagnosed by the observation of unintuitive patterns in the local distance-decay parameter estimates, especially if these patterns seem to be driven by location accessibility. Fotheringham also argued that spatial autocorrelation is simply a surrogate for multicollinearity. To this, Sheppard (1982) clarified that spatial autocorrelation in the residuals can also indicate that there is one of several types of misspecification that can result in biased parameter estimates. He argued that depending on model form spatial autocorrelation in the

locational attributes could indicate that relationships between variables have been misspecified. This is one of many potential sources of bias, though it should be noted that the diagnosis for this type of misspecification should be made by observing spatial autocorrelation in residuals and not in the independent variables. While Fotheringham (1982) agreed with Sheppard (1982), noting that spatial autocorrelation can be linked to misspecification bias, he also suggested that the spatial structure effect he describes in various origin-specific distance-decay parameter estimates arises due to a misspecification that is entirely separate of spatial autocorrelation. That is, it is an omitted variable misspecification that can occur even when there is no variation in the locational attributes, which is eventually identified as the accessibility of each destination to all other destinations (Fotheringham, 1982, 1983a).

2.2.5 Accounting for spatial structure effects in parametric spatial interaction models

Several methods have been proposed to account for spatial structure in spatial interaction models. These methods differ both in the way that they account for spatial structure effects and in the modeling assumption that is violated to give rise to such spatial structure effects. Before discussing the proposed methods, common violations are outlined.

2.2.5.1 Types of effects

An instructive overview of the various assumptions underlying ordinary least squares regression and the effects that arise when they are violated is provided by Sheppard (1984). Violations may include measurement error, high levels of

multicollinearity, spatially autocorrelated residuals, violation of normality assumptions, omitting important variables, and residuals that are correlated with the explanatory variable(s). Several generalizations to Sheppard's (1984) list of violations are proposed in order to include the breadth of work that ensued over the last few decades. First, the violation of non-normality is expanded to include the violation of any underlying distributional assumptions. This includes the assumptions used in Poisson regression, which have become popular for spatial interaction modeling. Second, aggregation error, which may be encapsulated within measurement error, is highlighted since spatial interaction models have often been specified using aggregated data and are therefore frequently prone to this specific type of error. The most common results of these different violations are biased and potentially inconsistent parameter estimates and/or biased variances. Since any or all of these violations can occur in a single study, it is of utmost importance to be clear about which one is being addressed. In the ensuing discussion of strategies for reducing the spatial structure effects, the underlying violations that are or could be corrected for are highlighted for each method where possible.

2.2.5.2 Omitted variables

Following the intense debate of the seventies and early eighties, one of the earliest and most complete theories and methods designed to deal with spatial structure effects was that of competing destinations (CD) (Fotheringham, 1983a). This theory posits that spatial structure is the "configuration of origins and destinations in a spatial system", but it is the configurations of destinations with respect to each other that can affect an individual's propensity to select a particular destination. To account

for this effect, we need to include within spatial interaction models a variable that captures the accessibility of each destination to all other competing destinations. This definition contrasts the earlier definition of Fotheringham (1981) in that it explicitly focuses on the configuration of locations and hence only secondarily involves location attributes (i.e., sizes). A novel reasoning in support of this ordering of priorities is that a spatial pattern amongst locations may be induced by either altering their spatial locations or altering the attributes of the locations.

This is demonstrated in figure 1 where a) there is a uniformly distributed set of locations that all have the same size; b) location sizes have been altered to create a pattern of south-west spatial clustering while maintaining a uniformly distributed set of locations; and c) locations have been shifted to create a pattern of south-west spatial clustering while holding location size constant. In scenario a) there no clustering due to physical location or size. In contrast, spatial autocorrelation amongst location size can likely be used to describe scenario b) where the clustering pattern is due to the distribution of the size attribute. However, in scenario c) common spatial autocorrelation statistics would be undefined since there is no variation in location size[1]. Instead, the clustering pattern is due only to the arrangement of the locations. Therefore, the effect of clustered locations is more fundamental and may exist even when no spatial autocorrelation can be measured.

The behavioral reasoning for the CD model enhancement is that spatial decision-making, such as location choice, often arises from a hierarchical two-stage or multi-stage decision-making process, where individuals first select a region or cluster of locations and then subsequently choose an individual destination from within that cluster. The

[1]This would also be true for scenario a), since there is also no variation in the location sizes.

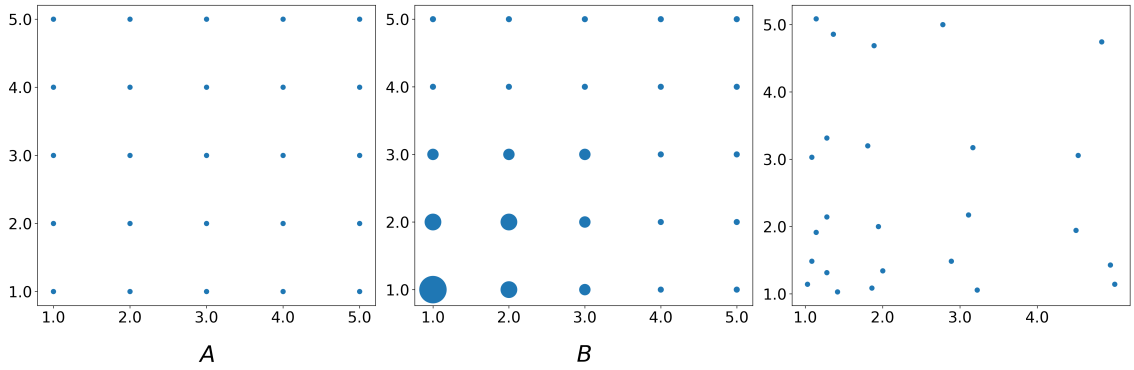


Figure 1: Three different sets of locations where a) there is a uniformly distributed set of locations that all have the same size; b) location sizes have been altered to create a pattern of spatial clustering while maintaining a uniformly distributed set of locations; and c) locations have been shifted to create a pattern of spatial clustering while holding location size constant.

effect is that as the accessibility of a particular destination, j, to all other potential destinations increases, j will experience greater competition from other destinations, and therefore the volume of flows to j would be smaller than predicted by a traditional spatial interaction model. Practically, this effect is captured by introducing a new variable into spatial interaction models to measure destination accessibility, A_{ij}, which has been shown to be consistent with maximum entropy and utility derivations (Fotheringham and O'Kelly, 1989), and can be thought of as the likelihood that other destinations are also considered along with destination j. For a production-constrained Poisson regression, for example, this results in

$$T_{ij} = \exp(k + \mu_i + \alpha \ln W_j - \beta \ln d_{ij} + \delta \ln A_{ij}) \tag{2.17}$$

$$A_{ij} = \sum_{\substack{k=1 \\ (k \neq i, k \neq j)}}^{n} \frac{W_k}{d_{jk}^{\sigma}} \tag{2.18}$$

where δ is the parameter corresponding to destination accessibility, A_{ij}, which is the sum of the attractiveness, W, at each alternative destination k weighted by its distance to each alternative destination d_{jk}, and σ is a parameter that controls the scale over

which destinations compete with each other, which in practice is often set to -1 or calibrated iteratively with the parameters in equation 2.17. Importantly, the set of locations defining the competing destinations need not be the entire set of locations, such that the definition of destination accessibility is likely unique for different contexts, and can even include locations that are not included in the original spatial interaction dataset. In fact, Fotheringham (1983a; 1983b) discusses how attraction-constrained and doubly-constrained models can be correctly specified (i.e. do not need to include A_{ij}) when every location in the system is both an origin and a destination and these locations are an accurate representation of all possible destinations available to each origin. However, unconstrained and production-constrained models are always misspecified if there is a relationship between spatial interaction and destination accessibility.

Whenever any of the spatial interaction models contain a competing destination type of misspecification, a failure to account for spatial structure results in origin-specific distance-decay parameter estimates that are biased upwards for accessible origins and biased downward for inaccessible origins. The strength of this bias is shown by Fotheringham (1984) to depend on the strengths of two relationships: that between the volume of flows and distance and that between distance from the origin to each destination and the accessibility of each destination. It is this bias that causes the unintuitive spatial patterns observed in origin-specific distance-decay parameter estimates, such as positive estimates for the most accessible origins (Fotheringham, 1981). Furthermore, the bias can be categorized in terms of the perception of destination clusters where there are competition effects (i.e., negative exponent on A_{ij}) or agglomeration effects (i.e., positive exponent on A_{ij}). For the former effects, the addition of a location to a cluster increases the attractiveness of the

cluster less than the attractiveness of the location itself, while for the latter effects, the addition of a location to a cluster increases the attractiveness of the cluster more than the individual attractiveness of the location (figure 2). The type of effect that arises is typically dependent upon the type of spatial interaction process being modeled, though competition is more often observed in empirical settings, hence the selection of the name the competing destination model. The proliferation of the competition effect may also be driven by the behavioral tendency of individuals to underestimate the overall attractiveness of large clusters. Compared to the competing destination model, the expected outcome according to a traditional spatial interaction model is that when an additional location is added to a cluster the attractiveness of the cluster is increased exactly by the attractiveness of the additional locations (the straight line in figure 2). If this is the case, then either there is no spatial structure misspecification or there are competition effects and agglomeration effect that are canceling each other out (Fotheringham, 1983b).

In fact, the precise nature of the potential bias of distance-decay parameters due to this type of spatial structure within an OLS regression has been defined by Fotheringham (1984) where bias may be due to

1. a relationship between A_{ij} and T_{ij}
2. a direct relationship between A_{ij} and d_{ij}
3. an indirect relationship between A_{ij} and d_{ij} due to a relationship between A_{ij} and W_j and a relationship between d_{ij} and W_j.

Bias source (3) can occur independently of bias source (2), since it can occur even when there is no relationship between A_{ij} and d_{ij}. In addition, if bias sources (2) and (3) do not occur, but source (1) does occur, then distance-decay parameters will

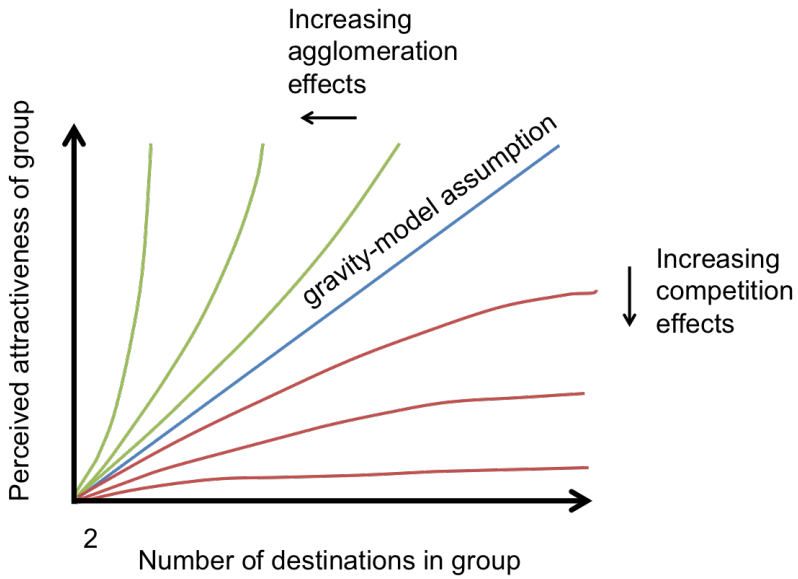


Figure 2: Impact of agglomeration and competition effects on spatial interaction in comparison to the expected flow volumes from a gravity-type spatial interaction model that does not consider competing destinations. This figure originally appears in (Fotheringham and O'Kelly, 1989).

not be biased; however, A_{ij} is still a relevant explanatory variable that can increase the accuracy of the model. These sources of bias can be similarly defined for the parameter estimates on other variables, though it is likely that the bias will be stronger for distance-decay whenever there is a stronger relationship between A_{ij} and d_{ij} than between A_{ij} and locational attributes (Fotheringham, 1984).

It becomes clear that the CD model is addressing a very particular misspecification that arises organically due to collinearity amongst variables that we would ordinarily include in most, if not all, spatial interaction models. Therefore, the violation it is attempting to correct for is the omission of a correlated spatially patterned variable. Baxter (1983; 1985) provides a more general analysis of bias using a different form of omitted variable misspecification that is less amenable to interpretation. While this general framework could be used to assess the consequences of different forms variable

misspecification, it does not provide strong empirical evidence for any particular form, and it is not clear how each form would arise.

Adopting the destination accessibility term, A_{ij}, is also advantageous in that doing so avoids the undesirable independence from irrelevant alternatives (IIA) property that exists in many spatial choice modeling frameworks. The IIA property assumes that all locations in the set of potential destinations that one can choose from are evaluated equally. This means that destinations are not perceived as clusters, but rather as individual locations (Fotheringham, 1986). Hierarchical choice models that are not free of the IIA property, such as the nested logit model, run into several issues for spatial choice problems. First, they require the choice set to be specified *a priori* by the analyst. Second, the set of alternative choices is assumed constant, which is problematic in spatial choice scenarios. For example, the IIA property implies that if choice a is a substitute for choice b and choice b is a substitute for choice c, then choices a and c are also substitutes for each other. However, this may be very unlikely if choice a and choice c are very far apart. Both of these issues may be avoided by using the CD model with A_{ij} as a measure of destination accessibility (Fotheringham, 1986, 1988). There has also been a great deal of additional research concerning choice set definition in spatial interaction models (Thill, 1992; Thill and Horowitz, 1997; Pellegrini *et al.*, 1997).

The CD model has been applied in many domains including urban modeling (Fotheringham, 1985; Fotheringham and Knudsen, 1986), the study of telecommunications flows (Guldmann, 1999), and crime location analysis (Bernasco, 2010), though it has enjoyed particular popularity within migration modeling (Ishikawa, 1987; Ishikawa, 1990; Fik *et al.*, 1992; Pellegrini and Fotheringham, 1999; Fotheringham *et al.*, 2000; Pellegrini and Fotheringham, 2002; Yano *et al.*, 2003; Fotheringham *et al.*, 2004;

Kalogirou, 2015), commuting-to-work research (Thorsen and Gitlesen, 1998; Gitlesen and Thorsen, 2000; Gitlesen *et al.*, 2010), and retail analysis (Fotheringham and Knudsen, 1986; Guy, 1987; Fotheringham, 1988; Pellegrini *et al.*, 1997; Birkin *et al.*, 2010). The underlying theory of the CD model has also been explored via simulation studies (Lo, 1991a; Fotheringham *et al.*, 2001), in association with cognition and spatial information processing (Hirtle and Jonides, 1985; McNamara, 1986; Curtis and Fotheringham, 1995; Fotheringham and Curtis, 1999), and it has been extended within other frameworks, such as central place theory (Fik and Mulligan, 1990) and trip chaining behavior (Bernardin *et al.*, 2009). While this review of work relating to the CD model is comprehensive it is certainly not exhaustive. It is also worthwhile to note that in the majority of studies using the CD model, locations are abstracted as points in space, though even when areal units are used, the CD model does not claim to account for any effects that arise from potential aggregation error.

In contrast to the CD model, there are some authors who have stressed that the explicitly spatial focus of the competing destination model overlooks other important factors (Gordon, 1985; Lo, 1991a; Lo, 1991b; Lo, 1992; Pooler, 1998; Hu and Pooler, 2002). For example, Gordon (1985) suggests that further work investigating the spatial patterns of distance-decay parameters should focus on functional and economic differences between locations rather than solely physical accessibility. Lo (1991a; 1991b; 1992) takes up this call and argues that spatial structure should be renamed to *destination interdependence*, where this interdependence is composed of physical aspects and economic aspects. The physical aspect, which Lo calls *locational substitutability*, is synonymous with the effects associated with the CD model, whereas *economic substitutability* is offered as an example of the economic aspects that might still be misspecified in CD models. Economic substitutability refers to consumer preferences

toward destination activities and services. If the degree that destinations provide activities that are substitutable, or conversely, that are complimentary, varies, and this is not accounted for, then spatial interaction models may produce biased parameter estimates. While it is demonstrated that the CD model can account for locational substitutability, thereby removing the spatial structure effects as they are theorized under the CD model, it is however shown that the CD model does not account for economic substitutability. Similarly, Pooler proposes a more general competition effect, termed *spatial influence* which also theorizes a hierarchical decision-making process, but where macro-level groups of destinations are based primarily on attributes (i.e., aspatial) rather than their locations in space. However, in the exposition of both the theories of economic substitutability and spatial influence, no generalizable framework or solution is proposed to account for the omitted variables that are causing models to be misspecified. Furthermore, neither of these alternative theories has yet to garner much empirical evidence. This is likely because these misspecifications are particularly context-dependent, whereas (locational) spatial structure effects can arise in any spatial interaction scenario. Consequently, several extensions have instead been proposed to account for more complicated location choice relationships without abandoning the CD model framework. For example, disaggregated spatial interaction models can include store brand and household type in regards to retail flows (Newing *et al.*, 2015) and variations of the CD model that separately account for agglomeration an competition effects (Bernardin *et al.*, 2009). Therefore, these criticisms of the competing destination model are likely either unfounded or overstated. This is perhaps best-illustrated by Hu and Pooler (2002) who assert that any spatial variation within local distance-decay parameter estimates means that the model is misspecified. In contrast, spatial variation could be due to variations in how distance

is perceived or due to aggregation error, which can occur even when the model is correctly specified. Ultimately, criticisms of the CD model are primarily calls for more in-depth model-building and more accurate data.

Some alternative specifications to explicitly account for omitted variables related to spatial structure have been put forth, though they are generally less developed compared to the CD model. One such example is given by Boots and Kanaroglou (1988), who use the principal Eigenvector of a binary contiguity matrix to derive a distance centrality measure. The inclusion of this measure in a nested logit model is statistically significant and improves model accuracy, especially in comparison to several other aspatial similarity measures. However, relatively little interpretation of this new variable, its parameter estimate, or the nature of the bias that results from its exclusion are provided. Furthermore, their modeling framework is that of the nested logit model and they do not discuss how the inclusion of their spatial structure measure relates to hierarchical location choice sets and the IIA property. Another example of an alternative spatial structure specification is that of network autocorrelation (Black, 1992), where a spatial autocorrelation statistic with a network-based definition of proximity is used to assess misspecification. This leads to several approaches to account for omitted variables, which include geographically-based categorical dummy variables or origin/destination accessibility terms (though not in exactly the same sense as specified in the CD model).

2.2.5.3 Variable functional form

Though the functional form of any of the terms within a spatial interaction model could be subject to functional form misspecification, it is the distance term that

receives the most attention due to the fact that the resulting distance-decay parameter estimates have often been interpreted as an indication of human behavior. It would be approximately thirty years before Johnston's argument was picked up by again by Tiefelsdorf (2003), who illustrates how spatial structure in local distance-decay parameter estimates can arise systematically due to conditional distance distributions *when the functional form of distance is misspecified.* Recall from section 2.2.2 that a power or exponential functional form of distance are the most popular because they arise from direct analogy or through the max-entropy derivation process. These two forms have also gained acceptance through empirical consensus where the exponential function is more appropriate for short term interactions such as intra-urban trips and the power function is more appropriate for longer distance trips such as migrations flows (Fotheringham and O'Kelly, 1989). Additionally, recall from section 2.2.4.2 about Johnston's argument that spatial variation in distance-decay parameter estimates can arise if the underlying data-generating process follows the intervening opportunities model but a spatial interaction model is calibrated on the data. This is essentially an extreme case of functional form misspecification, which Tiefelsdorf (2003) generalizes upon to show that any inconsistencies between the true data-generating functional form on distance and that which is specified in the model can result in distance-decay parameter estimates with spatial variation. Subsequently, it is recommended that the correct functional form of the distance term can be obtained by using the Box-Cox transformation. Since the Box-Cox transformation encompasses a spectrum of functions, depending on a parameter, q, an optimal parameter value is selected by maximizing the model fit as denoted by the likelihood ratio for the model. Tiefelsdorf (2003) claims that this allows the spatial structure effect that is caused strictly by distance to be accounted for, which cannot otherwise be accounted for by the CD

model, though there are several conceptual and technical problems with the evidence that is presented.

Most importantly, there is no guarantee that when the Box-Cox parameter is selected that the model does not compensate for another misspecification. For example, if the functional form of any other variable is misspecified, say the destination population, this could also result in spatial variation within the local origin-specific distance-decay parameter estimates depending on other relationships and mispecifications in the model. Following Tiefelsdorf's (2003) recommendation, one would then apply a Box-Cox transform and select an optimal q based on the model fit. However, it is possible that the optimization suggests an incorrect functional form for distance because it is also accounting for the misspecification of the functional form of destination population [2]. Generalizing this argument, it is essentially impossible to know for certain if any single effect is being accounted for by the Box-Cox transform in this context. Therefore, this solution neither guarantees that the spatial structure effects caused by conditional distance distributions are isolated, nor does it provide a means for diagnosing any particular misspecification.

There are also two technical problems with Tiefelsdorf's specification that invalidate his evidence against the CD model. First, his specification is only partially local, such that it produces local parameter estimates for distance but not for locational attributes. The second issue is that his specification includes both an origin-specific and a destination-specific parameter estimate for distance decay. Tiefelsdorf (2003) expresses that, "Most of the spatial structure discussion in interaction modeling so far focuses solely on origin specific distance decay parameters. Nevertheless, the

[2]The work of Tiefelsdorf (2003) was replicated in order to test this idea and the details are available in appendix A

destination specific distances decay parameters are just as meaningful." However, this new, and therefore, unsubstantiated, specification is not a good candidate for exploring spatial structure effects and seems to have been the root of some of the problems that Tiefelsdorf claims to illuminate and even potentially solve. For example, using an inter-state migration dataset for the continental U.S., Tiefelsdorf reports unexpected negative signs for origin and destination population attributes. In actuality, if a proper fully-local spatial interaction model is specified and the model is solely origin- *or* destination-specific, then the issue of reversed signs is ameliorated and should be seen as an artifact of improper model specification. In addition, a destination-specific focus typically has no behavioral meaning in the context of migration because individuals arriving at a destination do not choose their origin.

The potential insights provided by Tiefelsdorf (2003) are clouded further by several misconceptions. First, there is a discussion of the lack of monotonicity in the decreasing nature of distance-decay over longer distance as if it implies misspecification. However, distance-decay parameter estimates represent the effect of distance conditional upon other variables, such as origin and destination populations. Even where there are longer distances between locations, it is possible to have a less negative distance-decay parameter estimate if there are strong forces of propulsion and attraction. Next, it is hard to tell how extreme the spatial variation in the distance-decay parameter estimates in Tiefelsdorf's migration example are because of the centered dummy variable coding scheme and the way the estimates are mapped. Since the centered coding scheme generates a mean parameter estimate and then local parameter estimates that vary around the mean, it is necessary to calculate the combined effect from both the mean and local deviation. Instead, Tiefelsdorf only maps the local deviations, which may be very slight when compared to the combined effect. As a comparison, the spatial

variation that lead to the formulation of the CD model was so great, that it included positive values of distance-decay (Fotheringham, 1981)! The final misconception concerns details regarding the CD model. It seems that the specified accessibility term was actually an origin accessibility and not the accessibility from the destination to all other destinations. Though such a conceptualization of spatial structure is possible, it does not seem to be the desired intention.

While Tiefelsdorf correctly points out that functional form misspecification can cause some degree of spatial variation in local parameter estimates, little more can be contributed to the issue of spatial structure due to the various technical and conceptual shortcomings.

2.2.5.4 The return of spatial autocorrelation

Spatial autocorrelation, which was previously discussed in the context of spatial structure in SI, typically measures the association of observations over a two-dimensional (x, y) plane. One reason that spatial autocorrelation has become a dominant paradigm within quantitative geography for understanding geographic relationships is because it is simple to reduce many scenarios to two-dimensional spatial representations. For example, Curry (1972) theorizes the role of spatial autocorrelation amongst two-dimensional location attributes in spatial interaction models. However, spatial interaction data are typically four-dimensional (x_1, y_2, x_2, y_2), which is more complex to represent and derive relations between (Fischer and Griffith, 2008). Nevertheless, Griffith and Jones (1980) posit that spatial interaction itself is spatially autocorrelated without identifying a method of measuring association between flows, which is crucial to properly measure any type of autocorrelation. While much research

would eventually explore the nature of spatially correlated components of spatial interaction, it would be several decades before the focus shifted to directly developing concepts of association between flows themselves and methods to account for it. These efforts have been carried out primarily in the paradigms of spatial econometrics and eigenvector spatial filtering. However, it will be seen that recent work can largely be characterized by attempts to measure associations between flows in two-dimensional space rather than four-dimensions, whether those relations are determined by contiguity, distance, or another type of proximity. It will also be seen that these approaches contrast previous work that focuses on spatial structure in that they do not focus on local parameter estimates or a specific type of model misspecification.

2.2.5.5 Spatial econometric approaches

Spatial regression models, the workhorse of applied spatial econometrics (LeSage and Pace, 2008) are models that incorporate spatial dependence in (i) the the dependent variable, (ii) the error term, (iii) the independent variables, or (iv) some combination of (i-iii). The two spatial regression specifications that are most frequently used and studied in the spatial econometric literature are the spatial autoregressive model (SAR) and spatial error model (SE) (Halleck Vega and Elhorst, 2015), which satisfy (i) and (ii), respectively. The SAR model is given by

$$y = k + \rho My + X\beta + \epsilon \tag{2.19}$$

where y is an $n \times 1$ vector of observations on the dependent variable, M is the $n \times n$ spatial weights matrix, which defines the neighborhood of an observation, ρ is a spatial autoregressive parameter, X is an $n \times p$ matrix of observations on the p explanatory variables, β is the associated vector of explanatory variable parameters, and ϵ an $n \times 1$

vector of normal error terms. Since My is typically row standardized (or equivalently called the W-coding scheme), it is useful to think of it as a weighted average of neighborhood values (sometimes called the spatial lag), such that observations on the dependent variable at a location are also dependent upon observations of their neighbors, which is usually motivated by the theoretical equilibrium outcome of a process over space (Anselin, 2006). An example would be the price of real estate in one location, which is usually directly determined by the values of real estate around it. In this case, there is a clear unit of analysis and theoretical dependence process. In contrast, the SE model is not based on a theoretical process, and is given by

$$\begin{aligned} y &= k + X\beta + u \\ u &= \lambda Mu + \epsilon \end{aligned} \tag{2.20}$$

where k, X, β, and ϵ are as previously defined, and λMu is a spatially structured portion of the residuals that captures unaccounted spatial effects such as omitted variables, and measurement errors due to scale mismatch and aggregation. Therefore, the SE model is more suitable to help with practical data concerns than theoretical concerns. These more complex specifications will be biased and/or inconsistent when estimated by OLS and instead are typically estimated using maximum likelihood estimation, two stage least squares, and the general method of moments (Anselin and Rey, 2014). Finally, though less often utilized in applied spatial econometrics, a model with a spatial lag on the exogenous variables, X, (SLX) is given by

$$y = k + X\beta + MX\phi + \epsilon \tag{2.21}$$

such that M is now applied to X instead of the dependent variable, y, and ϕ is the associated spatial autocorrelation parameter (Halleck Vega and Elhorst, 2015).

Several early attempts have been made to extend the SE model to spatial interaction models. Brandsma and Ketellapper (1979) propose an error model where the error term

$u = \lambda Mu + \epsilon$ in equation 2.20 is expanded to $u = \lambda_i M_i u_i + \lambda_j M_j u_j + \epsilon$, where M_i and M_j are binary matrices that separately capture origin and destinations effects, respectively, and where spatial relations between flows can be defined by flows that share origins (destinations) or flows that share an origin (destination) and have adjacent destinations (origins). In contrast, Bolduc *et al.* (1989) propose a form of error dependence for flows where a single weight matrix is comprised of multiple additive terms each using distance based measures of proximity. Instead of separate origin and destination effects, a single spatial term is utilized with $M = (d_{r,l} + d_{s,t})^{-\theta_1} + (d_{r,t} + d_{l,s})^{-\theta_2}$ where M captures the direct distance effect and cross-distance effects between a flow from origin r to destination s and a flow from origin l to destination t, and θ_1 and θ_2 are additional (power function) distance-decay parameters to be estimated. The first term (i.e., direct effects) captures the additive effects from distance-based proximity between the origins and the destinations of the two flows, while the second term (i.e., cross-distance effects) captures the additive effects from distance-based contiguity between the origins of the two flows to the opposite destinations of the two flows. Though the direct distance effects are significant in an empirical example, the cross-distance effects are not, and the authors admit this type of spatial effects may be difficult to interpret. A more general specification is subsequently put forth by Bolduc *et al.* (1992) that incorporates the specifications of Brandsma and Ketellapper (1979) and Bolduc *et al.* (1989) and provides simpler interpretations of the autocorrelation parameters. This ultimately leads to a specification of the error term as $u = \lambda_i M_i u_i + \epsilon_i + \lambda_j M_j u_j + \epsilon_j + \lambda_{ij} M_{ij} u_{ij} + \epsilon_{ij}$ that has a separate independent spatial error term for origins, destinations, and origin-destination pairs, and the weight matrices are defined by distance-decay-based contiguity with estimable parameters. This model is estimated on a single realization of simulated data to demonstrate that estimation is feasible. However, the fixed and

known parameters used to simulate the data were not all replicated, which the authors argue is due to small sample size (i.e., 25 locations and 625 flow observations). Perhaps, more importantly, the distances used to determine proximity were all simulated from random uniform distributions, thereby skirting the complex issue of defining proximity between flows themselves. This model is also applied to an application of transportation flow modeling in Winnipeg where the term M_{ij} is defined using the concept of cross-distance effects put forth by Bolduc *et al.* (1989), but the estimation routine must be augmented such that all of the autocorrelation parameters and distance-decay parameters are identical in order to sufficiently calibrate the model (Bolduc *et al.*, 1995). All of these extensions of the SE model to spatial interaction seem to have either suffered from issues of interpretability or estimation or both.

Other attempts at modeling spatial structure effects in spatial interaction models using SE models have produced more encouraging results. Porojan (2001) includes a single spatial error term in a trade model where M is a binary contiguity matrix with non-zero entries denoting origin-destination pairs that share a contiguous border. While this specification results in more accurate interpretations for the parameter estimates, it is unclear why the spatial relationship encapsulated in M could not have been directly modeled using a binary indicator variable. Fischer and Griffith (2008) suggest a spatial interaction model of knowledge flows with the spatial error term defined in equation 2.20 but with $M = M_o + M_d$ for cumulative spatial effects between the origins and destinations of flows. Here, proximity between flows is defined by binary contiguity matrices with entries of non-zero entries for flows that share an origin (destination) and have adjacent destinations (origins). Lee and Pace (2005) model retail flows using a weighted additive spatial structure between origin nearest neighbors and contiguous destinations. Inclusion of this term significantly alters both

the magnitude and sign of many variables in the model, including distance, which becomes larger in magnitude compared to an OLS specification without a spatial error term. Finally, a Bayesian hierarchical framework has been used to incorporate spatially structured random effects for origins and destinations using either a Gaussian (LeSage and Llano, 2013) or Poisson (LeSage *et al.*, 2007) probability model. In both specifications, spatial structure is defined using first order contiguity. In the former, the distance-decay effect increases in magnitude, while in the latter, the distance-decay effect decreases when comparing these specifications to corresponding models that do not incorporate any spatial effects. While various spatial error terms have been suggested for use in a SE model of SI, there is no single specification that has emerged as superior or that accounts for a consistently identifiable spatial effect.

Recent work has also proposed an extension to the SAR model to accommodate flow data (LeSage and Pace, 2008; LeSage and Fischer, 2014; LeSage and Thomas-Agnan, 2015). To include a spatially dependent process in the unconstrained spatial interaction model, LeSage and Pace (2008) suggest the following specification

$$\begin{aligned} \ln T_{ij} &= \rho_i M_i y + \rho_j M_j y + \rho_{ij} M_{ij} y + k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij} + \epsilon \\ M_i &= I_n \otimes M \\ M_j &= M \otimes I_n \\ M_{ij} &= M_i \otimes W_j = M_i \otimes M_j = M \otimes M \end{aligned} \tag{2.22}$$

where $X\beta$ becomes $\mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij}$, M_i, M_j, and M_{ij} are spatial weight matrices that define neighborhoods from the perspective of origins, destinations, and origin-destination pairs, respectively, ρ_i, and ρ_j, and ρ_{ij} are the corresponding autoregressive parameters, I_n is an $n \times n$ identity matrix with non-zero entries on the diagonal representing n locations, and $\otimes$ denotes the Kronecker product. Note that here it is assumed that there is an equal number of origins and destinations and that

all origins are also destinations and, therefore, non-zero entries in M_{ij} are denoted by scenarios where both M_i and M_j are non-zero in the case of binary contiguity (see figure 3). This specification is motivated by the theory that the movements that people decide to make are based upon their knowledge of neighboring flows in a previous time period. Assuming that the exogenous variables are relatively stable over time and that the cross-section of flows may be taken as the steady-state equilibrium of a long-run process, then LeSage and Pace (2008) demonstrate mathematically that a SAR data-generating process may be an adequate representation for spatial interaction data. In this case, the specification given by equation 2.22 should be used and several restrictions on the $\rho's$ can provide different variations of spatial structure that is ultimately included in the model. Additionally, this model is argued for on the basis that using a SAR model along with a spatial lag of the explanatory variables (i.e., SLX model) can protect against biases that might arise due to omitted variables. While it is often not possible to tell which of the two underlying mechanisms is generating the data, LeSage and Fischer (2014) differentiate between endogenous and exogenous interaction effects. Endogenous effects are theorized to be caused by shared resources such as transportation infrastructure, whereby changes in shared resources cause reactions that diffuse potentially through the entire system. The SAR model allows for this type of feedback effects, which they also be called global spillovers, and can be thought of a particular instance of the feedback misspecification originally proposed by Fotheringham and Webber (1980). In contrast, exogenous effects are thought of local spillovers since they are not caused by changes in shared resources and feedback effects that propagate beyond immediate neighbors are not expected. Exogenous effects may be modeled not by using a SAR model but by extending the SLX specification in equation 2.21 to an spatial interaction model.

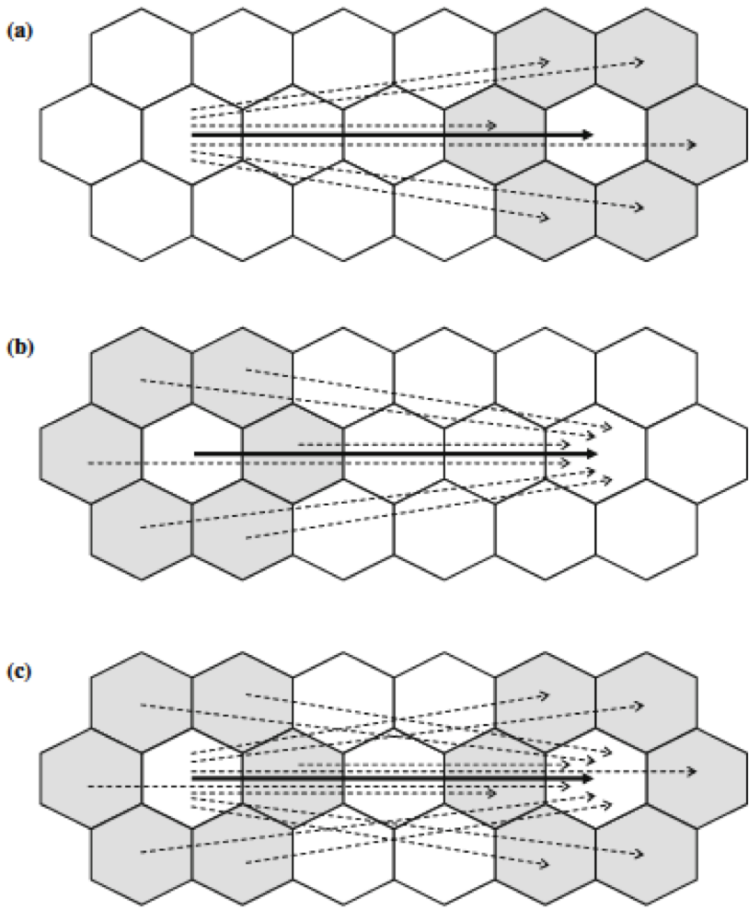


Figure 3: Definition of a) flows to nearby destinations (destination dependence); b) flows from nearby origins (origin-dependence); and c) flows between nearby origins and destinations (origin-destination dependence) according to LeSage and Pace (2008) and Chun (2008). Figure reproduced from Chun (2008).

Importantly, LeSage and Fischer (2014) also discuss that when locations are both origins and destinations and the same variable is used to represent both of them, that it is not possible to interpret how a change in a single origin attribute (destination attribute) would affect flows originating (terminating) from that origin (destination) without also considering how that change would also effect flows that terminate at that origin (originate at that destination). That is, a change in a single locational attribute can cause many changes in the volume of flows that a model would predict.

Therefore, scalar summary measures that capture the multiple changes of true partial derivatives, which are known as *effects estimates*, should be used rather than treating the regression parameter estimates as if they are partial derivatives. The effects estimates for the basic unconstrained spatial interaction model and the unconstrained SLX spatial interaction model are provided by LeSage and Fischer (2014) while those for SAR spatial interaction model are provided by LeSage and Thomas-Agnan (2015). These alternative interpretations should hold regardless of the underlying probability model or estimation technique; however, LeSage and Fischer (2014) claim that in the case where there is no spatial dependence in the endogenous or exogenous variables, using a typical interpretation of the coefficients as if they represent partial derivative is likely to have similar inferences to using the effects estimates.

Effects estimates have been shown to be stable to alternative specifications of spatial weights, even when they may result in different parameter point estimates, if there is high correlation between the specifications. High correlation between weight matrices occurs generally when the same scaling is employed (i.e., binary, row-standardized, etc.) even when the number of nearest neighbors, the distance band, or the distance-decay used to compose two different weights may vary, which is shown to be especially true for weight matrices based on higher order relations (LeSage and Pace, 2014). As a result, LeSage and Pace (2014) posit that it is a misconception that spatial regression is very sensitive to the specification of the spatial weight matrix. Instead, it is argued that the sensitivity of the point estimates should be seen as sign that the model is adjusting to accommodate changes in the spatial weight matrix, and is therefore well conditioned in terms of the stability of the effects estimates. However, these experiments were carried out for a set of areal units and it is unclear if they apply to spatial interaction models based on origin-destination based observations.

Since the focus of the work of LeSage and Pace (2008), LeSage and Fischer (2014), and LeSage and Thomas-Agnan (2015) is on the interpretation of spatial interaction parameter estimates, it is surprising how little is said here about the interpretation of the parameter estimate on the distance variable (i.e., distance-decay). LeSage and Thomas-Agnan (2015) acknowledge distance and spatial dependence may be competing to explain the same variation in the dependent variable when they observe that the distance-decay coefficient is smaller in magnitude in a SAR spatial interaction model than in a corresponding basic spatial interaction model. Similar results can be noted in additional applications of the SAR spatial interaction model (LeSage and Pace, 2008; de la Mata and Llano, 2013; Kerkman *et al.*, 2017), especially, with increasingly complex definitions of spatial structure (LeSage and Polasek, 2008), and for the SLX spatial interaction model (LeSage and Satici, 2013) as well. LeSage and Pace (2008) comment that it is not possible to compare the effects estimates for distance-decay for models with spatial lags (i.e., SAR model) to those without spatial lags because those with lags need the true partial derivatives to compute the actual effect. Curiously, the true partial derivative effect for distance-decay is otherwise not typically reported or discussed in these studies. In a study of air passenger data from Margaretic *et al.* (2017) who use a spatial Durbin model (i.e., SAR and SLX) with lags being based on either the origin or on the destinations, a counter-claim is provided that variables characterizing an origin-destination dyad, like distance, do not need the true partial derivatives for interpretation. Hence, it is not immediately clear how to interpret distance-decay in spatial regression models. Margaretic *et al.* (2017) also employ regional indicator variables that are significant with and without a spatial lag, though they admit there is no intuition for the level or sign of these effects. This denotes that there may be spatial heterogeneities in the spatial processes

underlying air travel and it may be more meaningful to use a model focused on spatial non-stationarity than spatial dependence.

Another theme that is not present in the spatial econometrics literature on spatial interaction models is that of constrained models. As already discussed, there is a history of using constraints (i.e., balancing factors or fixed effects on binary indicator variables) in spatial interaction models in order to ensure that the total inflows, total outflows, or both are preserved. Such preservations typically result in better model fit since more information is built into the model, and therefore, it is surprising that they have not been pursued in spatial econometrics. One reason for this may be that spatial econometrics is heavily based upon regression methods that assume a Gaussian probability model, where the constraints are not known to guarantee the preservation of flow totals as in Poisson regression (Arvis and Shepherd, 2013). A second reason is that there is far less work using Poisson regression for spatial interaction models within this literature. The existing work with Poisson spatial interaction models considers spatial dependence in various forms, but typically requires more complex estimators (Sellner *et al.*, 2013; LeSage and Satici, 2013; LeSage *et al.*, 2007). Still, these specifications do not use constraints and it is uncertain how such constraints would interact with the feedback effects implied by spatial dependence and the effects estimates that use true partial derivatives and scalar summary measures.

Many assessments of the general discipline of econometrics are available (Black, 1982; Hendry, 1980; Leamer, 1983). Concerns are often in terms of philosophical foundations or technical issues, which are in line with those outlined in section 2.2.5.1. Notably, Leamer (1983) concludes that some of the assumptions of econometric models are whimsical and where these models cannot be shown to be sufficiently insensitive to assumptions, then "we shall have to revert to our old methods (p. 43)",

where old methods may be interpreted as simpler or more general models with fewer assumptions. Similarly, there has been significant pushback against the SAR and SE models (McMillen, 2003; Pinkse and Slade, 2010; Partridge *et al.*, 2012; Gibbons and Overman, 2012; Corrado and Fingleton, 2012; McMillen, 2012; Halleck Vega and Elhorst, 2015) on the grounds that the models are plagued by identification issues and contain strong assumptions in order to justify their validity. Several additional perspectives are put forth to complement the use of these models, such as the use of natural experiments (Gibbons and Overman, 2012), incorporating stronger underlying theory for the model and the definition of the spatial weight matrix (Corrado and Fingleton, 2012), semi-parametric and nonparametric smoothing methods (McMillen, 2012), and to use the SLX model when there is no strong theoretical basis for the SAR model (Gibbons and Overman, 2012; Halleck Vega and Elhorst, 2015). Some advantages of the SLX model are that it has fewer issues with estimation and interpretation, is more flexible because M can be parameterized, is less likely to suffer from overfitting compared to more complex models, can rely on simpler tests for endogeneity and instrument sufficiency, and the local spillover's implied by the SLX model are typically easier to justify than the global spillovers implied by the SAR (Halleck Vega and Elhorst, 2015). Interestingly, the CD model with a single destination attractiveness variable and a Hansen-type accessibility term would be very similar to an SLX specification where $MX = S_{ij}$. However, this implies different spatial weight matrices where MX is a spatial averaging operator and S_{ij} is a distance-weighted summation operator. This simple, though foundational, difference perhaps gives rise to the very different interpretations of the associated parameter estimates where the SLX model focuses on spatial autocorrelation and the CD model arguably focuses on

more behaviorally rich concepts of agglomeration and competition in the context of spatial interaction and location choice.

2.2.5.6 Eigenvector spatial filtering

Eigenvector spatial filtering (ESF) is a technique that accounts for spatial autocorrelation based on the interpretation that the eigenvectors of a projected contiguity-based connectivity matrix [3] are the set of possible orthogonal and uncorrelated map patterns (Griffith, 1996; Griffith, 2011) given a particular definition of connectivity. Further, the first eigenvector, E_1, is the set of real numbers that produces the map pattern with the largest achievable Moran's I correlation coefficient (MC), the second eigenvector, E_2, is the set of real numbers that produces the map pattern with the largest achievable MC while remaining uncorrelated with E_1, and continues on such that E_n, achieves the largest negative MC and is uncorrelated with the preceding $(n-1)$ eigenvectors. The projected connectivity matrix, C, is most frequently defined as

$$(I - 11'/n)M_n(I - 11'/n) \tag{2.23}$$

where I is an $n \times n$ identity matrix, 1 is an $n \times 1$ vector of 1's, $'$ denotes the matrix transpose operation and M_n is the binary connectivity matrix for n mutually exclusive and exhaustive spatial units that partition the study space. While equation (2.23) is the most commonly found projection, others have been defined that ensure symmetric spatial relationships (Chun, 2008). In addition, M may be standardized using different

[3] Distance-based spatial weights have been used within the ESF technique, though it requires a distance cut-off which denotes the point at which all further relations become zero entries of the spatial weight. Furthermore, distance-based examples are based on research in the field of ecology and have not been employed in spatial interaction models (Borcard and Legendre, 2002; Legendre *et al.*, 2002; Borcard *et al.*, 2004; Dray *et al.*, 2006; Griffith and Peres-Neto, 2006; Blanchet *et al.*, 2008).

coding schemes (Boots, 1999; Chun, 2008). By selecting a subset of the eigenvectors derived from C and creating a linear combination, it is possible to produce a synthetic variable that is thought to potentially represent any missing spatially autocorrelated exogenous variables (Griffith, 2004), which can be included in a linear regression as follows (Chun and Griffith, 2011):

$$Y = X\beta + E\gamma + \epsilon \tag{2.24}$$

where Y is a dependent variable representing areal units, X is a set of explanatory variables, E is a set of selected eigenvectors, β and γ are coefficient vectors, and ϵ is a vector of normally distributed random errors. This specification has been shown to produce results where the error term does not violate independence assumptions (Griffith, 2000). ESF has also been proposed for accounting for positive spatial autocorrelation within auto-Poisson and auto-logistic model (Griffith, 2002; Griffith, 2004).

After the ESF framework was established, it was subsequently extended from spatial data aggregated to n areal units to spatial interaction flow data that occurs between n^2 pairs of origins and destinations for various types of spatial processes and in several geographical contexts (Griffith, 2007; Fischer and Griffith, 2008; Chun, 2008; Griffith, 2009b; Griffith, 2009a; Chun and Griffith, 2011; Griffith, 2011; Griffith and Fischer, 2013; Griffith and Chun, 2015; Griffith *et al.*, 2016)[4]. While this cluster of recent work might appear to indicate that a standard protocol has emerged for applying and interpreting the ESF methodology to spatial interaction data, a closer look at the literature shows that in actuality the paradigm is fraught with inconsistencies and ambiguities.

[4]In general, for these extensions, this implies that $X\beta$ becomes $\mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij}$

One theoretical inconsistency is that the primary motivations for using an ESF in a spatial interaction model is stressed to be either that flows are *a priori* dependent upon each other or that the ESF's can serve as a proxy for spatially patterned omitted variables or both. In the former motivation, this means that the proper way to account for any potential spatial autocorrelation amongst spatial interaction data is to use a SAR specification, which requires a theoretical explanation that is not typically provided. In the latter motivation, the potential autocorrelation can be remedied by using a SE specification or including the missing spatially patterned covariates. Interestingly, Tiefelsdorf and Griffith (2007) show that an ESF methodology can approximate a SAR or SE specification depending on the projection applied to the spatial weights matrix, where equation 2.23 is the projection that approximates the SAR specification. Somewhat contradictory, Fischer and Griffith (2008) demonstrate that an ESF spatial interaction model using the SAR-approximating projection can approximate a SE spatial interaction specification, though the projection corresponding to the SE model has not yet been applied within spatial interaction models. Therefore it is often unclear what the primary motivation for using an ESF framework is. In fact, much of the research using an ESF in a spatial interaction model tends to cite Curry (1972) and Griffith and Jones (1980), which has already been discussed as convoluting several modeling violations under the concept of spatial autocorrelation.

Furthermore, there are several issues with the concept of spatial autocorrelation amongst spatial interaction flows. In the ESF literature, spatial autocorrelation is often distinguished as the *local* distance effects in contrast to the *global* distance effects that are captured by distances between locations (Griffith, 2007, 2009a, 2011). Here, global and local have been emphasized because these concepts are only superficially engaged. Having an effect that occurs *locally* and *globally* implies that a process

represented by a single metric (i.e., the distance variable) occurs at two different scales and potentially varies over space. However, there has yet to be a local spatial interaction model, such as an origin-specific model, calibrated within the context of ESF's, which could be used to investigate process non-stationarity. In addition, geographically weighted techniques have recently been extended to model processes that occur at multiple scales (Fotheringham *et al.*, 2017), which could be used to shed light on the nature of scale in spatial interaction models rather than *a priori* assuming process stationarity.

Spatial autocorrelation in spatial interaction is an ambiguous concept because flows are more complex than simpler spatial units like points and areal units. Each flow is comprised of two or more locations and, therefore, the standard abstractions of spatial relationships are not sufficient. Consequently, in the ESF spatial interaction literature spatial autocorrelation is measured using Moran's I correlation coefficient, but with alternative definitions of the spatial weight matrix that consider spatial relationships between both origins and destinations. In the case that both the origin and the destination of a flow are proximal, the effect is typically taken to be either additive or multiplicative, with little theoretical justification. This definition of autocorrelation amongst flows has been designated *network autocorrelation* since a collection of interactions can be represented as a network and the work of Black (1992) is often cited as the conceptual foundation. However, the use of Black's network autocorrelation term is a misnomer in the context of the ESF spatial interaction literature. Black actually defines proximity in terms of network connectivity and not in terms of spatial proximity, which are two different representations of space. Further, the motivation for measuring network autocorrelation was for defining additional substantive geographical

variables such as regional indicators or accessibility terms, which Black demonstrated could successfully reduce the measured network autocorrelation to insignificant levels.

Defining and measuring spatial autocorrelation in spatial interaction is clearly non-trivial. This raises the question of whether or not methods based upon accounting for spatial autocorrelation, such as ESF, are sufficient extensions of the spatial interaction modeling framework. The ESF method may seem enticing since it generally results in increased model fit, but it has also been show to be prone to overfitting (Helbich and Griffith, 2016; Oshan and Fotheringham, 2016; Oshan and Fotheringham, 2017). A downside to an artificially high model fit is that it can mask the fact that other substantive covariates may be necessary, which hampers the further development of theory about spatial processes. Furthermore, Pace *et al.* (2013) demonstrate that if the true underlying data-generating process is a SAR model, an ESF may produce biased parameter estimates associated with some covariates even if it reduces bias in other parameter estimates. The degree to which an ESF may bias estimates is shown to be sensitive to the degree of spatial dependence (i.e., spatial autocorrelation captured by a spatial lag) in the explanatory variables, the degree of spatial dependence in the dependent variable, the number of eigenvectors used in the filter, and the spectra of the spatial weights matrix used in the filter, which is representative of the nature of the spatial structure of the study area. Similarly, Hodges and Reich (2010) and Paciorek (2010) demonstrate that a spatially correlated random effect, which an ESF is theorized to approximate (Fischer and Griffith, 2008), can compete with the substantive spatial effects (i.e., main effects) in a regression model. This indicates biased estimates and is shown to depend on the scale of the variability in the explanatory variables and the error term and the extent that the two are correlated. Griffith and Chun (2016) show that an ESF can help correct for omitted variable bias, as denoted by the

RESET statistic, when there are indeed omitted variables and no other known model violations. However, in empirical settings, where there are often multiple different types of violations, the RESET statistic indicates mixed results, and ultimately it is not clear what the ESF is attempting to account for. Furthermore, the ability of the ESF to account for omitted variables may be less effective with more variables and moderate multicollinearity. The work of Griffith and Chun (2016) also does not take into account the fact that the RESET statistic is known to be sensitive to spatial autocorrelation in the explanatory variables and error terms (Vaona, 2010). Thus, techniques using random effects like linear mixed models and ESF that aim to control for spatial autocorrelation may obfuscate several underlying model violations or misspecifications.

The ESF framework is dependent upon a number of specification decisions, many of which vary across existing research within spatial interaction modeling. For example, there is significant variation in how spatial relationships are defined (i.e., C and M matrices) and made operational (i.e., selecting a subset of eigenvectors E), the type of spatial interaction model the ESF is applied to (i.e., equations 2.2-2.5a), and the underlying probability model. Each of these issues is an important part of the modeling framework and changing any of them can have an impact on the model results. Further compounding the ambiguity of ESF spatial interaction models is that there is no substantial discussion of what the expected outcomes are. Table 1 captures the diversity of the ESF spatial interaction methodology specifications and their associated results, which presents the details of 22 spatial interaction models (with and without an ESF) that were extracted from 13 research articles, including data representing commuting, migration, patent citations, research collaboration, trade, and air travel. Any entries of *not reported* indicate that the necessary information

could not be found in a particular publication while bolded entries indicate that a particular detail of the ESF model was unclear and could not be concluded with certainty.

Table 1: Characteristics of Eigenvector spatial filtering methodologies applied to spatial interaction data

Source	Process	Scale (n)	spatial interaction model	Probability model	C	M	E	Pre-filter	Criterion	β effect	Significant	SE
Griffith (2007)	commuting	reported-3 (439)	UNC	Poisson	standard	contiguity	separate	> 0.25 MC	**minimize RSS or MC**	larger	not reported	not reported
Fischer and Griffith (2008)	patent citations	NUTS-2 (112)	UNC	log-normal	standard	contiguity	separate	> 0.25 MC	maximize likelihood	larger	no	larger
Fischer and Griffith (2008)	patent citations	NUTS-2 (112)	UNC	Poisson	standard	contiguity	separate	> 0.25 MC	maximize likelihood	larger	yes	larger
Chun (2008)	migration	states (49)	UNC	Poisson	symmetry-corrected	S-coded contiguity	product	not reported	minimize T statistic	smaller	no	smaller
Chun (2008)	migration	states (49)	PC	Poisson	symmetry-corrected	S-coded contiguity	product	not reported	minimize T statistic	smaller	no	smaller
Griffith (2009a)	commuting	NUTS-3 (439)	DC	Poisson	standard	contiguity	product	> 0.5 MC	**smallest p-value** (< 0.1)	smaller	yes	smaller
Griffith (2009b)	commuting	counties (254)	DC	Poisson	standard	contiguity	product	> 0.5 MC	not reported	smaller	yes	not reported
Chun and Griffith (2011)	migration	states (49)	UNC	log-normal	standard	contiguity	sum	> 0.25 MC	minimize AICc	larger	yes	larger
Chun and Griffith (2011)	migration	states (49)	UNC	Poisson	standard	contiguity	sum	> 0.25 MC	minimize Quasi-AICc	larger	yes	typically larger
Griffith (2011)	commuting	counties (73)	DC	Poisson	standard	contiguity	product	> 0.5 MC	**smallest p-value**	smaller	yes	larger
Scherngell and Lata (2012)	collaborations	NUTS-2 (255)	UNC	neg. binomial	standard	5 NN	separate	> 0.25 MC	not reported	smaller	yes	larger
Griffith and Fischer (2013)	patent citations	NUTS-2 (257)	DC	Poisson	standard	contiguity	product	> 0.5 MC	minimize AIC	larger	not reported	not reported
Patuelli *et al.* (2015)	trade	countries (64)	UNC	neg. binomial	symmetry-corrected	3 NN	separate	> 0.25 MC	**statistical sig.**	smaller	not reported	not reported
Griffith and Chun (2015)	commuting	counties (11)	DC	Poisson	standard	not reported	product	not reported	not reported	larger	not reported	not reported
Griffith and Chun (2015)	commuting	tracts (38)	DC	Poisson	standard	not reported	product	not reported	not reported	almost identical	not reported	not reported
Griffith and Chun (2015)	commuting	counties (73)	DC	Poisson	standard	not reported	product	not reported	not reported	smaller	not reported	not reported
Griffith and Chun (2015)	commuting	NUTS-1 (17)	DC	Poisson	standard	not reported	product	not reported	not reported	smaller	not reported	not reported
Griffith and Chun (2015)	commuting	NUTS-2 (40)	DC	Poisson	standard	not reported	product	not reported	not reported	smaller	not reported	not reported
Griffith and Chun (2015)	commuting	NUTS-3 (439)	DC	Poisson	standard	not reported	product	not reported	not reported	smaller	not reported	not reported
Griffith *et al.* (2016)	patent citations	NUTS-2 (257)	DC	Poisson	symmetry-corrected	8 NN	**sum**	$> +/- 0.25$ MC	**smallest p-value** (< 0.1)	larger	yes	smaller
Margaretic *et al.* (2017)	air travel	cities (279)	UNC	log-normal	symmetry-corrected	C-coded 4/5 NN	separate	> 0.25 MC	minimize AIC	larger	yes	smaller
Margaretic *et al.* (2017)	air travel	cities (279)	UNC	Poisson	symmetry-corrected	C-coded 4/5 NN	separate	> 0.25 MC	minimize AIC	larger	yes	smaller

- *UNC* = unconstrained; *PC* = production-constrained; *DC* = doubly-constrained
- *Standard* means the projection matrix given in equation 2.23 rather than one that forces symmetry
- The weight matrix M is assumed binary unless another coding scheme is noted

Since table 1 is organized roughly in chronological order, it is possible to detect some simple patterns. Initially the ESF methodology was applied to unconstrained spatial interaction models, though eventually the focus shifted primarily to doubly-constrained models. An associated trend is that originally there was a separate ESF for origin variables and destination variables; however, this was eventually abandoned in favor of a single ESF that is specified using a combination (i.e., sum or product) of origin proximity relationships and destination proximity relationships. This marks a shift in the primary motivation for using an ESF from correcting for spatial autocorrelation in the explanatory variables to spatial autocorrelation in flows themselves, though this distinction is not typically made. Further examining Table 1 reveals that outside of these patterns, *there is no standard protocol*. Indeed, almost every aspect of the ESF spatial interaction framework varies, including even the details that are ultimately reported in any given ESF spatial interaction application.

Perhaps the most varied aspect of the ESF-SI framework is the selection criterion employed to select a specific subset of eigenvectors. Forward or backward stepwise selection is always employed in the spatial interaction literature, and the selection criterion may involve directly optimizing the model fit, indirectly optimizing a model fit statistic, minimizing spatial autocorrelation, or finding all eigenvectors that are collectively statistically significant (Table 1). In addition, the collection of all eigenvectors is typically pre-filtered[5] so that only those with higher levels of positive spatial autocorrelation can be selected. In one recent case though, both negatively and

[5]A LASSO routine and a random effects variant of the ESF have been proposed that suggest more parsimonious methods for selecting a subset of eigenvectors (Seya *et al.*, 2015; Murakami and Griffith, 2015), though neither of them has been applied in the spatial interaction literature and therefore do not shed light on the variations found in existing ESF spatial interaction research. Moreover, Chun *et al.* (2016) show that an ideal number of eigenvectors is dependent upon the amount of spatial autocorrelation in model residuals and the size of the tessellation. A method for identifying an ideal number of eigenvectors is put forth, but has not been applied in the spatial interaction literature.

positively spatially autocorrelated eigenvectors are included (Griffith *et al.*, 2016). Hence, it is unclear which eigenvectors should be *a priori* filtered from the selection process and how sensitive the model results are to various filtering schemes, which is important because Chun *et al.* (2016) show that too many or too few eigenvectors can cause over- or under-correction by the ESF.

ESF spatial interaction model results are frequently deemed more intuitive than their non-ESF counterparts, though this seems inexplicable given how inconsistent some of the results are. In particular, the effect that adding an ESF has on the distance-decay coefficient estimate seems to vary extensively (Table 1). In some cases it results in a stronger distance-decay and in other cases it results in weaker distance decay. In addition, the standard errors for these estimates with an ESF are sometimes larger and sometimes smaller than the standard errors from a model without an ESF, though sometimes they are not reported at all. When the standard errors are reported, it is not always the case that the change in distance-decay is statistically distinguishable (at the 95% confidence interval) from the original distance-decay estimate. Together, this indicates that adding an ESF has an unpredictable effect on the estimated distance-decay and potentially obfuscates its interpretation rather than making it more intuitive.

Indeed, it is not possible to know what the expected behavior of an ESF should be on distance-decay without a well-developed theory and controlled simulations to test the theory, which is currently lacking (Patuelli *et al.*, 2015). Griffith and Chun (2015) do however demonstrate how distance-decay estimates vary with and without an ESF when the geographic scale and resolution are changed within an empirical example. Here the scale is the size of the region being analyzed and the resolution is the size and quantity of the areal units comprising a consistent study

region. When the scale is increased from the San Juan urban area to the San Juan metropolitan statistical area to the entire island of Puerto Rico, a doubly-constrained model without an ESF indicates that the distance-decay effect remains essentially stable while including an ESF indicates that distance-decay becomes less negative. The ESF distance-decay estimates are labeled superior, but how do we know what the expected pattern is supposed to be? In comparison, when the resolution is increased from German provinces to districts to kreises (larger to smaller areal units), a doubly-constrained model with or without an ESF indicates that the distance-decay effect becomes stronger, though the magnitude of all of the ESF models are smaller than those without an ESF. This means that the overall interpretation of the distance-decay estimates does not change with or without an ESF. Again, how can we know which scenario is more appropriate and based on what criterion? Overall, one is left wondering when exactly is the use of ESF appropriate in SI, what are the expected outcomes, especially on distance-decay estimates, and how exactly does this improve their interpretability? It is clear that a theory needs to first be defined and then tested in a rigorous simulation framework before the ESF spatial interaction framework can be accurately assessed.

2.2.5.7 Measurement error

A topic that is under-explored is the effect of measurement error in spatial interaction models. Since spatial interaction models can be thought of as a specfic conceptualization of log-linear regression or Poisson regression, we can borrow intuition from the breadth of existing work on measurement error. Here, the conventional wisdom is that measurement error on a variable within a simple linear regression

will bias the parameter estimate downwards towards the null, which is referred to as *attenuation*, and to inflate standard errors. However, the effects of measurement are dependent upon the nature of the error, the relationship between the measurement errors (and model error term), and the relationships between the explanatory variables in the model. For example, if there is correlation between a measurement error and the regression error term, this can have the opposite effect of attenuation whereby estimates are biased upward. Surprisingly, even the parameter estimates of variables that are measured without error can be biased too when they are correlated with variables that include measurement error. These biases may be further exacerbated due to multicollinearity between the explanatory variables. Thus, in more realistic scenarios of multiple regression, with collinear regressors, and potentially correlated errors, it is possible to have biased parameter estimates and precision of any nature, which can ultimately lead to real effects being hidden (i.e., false negatives), the observed data having relationships that are not present in the error-free data (i.e., false positives), and reversed signs on coefficient estimates in comparison to estimates from data with no measurement error (Caroll *et al.*, 2006). Even more sobering, is that Le Gallo and Fingleton (2012) demonstrate that in the case of spatial dependence in an explanatory variable with measurement error, and where there is a spatially correlated error term (i.e., omitted variable), that using an SE model can produce significantly more bias in parameter estimates than ignoring the regression error dependence and simply using an OLS regression. Clearly, measurement error can cause serious malaise and care should be taken to reduce it or account for it.

There are several general approaches for correcting for bias due to measurement error, which tend to require varying amounts of prior information about the nature of the measurement error, though the focus is not typically on spatial models (Caroll

et al., 2006). More recently, research that concentrates explicitly on spatial regression models has shown that measurement error in regressions using Gaussian errors causes attenuation with a strength that is in part influenced by the spatial correlation in the explanatory variables and in the error term (Li *et al.*, 2009; Huque *et al.*, 2014, 2016). Unfortunately, this research is all based on the case of a single explanatory variable and does not include the case of multiple variables or omitted variables. Furthermore, this work has not yet been extended to non-Gaussian probability models and therefore does not directly apply to Poisson regression and other distributions used for modeling counts. Therefore, the effects of measurement error in the context of spatial interaction models, which imply the use of several explanatory variables and have already been demonstrated to often suffer from omitted variable biases, is an interesting topic for explorarion.

2.3 Non-parametric spatial interaction models

A recent trend in spatial interaction modeling is the use of non-parametric techniques, which means that no parameters need to be estimated and/or there are no underlying distributional assumptions. The lack of parameters means the primary focus of these models is predicting spatial interaction rather than explanation and includes neural network spatial interaction models and "universal models".

2.3.1 Neural network spatial interaction models

Neural network (NN) spatial interaction models are an entirely separate framework for modeling spatial interaction from those previous introduced, which draw on an

analogy to neurons in biological systems (Miller, 2009). They are an attractive tool due to their universal ability to 'learn' the functions that generate observed data without any *a priori* assumptions about a parametric probability model. That is, given sufficient input data, and an appropriate neural network methodology, it is possible to approximate any data-generating process, whether it be linear or non-linear in nature. Therefore, NN's have been applied to a wide variety of classification and prediction problems. Openshaw (1993) was the first to propose the use of neural networks to model spatial interaction flows, which are often noisy, non-linear, and may vary from place to place. Consequently, NN spatial interaction models have been used in various contexts, such as predicting telecommunication traffic, journey-to-work trips, and commodity flows (Fischer, 2002; Mozolin *et al.*, 2000; Black, 1995) and generally boast higher accuracy than classic models based on max-entropy or utility theory.

There are generally three steps involved in building a NN spatial interaction model. First, a feed-forward NN architecture must be specified. For spatial interaction models, this has generally consisted of a two-layer system (Openshaw, 1993; Black, 1995; Mozolin *et al.*, 2000; Fischer, 2013), as demonstrated in figure 4. The first layer is comprised of the input weights, which consist of connections between each input node and h hidden nodes. One input node is specified for each of the model explanatory variables and h is selected through experimentation. For spatial interaction modeling, this typically results in three input nodes that correspond to an origin variable, a destination variable, and a variable to represent the 'cost' to interact from each origin to each destination. The second layer consists of the output weights, which are the connections from each hidden unit to a single output node. For each data point

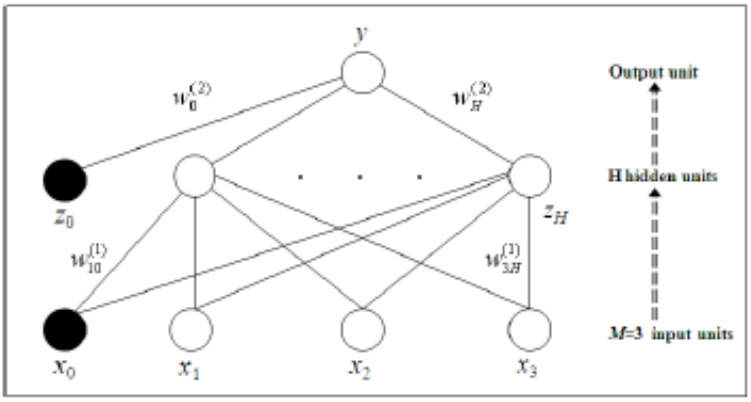


Figure 4: This figure was adapted from Fischer (2013). The first set of nodes, labeled x_m represent the model inputs, the second set of nodes, labeled z_h, represent the h hidden units, and the output node is labeled y. The nodes that are filled in black and subscripted with zero are the 'bias' units. Weights between all units are represented by lines. Weights in the first layer include those from input units to hidden units $w_{mh}^{(1)}$, and those from the bias unit to the hidden units, $w_{ho}^{(1)}$. Weights in the second layer include those from hidden units to the output unit, $w_h^{(2)}$, and those from the bias unit to the output unit, $w_o^{(2)}$.

that is fed into the neural network, a linear combination of the weighted explanatory variables, a_h, is determined at each hidden node

$$a_h = \sum_{m=1}^{M} w_{mh}^{(1)} x_m + w_{ho}^{(1)} \tag{2.25}$$

where M is the number of input explanatory variables, h is the number of hidden units, $w_{mh}^{(1)}$ represents the weight associated with a connection from input node m to hidden node h (i.e. connections in layer 1), x denotes an explanatory variable, and $w_{ho}^{(1)}$ is an additional weight that is considered at every hidden unit, often called the 'bias'. This quantity, a_h, is also called the input activation and is transformed such that

$$z_h = \phi(a_h) \tag{2.26}$$

where ϕ is the transfer function and z is the final value at each hidden node. A single linear combination of the values from all of the hidden nodes, called the output activation, is then computed

$$a_o = \sum_{h=1}^{H} w_h^{(2)} z_h + w_o^{(2)} \tag{2.27}$$

where $w_h^{(2)}$ represents the weight associated with the connection from each hidden node to the output node (i.e. connections in layer 2) and $w_o^{(2)}$ is an additional bias weight. Finally, this output activation is transformed via transfer function ψ to produce the model output, y, such that

$$y = \psi(a_o). \tag{2.28}$$

The input transfer function ϕ and output transfer ψ must be continuous and differential where ϕ is often a sigmoid function and ψ is often either sigmoid or quasi-linear. It is the nature of the transfer functions that introduce non-linearity to the approximated data-generating function (Fischer, 2013).

The second step focuses on deriving the values for the weights, also called network 'training'. Input data are first split into a training set, a validation set, and a testing set. Then training is carried out by adopting an objective function, also known as a loss function, that assesses the ability of a set of weights derived using the training data to predict the validation data set. By minimizing the loss function, it is possible to obtain an optimal set of weights. In the spatial interaction literature, loss functions are generally based on either a maximum likelihood framework or a least-squares framework (Fischer, 2002, 2006). Optimization of the loss function has been carried

out using gradient search methods or heuristics such as a genetic algorithm, each of which have different computational and accuracy trade-offs (Fischer and Hlavackova, 2004).

In order to find a level of model complexity that is generalizable and is therefore not overfit to the training data, networks with a different number of hidden nodes, h, are trained and then evaluated on the test data. This can be done simply by choosing the number of hidden nodes that produce the lowest prediction error rate on the test dataset or using more in-depth methods such as 'early stopping' or 'regularization', which seek to achieve a compromise between low error rates and generalizability (Fischer, 2013). The true prediction capabilities of the model that is finally selected may then be assessed using new data. One issue within the NN framework is how to spit data into training, validation, and testing sets. A bootstrapping technique that uses resampling with replacement is suggested over an exhaustive splitting of the data or a one-time random sample based on resampling's ability to produce standard errors and confidence intervals (Fischer and Reismann, 2002; Fischer, 2013).

Constrained variants of spatial interaction models were originally proposed within the NN framework using a two-step procedure that consisted of applying an adjustment procedure after using the basic network architecture proposed in figure 4 (Openshaw, 1993). Later, an entirely new architecture was developed to accommodate singly-constrained models within a single step (Fischer *et al.*, 2003). Instead of using the typical two-layer network, a three-layer network is adopted so that there are two layers containing sets of hidden nodes. The first set of hidden nodes uses a multiplicative combination rather than a linear combination (i.e., additive) of the weighted observations, though the second set of hidden nodes, which introduces the constraints, are additive. This architecture is then repeated for each destination in

the case of a production-constrained model, where the results from each network component are finally combined into the third layer to produce the final output. Empirical results of this new constrained neural network spatial interaction model indicate that it achieves higher out-of-sample prediction accuracy compared to the two-step constrained NN spatial interaction model and the classic constrained spatial interaction model.

Several studies have compared the predictive capabilities of NN spatial interaction models to classic spatial interaction models. While Openshaw (1993), Fischer and Gopal (1994), and Black (1995) have reported increased accuracy using NN spatial interaction models, Mozolin *et al.* (2000) demonstrate that when the test data pertains to a time period other than the time period used for the training and validation data, that classic spatial interaction models outperform NN models. This result may be due to NN models overfitting to the training data and changes in the data-generating processes over time. Importantly, Mozolin *et al.* (2000) do note that if data for the same time period is used for training, validation, and testing, that the NN spatial interaction methods do indeed provide increased accuracy over classic spatial interaction models. In the context of big data, where data with finer temporal resolutions are becoming available, the definition of the temporal resolution for a process therefore becomes increasingly prevalent.

In recent years, research pertaining to NN spatial interaction models has been limited. At the same time, more general research into NN's has blossomed and their use for machine learning tasks, such as feature extraction and prediction, is ubiquitous across industry and academia. It is likely that for most problems it would be possible to develop an advanced NN architecture with better predictive accuracy than any

parametric model, including spatial interaction models. However, this task is beyond the scope of this research and NN's will not be further investigated.

2.3.2 Universal models

The often self-described universal models take their name from the fact that their proponents suggest that they can be applied to many different types of spatial interaction phenomena (i.e., migration, commuting, commodity flows, etc.) or in various study regions using a constant set of underlying assumptions. One of the earliest universal models is the so-called "Radiation" model (Simini *et al.*, 2012), which earns its title from an analogy to radiation and absorption processes that yields

$$T_{ij} = T_i \frac{v_i w_j}{(v_i + s_{ij})(v_i + w_j + s_{ij})} \tag{2.29}$$

where T_{ij} is the predicted flows between i and j, v_i and w_j represent the population at i and j, respectively, T_i is the total number of flows starting at i, and s_{ij} is the sum of destination attractiveness for all locations between i and j (Simini *et al.*, 2012). The radiation model contrasts gravity-type models in that it does not explicitly consider distance in the computation of SI. Rather, it only uses the distance between an origin and a candidate destination as a radius from the origin to define alternative destinations. The sum of the attractiveness for all alternative destinations, usually given by their populations, is then interpreted as the deterrent of interaction compared against the attractiveness at the candidate destination, which makes the radiation model similar to Stouffer's intervening opportunities model (Stouffer, 1940; Stouffer, 1960). The derivation of the radiation model was subsequently updated to account for the fact that spatial interaction systems are finite in nature. This results in a

renormalization of 2.29 that yields

$$T_{ij} = \frac{T_i}{1 - \frac{v_i}{V}} \frac{v_i w_j}{(v_i + s_{ij})(v_i + w_j + s_{ij})} \tag{2.30}$$

where V is the total population at all of the locations in the system.

Several efforts have been made to compare the predictive capabilities of the radiation model to a variety of gravity-type spatial interaction models (Simini *et al.*, 2012; Masucci *et al.*, 2012; Lenormand *et al.*, 2016) with two important conclusions. First, the gravity model outperforms the radiation model at smaller scales, implying the radiation model is not truly universal (Masucci *et al.*, 2012). Yang *et al.* (2014) and Kang *et al.* (2015) suggest reformulating the radiation model with the addition of estimable parameters to overcome this weakness, thus further highlighting the non-universality of the radiation model. The second conclusion is that when a comparison between the radiation model and its proper gravity-type model counterpart is made, the gravity model performs better, likely due to the flexibility provided by its estimated parameters and functional form of distance-decay (Lenormand *et al.*, 2016). This is in contrast to an initial claim of superiority of the radiation model over a rudimentary gravity model (Simini *et al.*, 2012), which does not provide a level playing field for comparison.

While the radiation model has perhaps received the most attention (see also Lenormand *et al.*, 2012; Liang *et al.*, 2013; Ren *et al.*, 2014; Yang *et al.*, 2014), other universal models have also been developed. The population-weighted opportunities model (PWO) was designed as a variant of the radiation model specifically for finer scale intra-urban flows (Yan *et al.*, 2013). By recognizing the relatively higher mobility of populations within cities, it assumes that the number of trips between an origin and destination is a function of the attractiveness of the destination compared to all other destinations, striking a resemblance to the competing destinations model.

Further, a destination's attractiveness is assumed to be inversely proportional to the total population between the destination and origin, which yields

$$T_{ij} = T_i \frac{w_j(\frac{1}{S_{ji}-\frac{1}{M}})}{\sum_{k\neq i}^{N} w_{j_k}(\frac{1}{S_{ji_k}-\frac{1}{M}})} \tag{2.31}$$

where T_{ij}, T_i, w_j are as previously defined, S_{ji} is the sum of population found between j and i, N is the total number of locations, and M is the total population in the city. The model is essentially a ratio of the attractiveness of destination j from origin i to the sum of the attractiveness of all other destinations from i. It is important to note that the distance between i and j is still used as a radius to compute the sum of the population between i and j, but now the radius is centered on j rather than i, which is denoted by the reversed subscript in S_{ji} when compared to the radiation model.

A universal model of commuting networks (UMCN) has also been proposed (Gargiulo *et al.*, 2011; Lenormand *et al.*, 2011; Lenormand *et al.*, 2012). Rather than using an analytical solution, an agent-based approach is taken to allocate trips progressively according to probabilities that increase with the number of commuters that arrive at a destination and decrease with increasing distance. That is, the algorithm computes the origin-destination matrix iteratively where probabilities are updated after each iteration. The UMCN has been compared to the doubly-constrained gravity-type model because it incorporates information regarding the total inflows and total outflows into the algorithm, however, it does not respect the total inflow and outflow constraints in the predicted flows. After initializing the the origins and destinations with the total outflows (S_i^{out}) and inflows (S_j^{in}), an origin (i) and destination (j) are each selected at random, and then a trip is allocated between them with the probability

$$P_{ij} = \frac{S_j^{in} e^{-\beta d_{ij}}}{\sum_{k=1}^{N} S_k^{in} e^{-\beta d_{ik}}} \tag{2.32}$$

where distance is given by d_{ij} with an exponential function of distance-decay and β is the familiar distance-decay parameter. For every trip that is allocated, an inflow and outflow trip is subtracted from the corresponding S_i^{out} and S_j^{in}, and this continues until $S_i^{out} = 0$ for all i. Rather than calibrating β, Lenormand *et al.* (2012) suggest inferring it from a relationship between the average surface area of the areal units of the study area (ψ) and previously calibrated distance-decay parameters for 80 study regions of varying scales, which is given by

$$\beta = 0.000315\psi^{-0.177}. \tag{2.33}$$

Using this relationship makes it is possible to predict an origin-destination trip matrix using only the total inflows, total outflows, and distances. Unfortunately, this relationship only applies when considering the average areal unit surface area and is therefore not useful for making local (i.e., origin-specific) predictions. It may also be difficult to apply the UMCN when origin and destination locations are points instead of areal units.

In contrast to NN spatial interaction models, which require data-intensive model training, universal models carry out prediction of spatial interaction without the use of any prior origin-destination data, and therefore may be useful in data scarce scenarios. However, neither universal models nor NN spatial interaction models permit inference and therefore do not provide a framework for model building, hypothesis testing, or regional comparisons. That is, they do not facilitate the investigation of the spatial processes that generate spatial interaction, which is a major limitation. Furthermore, the increasing volume and availability of different types of spatial interaction data suggests that the applicability of universal models may remain limited in scope. A final drawback to the universal spatial interaction models paradigm is that it is not possible

to estimate intra-zonal flows with radiation models (Lenormand *et al.*, 2016), whereas various extensions have been proposed to do so within gravity-type spatial interaction models (Kordi *et al.*, 2012; Tsutsumi and Tamesue, 2012). Overall, non-parametric spatial interaction models are not pursued in this research, since the focus is explicitly on the substantive interpretation of parameter estimates that are either lacking or only of secondary interest in these models.

2.4 Model Assessment

In order to evaluate the fit of spatial interaction models, it has been recommended that a variety of statistics be used (Knudsen and Fotheringham, 1986). Therefore, several metrics are introduced here to be used in empirical work. For the log-normal regression specification, it is popular to utilize the coefficient of determination (R^2), though this statistic is not available within the GLM framework typically used to carry out Poisson regression. In replacement of the R^2 statistic, a pseudo R^2 can be used that is based on the likelihood function (McFadden, 1974),

$$R^2_{pseudo} = 1 - \frac{\ln \hat{L}(M_{full})}{\ln \hat{L}(M_{Intercept})} \tag{2.34}$$

where $\hat{L}$ is the likelihood of an estimated model, M_{full} is the model including all explanatory variables of interest, and $M_{Intercept}$ is the model with only an intercept (i.e., no covariates). Like the R^2 statistic, the pseudo version is at a maximum at a value of 1 with higher values denoting better model fit. To account for model complexity, there is also an adjusted version of this statistic,

$$R^2_{adj-pseudo} = 1 - \frac{\ln \hat{L}(M_{full}) - K}{\ln \hat{L}(M_{Intercept})} \tag{2.35}$$

where K is the number of regressors. If model fit does not sufficiently improve, then it is possible for this measure to decrease as variables are added, signaling that the additional variables do not contribute towards a better model fit. Henceforth, these pseudo R^2 statistics are referred to solely as R^2 and adjusted R^2. Another model fit statistic that can be used that also accounts for model complexity is the Akaike information criterion (AIC),

$$AIC = -2 \ln \hat{L}(M_{full}) + 2K \tag{2.36}$$

where lower AIC values indicate a better model fit (Akaike, 1974). This statistic is grounded in information theory, whereby the AIC is an asymptotic estimate of the information that is lost by using the full model to represent a given theoretical process.

The R^2 and AIC are designed for model selection, which means they should not be used to compare between different spatial systems. One solution to this issue is the standardized root mean square error (SRMSE),

$$SRMSE = \frac{\sqrt{\frac{\sum_i \sum_j (T_{ij} - \hat{T}_{ij})^2}{n*m}}}{\frac{\sum_i \sum_j T_{ij}}{n*m}} \tag{2.37}$$

where the numerator is the root mean square error of the observed flows, T_{ij}, and the flows predicted by the model, $\hat{T}_{ij}$, and the denominator is the mean of the observed flows and is responsible for standardization of the statistic. Here, $n * m$ is the number of origin-destination pairs that constitute the system of flows. A SRMSE value of 0 indicates perfect model fit, while higher values indicate decreasing model fit; however, the upper limit of the statistic is not necessarily 1 and will depend on the distribution of the observed values (Knudsen and Fotheringham, 1986).

One final fit statistic, a modified Sorensen similarity index (SSI), is reviewed here because it has become increasingly popular in some spatial interaction literature that deals with universal non-parametric models (Lenormand *et al.*, 2012; Masucci *et al.*,

2012; Yan *et al.*, 2013). Using the same symbol definition from the SRMSE, the SSI is defined as,

$$SSI = \frac{1}{(n * m)} \sum_i \sum_j \frac{2min(T_{ij}, \hat{T}_{ij})}{T_{ij} + \hat{T}_{ij}} \tag{2.38}$$

which is bounded between values of 0 and 1 with values closer to 1 indicating a better model fit.

2.5 Moving forward

This chapter provided a contemporary history of spatial interaction models that includes basic gravity models, the 'family' of entropy-maximizing gravity-type spatial interaction models, local models, the intervening opportunities model, the issue of spatial structure in spatial interaction and several methods for accounting for it, non-parametric models such as 'universal' models and neural network spatial interaction models and, finally, metrics for assessing spatial interaction models. From the breadth of this review it is clear that spatial interaction models have a rich history as a useful tool in applied spatial analysis and quantitative geographic thinking at large.

Several important conclusions may be drawn from this comprehensive review of the literature. First, distance-decay is a quintessential concept within human geography and much energy has been put forth to understand the intricacies associated with it. This may be most noticeable by the very active debate over distance decay in spatial interaction models in the 70's and 80's. A relative dearth of work on spatial interaction followed this period, but interest has increased in the last decade due to new data sources and increased computational power. Second, while local models and spatial non-stationarity were originally used to diagnose problems arising from spatial structure, there has been little-to-no research leveraging local spatial interaction

models in recent years. Several general frameworks have been proposed to account for spatial structure, though these newer concepts that focus on spatial autocorrelation and spatial dependence have virtually ignored the possibility of spatial non-stationarity. Related to this, these newer spatial analytical frameworks have also been relatively disinterested in interpretations of distance-decay and human behavior. A final takeaway is that simulations have been critical in building an understanding of spatial interaction models, yet few simulations have been carried out to verify newer model specifications, which have been identified to be ambiguous and produce inconsistent interpretations. Therefore, it is important to compare the primary frameworks incorporating spatial structure using simulated data to evaluate how their behavioral interpretations differ.

As previously mentioned, a renewed interest in spatial interaction models is owed to the availability of data and computational power in the age of 'big data'. On the one hand, new data has created opportunities to investigate new spatial interaction processes and build more intricate models. On the other hand, increased computational power has made it possible to build models that incorporate data with higher spatial and temporal resolutions and to apply more complex statistical techniques. In the next chapter, these themes are explored with a review of spatial interaction data within the context of era of big data.

Chapter 3

SPATIAL INTERACTION IN THE ERA OF BIG DATA

3.1 Introduction

Data representing spatial interaction are essential for studying a wide spectrum of geographic phenomenon, such as the location of services, product demand, transportation trends, and demographic dynamics. Hence, spatial interaction is a core concept within geographical research. While it has a long history, it was not until the mid-20th century that the processes underlying spatial interaction became of widespread interest to regional scientists and geographers. Following a relative 'trough' in spatial interaction research over the past few decades, there has been a renewed interest in human movement under the banner of 'human mobility'. This is primarily due to the widespread availability of spatially and temporally disaggregate mobility datasets from sources such as automated transportation systems, mobile phone records, GPS trajectories, and social media, often described under the umbrella of 'big data' (Arribas-Bel, 2014). However, this new thrust of research has moved away from trying to understand processes and tends to focus on predicting the movement of individuals (Song *et al.*, 2010b; Lin *et al.*, 2013; Pirozmand *et al.*, 2014; Do and Gatica-Perez, 2014) or establishing regularities (Brockmann *et al.*, 2006; González *et al.*, 2008; Han *et al.*, 2009; Bazzani *et al.*, 2010; Song *et al.*, 2010a; Liang *et al.*, 2012; Wang *et al.*, 2014). In contrast, spatial interaction models seek to explain and predict *aggregate* movements or flows. Their aggregate nature provides an important tool that can avoid privacy issues associated with individual-based data. At the same time, spatial

interaction models have been linked to techniques and theories for understanding individualistic behavior (Anas, 1983; Fotheringham, 1986) such that aggregate data can be used to gain insight into individuals' behavior. Furthermore, since spatial interaction models consider the attributes of a place that make it attractive as a destination, against the costs that must be overcome to travel to it, they are a key tool for understanding decision-making processes associated with movement. Knowledge of location choice processes is important for policy development, which can be useful on its own, as well as a factor within other regional models, such as land-use/land-change, market analysis, or location allocation. Surprisingly, there has been little focus on exploring the role of 'big' datasets within the spatial interaction modeling paradigm. Compared to more traditional spatial interaction data (i.e., the decennial census), it remains largely unknown whether or not these new data sources afford new insights or if they have severe limitations as sources of information about movement processes. Therefore, exploring big spatial interaction data and incorporating them into the spatial interaction modeling framework is a crucial task necessary for modernizing the geographical sciences toolkit, especially applied to urban areas, which are increasing in number, density, and importance (United Nations and Department of Economic and Social Affairs, 2014). As such, this chapter will provide a synthesis of spatial interaction data in the context of big data, outline deficiencies of some analysis methods that are employed instead of spatial interaction models, and discuss previous research pertaining to two newer sources of urban transportation data – bike-share cycling trips and taxi trips – that can potentially be used to calibrate spatial interaction models.

3.2 Defining 'big' data

Under the umbrella of 'big data' many new forms of spatial interaction data have become available. To understand these new sources, it is important to first comprehend the basic tenants of big data and then to relate them to features that are particular to spatial interaction. Defining big data, either physically or conceptually, has been problematic due to its interdisciplinary nature. Each discipline has its own idea of what should be considered big data and, therefore, what constitutes interesting research questions associated with it (Diebold, 2012, Ward and Barker, 2013). Nevertheless, many general definitions have been proposed for big data. Ward and Barker (2013) deduce three characteristics common to most definitions of big data: size, complexity, and the technologies needed to store an analyze it. They note that most previous definitions have included at least one or two of these aspects, if not all of them.

In the case where complexity is the main focus, big data need not be very large at all. Instead, the defining characteristic is that there is uncertainty regarding the ability of current tools to accommodate the analysis of the data source (Batty, 2015). This idea resonates with recent research on human mobility and new data sources. For instance, there has been much interest in how social media can be used to enhance our understanding of human movement (Kruger *et al.*, 2014; Wu *et al.*, 2014; Barchiesi *et al.*, 2015; Li *et al.*, 2015). More specifically, some researchers have directly tested movement data extracted from social media sources, such as Twitter (Lovelace *et al.*, 2014; Llorente *et al.*, 2015), or Foursquare (Noulas *et al.*, 2012), within spatial interaction models; however, these studies have only considered basic spatial interaction models. There have also been explorations of new forms of transportation data resulting from automated systems and GPS tracking such as bike trips (Wood

et al., 2011; Mooney *et al.*, 2010; Froehlich *et al.*, 2009; Hampshire and Marla, 2012; Beecham and Wood, 2013), and taxi rides (Peng *et al.*, 2012; Wang *et al.*, 2015; Gong *et al.*, 2015; Ferreira *et al.*, 2013). Similarly, these transportation-based spatial interaction data sets have only been utilized within basic gravity models (Goh *et al.*, 2012; Zaltz Austwick *et al.*, 2013; Yue *et al.*, 2012). Therefore, contemporary spatial interaction modeling approaches that capture more complex spatial processes have not been applied to these new data sources.

Another helpful big data definition, which is one of the earliest and most popular, names three facets of a dataset that may become large: volume, velocity, and variety (Laney, 2001). Therefore, this is a size-centric definition of big data. Previously, Lovelace *et al.* (2015) discussed the three V's of big data in the context of spatial interaction modeling. However, their purpose was to support the introduction of a fourth V, veracity, which is concerned with consistency between datasets and models, rather than with issues arising from the original three V's. Surprisingly, there has yet to be a formal extension of the three V's to spatial interaction data with a focus on challenges that directly arise from larger data sets. Analyzing the three V's of spatial interaction data may require us to reconsider how we define place-based attributes (origins and destinations), spatial coverage (scale and resolution) and the temporal resolution (frequency of observations) for any given model.

3.3 Properties of 'big' spatial interaction data

3.3.1 Volume

In association with spatial interaction, the first attribute, volume, refers to the spatial coverage (scale and resolution) of a data set. Historically, spatial interaction data has been available through surveys like the Unites States Census Bureau where data are collected and aggregated for a limited number of geographic locations (by state, county, tract, etc.). In contrast, big spatial interaction data have a much higher coverage such that there is no limit to the number of locations where data may be collected within any study area. This shift is primarily due to the recent explosion of sensor technology that provides cheap and efficient data collection. For instance, GPS trajectories, which have a precise x-y coordinate as a starting and ending location, can be aggregated to an endless possibility of areal units at any scale. Importantly, increased spatial coverage may become problematic since the number of possible unique interactions increases exponentially as n^2 when there are n locations serving both as origins and destinations.

One issue that arises is that larger sets of locations often result in many origin-destination pairs where no movements occur, thereby resulting in zero-inflated interaction matrices (Wilson, 2010a). For example, transit infrastructure, urban form, and personal preference give rise to the hierarchical, heterogeneous, and hub-focused nature of urban transportation (Roth *et al.*, 2011; Zhong *et al.*, 2014). More generally, network science has shown that many social activities result in patterns of preferential attachment, whereby popular locations become even more popular, often resulting in an uneven and asymmetric distribution of events over a network (Barabasi and Albert,

1999; Newman, 2013). Such zero-inflated datasets are problematic for the log-linear model calibrated via ordinary least-squares regression due to the issue of taking the logarithm of a zero observation, as well as potentially causing estimates to be biased and inconsistent (Flowerdew and Aitkin, 1982; Fotheringham and O'Kelly, 1989; Santos Silva and Tenreyro, 2006). When the zero-inflation is mild, these problems may be alleviated by adopting a Poisson model and using maximum likelihood estimation (Flowerdew and Aitkin, 1982) or by using a selection model (Linders and De Groot, 2006). Further problems of zero-inflation, which usually result in overdisperison, can be overcome by using a modified Poisson estimator such as a negative binomial model or specialized zero-inflated models (Burger *et al.*, 2009). The negative binomial model accommodates scenarios where the conditional variance is larger than the conditional mean, thereby violating the Poisson assumption of equidispersion. However, the Poisson and negative binomial models may still under-predict the number of zero flows when there are many, indicating that a zero-inflated extension may be necessary. Zero-inflated extensions provide a more sophisticated way to deal with zero flows such that the model can distinguish between flows with exactly a zero probability of occurrence, flows with a non-zero probability that have not occurred, and flows that have a non-zero probability that have occurred. Therefore, zero-inflated models can consider two kinds of theoretically different zero flows: those that could never occur (zero probability of occurrence) and those that have not yet occurred but could (non-zero probability of occurrence). That is, the zero-inflated Poisson and zero-inflated negative binomial models allow for the possibility to detach the interaction probability from the interaction volume (Farmer, 2011). Methods to deal with zero flows have been tested in various studies (Abdmoulah, 2011; Philippidis *et al.*, 2013; Tran *et al.*, 2013), though it is not yet clear that any single model is superior over all others.

Even large datasets that are not zero-inflated may complicate analysis. At the most basic level, spatial interaction's multidimensionality makes it impossible to efficiently visualize flow systems as the number of observations increases. Many techniques have been proposed to visualize spatial interaction (Ghoniem *et al.*, 2004; Holten and Van Wijk, 2009; Xiao and Chun, 2009; Rae, 2009; Wood *et al.*, 2010; Wood *et al.*, 2011; Rae, 2011; Sander *et al.*, 2014), though each requires simplifications of the data such that some information is lost. Further problems can arise in models that require computations on large matrices, such as determinants or eigenvectors, which can be problematic to calculate as the number of origin-destination pairs increases (Chun, 2008; Griffith, 2009a; Bivand *et al.*, 2013). They may require hours and days to complete a single calibration routine and in the worst cases, some models may be altogether intractable given limited computer resources. Importantly, these difficulties may arise using only moderately large datasets (Batty, 2015), and while this is not a new issue in spatial interaction modeling it is becoming increasingly important as larger volume data becomes available.

3.3.2 Velocity

If volume refers to spatial coverage, then velocity naturally defines the temporal resolution at which events are recorded. Traditional spatial interaction data are aggregated over many months or years. In contrast, automated data collection enables movements to be recorded as they occur such that they can then be aggregated daily, hourly, or even on a minute-by-minute basis. As a result, big spatial interaction data provides the means to explore the dynamics of human behavior. There is a tradition of using dynamic spatial interaction models (Harris and Wilson, 1978;

Fotheringham and Knudsen, 1986; Nijkamp and Reggiani, 1988; Clarke *et al.*, 1998; Wilson, 2010b) where the focus is on how one model component responds to changes in other components. In some cases, the output of one model run may become the input to a subsequent iteration of the model. These dynamic models typically implicitly include time, whereby each model iteration represents one unknown time step. While dynamic models allow us to investigate the possible evolution of a system, they neither expose temporal trends (i.e., rush hour in commuting), nor do they shed light on how human behavior changes over time. That is, they are particularly focused on the urban structure rather than the actors that navigate through it. Another trend is the adoption of spatial interaction models using panel data (Kim and Cohen, 2010; Chun and Griffith, 2011; Metulini, 2013; Patuelli *et al.*, 2013) to include the effects of temporal autocorrelation that may occur, typically across several years of data. Similarly, this method does not facilitate an analysis of how human behavior changes over time. Fortunately, new high velocity time series data that describes spatial interaction, as well as places, should make it possible to extend analytical capabilities to capture temporally dynamic trends. In addition, such data could provide a means to synchronize dynamic spatial interaction models with empirical data making it possible to validate policies and test planning strategies.

3.3.3 Variety

Finally, variety can be characterized by the breadth of information available to describe a particular phenomena. For spatial interaction, this includes both new types of movement data as well as new variables to describe locations, which are collected most often within high density urban environments. On the one hand, preprocessing

and aggregation can be used to infer spatial interaction from GPS trajectories, social media, and mobile phone records. While much current research illustrates the potential within these sources, effort has only recently been put forth to examine how they can be used to further understand spatial behavior (Sila-Nowicka, 2016; Siła-Nowicka *et al.*, 2016) or how they fall short in the context of spatial interaction modeling (Lovelace *et al.*, 2015). On the other hand, open data portals and web-services provide new variables representing the density of human activities and descriptions of the urban environment, and may serve as proxies to the classic place-based variable of census population estimates. For example, at any given time, in a predefined geographic area, the number of social media check-ins may be correlated with the attractiveness of the area or the number of subway entrances may be used as a local indicator of an area's propulsiveness. Since many of these place-based variables are also collected in real-time, fluctuations in their values over time may be able to explain temporal patterns in spatial interaction, which ties into the aforementioned concept of velocity. Therefore, much work still needs to be done to demonstrate which data sources are effective representations of spatial interaction and can efficiently be linked to specific locations. Another related challenge that remains is a means to tie together individual-based real-time sensed data and cross-sectional population-based census data aggregated at larger areal units (Batty, 2015; Romanillos *et al.*, 2016).

3.4 Alternative analytical methods for spatial interaction data

In this section, various methods used for the analysis of spatial interaction data are discussed. However, many of them have limitations compared to spatial interaction

models that make them less desirable for investigating behavior associated with spatial processes.

3.4.1 Visualization

Many methods have been proposed for visually exploring movement data. However, because movement data have an origin, a destination, a direction, a length, a time component, and sometimes a flow magnitude, it is often difficult to include all of the possible information encoded in movement data. Therefore, in order to detect spatial patterns, typically some form of dimensionality reduction is necessary. For example, Beecham and Wood (2013) study differences in cycling behavior between genders and Beecham *et al.* (2014) study commuting behavior of cyclers by subsetting the larger dataset and then using curved vectors and density metrics to identify important flow patterns Wood *et al.* (2010, 2011). Similarly, important aspects of the flow data may be extracted by abstracting the data as a graph and exploiting characteristics of the network or by treating the movements as two sets of points of space and using point-based clustering techniques (Guo, 2007, 2009; Guo *et al.*, 2010, 2012; Koylu and Guo, 2013; Guo and Zhu, 2014; Zhu and Guo, 2014). For these techniques, some of the less frequent trips become essentially invisible in order to detect some prevalent patterns and this tradeoff may manifest itself differently depending on the generalization method employed. Murray *et al.* (2012) propose normalizing the origins (destinations) and focusing only on the destinations (origins), which maintains all of the individual movement vectors but partially generalizes the geographic context. Another trend in flow visualization, particularly for studying migration, is that of circular plots such as cord diagrams and kriskograms (Xiao and Chun, 2009; Sander

et al., 2014; Charles-Edwards *et al.*, 2015). In this instance, it may be possible to include all potential movements in the visualization, but the data becomes completely detached from its geographic context. Finally, interactive visualization software has also been developed for analyzing flows that may include combinations of techniques, both spatial and aspatial (Yan and Thill, 2009; Rae, 2009; Boyandin *et al.*, 2011; Rae, 2011). While interactive techniques typically allow for more dimensions of the data to be explored, it is ultimately left up to the analyst to have some prior knowledge of which dimensions to interact with (Rae, 2009). Though many techniques are available for visualizing spatial interaction data, consideration of the size of the data, the specific type of spatial interaction process, spatial context, and primary research questions will ultimately play a factor in what defines a useful visualization.

3.4.2 Spatial dependence and global autocorrelation

Spatial dependence is a general term that refers to any underlying spatial processes or spatial effects that would result in proximal observations being related to each other. Measures of spatial autocorrelation, such as Moran's I and Geary's C are typically used to detect and diagnose the presence of such dependencies. In this section, global measures of autocorrelation that have been proposed to detect dependencies in spatial interaction data are reviewed. Global measures typically result in a single metric that indicate the level and nature of dependence, which contrasts local measures that provide a metric for every location or even every observation in a sample. They are sometimes mentioned in the context of spatial interaction modeling because they have been suggested as a criterion for the selection of eigenvectors within the eigenvector

spatial filtering methodology (Tiefelsdorf and Griffith, 2007; Griffith, 2007; Chun, 2008).

An initial measure of spatial dependence among spatial interaction data is put forth by Black (1992) who suggests the term network autocorrelation since spatial interaction can be represented by physical or abstract networks. Black demonstrated how network autocorrelation may be captured in data or in model residuals by specifying a network-based weight matrix within a Moran's I spatial autocorrelation statistic. Spatial associations between flows are defined by

$$M(ij,kl) = \begin{cases} 1, & \text{if } i = k \text{ or } j = l \\ 0, & \text{otherwise} \end{cases} \tag{3.1}$$

where ij and kl are two flows between origins (i or k) and destinations (j or l), which are deemed neighbors if they share either an origin or a destination ($i = k$ or $j = l$). Then a Moran's I can be calculated for the spatial interaction system as

$$I = \frac{n}{\sum_{ij}\sum_{kl} M_{(ij,kl)}} \frac{\sum_{ij}\sum_{kl} M_{(ij,kl)}(x_{ij} - \bar{x})(x_{kl} - \bar{x})}{\sum_{ij}(x_{ij} - \bar{x})^2} \tag{3.2}$$

where x denotes the magnitude of a flow, $\bar{x}$ denotes the mean of all of the flows, and n is the number of flows. Similar to a contiguity-based Moran's I, higher positive values indicate increasingly positive autocorrelation while more negative values indicate stronger negative spatial autocorrelation. Significance testing is initially based upon the typical assumptions of either normality or randomization and then using the associated analytically defined variance of I to compute a z score. The feasibility of the network Moran's I is demonstrated within an abstract flow network in a migration context where the residuals of a spatial interaction model are diagnosed with significant network autocorrelation, which is then remedied using either regional dummy variables or accessibility variables. This network autocorrelation statistic is

then subsequently used for an exploratory analysis of automobile crashes along the physical links of a highway system (Black and Thomas, 1998). As part of this analysis, it is shown through random permutations of values across the network links of a simulated dataset that the network Moran's I follows a normal distribution. However, the simulated data were crash events constrained to the network rather than an actual origin-destination flows that are constrained in geographic space.

The term network autocorrelation was more recently co-opted by Chun (2008), Chun and Griffith (2011), and Griffith and Chun (2015) who use it to refer more generally to any conceptualization of spatial dependence within abstract networks of spatial interaction. Instead of using the above definitions of proximity based on nodes on the network itself, Chun (2008) and LeSage and Pace (2008) defined origin-destination contiguity spatial weights for spatial interaction data from the perspective of origin, the perspective of destinations, and the perspective of origins and destinations, which are denoted in equations 3.3 - 3.5 below

$$M(ij,kl) = \begin{cases} 1, & \text{if } i \text{ and } k \text{ share a border} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

$$M(ij,kl) = \begin{cases} 1, & \text{if } j \text{ and } l \text{ share a border} \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

$$M(ij,kl) = \begin{cases} 1, & \text{if } i \text{ and } k \text{ or } j \text{ and } l \text{ share a border} \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

where ij and kl are two flows between origins (i or k) and destinations (j or l).

There are several problems associated with these global measures of spatial dependence for carrying out analyses of spatial interaction data. First, these methods

are univariate such that they only include flow data and do not include any location or separation variables, which are clearly important aspects of spatial interaction. Second they can only capture a single source of dependence and they do not shed any light on the contextual nature of the dependence. Finally, inference for these methods should likely be carried out using computationally expensive permutations method or other approximation methods, since analytical distributions associated with these new types of proximity weights have yet been defined (Tiefelsdorf, 2002).

3.4.3 Spatial autocorrelation in vectors

In contrast to the previous measures of aggregate spatial autocorrelation, Liu *et al.* (2014) have developed a variation of Moran's I to measure spatial autocorrelation in individual movement vectors. This statistic is enticing because it should theoretically be bereft of aggregation bias, since the observations are assumed to be disaggregated. Instead of comparing the spatial association of the volume of an origin-destination flow to all others, it seeks to compare the direction and magnitude of a single vector to all others, which can be computed as

$$I = \frac{n}{\sum_i \sum_j M_{ij}} \frac{\sum_i \sum_j M_{ij}(u_i u_j + v_i v_j)}{\sum_i (u_{i^2} + v_i^2)} \tag{3.6}$$

where $u = (x^D - x^O) - (\bar{x}^D - \bar{x}^O)$, $v = (y^D - y^O) - (\bar{y}^D - \bar{y}^O)$, the O and D superscripts denote whether the planar coordinates x and y are associated with origins or destinations, and M_{ij} is a spatial weight matrix used to define spatial association amongst flows. Within the original conception of this vector-based Moran's I, distance-decay weights are utilized to define proximity between either vector origins or vector destinations, though the authors claim any theoretically justifiable weight could be used just as well.

One potential issue with this measure is hypothesis testing because it is more complex to simulate random permutations for two-dimensional vectors. Specifically, it has been shown, using simulations, that different assumptions about the boundary of the study area result in different empirical distributions of the test statistic. Unfortunately, only a single empirical distribution is simulated for each scenario and consequently no analysis of the potential power (i.e., susceptibility to false negatives) or size (i.e., susceptibility to false positives) of the test statistic is provided. Further simulation experiments were carried out in appendix B and it was discovered that the two permutation-based methods proposed to conduct significance tests for this statistic do not behave as described by Liu *et al.* (2014). Method (1) preserves the distribution of vector magnitudes, the distribution of vector directions and the set of origin points, though it does not preserve the set of destination points. In contrast, method (2) preserves both the set of origin points and destination points, but it preserves neither the distribution of magnitudes nor the distribution of directions. In appendix B it can be seen that method (1) is too liberal and tends to reject the null hypothesis whether there is a pattern present or not and that method (2) is too indecisive and only rejects the null hypothesis by random chance, whether there is a pattern present or not. Therefore, no sufficient permutation method can be defined and this method does not seem to be useful for analyzing spatial interaction data.

3.4.4 Spatial cluster detection and local autocorrelation

A recent focus of exploratory data analysis has been the extraction of clusters of spatial interaction, which is useful for identifying important events or delineating regions. Since the goal of spatial clustering is to find sets of features that are close

together in geographic space, these methods are very similar in nature to methods that focus on spatial dependence (i.e., global spatial autocorrelation). However, a major difference is that clustering focuses on extracting specific features whereas spatial dependence measures indicate whether or not clustering occurs more generally. The breadth of available spatial clustering techniques may be classified by several characteristics. First, flow clustering research may classified by whether or not the aim is to identify clusters of individual flow events (Lu and Thill, 2003, 2008; Guo *et al.*, 2010, 2012) or aggregated flow events (Berglund and Karlström, 1999; Yamada and Thill, 2010; Guo, 2009; Guo and Zhu, 2014; Zhu and Guo, 2014) or both (Tao and Thill, 2016). Second, clustering methods may be categorized by how they define spatial clusters. This is typically done using features of a network (Guo, 2007, 2009; Guo *et al.*, 2010), by treating the origins and destinations as two separate sets of points in space and defining clusters among the sets of points (Lu and Thill, 2003, 2008; Guo *et al.*, 2012; Guo and Zhu, 2014; Zhu and Guo, 2014), or by using higher dimension measures of proximity amongst flows (Tao and Thill, 2016). This latter methodology is particularly novel and insightful, since it jointly leverages all of the information about each flow, including flow length and direction. Tao and Thill (2016) suggest the term *flow distance*, which is defined between flows T_{ij} and T_{kl} as

$$d_{ij,kl} = \sqrt{(x_i)^2 + (x_j)^2 + (y_i)^2 + (y_j)^2 + (x_k)^2 + (x_l)^2 + (y_k)^2 + (y_l)^2} \tag{3.7}$$

where x and y are the coordinates of the points or centroids that define the origins (i, k) and destinations (j, l). Kordi and Fotheringham (2016) use this definition of distance to specify spatially weighted interaction models and Tao and Thill (2016) discuss how this specification can be enhanced with tuning parameters that shift the focus in favor or either origins or destinations.

A large number of flow clustering techniques are actually local versions of global

spatial autocorrelation statistics or they were extended from previously defined local statistics. Therefore, a final way of classifying clustering techniques is by the type of statistic that has been extended. One of the earliest extensions is by Berglund and Karlström (1999) who extend the Getis-Ord G_i^* statistic to the G_{ij}^* to accommodate flows. Similarly, Yamada and Thill (2010) also provide a flow-based extension to the Getis-Ord statistic, as well as a local flow-based Moran's I extension. They demonstrate the importance of accounting for the structure of the network and the underlying population of the network for carrying out significance testing. Finally, Tao and Thill (2016) extend the local K function to accommodate flows and define a novel adaptive subsetting routine for significance testing that considers the scale of the flows under analysis. The nature of these localized flow statistics is such that a test is carried out for each observation, which means they are computationally intensive. Furthermore, when they are extended to flows, there are potentially n^2 statistics for each flow to interpret, rather than n statistics for each location, which is more difficult to map and evaluate. Since cluster detection is not a core theme within this research, these techniques will not be further engaged.

3.5 Cycling and taxi trips as spatial interaction

3.5.1 Cycling and bike-sharing schemes

In the past decade there has been an increasing interest in understanding cycling trends. This has been primarily due to the promise of cycling as an alternative mode of transportation that is environmentally friendly, decreases road congestion, and promotes physical activity. Furthermore, cycling may provide a link between

less connected nodes of public transportation systems, thereby decreasing overall commuting times (Jäppinen *et al.*, 2013). It is no surprise then that cycling has been increasing around the world (Shaheen *et al.*, 2010; Fishman, 2016a,b), especially in central cities, gentrifying neighborhoods, near central business districts, and around universities in the context of North America (Pucher *et al.*, 2011).

The majority of research pertaining to cycling has been concerned with identifying the traits of cyclists and their urban environment that characterize high up-take in the choice to cycle to work. By isolating the most important factors, policies can be developed that may be helpful in converting more of the population to become cyclists (Handy *et al.*, 2014). Within these studies, some results show that physical aspects of the urban environment, such as slope, terrain, safety conditions, and travel time are the most important (Rietveld and Daniel, 2004; Wardman *et al.*, 2007; Heinen *et al.*, 2011). Therefore, safety and convenience are more important to individuals, rather than the long term incentives, such as health and environmental sustainability (Gatersleben and Appleton, 2007; Fishman, 2016a). In contrast, a number of other researchers argue that the attitudes and perceptions of cyclists are more important for promoting more individuals to cycle to work (Handy and Xing, 2011; Daley and Rissel, 2011; Forsyth and Krizek, 2011; Guinn and Stangl, 2014). It is obvious then, that a wide range of factors may be important in the decision to cycle as a form of transportation. Additionally, it has been shown that factors may differ within and across populations. For example, it was found that the rate that individuals commute to work by bike is often heterogeneous across different ethnicities, and genders (Gardiner and Hill, 1997; Rietveld and Daniel, 2004; Beecham and Wood, 2013). Another study employed a spatial regression framework to demonstrate that bike usage in different spatial regimes, or regional clusters of municipalities, may be

explained by different factors (Vandenbulcke *et al.*, 2011). Research also shows that individuals' perceptions and attitudes may change depending upon on the length of the journey under consideration (Heinen *et al.*, 2011). These conclusions indicate that there are likely spatial heterogeneities within the decision-making process of whether or not to cycle that should be further explored.

The wide-spread establishment of bike-sharing schemes around the world has encouraged individuals to cycle by providing affordable access to bikes without the overhead of buying expensive equipment or the need to worry about theft (Bachand-Marleau *et al.*, 2012; Fishman, 2016a). These programs consist of stations where bikes can be automatically checked in and out, thereby creating a detailed log of bike usage throughout a given city. The open nature of bike-share trip data, which is typically released for free by the governing municipality or program sponsors, has resulted in much recent research about cycling (O'Brien *et al.*, 2014; Romanillos *et al.*, 2016).

Some of these efforts have sought to understand usage patterns within bike-sharing schemes. Bachand-Marleau *et al.* (2012) surveyed bike share participants and found that the primary factor affecting a participant's frequency of usage is the proximity of a station to their origin location. Another result of their study was that participants often combined cycling with other modes of transportation, with the combination of bicycle and metro being the most frequent. These findings are echoed by Padgham (2012) who found that there was a strong correlation between the number of long-distance bike-share journeys in London that start and end at a bicycle station and the number of passengers entering or exiting the nearby subway stations; however, they also found that this relationship is very weak for shorter journeys. Similarly, Martens (2004) concluded that in the Netherlands, the closer the bicycle parking stands are to public transport hubs, the more likely cycling is to be incorporated into urban

transport. Finally, Hampshire and Marla (2012) reported that the number of stations, the population density, and the size of the labor market are the factors that best explain trip generation and attraction in bike-share usage in sub-city districts in Barcelona and Seville while Faghih-Imani *et al.* (2014) use a multilevel modeling approach to provide evidence that the weather, built environment, cycling infrastructure, and time of the day are all important factors in explaining usage at stations.

There is also no shortage of research that seeks to leverage the data-rich nature of the automated bike-sharing schemes in order to better understand them and make them more efficient. A major problem in bike-sharing schemes is that they can become unbalanced - heavy commuting patterns may deplete bikes at some stations - and are no longer reliable for the users. As a result, Shu *et al.* (2013) and Schuijbroek *et al.* (2013) focused on optimizing the spatial layout of bike stations to minimize unbalancing. The former analyzes the location and capacity of stations, while the latter incorporates optimal routes for re-stocking vehicles within the overall optimal strategy. Similarly, there has been a series of attempts to extract spatio-temporal usage patterns (Froehlich *et al.*, 2008; Borgnat *et al.*, 2009; Mooney *et al.*, 2010; Vogel *et al.*, 2011), usually with the goal of developing a model to predict future usage (Froehlich *et al.*, 2009; Borgnat *et al.*, 2010; Yufei *et al.*, 2014; Yoon *et al.*, 2012). Collectively, these studies demonstrate that there are indeed spatial-temporal patterns within bike-share usage.

Much of the general cycling research discussed above is not concerned with analyzing the distribution of bike trips over geographic space because either the study was not primarily geographical in nature or because the origin-destination information was not available to the researchers. As a result, the literature is overwhelmingly concentrated with insights that help to answer the questions, "who cycles" and "how

many people cycle?", rather than "where do people cycle?" or "why do cyclists prefer certain locations?". Even when origin-destination information is available, as is the case for bike-share cycling data, spatial interaction models that can answer these latter questions are not typically utilized. One exception is a study by Zaltz Austwick *et al.* (2013) who employ a series of spatial and network methods to analyze bike share origin-destination trip data for five cities. Their study included the use of a simple gravity model from which they visualized the model residuals as indicators of locations where flows do not conform to the basic tenets of the gravity model. Consequently, it was found that trips connected to stations in the centralized areas of cities were fewer in number than expected by the model. The reason cited for this observation is that there is a relatively high density of bicycle stations in these areas (i.e., high accessibility), which may diffuse the incoming bicycle traffic over a larger number of origin-destination routes. This may be analogous to the competition forces that are commonly accounted for within the competing destinations model (Fotheringham, 1983a) and therefore warrants further research.

3.5.2 Taxi trips

While there seems to be less substantive research into taxi usage itself compared to cycling, the relatively recent addition of GPS technology into taxi fleets has transformed taxi trips into an increasingly popular transportation dataset. This is probably due to three reasons. First, the data become more accessible. Many taxi datasets used to be available only to the personnel of municipal organizations that manage them. If a researcher or analyst wanted to use the data they needed to have an inside collaborator or go through the process of filing a data request if the

municipality has open data and transparency laws, such as New York State's freedom of information law (FOIL) (Whong, 2014). These kinds of requests have resulted in the general release of urban transportation data in machine readable format and through easily searchable data portals (TLC, 2017; NYC, 2017). Second, is the rise of the civic hacker. Civic hackers have led the way in investigating new datasets, such as taxi trip data, and providing tools for harnessing these massive databases that can be difficult to manage (Schneider, 2015; Shekhar, 2017). While the exploratory analysis of civic hackers may seem superficial to some extent, the often elegant visualizations they produce have received a lot of media attention, which ultimately promotes further research. Thirdly, taxi trips have become popular in applied work because they provide a dataset with high temporal frequency and extensive geographical coverage. For an example that illustrates the culmination of these three factors, see figure 5, which is a visualization of taxi pickups in New York City produced by a civic hacker using open data and where the density and coverage of the data are so great it is possible to see most of the city's street network despite the fact that the data are comprised of points.

A wide array of applied research has been carried out using taxi trip datasets. Under the banner of human mobility, methods have been devised for describing the collective nature of taxi trips (Liang *et al.*, 2012; Liu *et al.*, 2012; Peng *et al.*, 2012; Zheng *et al.*, 2015). Several interesting observations result from this work. First, taxi trajectories tend to exhibit strong daily rhythms and different patterns exist for different travel purposes (Liu *et al.*, 2012; Zheng *et al.*, 2015). Second, urban shape influences the direction and distance distributions of trajectories instead of being uniformly distributed, which leads Liu *et al.* (2012) to conclude that, given an origin and destination, the probability that a trip occurs between them depends

Figure 5: Visualization of taxi pickup locations in New York City from 2009 to 2015. This image was reproduced from http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/

on their populations and the distance between them. This conclusion encapsulates the assumptions underlying basic spatial interaction models, such as the gravity model. Other applications using taxi trip data include inferring trip purpose based on spatial and temporal patterns (Gong *et al.*, 2015), community detection (Liu *et al.*, 2015), predicting congestion (Wang *et al.*, 2015), and event detection using time-series decomposition (Zhu and Guo, 2017). Finally, various researchers have created visualizations to capture trends and anomalies associated with the spatial, temporal and more semantic characteristics of taxi trajectories (Ferreira *et al.*, 2013; Savage and Vo, 2013; Chu *et al.*, 2014). In particular, Ferreira *et al.* (2013) and Savage and Vo (2013) employ interactive visualization software to visualize New York City taxi data and Ferreira *et al.* (2013) discuss how trips cluster around major transportation hubs, with overall patterns changing during holidays, extreme weather, or major events that lead to road closures.

Similar to bike-share trip data, this previously discussed research does not typically include a spatial interaction modeling framework with the exception of Yue *et al.* (2012) who process taxi trajectories into aggregate trips to several shopping centers in the city of Wuhan. They then effectively calibrate Huff-style gravity models (Huff, 1963, 1964) that predict the percentage demand at locations using shopping center size as a proxy for attractiveness against the cost of distance needed to arrive at the shopping center. However, this is a rudimentary spatial interaction model and Yue *et al.* (2012) conclude that "Additional research is needed to identify shopping center attractiveness factors and a proper spatial interaction model to better depict the relationships". Therefore, more research regarding the appropriateness of taxi trips in spatial interaction models is needed.

3.6 Moving forward

This chapter covered a review of spatial interaction data in the era of 'big' data. First, it outlined several definitions of big data, especially pertaining to features of spatial interaction data and models. Next, it reviewed issues with some alternative spatial interaction analysis methods other than spatial interaction models. Overall, these techniques are limited in their ability to analyze spatial interaction data and this highlights the need to calibrate spatial interaction models to understand spatial processes and behavior. Finally, it explored two newer types of spatial interaction data – bike-share trips and taxi trips – that fit many of the definitions of big data and have been used surprisingly scarcely within spatial interaction models. Consequently, these two data sets will be tested in several spatial interaction modeling frameworks in the subsequent chapters. The primary interests are a) the extent that they can be employed in spatial interaction models and b) what can be gained from their increased spatial and temporal resolutions.

The next chapter will present the details of the data that will be analyzed throughout this research. This includes the basic spatial and temporal trends of bike-share trips and taxi trips in New York City, as well as locational attributes to describe destination attractiveness. Both traditional census variables and alternative variables available through open data portals will be considered for locational variables.

Chapter 4

SPATIAL INTERACTION IN NEW YORK CITY

4.1 Introduction

This chapter introduces the empirical data associated with New York City (NYC) that will be used throughout this research. This includes an overview of the study area, movement datasets, and several locational attributes that are used to measure the propulsiveness of origins and the attractiveness of destinations. Movement data and locational attributes are necessary to calibrate spatial interaction models to obtain parameter estimates and ultimately analyze spatial behavior. As a result, more detailed and diverse data may allow us to carry out a more in-depth analysis, which can lead to a better understanding of movement processes. For each type of data introduced, the method of collection and preparation will be outlined, visualizations of spatial and temporal features will be provided, and any potential issues or limitations in the context of spatial interactions models are discussed.

4.2 Study Area

NYC is the most populous city in the United States with an estimated population of 8,175,133 in 2010. It is also the most dense urban metropolis in the United States with a relatively small metropolitan area of approximately 302.64 square miles in 2010 (population density of 27,012.5 per square mile)[6]. The city is divided into five

[6] $https://www.census.gov/quickfacts/fact/table/newyorkcitynewyork/PST045216$

boroughs – Manhattan, Brooklyn, Queens, The Bronx, and Staten Island – that each have different demographics and geographical features (figure 6).

Manhattan (figure 7) is perhaps the most famous of the five boroughs due to its historical role as the center of the city and as a global financial center. It is also home to several famous attractions such as Central Park, Time Square, and The Empire State Building. While Manhattan is an island, it is connected to the Bronx to the North via bridges and rail transit, to Brooklyn and Queens to the east via bridges and rail transit, and to Staten Island to the south via pedestrian ferry services. Though there is no direct road connection between Manhattan and Staten Island, they are linked indirectly through roads and bridges via Brooklyn. Manhattan is also connected to New Jersey in the West via bridges and rail; however, this research will be limited to spatial interaction within NYC and interactions to and from New Jersey will not be considered.

Brooklyn and Queens (figures 8 and 9) contain the majority of the City's population with approximately 2,504,700 and 2,230,722 residents, respectively. Much traffic runs through these two boroughs due to commuters traveling to Manhattan for work or people moving to and from the two airports (JFK and La Guardia). Manhattan and the Bronx (figure 7 and 10) are the next two highest populated boroughs at approximately, 1,585,873 and 1,455,720 residents, respectively. Finally, Staten Island (figure 11) is the least populated borough with approximately 476,015 residents, which is also an island and is relatively isolated in terms of necessary travel distance and available transportation options in comparison to the other boroughs[7].

Throughout this research the spatial units used for analysis will be the 2010 census

[7] *http : //www1.nyc.gov/site/planning/data − maps/nyc − population/current − future − populations.page*

tracts from the United States Census Bureau. Census tracts are relatively homogenous in terms of demographics and contain an average of about 4,000 inhabitants[8]. They are not the smallest available division of population information for the United States, though they do provide a much finer scale of analysis than the county and state boundaries that are often employed in spatial interaction modeling (see for example (Chun, 2008; Griffith, 2009b; Chun *et al.*, 2012)). In addition, smaller units of analysis, such as census block groups and census blocks are too large in number for efficient computation, since they result in billions of origin-destination pairs that comprise observations of potential trips and the vast majority of these will be zero. In contrast, the 2,166 census tracts of NYC, displayed in figure 12, result in 4,691,556 potential origin-destination observations. It can be seen that Manhattan, with the most dense population, tends to consistently have the smallest census tracts and Staten Island, with the least dense population, tends to consistently have the largest census tracts.

[8] $https://www.census.gov/geo/reference/gtc/gtc_ct.html$

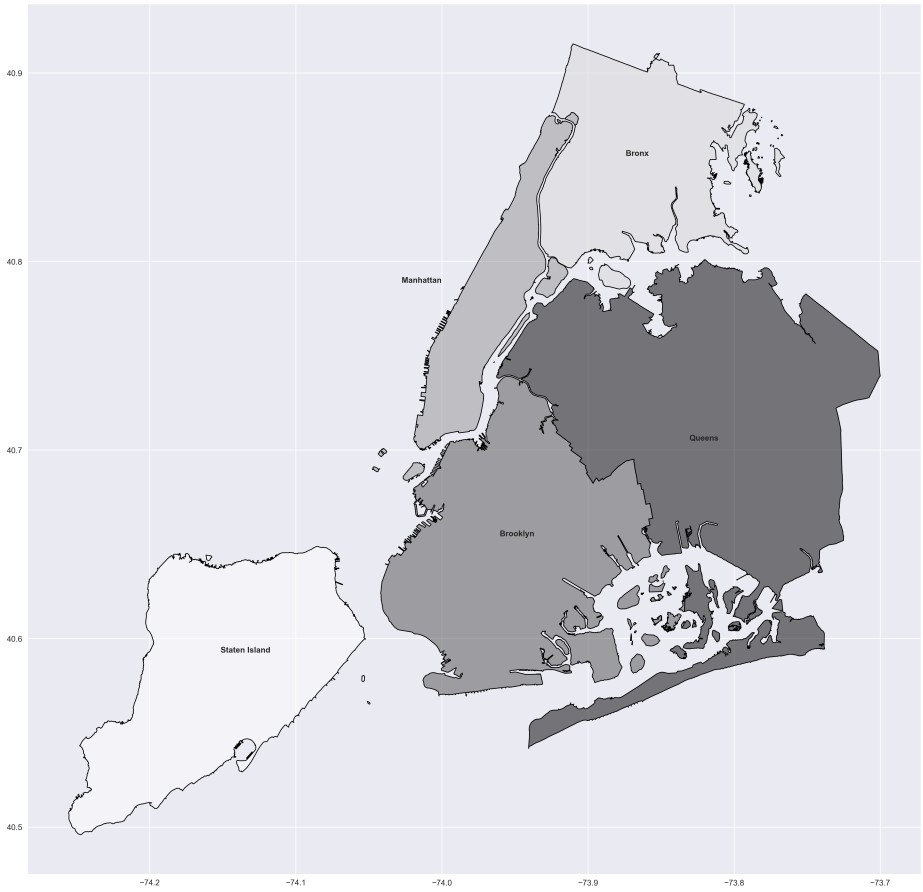


Figure 6: The five boroughs of New York City

Figure 7: Manhattan borough

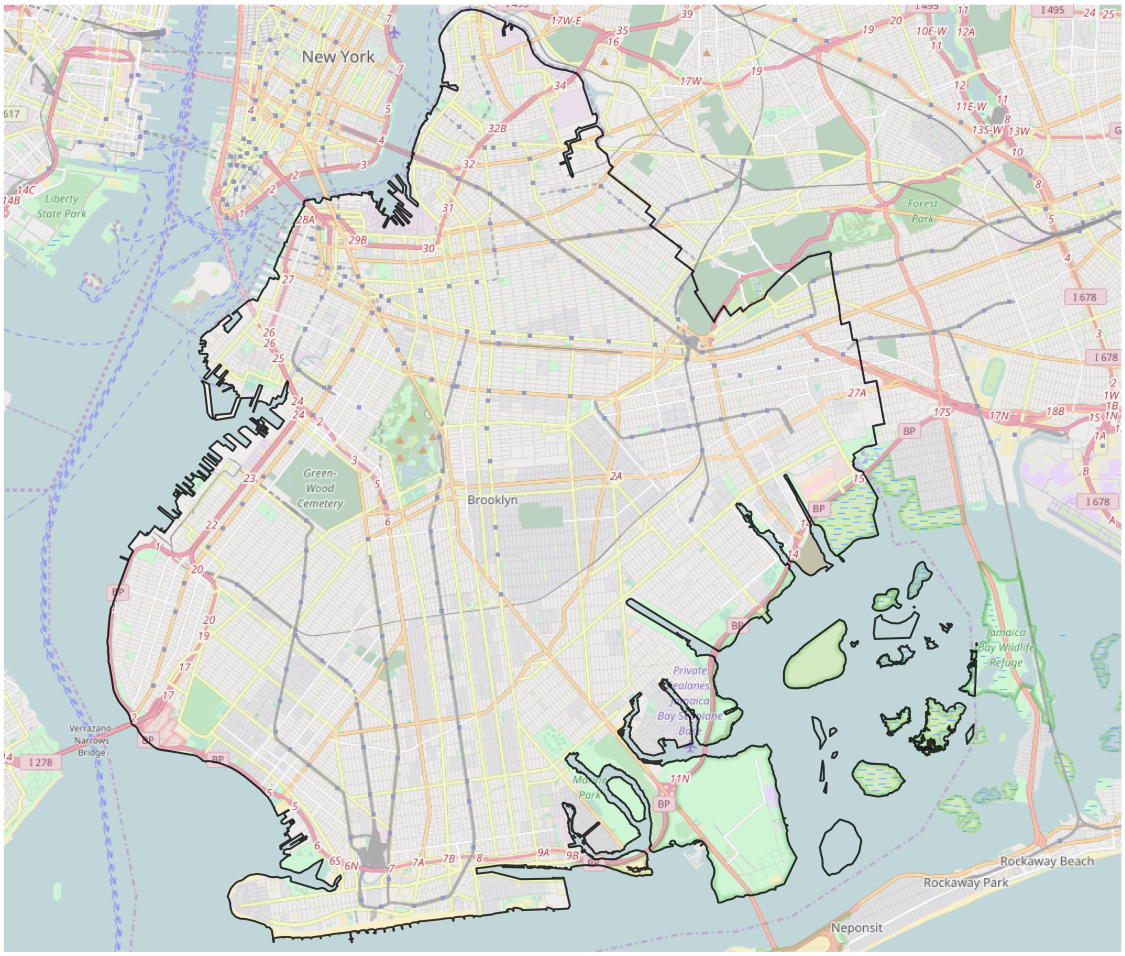


Figure 8: Brooklyn borough

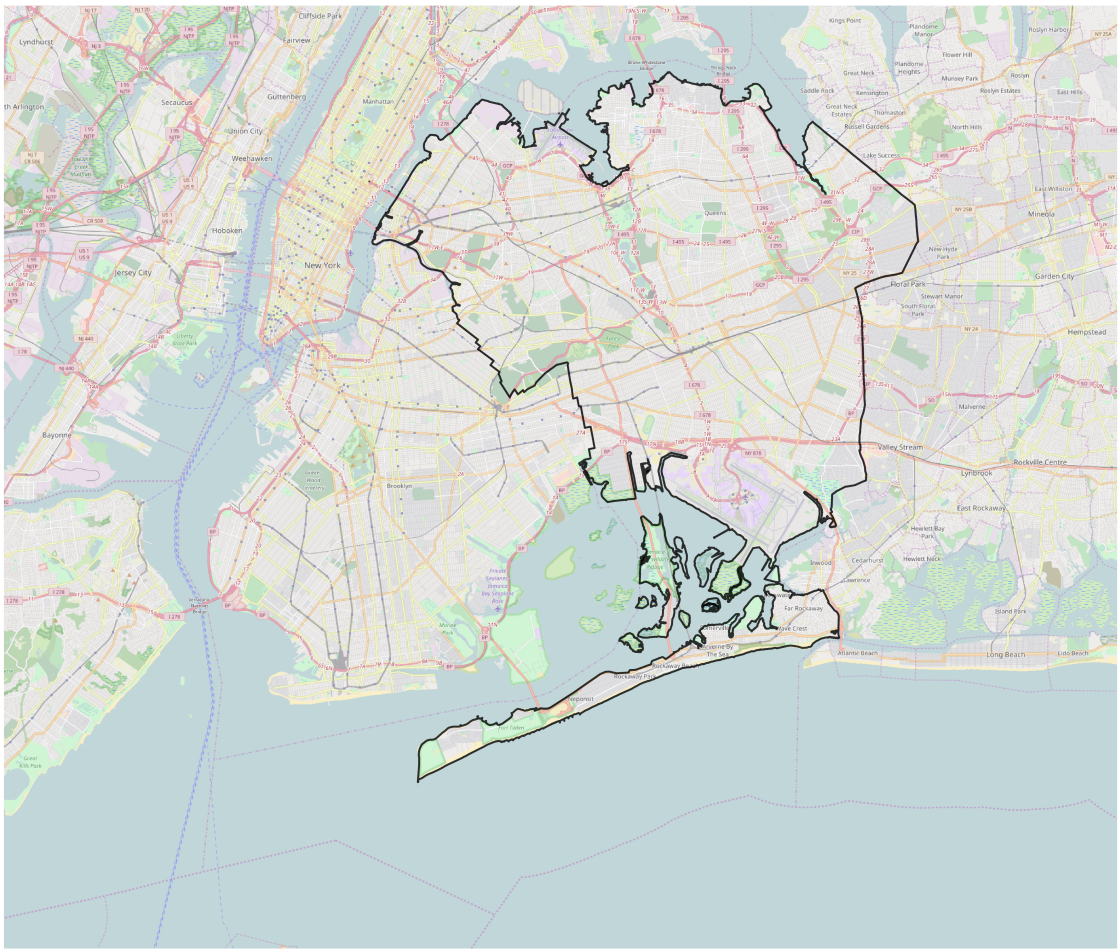


Figure 9: Queens borough

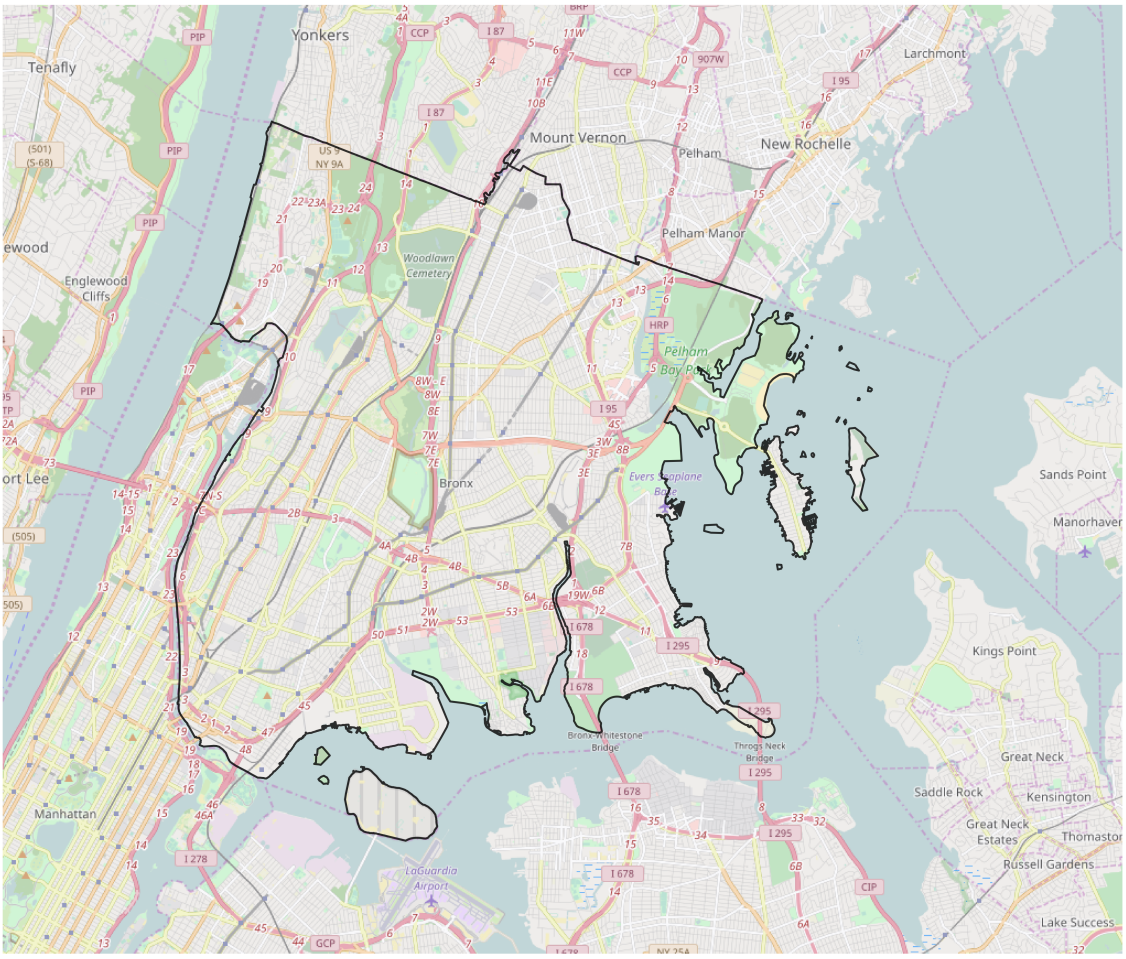


Figure 10: The Bronx borough

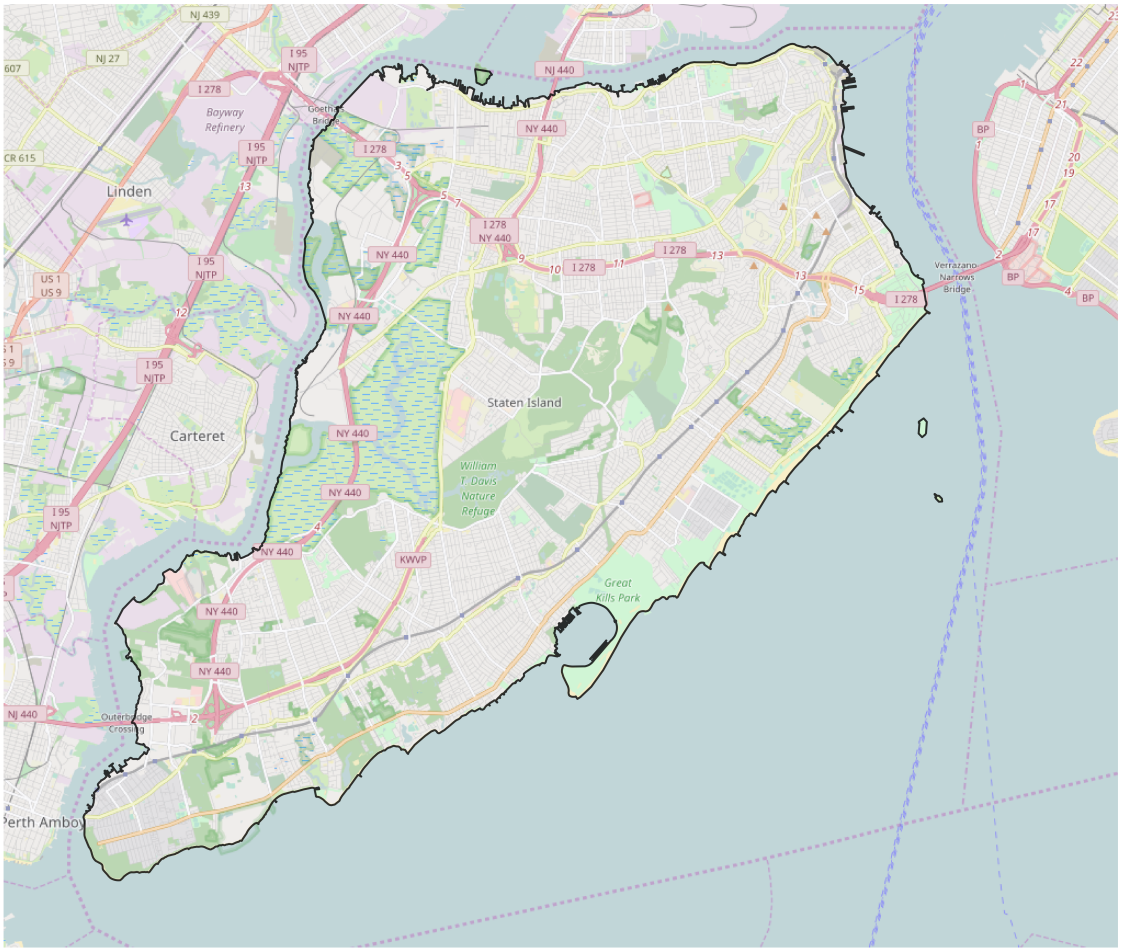


Figure 11: Staten Island borough

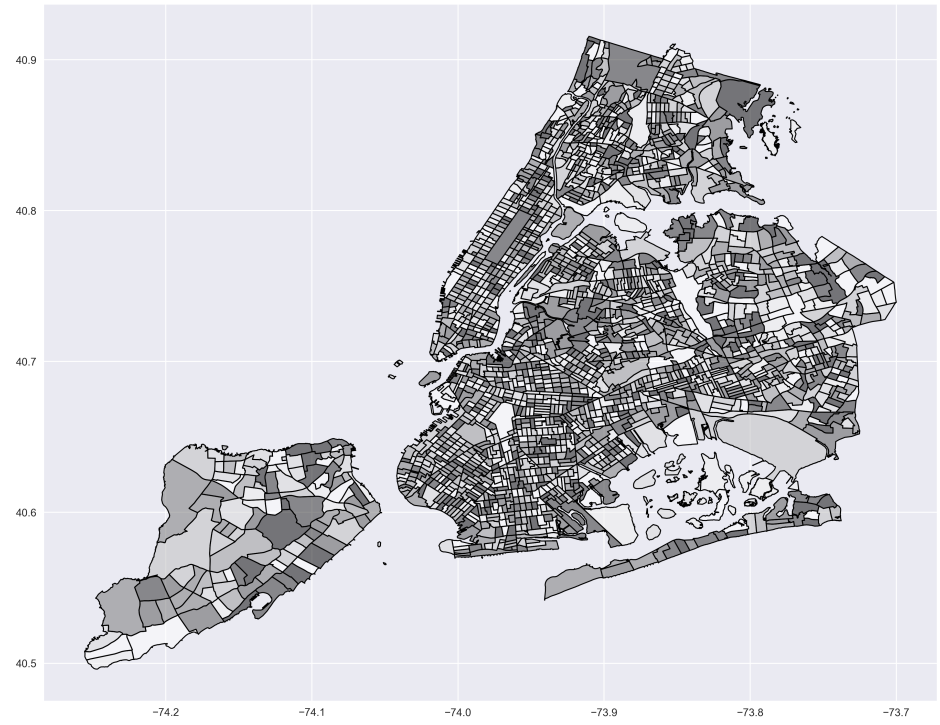


Figure 12: Census tracts in New York City.

4.3 Spatial Interaction Data

Three spatial interaction datasets will be utilized. First, CITI bike-share cycling trips and taxi trips will be presented, which both represent new forms of spatial interaction data that are available at finer spatial and temporal resolutions than traditional spatial interaction data. Second, commute-to-work survey data from the census are presented, which is considered a traditional spatial interaction data source. The census data do not contain any time attribute and is therefore limited to a single temporal aggregation of several years. In contrast, the bike and taxi trips can be flexibly aggregated to any temporal unit. All data in this work has been limited to the time frame from June 2014 through May 2016 to keep the number of data rows that need to be processed from becoming too large. Within this range, temporal units of months, weeks, days, and hours are explored.

4.3.1 CITI Bike-Share Trips

The CITI bike-share trip data[9] for the time period of this work consists of approximately 20 million bike trips. The number of stations in the system at the beginning of the study (May 2015) was approximately 250 but increased to approximately 500 by June 2016. Station locations can change slightly over time as the system is continuously monitored and optimized, though these changes tend to be very minor, such as a relocation of a station to a different corner of the same intersection. In the situation where a station has been relocated during the period of the study, an average of the two sets of coordinates is taken as the location. In one extreme case,

[9]https://www.citibikenyc.com/system-data

a station was moved across the city while maintaining the same station id (station 3016), and this station was excluded from the analysis. Each bike trip consists of an origin station, a destination station, trip duration, a start time, an end time, cyclist gender and whether or not the user is a subscriber or a one-off user. Since station coordinates are provided as a point in space (latitude/longitude) and the trip start time provides information up to the precise second, this data can be aggregated to any larger spatial or temporal unit. Aggregating this data into census tracts results in trips between 247 census tracts, which can be analyzed both spatially and temporally.

The data can be viewed as a time series and can be sampled monthly, weekly, or daily (figure 13) over the two year period. First, by sampling the data monthly (figure 13 top), it is possible to extract distinct seasonal weather trends where there are more bike trips in the warmer months and fewer trips in the colder months. It is also possible to observe a general trend of increasing usage, which may be attributed to the expansion of the number of bikes and stations and increasing popularity of the bike-sharing system. Next, by sampling the data weekly (figure 13 middle), more detail is added and small spikes or drops in the number of trips are apparent, which might be associated with fluctuations in the weather, such as the presence of precipitation, as well as the effects of public events or holidays where people are either more stationary or out of town. The big dips in the first and last weekly observation are artifacts of the data, since the data were subset using daily break points and these weeks might not have seven days of data. Finally, by sampling the data daily (figure 13 bottom), even more detail is added to the time series and the trends found in the weekly series are enhanced. Though there are relatively few large spikes above the average usage trend, there are several large dips that are likely related to inclement weather. In fact, the dashed red line corresponds to a blizzard that occurred January 22nd-24th and

resulted in approximately 27.5 inches of snow [10]. This blizzard, along with typically cold and windy weather, resulted in January having the fewest rides for the Winter of 2016 (figure 13 top). It can be seen that the rides actually continued to decline after the storm initially started on the $22nd$, since the bike-share program was closed for safety reasons and remained closed for several days while the snow was removed from the roadways. This closure is evident in the daily sampling (figure 13 bottom) where there is a discontinuity in the series.

Additional temporal trends can be extracted by aggregating the data by the day of the week or the trip start hour of the day (figures 14 - 15). If the trips are aggregated solely by the day of the week (figure 14), it can be seen that on average there are fewer bike trips made on the weekends than during the week, which is likely due to weekday commuting trips. Furthermore, if the trips are aggregated solely by the hour of the day (figure 15), then the intra-daily commute patterns become clear. Average usage tends to increase starting around 5:00am through 5:00pm with spikes around 8:00am, and 5:00pm for the morning and evening rush hour commutes, respectively. The average number of trips then declines from 6:00pm until 4:00am. The trips can also be aggregated by grouping them by the hour of the day and the day of the week (figure 16), which show some additional interesting patterns. Here, it can be seen that weekend trips follow a different hourly trend than weekend days trips, which still exhibit morning and afternoon rush hour peaks. Instead, weekends trips tend to more gradually increase throughout the morning and early afternoon with one smooth peak around 2:00pm in the afternoon and a gradual decrease in trips throughout the evening, though with more late-night trips. These trends indicate that trips over the weekend are associated with leisure activities.

[10] https://www.weather.gov/media/okx/Climate/CentralPark/BiggestSnowstorms.pdf

The bike stations are currently only available in portions of Manhattan, Brooklyn and Queens with the highest number of stations in Manhattan, the second highest number of stations in Brooklyn, and only a few stations in Queens (figure 17 top left). In addition, Manhattan stations tend to have a higher capacity of bike docks than stations in Brooklyn and Queens (figure 17 top right). It is also possible to see that although station use is high throughout Manhattan, there is exceptionally high bike dock capacity along Broadway Avenue, which runs from the northwest to the southeast. However, this trend is not apparent when the number of bike docks is aggregated to census tracts (figure 17 bottom left). Some of the larger tracts stand out as having the highest number of total docks, though this trend is removed when the total capacity in each tract is normalized by the area of the tract (figure 17 bottom right). Comparing the bottom left and bottom right maps of figure 17, it is apparent that several larger tracts are the least dense in terms of the number of docks. For example, in Manhattan, Central park (long tract in the center) and the Chelsea Piers (west coast) exhibit this feature. These are all places that receive some traffic due to industrial infrastructure and leisure activities, but have low residential populations.

Similar trends are noticeable when visualizing the total outflows and inflows (figure 18 top left and right) and the density of outflows and inflows (figure 18 bottom left and right). Comparing the totals (top) to the density (bottom), large tracts such as Chelsea Piers and Central Park again have some of the highest totals, but much smaller densities. Moving to densities from total counts of bike trips, also has the effect of highlighting the high traffic in some smaller tracts. Interestingly, it is possible to see the northwest to southeast trend along Broadway Avenue in the density of outflows and inflows, which corresponds to the number of docks available at individual stations. Overall, these trends seem to be very similar for both outflows (left) and

inflows (right); however, if the data are subset for only trips that start during the primary morning commuting hours (6:00 am - 10:00 am) additional patterns emerge (figure 19). While outflow density (figure 19 bottom left) seems to tend towards being randomly distributed, there is a clear pattern of a higher density of trips ending (inflow) (figure 19 bottom right) on the east side of Manhattan where there is known to be many skyscrapers serving as office space. Furthermore, there is a clear cluster of low density trip ends on the most easterly portion of Manhattan. This area is known as Alphabet City, since it consists of several avenues, each denoted by a single letter, and is known to be primarily a residential neighborhood where few commuting trips would be expected to terminate.

Several variables may be used as a proxy for cost associated with making a trip between a particular origin and destination (figure 20). Costs in spatial interaction models are often theorized to be related to the distance traveled, the most basic of which is the Euclidian distance between locations (figure 20 left), which is also sometimes called the "straight-line" distance or "as-the-crow-flies" distance and alludes to the fact that it is the shortest possible distance between two points. Since the underlying road network and urban environment are not accounted for within this type of distance, it is the simplest to compute, but typically underestimates distance, and therefore, the underlying cost. A synthetic distance that is created by routing a cyclist over the transportation network may also be used as a more realistic distance in lieu of physically recording the distance as each trip occurs. To do this, Mapzen's *Matrix* routing service[11] was employed on the bike station coordinates (figure 20 middle). This service takes coordinates as input and returns the distance needed to travel on the transportation network provided by OpenStreetMap's database of volunteered

[11]https://mapzen.com/documentation/mobility/matrix/api-reference/

geographic information. To determine the overall route and specific transport network links utilized, the routing service has a series of "costing" parameters, such as the preferred type of roads. For this research, the default parameters were used, which tend to favor roads with bike paths and bike access, though it is not limited to them. This has the effect of generating more realistic trip distance, and comparing them (figure 20 middle) to the Euclidian distances (figure 20 left), it can be seen that they have a similar distribution shape, but that the network-based distances are typically longer than the Euclidian distances. This confirms that Euclidian distance often under-estimates the distance required to travel between two locations in an urban environment. A limitation of these distance measures and the bike data more generally, is that full journeys likely to start at residences and end at places of work or leisure rather than stations, which means some of the trip information is not directly available.

In addition to distance, time is frequently used as a proxy to distance or cost. Though the time of each bike trip is recorded as the difference between when the bike is checked out of a station and when it is checked back in to a station (figure 20 right), this attribute cannot be employed in spatial interaction models, since there is only trip duration available for those origin-destination routes have been observed and it would be inappropriate to assign a trip duration of zero to origin-destination routes with no observations. Therefore, analysis will be limited to the two distance measures previously introduced.

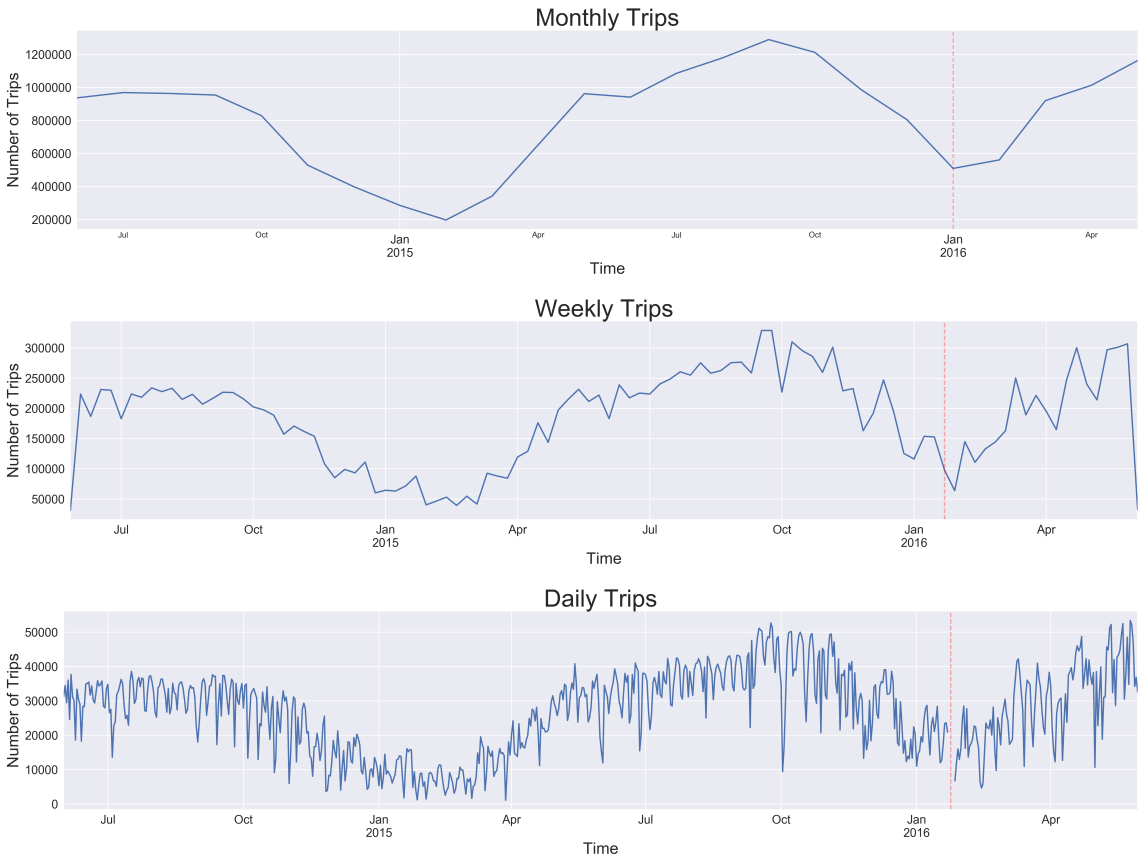


Figure 13: Bike trips by month (top), by week (middle), and by day (bottom) throughout the case study time period.

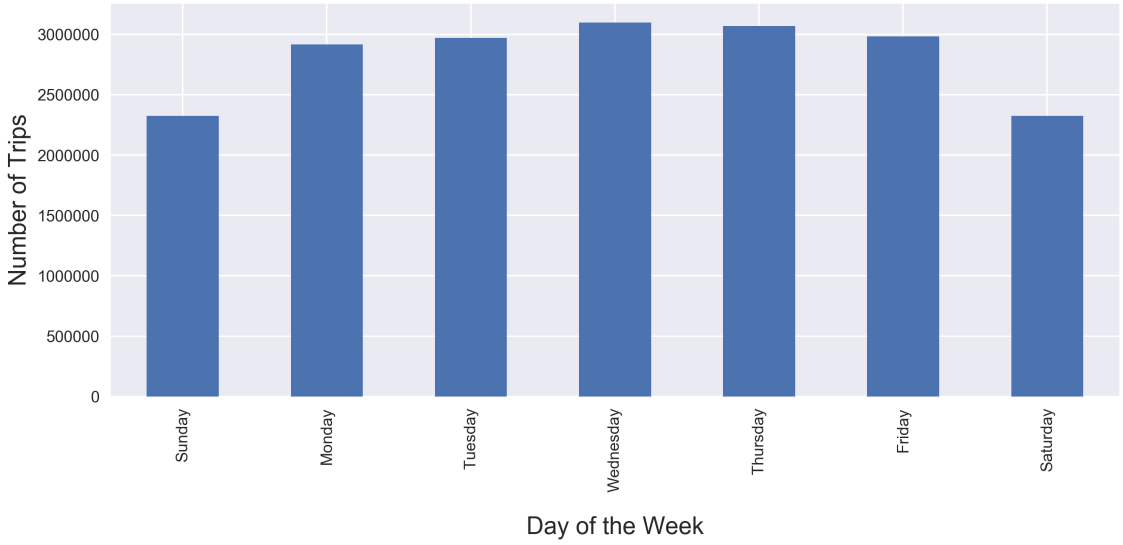


Figure 14: Bike trips by the day of the week.

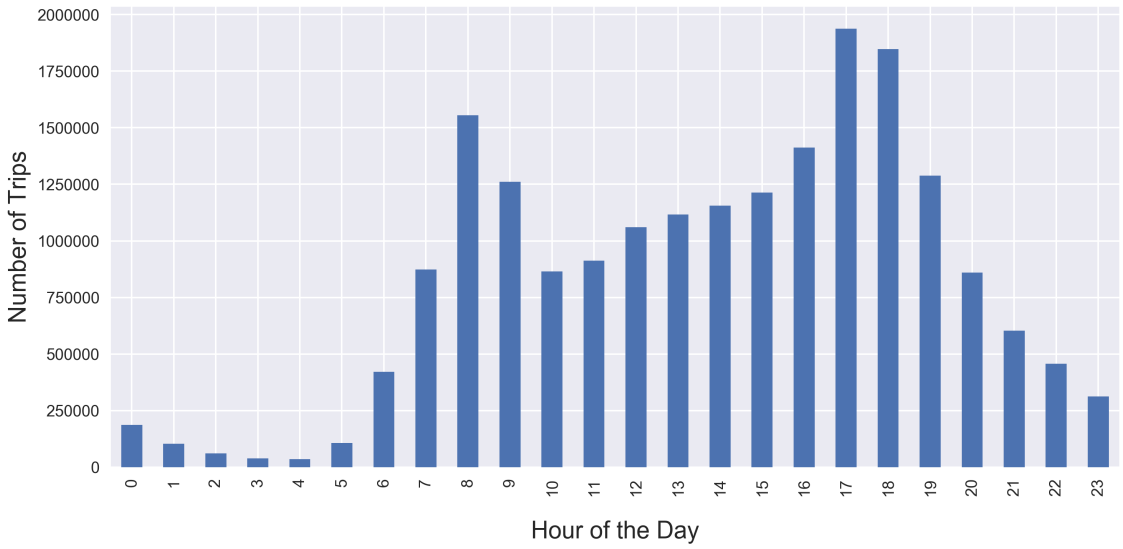


Figure 15: Bike trips by the hour of the day.

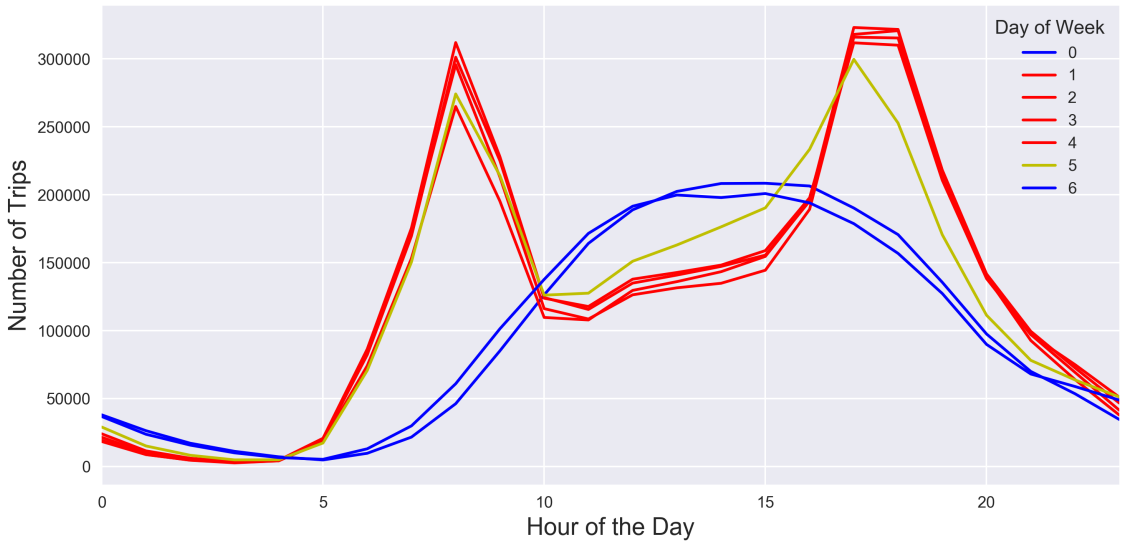


Figure 16: Bike trips by the day of the week and the hour of the day. Monday-Thursday trips are red, Friday trips are in yellow and Saturday and Sunday trips are in blue.

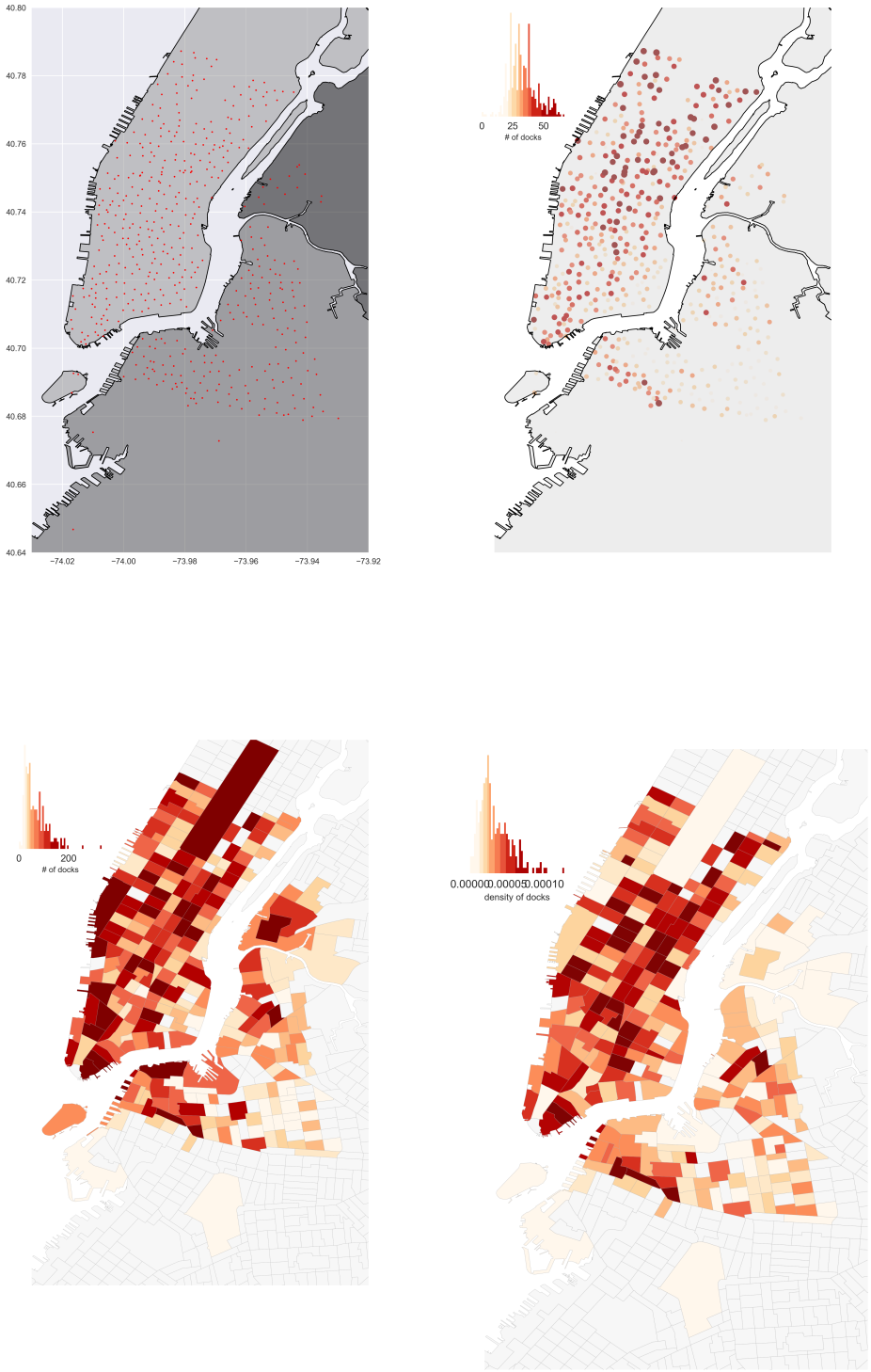


Figure 17: Bike station locations (top left), the number of bike docks located at each station (top right), the number of bike docks located in each census tract (bottom left) and the density of bike docks in each census tract (bottom right).

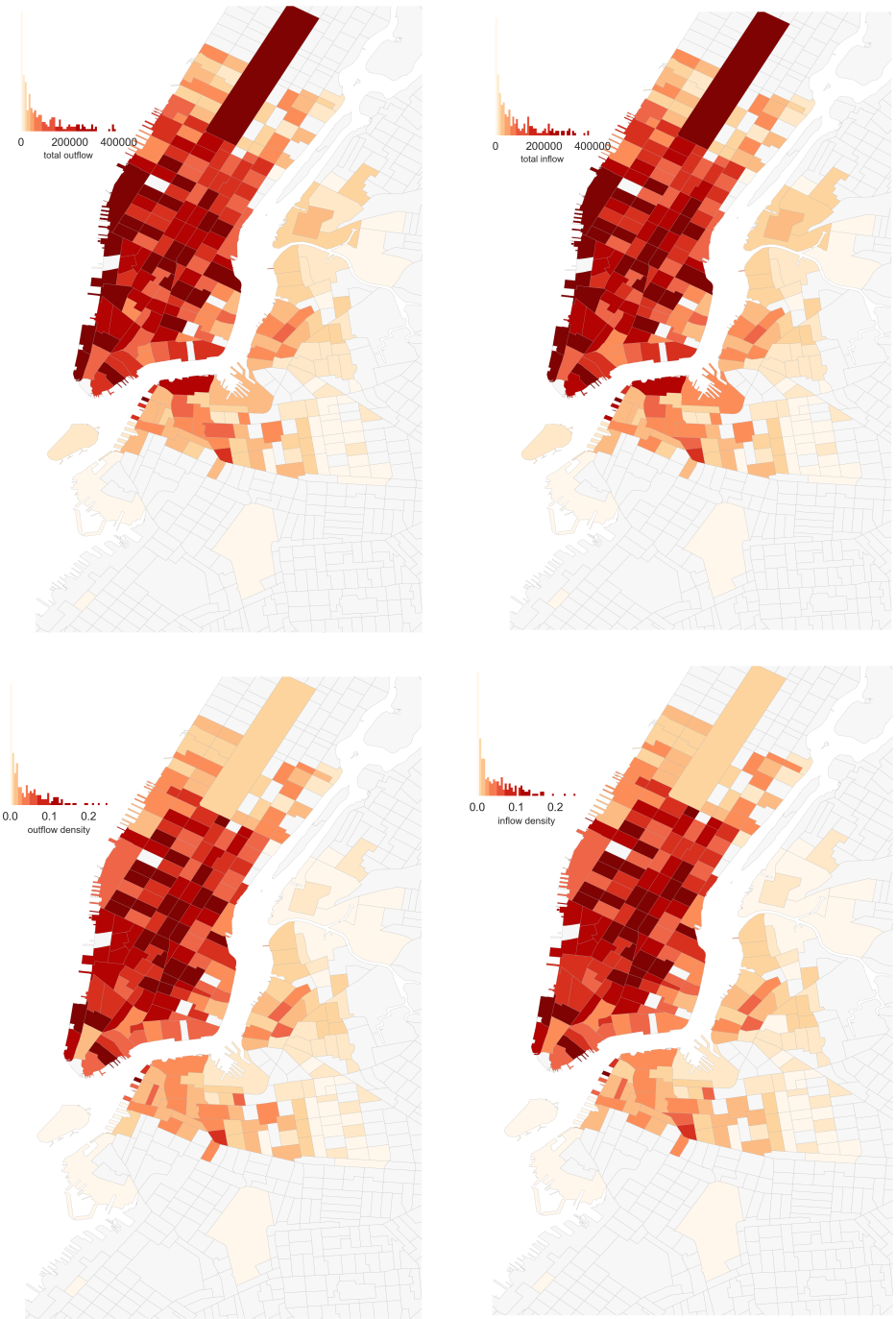


Figure 18: The number of bike trips that start in each census tract (top left), the number bike trips that end in each census tract (top right), the trip start density in each census tract (bottom left) and the trip end density in each census tract (bottom right).

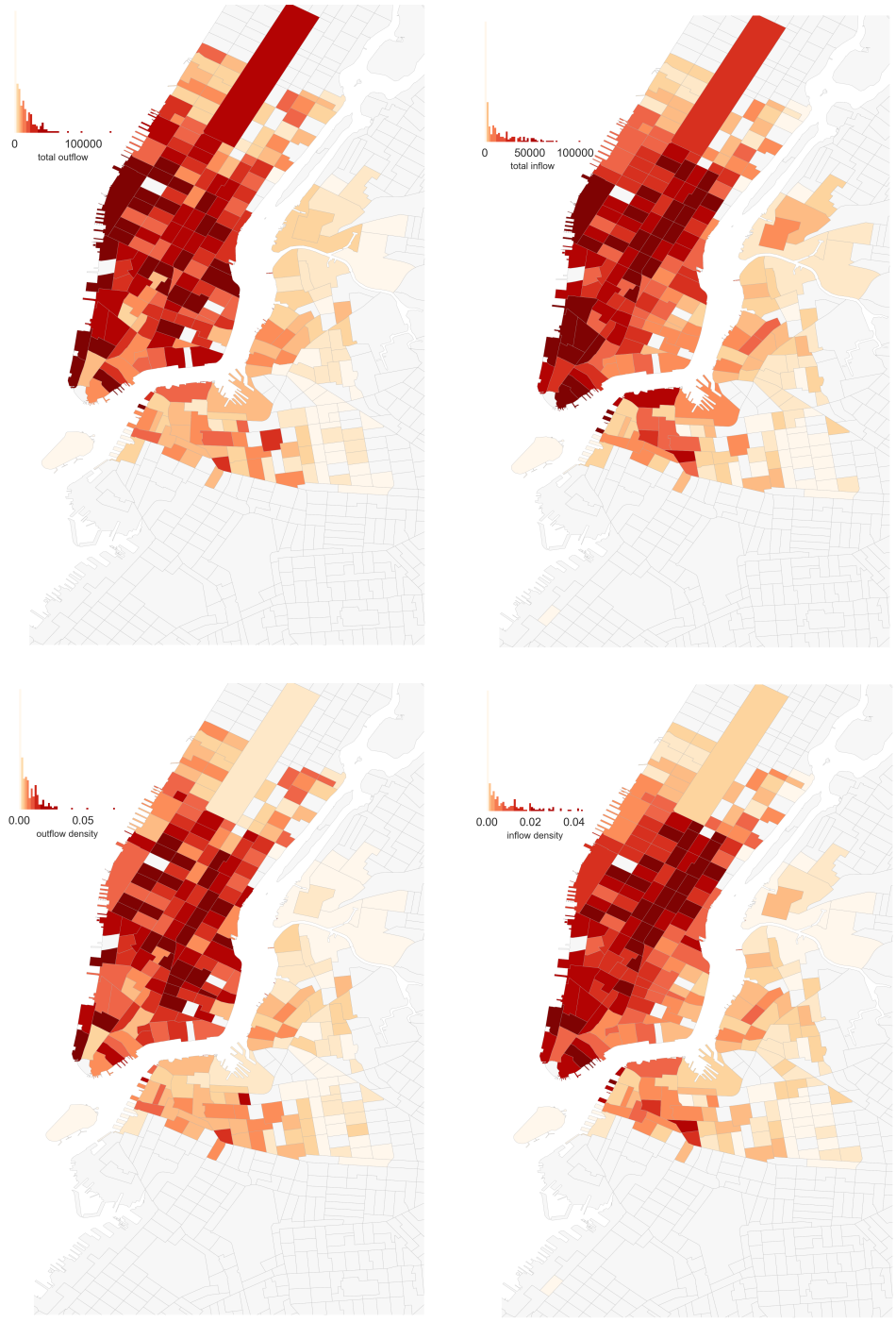


Figure 19: The number of morning commute bike trips that start in each census tract (top left), the number morning commute bike trips that end in each census tract (top right), the morning commute trip start density in each census tract (bottom left) and the morning commute trip end density in each census tract (bottom right).

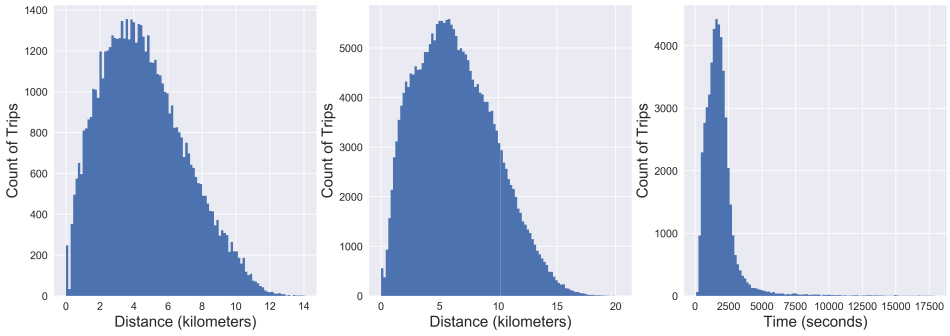


Figure 20: Variables that are a proxy for the cost to travel by bike between census tracts. This includes Euclidian distance in kilometers between centroids (left), synthetic network-based travel distance also in kilometers (middle), and trip duration in seconds (right).

4.3.2 Taxi Trips

There are currently over 1 billion taxi trips[12] now available in NYC from January 2009 to the present. For the study period of June 2014 to June 2016, this amounts to approximately 320 million trips. These trips are reported with the precise latitude and longitude for the trip pick-up and drop-off point, start time, end time, trip distance, trip cost, and passenger count. Similar to the bike data, these trips can be aggregated to census tracts and to various temporal units.

Viewing the taxi trips as a time series (figure 21), several trends that contrast the bikes can be determined. First, sampling the data monthly (figure 21 top), we can see there is a decrease in taxi trips over time rather than in increase. This may be attributed to competition with the new bike-sharing system, as well as competition will ride-sharing services, such as Uber, that have become popular. It can also be seem from the monthly time series that the seasonal trend is much less apparent. Moreover, trips seem to be at a low during the winter when the weather is harsh, but also during the summer months when many people opt to cycle or walk instead. Peaks in the fall and spring may also be weather related since the weather is not harsh enough to deter travel, but precipitation might drive individuals to choose a taxi over other options. Second, the weekly sampling of the series (figure 21 middle) adds more detail and shows agreement with an overall trend that is less prone to seasonality effects. Thirdly, sampling the data daily (figure 21 bottom) adds even more detail. Of note is that there are similar tends in both 2015 and 2016 of big dips in taxi trips right before and after the New Year when people are observing holidays. Importantly, the effects of the blizzard in late January are visible in a similar manner to the bike trips data.

[12]http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

The one difference is that there is no discontinuity in the taxi trip data set since taxi service was not explicitly discontinued. However, the number of taxi trips drastically decreases to the lowest point in the entire series. Also similar to the bike data, the weekly taxi data shows artificial lows near the beginning and end of the series, which is also an artifact of the data processing.

Differences in taxi trips compared to the bike trips can also be discerned when the data are aggregated by day of the week or hour of the day (figures 22 and 23). Using the former aggregation, there is a small increase in the number of taxi trips each day from Monday to Saturday, with a decline from Saturday back to Monday. The overall trend is that there is a spike in the number of trips on Friday and Saturday when many people engage in evening leisure activities. Using the latter aggregation, it can be observed that the number of taxi trips begins increasing around 6:00am with the beginning of the morning commute. However, the number of trips increases until about 9:00am where the number of trips plateaus until the evening commute around 5:00pm, rather than declining. The taxi trips are also different from the bike trips in that the number of taxi trips remains much higher throughout the evening and does not significantly decrease until the latest hours of the night. These trends can be corroborated and expanded upon by aggregating by both the day of the week and the hour of the day (figure 24). As with the bike trips, the taxi trips have much different weekday trends compared to weekends when the majority of taxi trips take place later in the day. In particular, we can see that Friday and Saturday tend to have an additional evening peak probably associated with nightlife activities. Of note is that the number of trips on Fridays tend to mirror those of other weekday mornings, while having an evening trend more closely related to Saturday. On the contrary,

Sunday tends to have a similar number of trips as Saturday in the morning, but is more similar to the weekdays in the evening.

The total trip outflow and inflow for each census track (figure 25) shows that the highest volume of taxi trips occurs in Manhattan and the areas of Brooklyn, Queens, and the Bronx that are closest to Manhattan. Overall, the distribution of trip counts is more skewed than the bike data, with many more census tracts having a relatively smaller number of trips compared to a much smaller group of tracts with very large trip counts. Normalizing trip counts by census tract area moves some of the larger census tracts from the extreme (i.e., large counts) end of the distribution toward more moderate values (figure 26). This includes important places like Central Park and JFK airport (circular tract in the southeast) and many less distinguished tracts in eastern Queens and Brooklyn and the northern part of the Bronx. Whether visualizing counts or densities, there does not seem to be any strong differences between the outflows (left) and the inflows (right). Unlike the bike data, no clear patterns emerge, when the data are limited only to trips associated with the morning commute (figures 27 and 28).

Similar to the bike trips, we can employ a simple Euclidian distance between census tracts (figure 29 top left), though Mapzen's *Matrix* service is prohibitively expensive to deploy for the more than 4 million possible origin-destination routes associated with the approximately 2100 census tracts where taxi trips can occur. The taxi trip observations also include the distance traveled during each trip (figure 29 top right), the total fare charged (figure 29 bottom left) to the customer and the duration of the trip (figure 29 bottom right). The trip fare is determined by a combination of pre-determined fees, time of the trip and the distance traveled. Spikes in the taxi fare distribution are likely due to the addition of pre-determined fees, such as

Figure 21: Taxi trips by month (top), by week (middle), and by day (bottom) throughout the case study time period.

tolls. It is also possible to see a large spike at 50 dollars that is likely associated with taxi trips to the airports, which are charged as a flat rate. However, like the bike trip duration, these three included measures of separation are only available for origin-destination routes where trips have been observed and not for the entire matrix of potential origin-destination routes and therefore not be effectively employed in spatial interaction models.

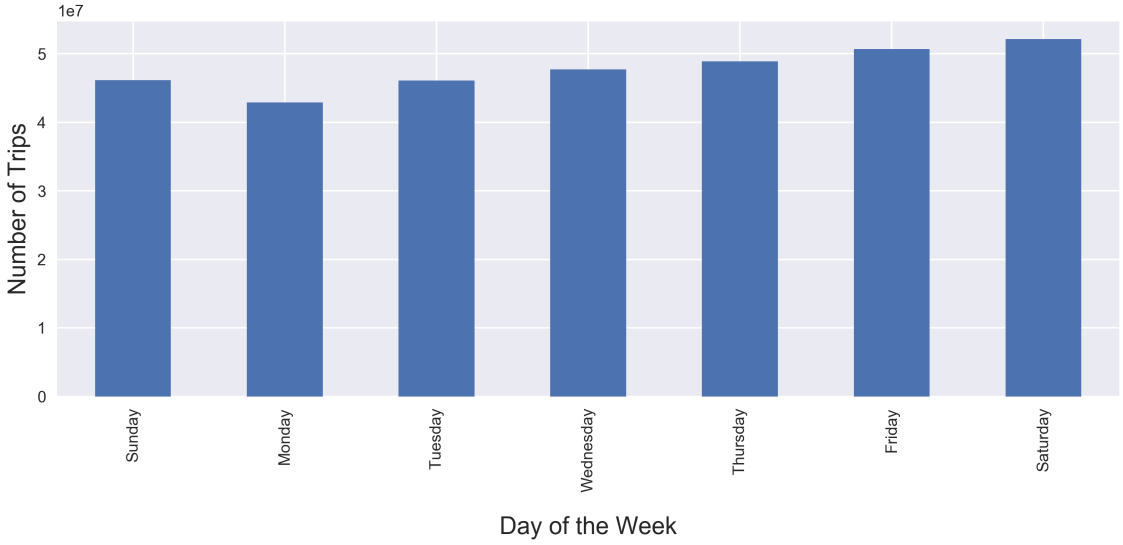


Figure 22: Taxi trips by the day of the week.

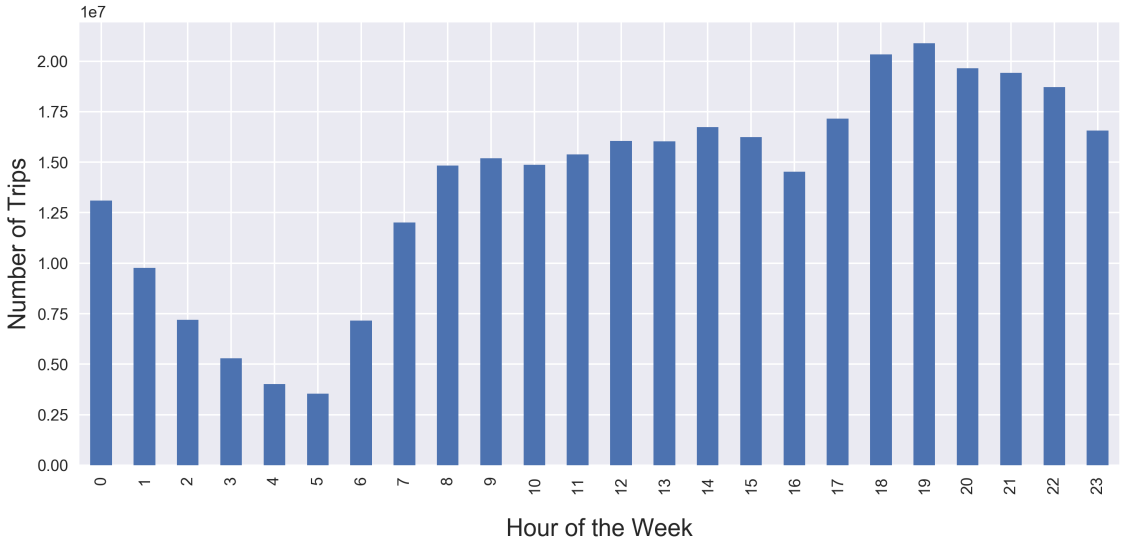


Figure 23: Taxi trips by the hour of the day

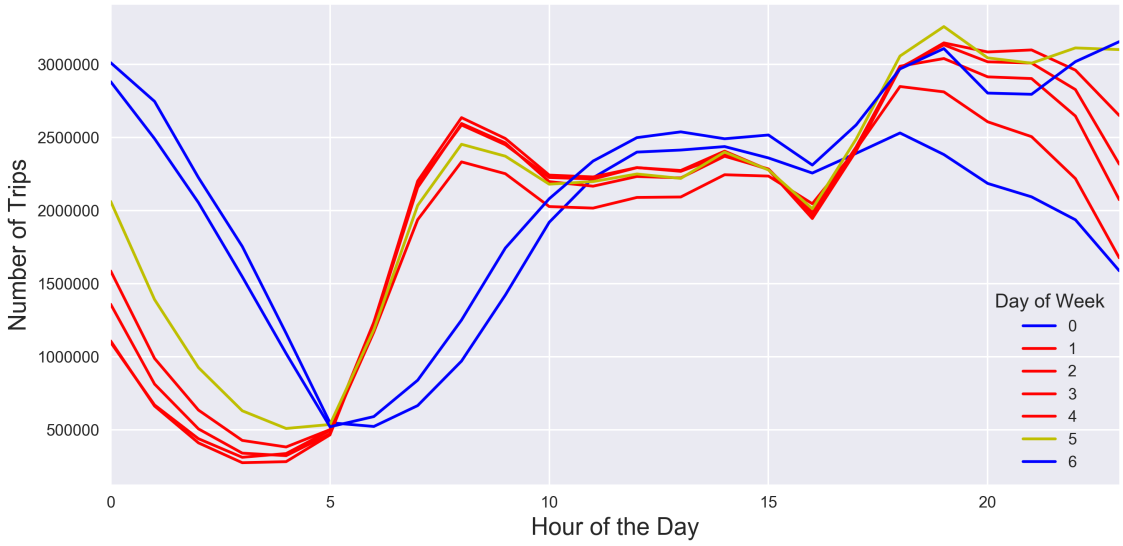


Figure 24: Taxi trips by the day of the week and the hour of the day. Monday-Thursday trips are red, Friday trips are in yellow and Saturday and Sunday trips are in blue.

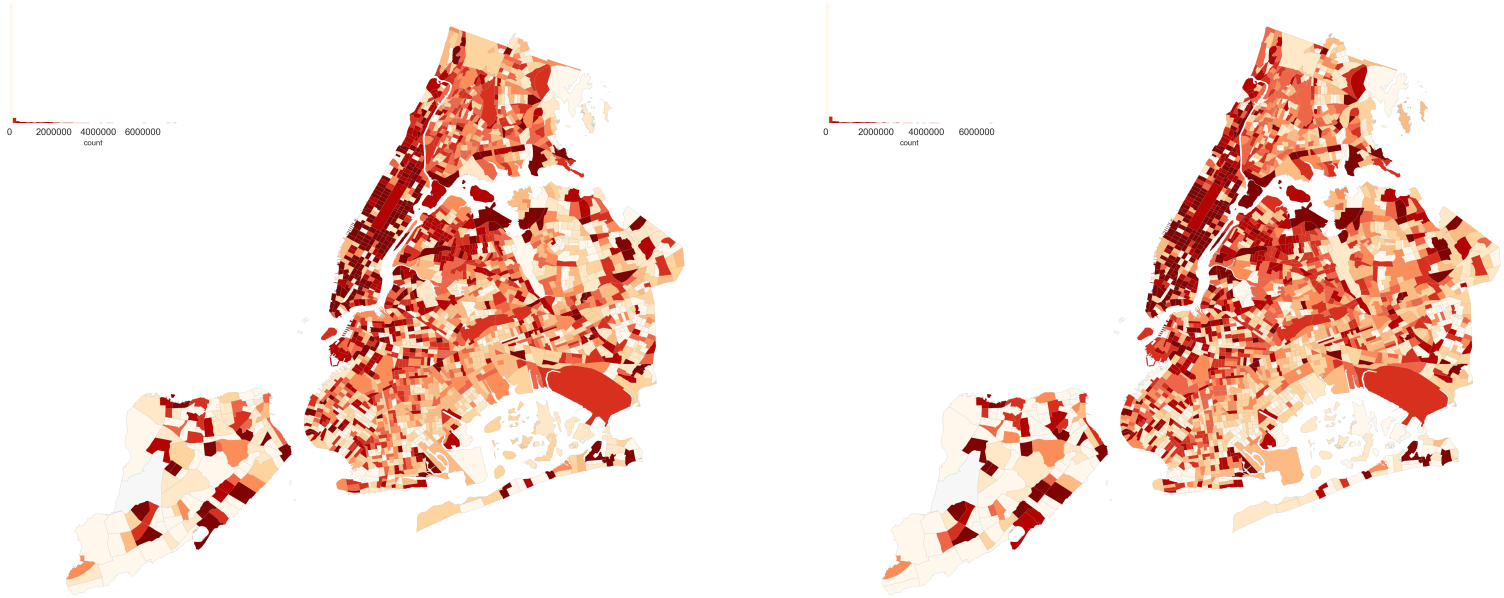


Figure 25: The number of taxi trips that start in each census tract (left), the number taxi trips that end in each census tract (right).

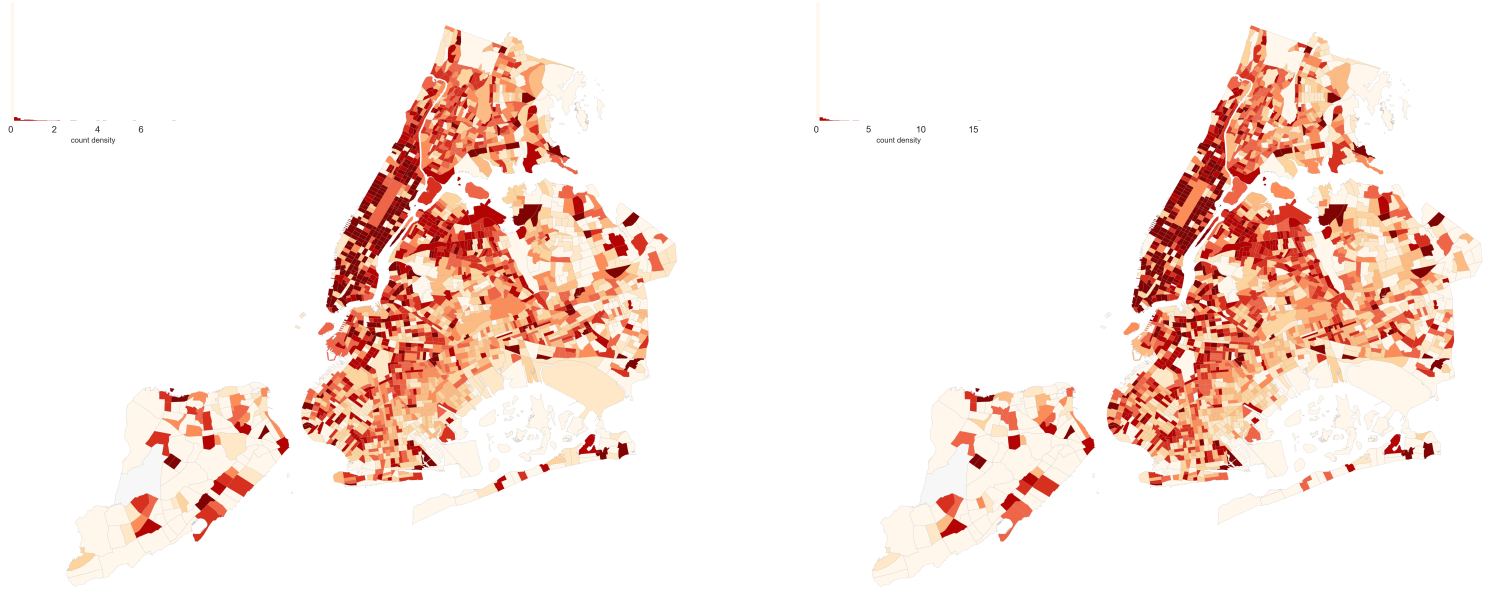


Figure 26: The trip start density in each census tract (left) and the trip end density in each census tract (right)

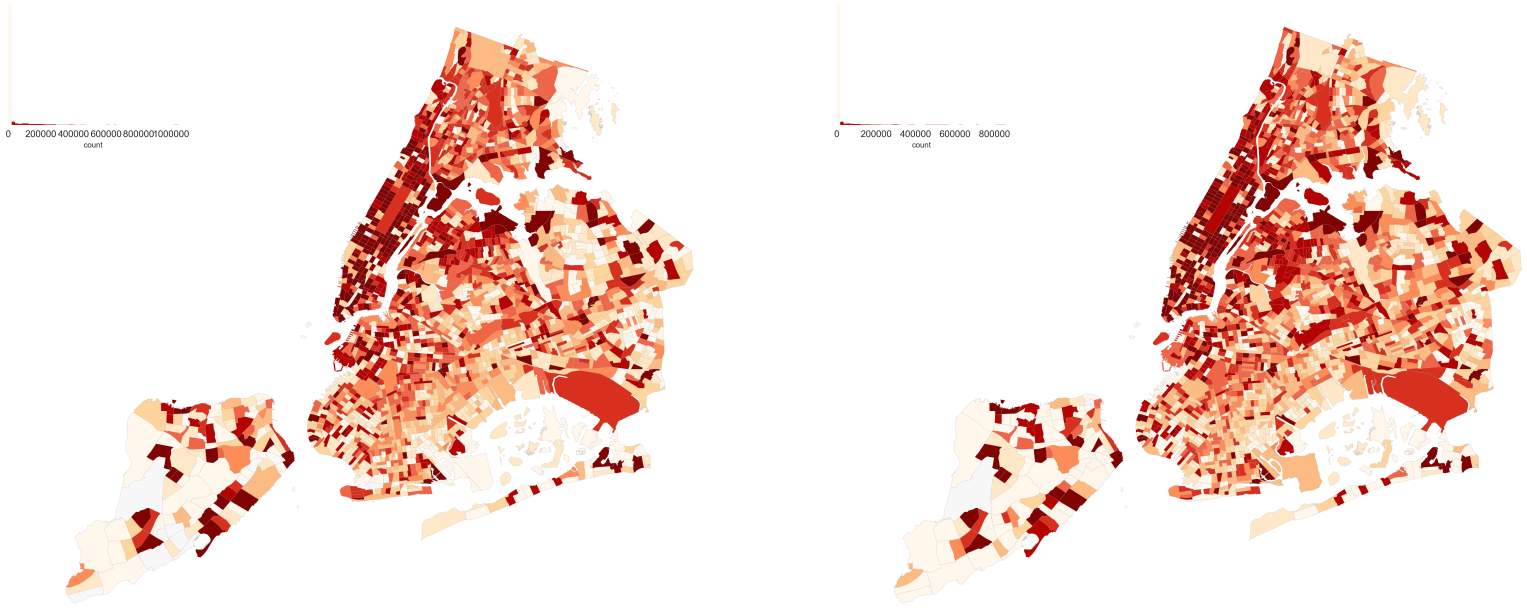


Figure 27: The number of morning commute taxi trips that start in each census tract (left), the number morning commute taxi trips that end in each census tract (right).

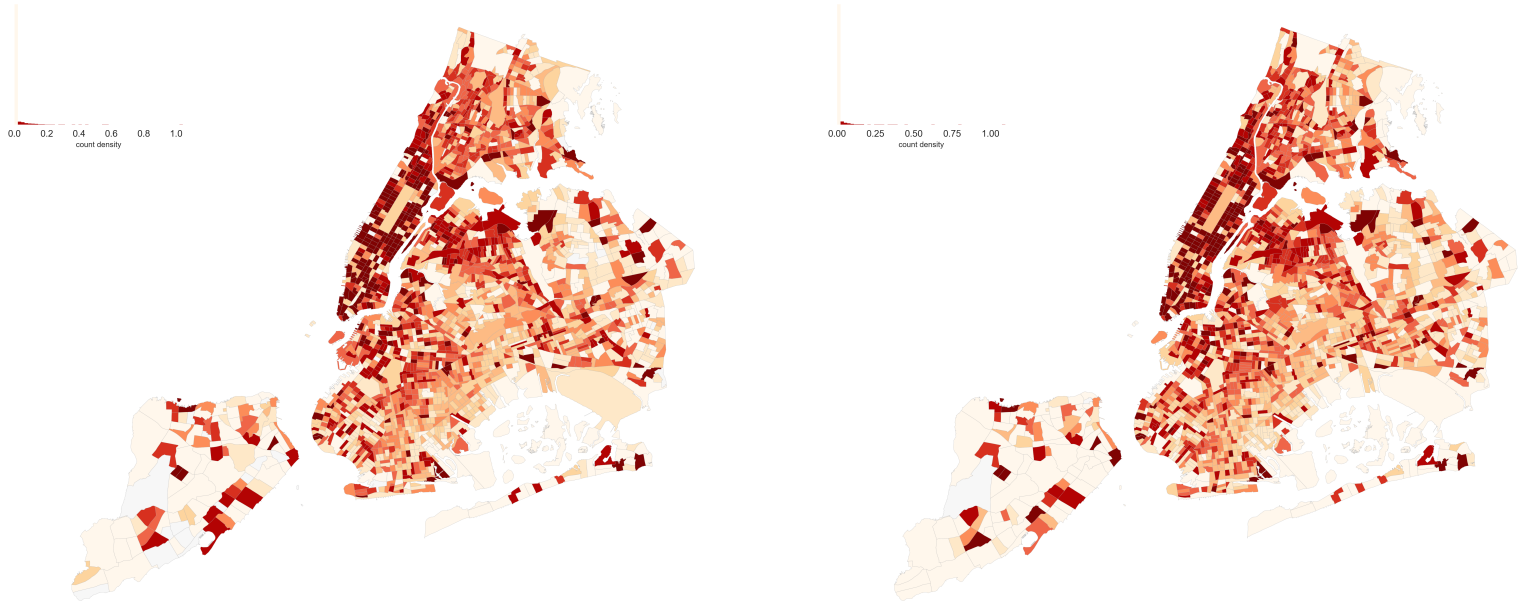


Figure 28: The density of morning commute trip starts in each census tract (left) and the density of morning commute trip ends in each census tract (right)

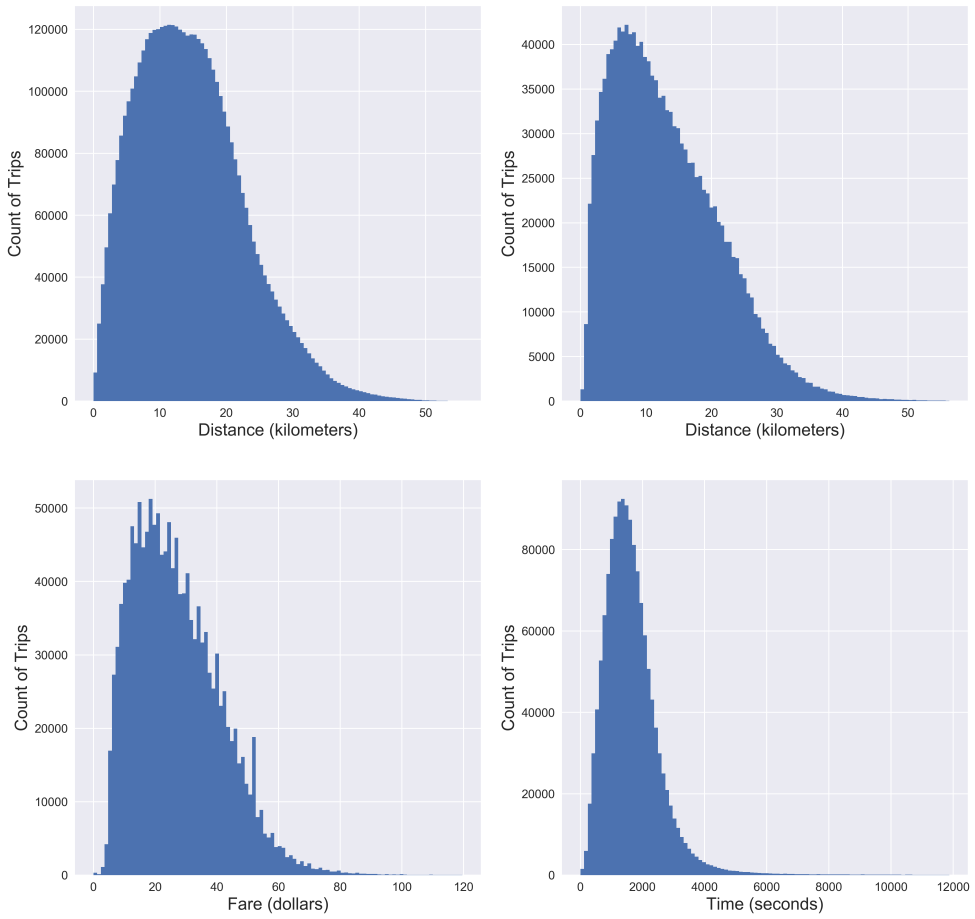


Figure 29: Variables that are a proxy for the cost to travel by taxi between census tracts. This includes Euclidian distance in kilometers between centroids (top left), actual road distance also in kilometers (top right), trip fare in dollars (bottom left), and trip duration in seconds (bottom right).

4.3.3 Census Commute-to-work Survey

Before the recent explosion of big data, movement data that represent urban mobility and commuting were limited to traditional surveys collected by the United States Census Bureau. These datasets are aggregate in nature, often reflecting information associated with multiple years and aggregated to the state, county, or tract, spatial resolution. One example is the aggregate commute-to-work flows that are available at tract resolution from the Census Transportation Planning Products[13] (CTPP) unit of the Census Bureau, which is based on the 2006-2010 American Community Survey (ACS).

After selecting the flows that occur from residences in tracts within NYC and to workplaces in tracts within NYC, the spatial distribution of the outflows and inflows (figure 30 and 31) can be visualized in a manner similar to the bike and taxi trips, though some noticeable differences are evident. Since this commute-to-work data are specifically about a single process, assumptions about origins and destinations are built into the data. For example, the outflows show different patterns than the inflows, such as many trips arriving at JFK airport but no trips originating from there. This is because there are no residences in the tract that correlates to JFK airport to provide survey data, whereas taxis are picking up and dropping off customers art the airport at all hours of the day. A similar argument holds for Central Park where there are no residences, though bike trips may occur there throughout the day. Additional effects of different types of locations serving as origins and destinations can be seen in the histogram legends of figures 30 and 31. The outflows tend to have a more diverse distribution of values, since residents live in and commute from almost every

[13] http://ctpp.transportation.org/Pages/5-Year-Data.aspx

census tract. In contrast, the inflows tend to be bimodal with tracts either receiving a relatively high number of commuters (e.g., tracts in Manhattan) or a relatively small number of commuters (e.g., areas farther from Manhattan). Finally, normalizing the flows by tract area, many large tracts have a relatively smaller intensity of commutes, such as those in Staten Island (figure 31).

One advantage to commute-to-work survey data is that it can be expected that there is less noise in the data, such as flows that do not occur based on commuting related decision-making processes. However, at the same time, there are several restrictions associated with these data. First, these trips are restricted to the tract spatial resolution. This means that the only cost variable that can be defined is Euclidian distance between the tracts (left in figure 20 and top left in figure 29) because we have no additional information on the precise begginning and end of each trip or the trip duration. Second, there is no temporal component to these trips. It is not possible to examine how commuting varies over the course of a day or week, nor is it possible to identify evolving patterns over, weeks, months, and years. Therefore, this census flow dataset will be compared to the newer taxi and bike flow datasets to help evaluate any potential or pitfalls associated with each data source.

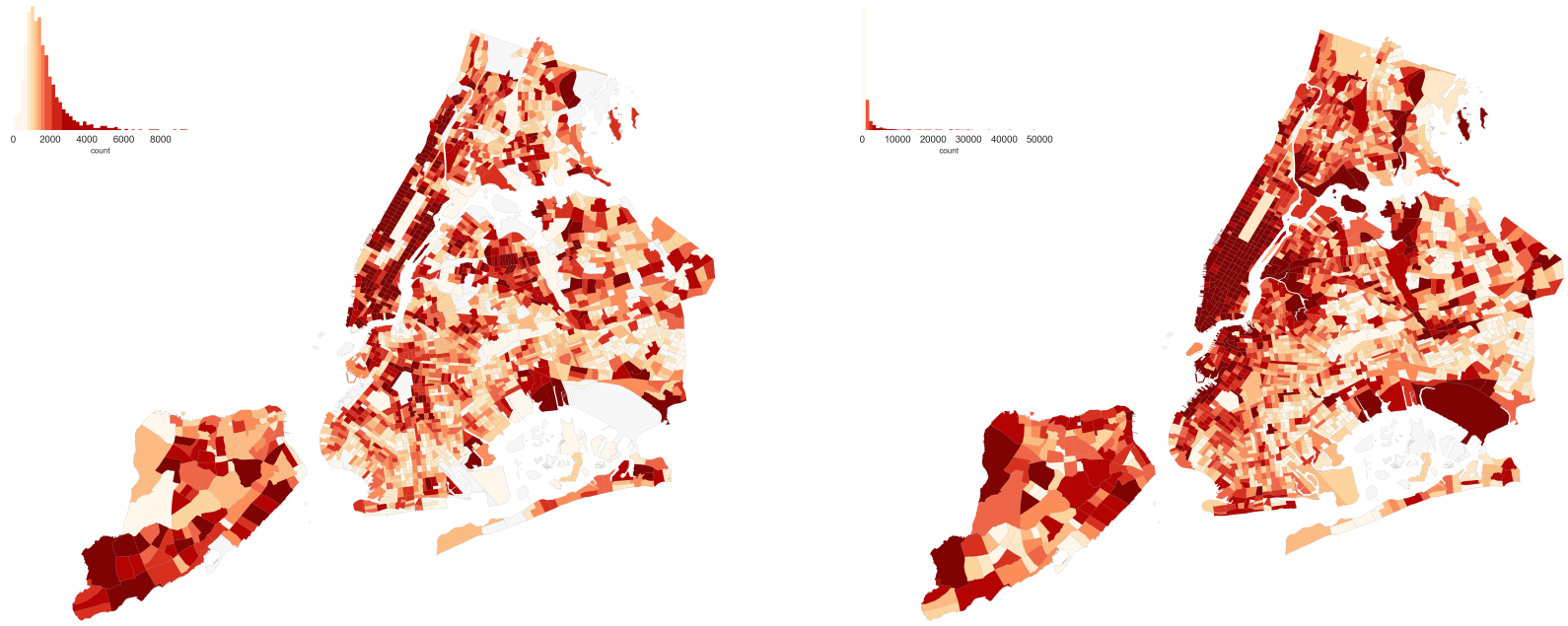


Figure 30: The number of commute-to-work trips that start in each census tract (left) and the number commute-to-work trips that end in each census tract (right).

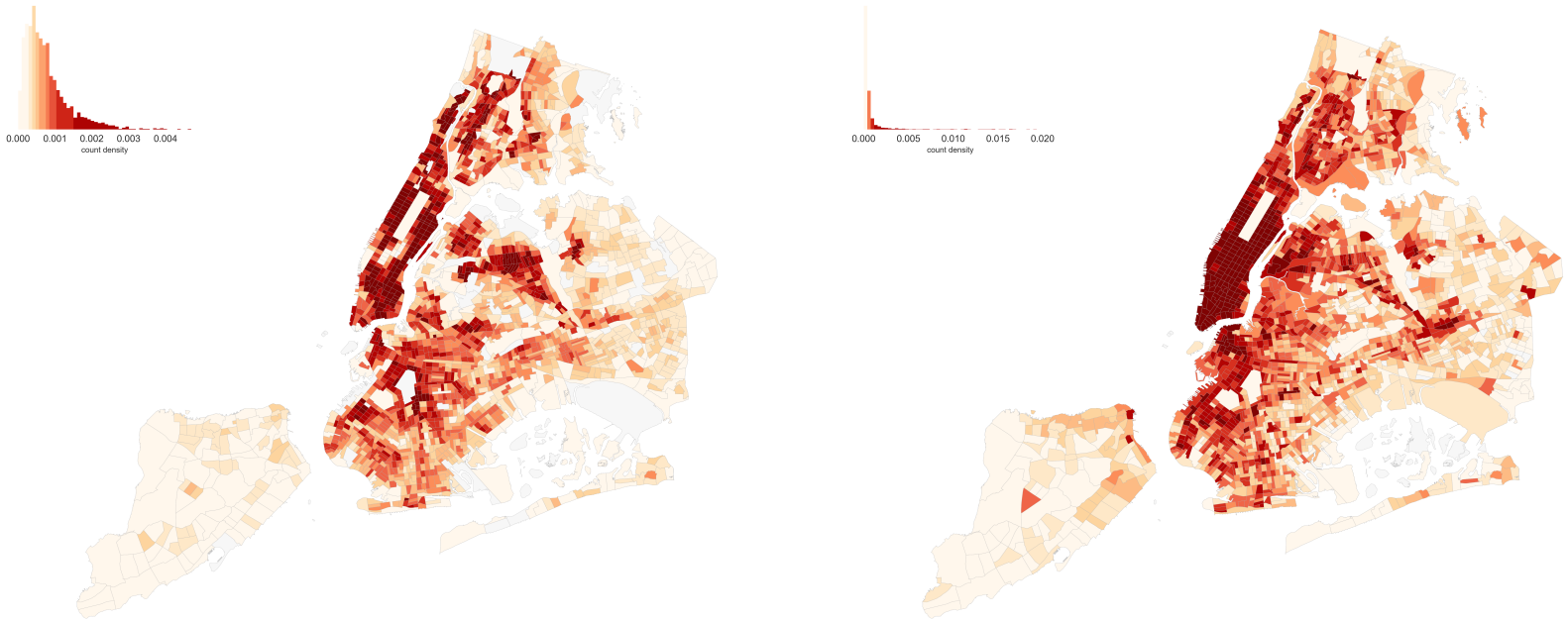


Figure 31: The density of commute-to-work trips that start in each census tract (left), the density of commute-to-work trips that end in each census tract (right).

4.4 Location Attributes

In order to understand what make locations attractive, data that describe the benefits or disincentives associated with each location are needed. Classically, these attributes are limited to data from the United States Census Bureau. However, in the era of 'big data' many additional variables are now also available. In this section, variables will be presented that will be used as locational data and includes traditional attributes, such as population and number of jobs, as well as newer data sources, such as points of interest (POI's) from OpenStreetMap[14] and the municipal government open data portal[15].

4.4.1 Census Variables

The 2010 decennial census and the American Community Survey (ACS) of the United States Census Bureau are extremely valuable repositories of socio-economic data that are frequently employed in spatial modeling. They provide data on a wealth of socio-economic variables, such as population, income or race, which are generally available at the census tract resolution. Due to the sampling frequency of these data sets they are representative of *slow dynamics* in that they are only sampled once a year or less frequently and are not expected to change much between samples. It should also be noted that all census data are estimates and therefore contain various levels of uncertainty; however, incorporating this uncertainty is not a topic of this research. For this research, the following variables have been collected for each census

[14] $https://www.openstreetmap.org/$

[15] $https://opendata.cityofnewyork.us/$

tract: population count, housing unit count, average income, the percentage of people living in poverty, and the number of jobs.

The spatial distribution of census tract population shows, as expected, that tracts in Manhattan consistently have very high populations (figure 32 top). Unexpectedly, it can also be seen that high population counts are recorded for Staten Island and eastern portions of Queens and Brooklyn. However, normalizing the population by the area of the census tract indicates that these high counts are mostly due to these tracts being larger in size (figure 32 bottom). Similar trends are apparent for the number of housing units and the density of housing units (figure 33). This suggests that these two datasets will be collinear and only one should be used in each form (i.e., counts or density).

Unsurprisingly, average income and the percentage of people living in poverty tend to have an inverse relationship where areas with high income have low poverty (figure 34). Several trends emerge by visualizing these two variables. First, the southern portion of Staten Island tends to have higher income and lower poverty. Second, income is higher and poverty is lower in Manhattan with the exceptions of Alphabet City, which is known to contain low-income housing, and north of Central Park, which contains areas known for poverty, such as parts of Harlem. In the Bronx there is typically higher poverty except for an enclave all the way in the northwest and northeast. Finally, the areas of Brooklyn and Queens closest to Manhattan tend to exhibit higher income levels, though as you move east to the central areas of Brooklyn and Queens, the tracts tend to have higher poverty and lower income. These are also the areas that do not typically have high transportation access. Interestingly, if you continue moving east towards long island, the poverty in Queens tends to decrease and income tends to increase.

The final census variable is the number of jobs in each census tract. It is clear that Manhattan and areas closer to Manhattan have a higher number of employment opportunities (figure 35 top). However, normalizing the number of jobs by census tract area, it can be seen that large tracts that contain important features like Central Park and JFK international airport are outliers and do not contain a high number of jobs relative to smaller tracts (figure 35 bottom). The distribution of jobs is heavily skewed with most tracts having relatively few jobs compared to a minority of tracts with a very high number of employment opportunities. This can be further investigated by filtering out some of the tracts with the highest number of jobs (figure 36 top) where most of these tracts are either in Manhattan or very close to Manhattan. One outlier is the tract containing JFK airport, but this pattern does not exist when considering job density (figure 36 bottom).

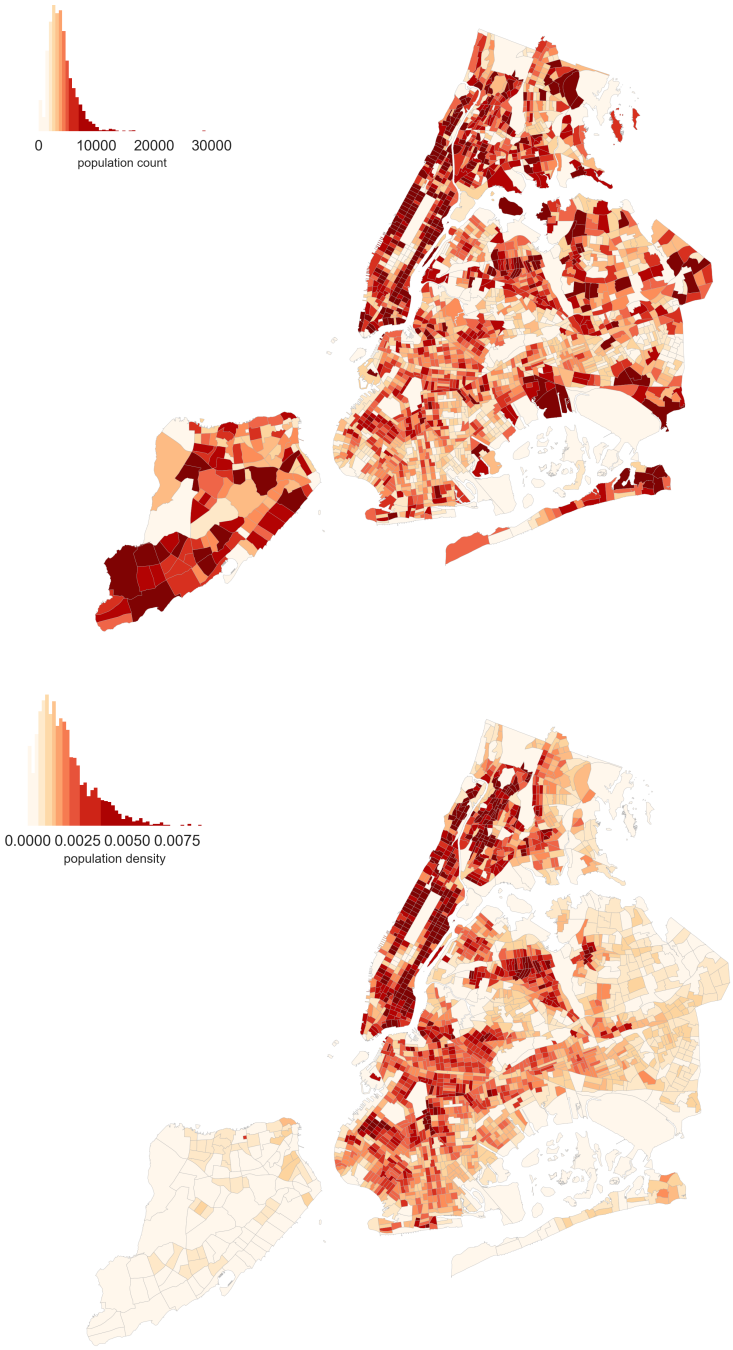


Figure 32: Distribution of population by census tract (top) and distribution of population density by census tract (bottom).

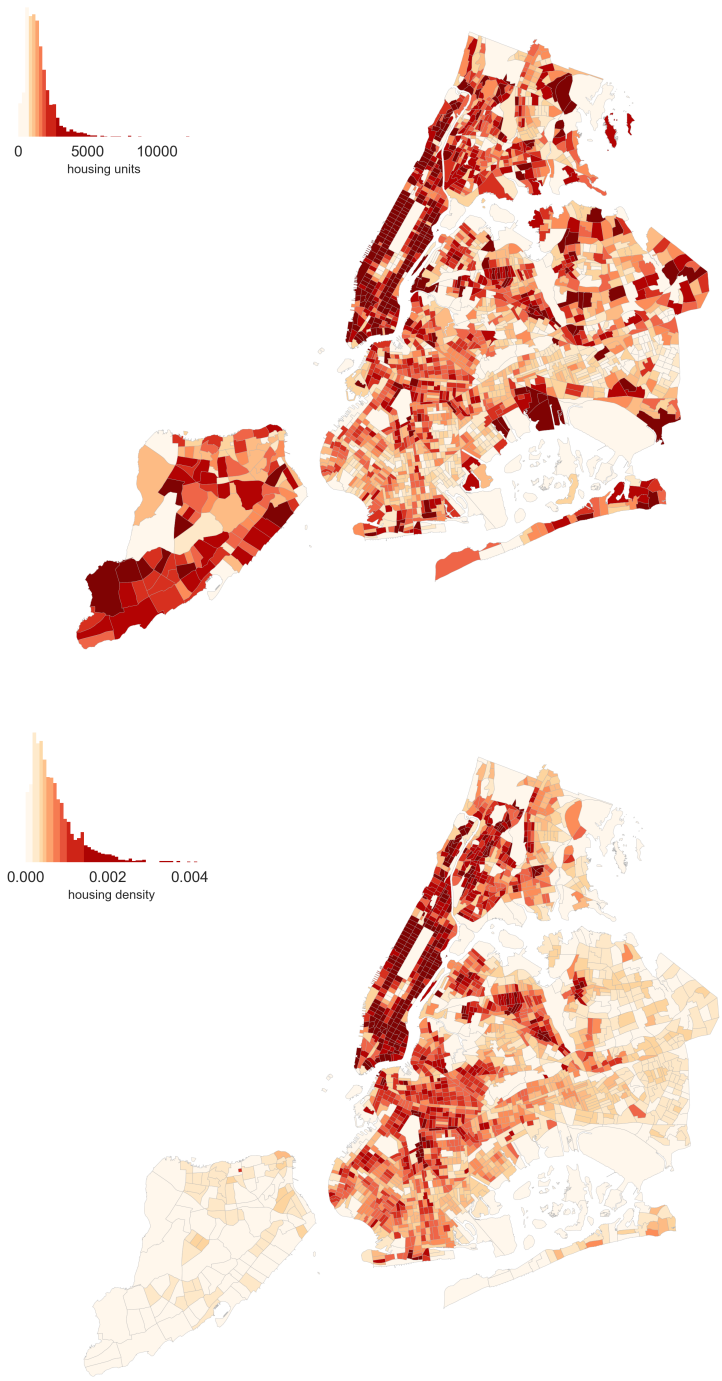


Figure 33: Distribution of housing units by census tract (top) and distribution of housing units density by census tract (bottom).

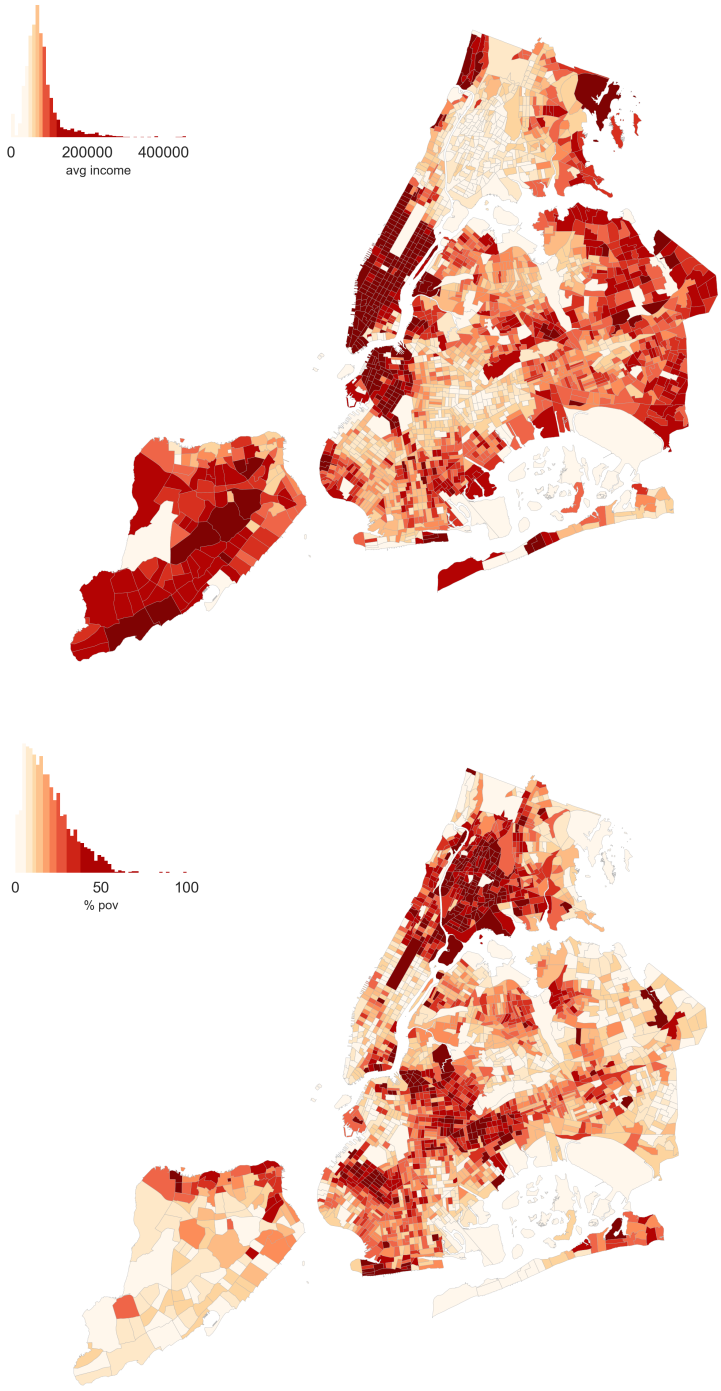


Figure 34: Distribution of average income by census tract (top) and distribution of percentage of households in poverty by census tract (bottom).

Figure 35: Distribution of jobs by census tract (top) and distribution density of jobs by census tract (bottom).

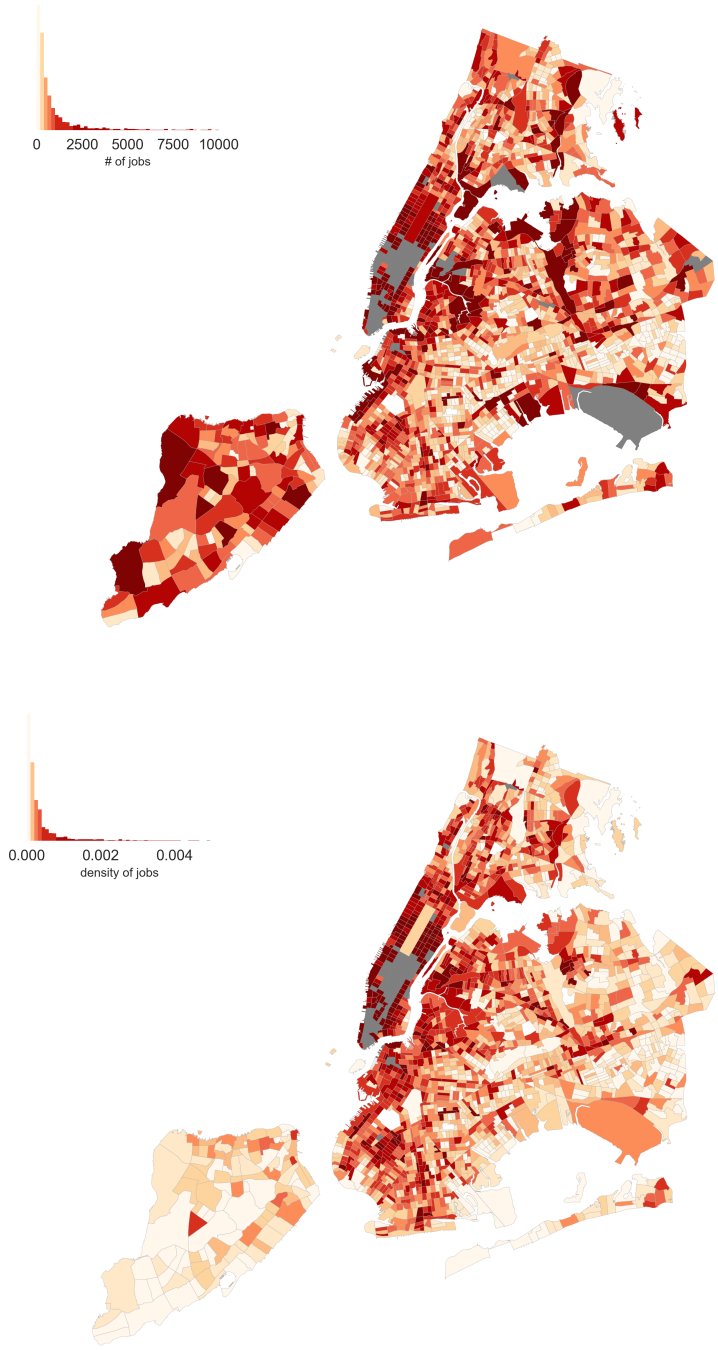


Figure 36: Distribution of filtered jobs by census tract (top) and distribution of density of filtered jobs by census tract (bottom).

4.4.2 Urban Environment

A variety of spatial data sets that can be used to describe the urban environment and may serve as proxies to the attractiveness of a location are available via the NYC open data portal[16]. For instance, building footprints, which include the number of floors in each building, may be used to compute the total building square footage in each census tract to represent the general opportunities available (figure 37). Here it can be seen that midtown and downtown Manhattan, along the portion of the outer boroughs that are closes to Manhattan, have the building square footage.

Another way to capture different aspects of the built environment is through a database of points-of-interest (POI) for NYC, which can be obtained through the OpenStreetMap[17] project that collects and stores volunteered geographic information. Figures 38 - 44 represent various types of POI's that include bars, cafes, restaurants, shops, tourist destinations, museums, and colleges and universities. Each of these variables may also help define what makes a census tract attractive, however there are several issues with these data. First, the data are available as a set of points which need to be aggregated to census tracts. Second, and more importantly, these data may suffer a reporting bias, since most people do not spend time volunteering geographic information. It is likely that there are more POI's reported where there are more younger communities that are more technology savvy. Thirdly, volunteered information may have lower accuracy than more official sources. Therefore, this

[16]https://nycopendata.socrata.com

[17]http://wiki.openstreetmap.org/wiki/Planet.osm

Figure 37: Distribution of total building square footage by census tract locations.

research will investigate the usefulness of this type of data in spatial interaction modeling.

Many of these of individual types of POIs are relatively sparse. For example, museums (figure 43), and places of higher educations (i.e., colleges and universities) (figure 44), were so sparse that they were more easily visualized using binary indicators for tracts that contained a POI (in red) or did not. Therefore, a single composite variable composed of all POI's (figure 45) was created and will be employed in empirical modeling in a later chapter.

Figure 38: Distribution of bar locations (top) and distribution of number of bars by census tract (bottom).

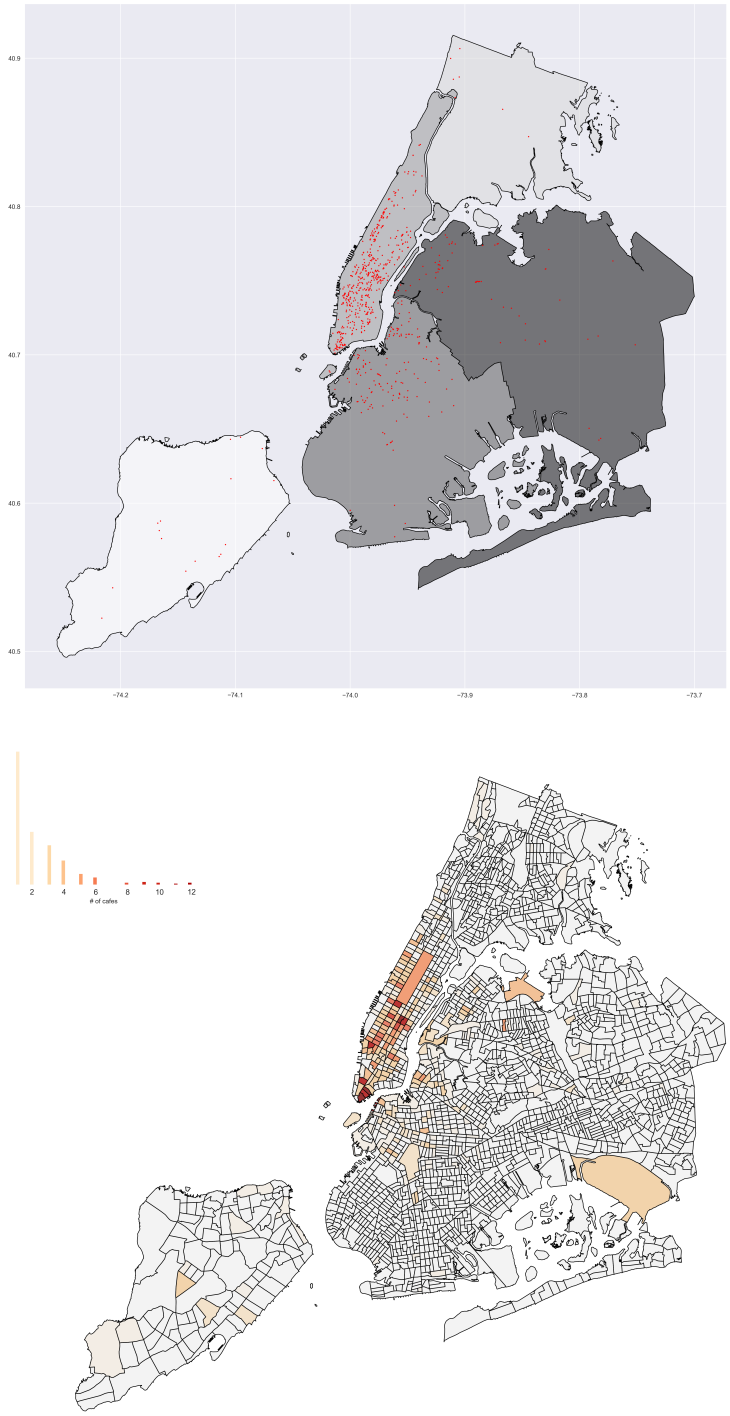


Figure 39: Distribution of cafe locations (top) and distribution of number of cafes by census tract (bottom).

Figure 40: Distribution of restaurant locations (top) and distribution of number of restaurants by census tract (bottom).

Figure 41: Distribution of shop locations (top) and distribution of number of shops by census tract (bottom).

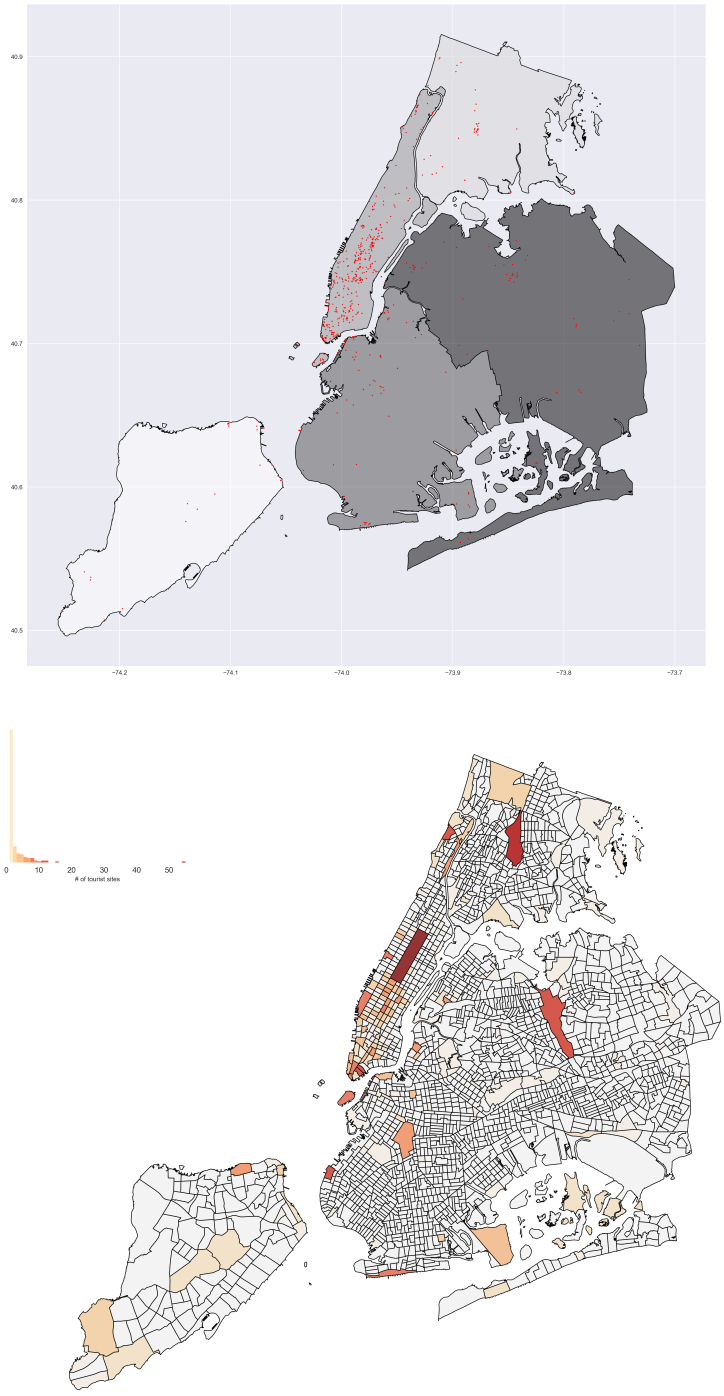


Figure 42: Distribution of tourist site locations (top) and distribution of number of tourist sites by census tract (bottom).

Figure 43: Distribution of museum locations (top) and census tracts containing at least one museum in red (bottom).

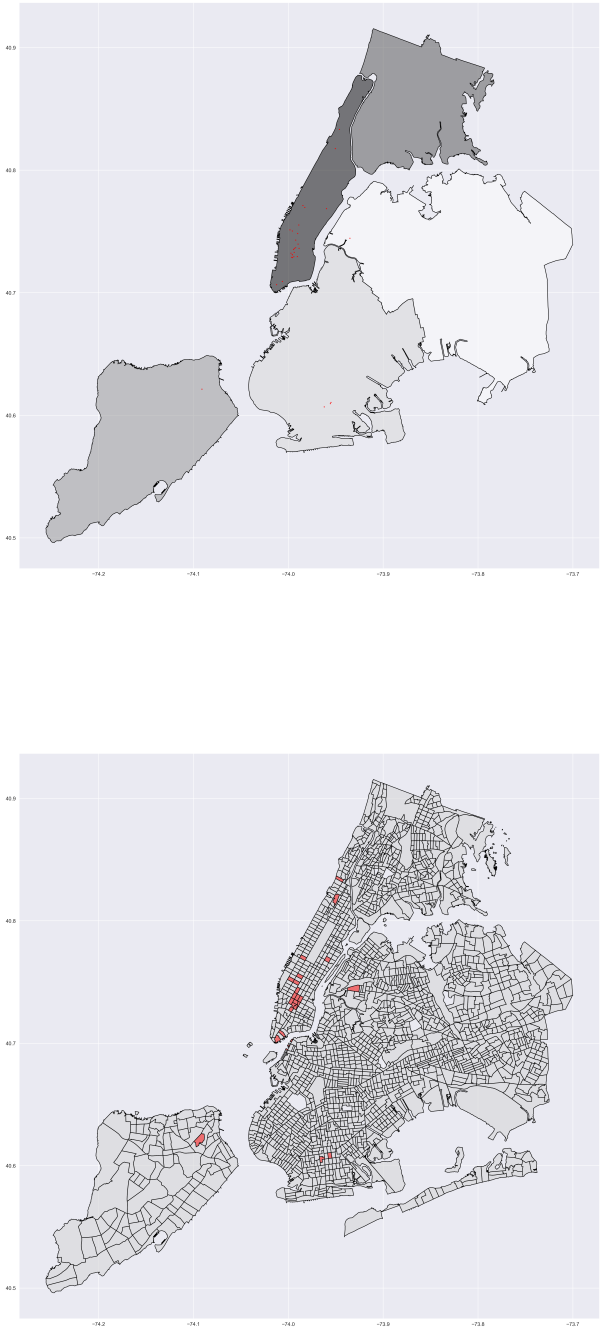


Figure 44: Distribution of college and university locations (top) and census tracts containing at least higher education POI in red (bottom).

Figure 45: Distribution of number of all POI's by census tract locations (top) and distribution of density of all POI's by census tract (bottom).

Sensor data collected can also potentially be used to represent the urban environment. While sensor data are traditionally hard to acquire due to their proprietary nature and large size, they are becoming more available with the spread of open and smart cities (Arribas-Bel, 2014). One example is the subway turnstile data[18], which logs the number of individuals that enter and exit each particular subway station and could serve as an alternative proxy to the attractiveness of a destination.

The location of subway stations throughout the city is given in figure 46. The spatial distribution of entrances and exits aggregated over time is given in figure 47 where it can be seen that subway usage corresponds to the tracts with more stations and the highest usage is associated with Manhattan. Furthermore, the histograms for the total entrances and exits show that most stations experience relatively low usage compared to a minority of stations that experience very high usage. In fact, visualizing the logarithm of subway usage indicates that there are essentially two distributions of subways: those with a lower level of usage and those with a higher level of usage (figure 47 legend).

In particular, the number of subway entrances and exits may be an indicator of destination attractiveness that changes over time. First, the number of subway trips that start at a location may be indicative of the attractiveness of the subway station as a feature of a location. For example, individuals may make more bike trips to tracts that also have a subway station so that they can leverage multi-modal transit routes. Second, the number of trips that end at a location may be a proxy for other attractions available there, since locations where more subway trips end are likely correlated with areas of high residence, more office space, or a cluster of leisure activities.

---

[18]http://web.mta.info/developers/turnstile.html

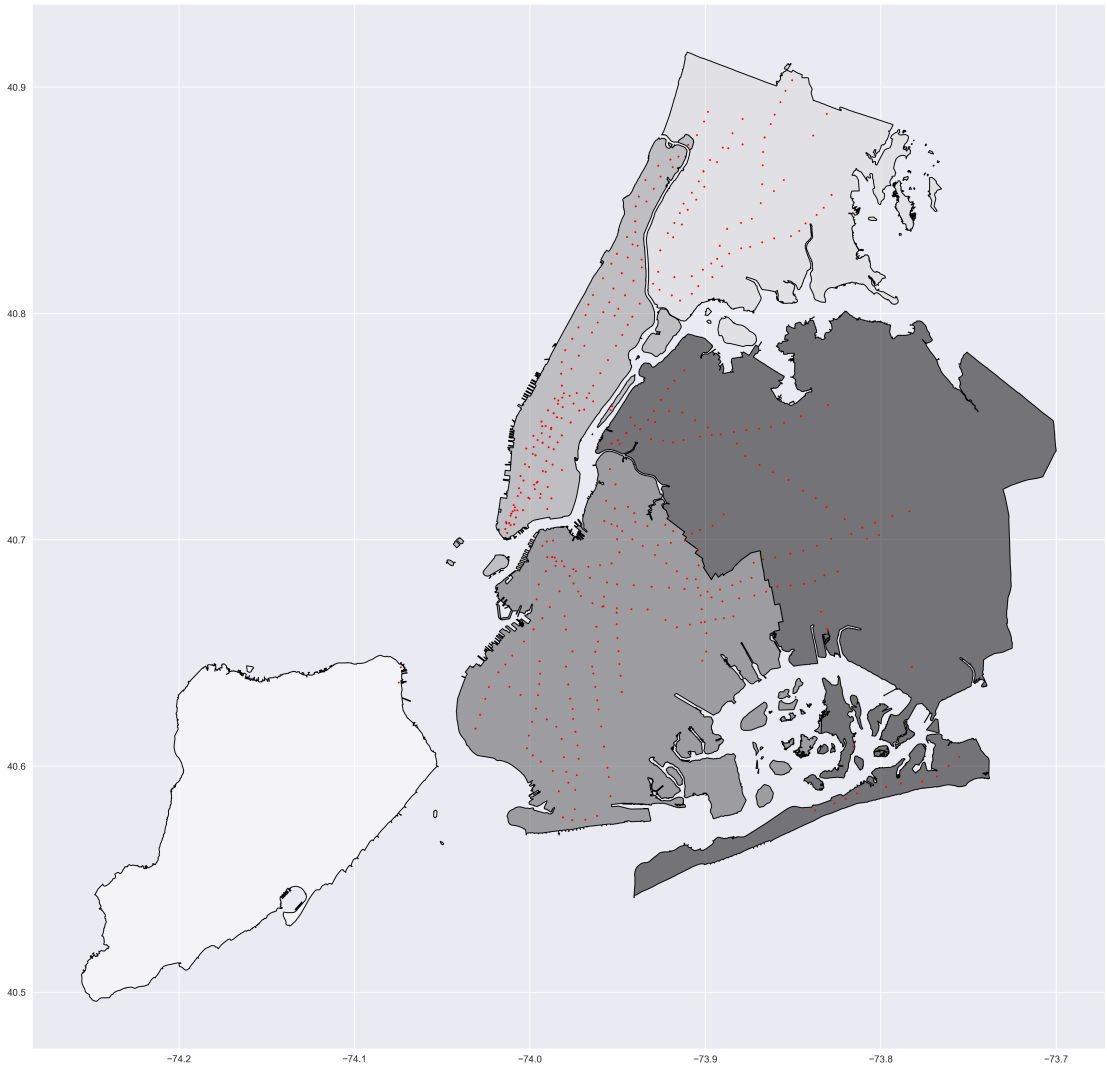


Figure 46: Location of subway station entrances and exits.

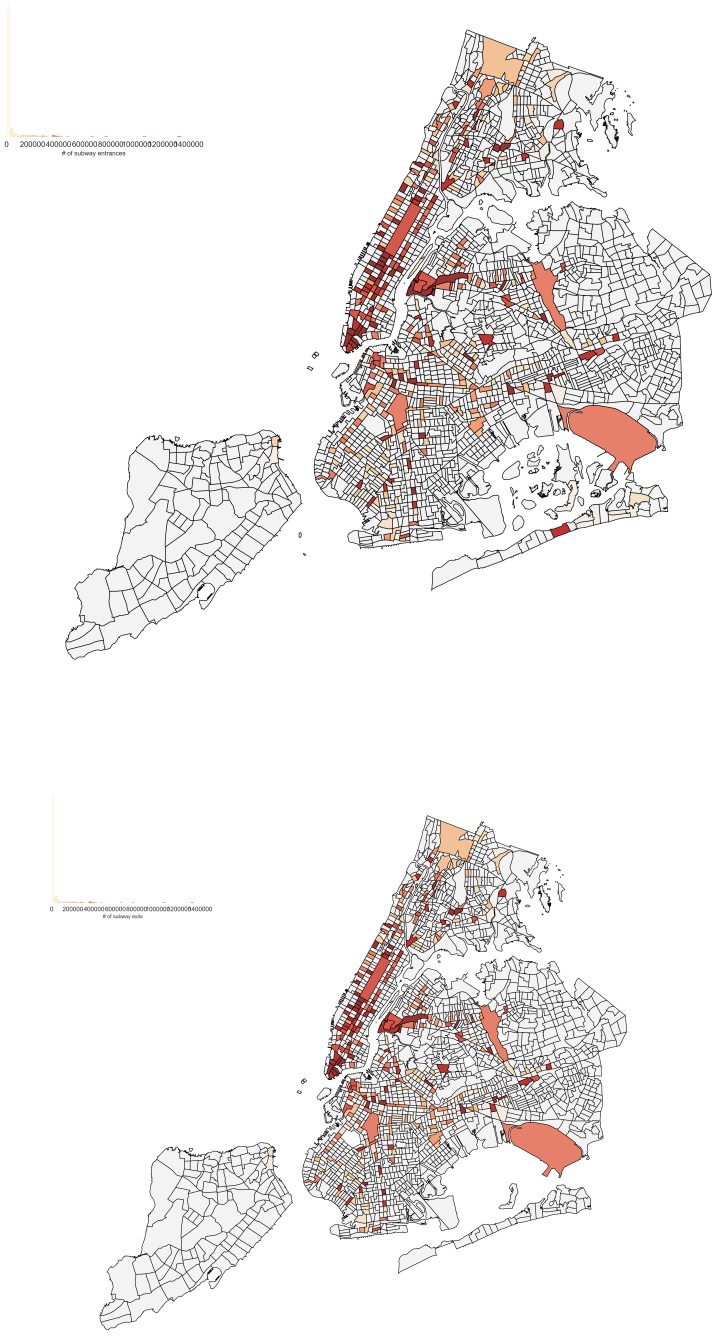


Figure 47: Distribution of the number of aggregate station entrances (top) and the number of aggregate station exits (bottom)

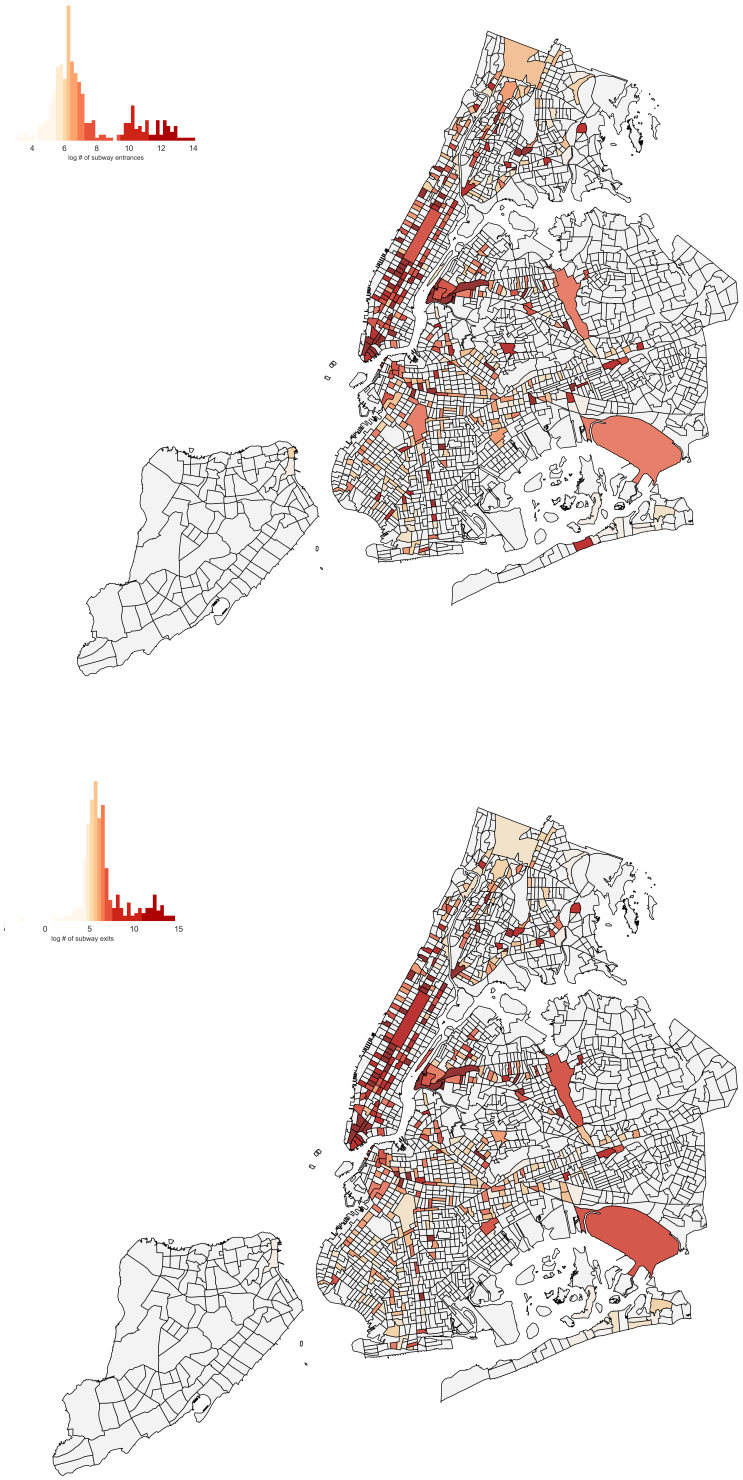


Figure 48: Distribution of the number of logged aggregate station entrances (top) and the number of logged aggregate station exits (bottom)

4.5 Moving forward

An array of data sources has been introduced in this chapter that are employed throughout the remainder of this research. First, several types of spatial interaction data were examined. These include newer sources, such as bike trips and taxi trips, and traditional survey data from the Census. A major difference between these sources is in the available temporal and spatial resolution associated with each dataset. Next, location data to represent the attractiveness of destinations was explored and included traditional census variables, such as population, as well as non-traditional variables, such a s POI variables from OpenStreetMaps. The advantage of the census data is the extensive coverage provided, while the advantage of the POI variables is that they provide a more diverse set of destination attractions. These data sources provide a mix of traditional and newer datasets for modeling urban spatial interaction at increasingly finer spatial and temporal scales.

The following chapters will utilize these empirical data sources, as well as simulated data, to evaluate several concepts and spatial interaction model specifications. In conjunction with this usage, the limitations of the various data on propulsiveness and attractiveness will be examined throughout this research. The next chapter will begin the analysis using controlled simulations to consider the bias that may occur in parameter estimates when spatial structure is not properly accounted for.

Chapter 5

A SIMULATION-BASED INVESTIGATION OF SPATIAL STRUCTURE EFFECTS IN SPATIAL INTERACTION MODELING

5.1 Introduction

Before calibrating models on empirical data and interpreting the resulting parameters estimates, it is prudent to investigate the results of calibrating various types of spatial interaction models on simulated data with well-known properties. Chapter 2 outlined the corpus of theories on spatial structure effects in spatial interaction modeling and demonstrated that there are several perspectives that are often incongruent with each other. In particular, these perspectives have led to the three[19] modeling frameworks that were outlined in chapter 2 to account for spatial structure: competing destinations (CD), spatial lag of explanatory variables (SLX), and spatial autoregressive processes (SAR). Though each framework claims to account for spatial structure effects, they entail different explanatory frameworks, different methodologies, different abstractions of space, and different underlying data-generating processes (DGP's). Thus, simulations will be used to compare and contrast the ability of these frameworks to capture spatial structure effects and also to evaluate the extent that they overlap or depart from each other.

The on-going debate about spatial structure in spatial interaction may be char-

[19]Additional frameworks were also reviewed, such as using a Box-Cox transformation or using eigenvector spatial filtering, but they were demonstrated to be insufficient. The Box-Cox method was shown to be inadequate using empirical experiments (see appendix A) and through a comprehensive literature review the eigenvector spatial filtering method was shown to be ambiguous and a-theoretical.

acterized by varying definitions of what is meant by spatial structure. While the earlier literature (i.e., 70's and 80's) is focused on the physical organization of the spatial interaction system (Fotheringham and Webber, 1980; Griffith and Jones, 1980; Fotheringham, 1981, 1983a), more recent work has focused on the concept of spatial autocorrelation (Griffith, 2007; Chun, 2008; LeSage and Pace, 2008). In this dissertation, the earlier definition of spatial structure will be adopted and the primary focus will be on the effect of clustering of locations in space. Data are simulated from the perspective of individual locations abstracted as points rather than from the perspective of aggregate areal units. Abstracting locations as points is more intuitive because individuals often choose individual places of employment, residences, or cities rather than broader, more aggregate spatial units. Furthermore, simulating data from the perspective of individual locations allows the effects of spatial structure to be assessed without the obfuscation of aggregation bias, which is a type of measurement error. Therefore, the simulations carried out in this chapter will consider point locations with varying degrees of spatial clustering.

Though aggregation may obfuscate variables and relationships between variables, it is an inevitable reality of applied empirical modeling. This is particularly true in spatial interaction where the flow data or the locational attributes may only be available at the aggregate level. Hence, it is also of interest how the effects of spatial structure are affected by aggregation and the extent that each of the three frameworks are robust in the presence of aggregation bias. Therefore, simulation experiments will also be carried out that aggregate the data to various scales to better understand the effects of aggregation bias in the context of each framework and each spatial structure scenario.

In the subsequent sections of this chapter, the simulation design is first described, followed by the results, and finally, some conclusions.

5.2 Simulation design

Simulated data with controlled and known features are needed to understand the effects of spatial structure and aggregation on spatial interaction models and to compare and contrast model specifications. This section outlines a methodology for simulating data and calibrating models that will be used to explore these issues.

5.2.1 Spatial structure

In order to control the effects of spatial structure, it is necessary to generate data with varying degrees of clustering in the locations. To do so, three scenarios were formulated (figure 49). The first scenario is comprised of 144 locations uniformly distributed across a 12 by 12 lattice in a 480 by 480 window and is representative of the situation where there is no spatial structure (figure 49 top). While the distribution of locations (i.e., opportunities) rarely manifests in this pattern[20], this pattern represents the extreme circumstance of no spatial clustering. The second scenario is comprised of 144 locations that are randomly distributed in a 480 by 480 window using a Poisson point pattern process. Since points are distributed at random, some of them will cluster together while others will be isolated so that this scenario represents mild spatial structure (figure 49 middle). Finally, the third scenario is comprised of 100

[20]Even in cities that use regular street grid, different types of locations, such as apartments of offices, are heterogeneously distributed across the grid

locations that are generated with a Poisson cluster point process and 44 randomly distributed points in a 480 by 480 window (figure 49 bottom). The Poisson cluster point process is formed by choosing 5 random centroids in the window and then generating 20 points within a 35 unit radius of each centroid. This scenario is indicative of strong clustering, where a few regions contain most of the locations and other regions have few or none. Such a pattern could be common in residences where housing communities are developed as a single project, the distribution of cities and towns across a country, shops within shopping centers, or employment opportunities at a collection of industrial facilities where agglomeration forces are present.

After the locations have been generated, populations to represent location propulsiveness and attractiveness are assigned by drawing values from a random uniform distribution with a minimum value of 50,000 and a maximum value of 5,000,000. These limits were chosen with the logic that they span several orders of magnitude that closely resemble small, medium, and large cities that could be found within the United States. More importantly, a random uniform distribution was chosen even though a power distribution with few large values and many small values would be more realistic. This is because values drawn from distributions other than a uniform distribution would have likely resulted in clusters of locations with similar populations and this could convolute the effect of spatial structure with those of spatially autocorrelated locational attributes. Sets of locations and their associated populations were generated 100 times so that the variation within each scenario can be examined throughout the experiments. These location systems were then used as input for the DGP's of the three spatial interaction modeling frameworks in order to simulate flows between the locations.

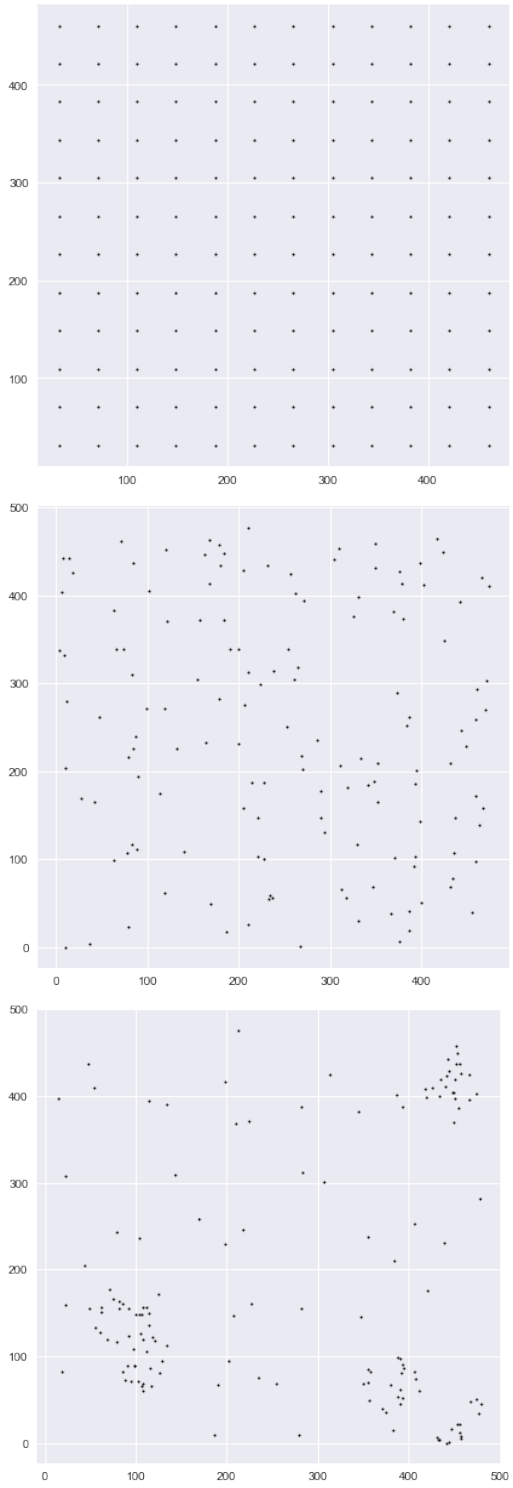


Figure 49: An example uniform points (top), random points (middle), and clustered points (bottom).

5.2.2 Data-generating processes

Separate sets of flows were generated using each of the three spatial interaction modeling frameworks, as well as from a null model that does not incorporate spatial structure. The DGP's for each of the four models are explained in detail in chapter 2 and are briefly reviewed here, along with their substantive parameterizations for the simulation experiments.

First, for the null model, an unconstrained gravity-type spatial interaction model is used, which assumes *a priori* that there is no relationship between the number of destination locations in a cluster and the number of flows to destinations. That is, spatial structure of locations does not impact the number of flows generated between locations[21]. The unconstrained gravity-type spatial interaction model, which is henceforth referred to as the gravity model, is given in its multiplicative form as,

$$T_{ij} = kV_i^{\mu}W_j^{\alpha}d_{ij}^{\beta} \tag{5.1}$$

where T_{ij} is the number of flows between origin i and destination j, k is a scaling parameter, which is equivalent to the intercept in a log-linear regression form, V_i is one or more variables representing origin propulsiveness, W_j is one or more variables representing destination attractiveness, d_{ij} is a variable that represents the separation or cost to travel between an origin and destination, and k, μ, α, and β are parameters to be estimated. Here, V_i and W_j are simplified to consider only a single variable each, which are specified as the origin population and the destination population, respectively. The parameters μ and α that are associated with V_i and W_j are set to 1 to indicate the positive relationships that would be expected between origin

[21]This holds for any of the family of gravity-type spatial interaction models and not just the unconstrained variety.

propulsiveness and the number of flows and destination attractiveness and the number of flows. In contrast, the distance-decay parameter, β, is set to -1 to indicate the negative relationship associated with increased costs to travel over longer distances. In this research, distances between locations are computed as the Euclidian distance between each origin and destination. Finally, k, is set to 1, which does not scale the flows and is therefore not of interest throughout the simulation experiments and results. By plugging in the populations, distances, and parameter values into equation 5.1, flows that follow the gravity model are generated. To add a stochastic element and to ensure the same number of flows in each realization of simulated flows (i.e., normalization), the flows were converted to probabilities that were used to generate a random realization of flows from a multinomial distribution with 5,000,000 observations. Probabilities were obtained using the following formula,

$$p_{ij} = \frac{T_{ij}}{\sum T_{ij}} \tag{5.2}$$

which is simply the division of the number of flows between each origin-destination pair by the total number of flows in the system. Of note, is that this simulation process implies for intra-zonal flows that a value of zero distance is raised to a negative power[22], which is undefined. To overcome this limitation, these intrazonal flow values were set to zero rather than computing the logarithm of zero-valued distances, which indicates intra-zonal flows are not part of the substantive model and will be revisited later.

The first specification that incorporates spatial structure is the CD model, which is given in its multiplicative form as,

$$T_{ij} = kV_i^{\mu} W_j^{\alpha} d_{ij}^{\beta} A_{ij}^{\delta} \tag{5.3}$$

[22]This is equivalent to a division denominator of 0 and similarly to taking the logarithm of zero in the log-linear form

where the symbology is the same for the gravity model with the addition of an accessibility term, A_{ij}, that captures the accessibility of destination j in relation to all other destinations and its associated parameter, δ, which was set to -1 in this simulation experiment. This negative parameter indicates the presence of competition effects whereby destinations that are more clustered receive fewer flows than would be expected in comparison to the gravity model. Following Fotheringham (1983a), accessibility is computed as,

$$A_{ij} = \sum_{\substack{k=1 \\ (k \neq j)}}^{n} \frac{W_k}{d_{jk}^{\sigma}} \tag{5.4}$$

where k denotes each alternative destination other than j, and which is essentially a distance-weighted sum of destination attractiveness that is moderated by σ. This parameter is a type of distance-decay that can be defined empirically through a grid search or via iterative substitution with β, though here is was always set to -3. A large negative σ indicates strong distance-decay, and therefore that the definition of clusters is a relatively local phenomena. Once CD flows are computed using equation 5.3, they are similarly normalized using their respective probabilities and a multinomial distribution.

Next, flows were generated using an SLX framework, which is the second framework that considers spatial structure. Instead of the general SLX specification that considers a spatial lag of each explanatory variable, only a simplified specification that considers a spatial lag of destination population is considered, which is given in its multiplicative form as,

$$T_{ij} = kV_i^{\mu} W_j^{\alpha} d_{ij}^{\beta} M W_j^{\phi} \tag{5.5}$$

where M is a row-standardized spatial weight matrix that denotes neighbors (i.e., spatial lag operator) of j and effectively computes the average of j's neighboring destination populations. This simplified SLX specification is considered because it

would seem illogical to conceive of a spatial lag of distance and only considering a spatial lag of destination populations provides a clearer means of comparison of the SLX and CD frameworks. Since locations are abstracted as points, it is natural to define neighbors using a k nearest neighbors with k set to 8. This parameterization of M was chosen because it implies a relatively local definition of neighbors and is equivalent to a queen definition of contiguity neighbors for uniformly distributed locations if each locations is aggregated to a a unique areal unit. The parameter ϕ was set to 3 in these simulations, which corresponds to strong positive spillovers or indirect effects that imply the importance of neighbor values and would generally be associated with spatial autocorrelation that cannot otherwise be accounted for by the explanatory variable that is not lagged. While M may also be defined using a parameterized distance-decay (Halleck Vega and Elhorst, 2015), like A_{ij} in the CD framework, this is not typically done in the spatial interaction literature and is not considered here. Similar to the gravity flows and CD flows, after the SLX flows were generated using equation 5.5 they were then normalized using their respective probabilities and a multinomial distribution.

A SAR spatial interaction specification was used as the third framework for incorporating spatial structure. Instead of using a SAR spatial interaction model with three autoregressive terms – one for the origins, one for the destinations, and one for a combination of origins and destinations – a simpler specification was adopted that only uses a single autoregressive term that is defined using a combination of origin spatial contiguity and destination contiguity. This simpler specification is part of a family of SAR spatial interaction models defined by (LeSage and Pace, 2008). They argue that such a specification is necessary because flows may influence nearby flows. It is not possible to express this specification in a multiplicative form similar to the gravity,

CD, and SLX models because the SAR spatial interaction model is an extension of a Gaussian process model and a Gaussian process is a log-linear approximation for the distributions of discrete flows that are typically observed. Therefore, the model specification is given in log-linear form as,

$$\begin{aligned} \ln T_{ij} &= \rho_{ij} M_{ij} y + k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij} + \epsilon \\ M_i &= I_n \otimes M \\ M_j &= M \otimes I_n \\ M_{ij} &= M_i \otimes W_j = M_i \otimes M_j = M \otimes M \end{aligned} \tag{5.6}$$

where the symbology is the same as for the SLX model, though now M represents a spatial weight that is constructed based on the 8 nearest neighbors of both origins and destinations. Two additional differences should also be noticed. First, T_{ij}, V_i, W_j and d_{ij} are now logged, which means this specification produces the logarithm of flows. Second, an error term, ϵ, has been included here because it is intrinsic to the DGP, which can be made more clear by expressing the DGP in matrix form as,

$$\ln T_{ij} = (I - \rho_{ij} M_{ij})^{-1} + (Z\zeta + \epsilon) \tag{5.7}$$

where $(I - \rho_{ij} M_{ij})^{-1}$ is the spatial autoregressive operator and $Z = \{V_i, W_j, d_{ij}\}$ in their logged form and $\zeta = \{\mu, \alpha, \beta\}$. The error term, ϵ, is drawn from a random normal distribution with a mean of 0 and variance equivalent to one quarter of the variance of $Z\zeta$[23]. This ties the variance of the error term to the variance of the flows before they are transformed by the spatial autoregressive operator and has the desired effect of maintaining a small and proportional stochastic component across realizations of the SAR DGP. Flows produced by the SAR DGP using equation 5.7

[23]Intra-zonal origin-destination pairs imply the logarithm of zero, and these values were forced to be zero after taking the log of inter-zonal values, which is commensurate with the previous DGP's.

cannot be normalized using a multinomial distribution like the previous frameworks because the SAR DGP produces a continuous variable. Therefore, the logged flows were transformed using the exponential operator and then normalized by computing the probabilities and multiplying them by the desired number of 5,000,000 total flows.

After 100 realization of flows have been simulated between the locations abstracted as points, for each DGP, and within each of the three scenarios, the data can then also be aggregated to areal units that represent coarser scales of analysis.

5.2.3 Aggregation and scales of analysis

It is common for spatial interaction data to be reported in aggregate for areal units, such as municipal boundaries or for spatial interaction data to be aggregated to areal units because origin propulsiveness and destination attractiveness proxies are only available at a coarser scale. While the effects of aggregation on regression variables defined using distance have been studied and are understood in regression models more generally, they have not been explored in the context of spatial interaction and spatial structure. Therefore, the data generated between point locations was aggregated to two grids that represent increasingly coarse scales. The first scale was generated using as 24 by 24 grid composed of 576 areal units and the second scale was generated using a 12 by 12 grid composed of 144 areal units. The simulated flows were then aggregated[24] amongst the 331,776 origin-destination pairs for the first scale and the 20,736 origin-destination pairs for the second scale, while the location populations were aggregated to the 576 areal units for the first scale and 144 areal units for the second scale. An example of the spatial distributions of aggregated

[24]Here aggregation implies a summing of all observations from a lower scale.

locations for each scenario (uniform points, random points, and clustered points) is provided for the two scales in figures 50 and 51, respectively. The aggregation of the uniform points is peculiar in that aggregating to the 24 by 24 grid results in a perfect checkerboard[25] pattern with red values indicating a single location and grey values indicating none, and aggregating to the 12 by 12 grid results in no pattern since each areal unit contains a single point.

The inter-location distances for origin-destination pairs were re-calculated between centroids of the areal units for each scale. In addition, the spatial structure components of each model were also recomputed using coarser level measurements. For the CD model, this entails computing the accessibility term using centroid distances and aggregate populations. In contrast, contiguity based weights are typically used in the spatial interaction literature for the SAR and SLX models and therefore a queen definition of contiguity amongst areal units was used here for both aggregate scales.

After completing the aggregation process, there was a total of 3,600 datasets (100 realizations x 3 scales x 4 DGP's x 3 spatial structure scenarios) ready for model calibration in order to obtain parameter estimates and evaluate the various frameworks and scenarios.

5.2.4 Model calibration

Parameter estimates were obtained for each dataset using four model specifications – the three frameworks that account for spatial structure and the gravity model that

[25]Here, 8 nearest neighbors has to be used for defining a spatial weight matrix since the checkerboard pattern implies areal units that contain a location do not have any contiguous neighbors.

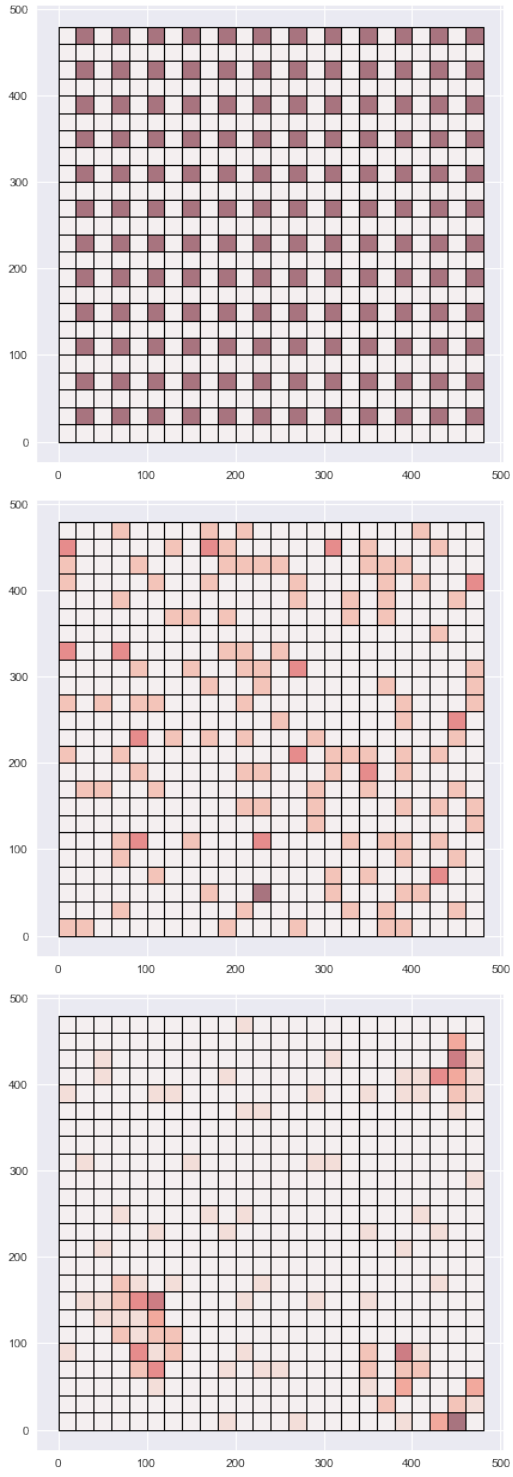


Figure 50: An example of uniform points aggregated to the 24 by 24 grid (top), of random points aggregated to the 24 by 24 grid (middle) and of clustered points aggregated to the 24 by 24 grid (bottom).

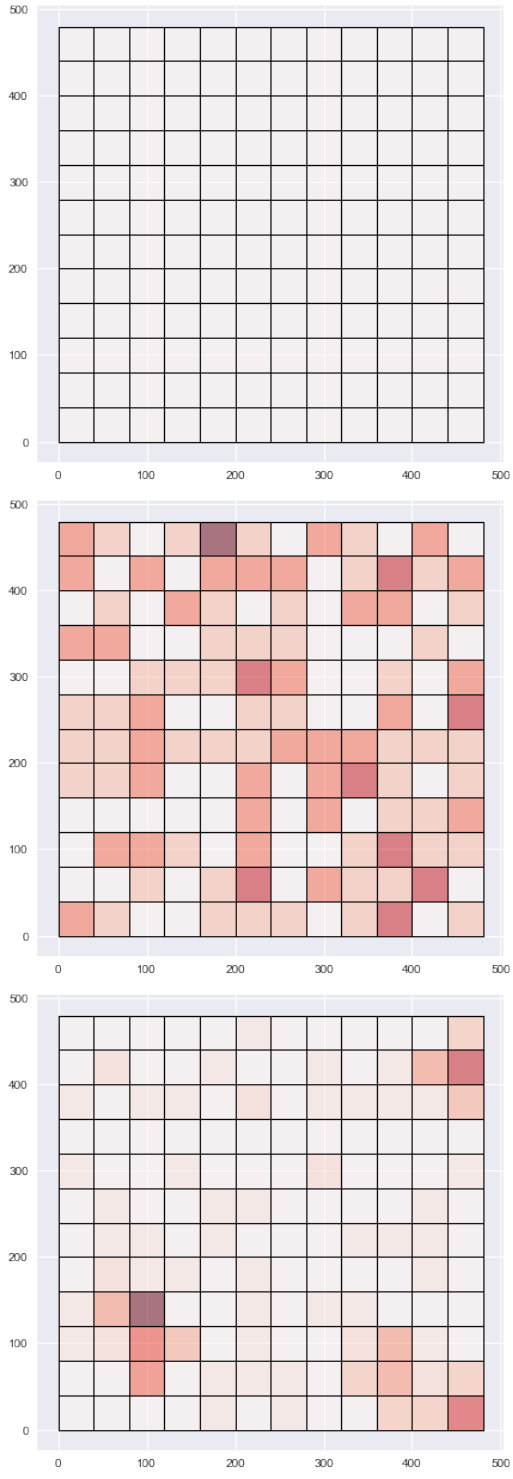


Figure 51: An example of uniform points aggregated to the 12 by 12 grid (top), of random points aggregated to the 12 by 12 grid (middle) and of clustered points aggregated to the 12 by 12 grid (bottom).

serves as a null specification. When employing the gravity model, CD model, and SLX model specifications, Poisson regression was used for calibration after filtering out the intra-zonal observations that correspond to forced zeros and then taking the logarithm of the explanatory variables. However, an exception to this was when these models were calibrated on data generated from a SAR DGP. Instead, a Gaussian regression (i.e., ordinary least squares) was instead employed since the SAR DGP simulates a continuous variable rather than discrete counts. In these instances, the flows were logged in addition to the explanatory variables. Finally, when employing the SAR model specification, the general method of moments[26] was used for calibration. Both the flows and the explanatory variable are still logged, though a notable difference between the SAR and gravity models is that the SAR model requires the inclusion of intra-zonal flows to maintain the necessary dimensions within the data. As a result, the same strategy used to generate the data of setting intra-zonal observations to zero after taking the necessary logarithms was also used here[27]. Once calibration of the models was complete, the resulting parameters estimates were collected and compared to their expected values. Of note is that the interpretation of LeSage and Thomas-Agnan (2015) that requires scalar summary measures is not considered here because these are only for the effects associated with locational attributes, and the distance-decay is the primary interest when evaluating spatial structure effects[28].

[26]More specifically, the PySAL GM_lag function was used.

[27]In this case, there is no need to define intra-zonal distance because the number of flows is set to zero regardless of the values of the explanatory variables

[28]In fact, code to compute the scalar measures for the SAR model was developed and tested on some of the simulated data. It was found that origin and destination parameter estimates from the scalar summary estimates were biased in the scenario where the conventional origin and destination parameter estimates were not (i.e., a SAR model was calibrated on data derived from a SAR DGP, the data was not aggregated and the correct spatial weights matrix was specified). In addition, it

5.3 Results

Since the results include parameter estimates from 14,400 models estimated on the 3,600 datasets, it is challenging to succinctly display the results in their entirety. Therefore, only subsets of the results will be presented here, though the full results are available in appendix C. First, results are presented for each of the scenarios (uniform, random, clustered) at the point scale in order to get a baseline understanding of model estimation quality and the effects of spatial structure. Subsequently, results regarding the effects of aggregation are presented.

5.3.1 Baseline results and spatial structure effects

Baseline results are presented in figures 52, 53, and 54, which focus on the parameter estimates obtained at the scale at which the data were generated (i.e., points). Several important trends can be extracted from these figures that provide boxplots of the 100 parameters obtained for each circumstance. First, the model specifications are validated because little-to-no bias or variation occurs whenever a model specification is calibrated on data from its respective DGP (figure 52), regardless of whether or not the points have spatial structure (i.e., clustering). This is clear, since the boxplots in figure 52 are always shrunk to the mean (red line), which is centered on the true parameter values (blue star).

Second, an analysis of false positives is possible by examining figure 53. This figure presents the results from calibrating each model specification on data generated from a

would have required weeks to compute these scalar summary estimates for all of the simulated data. For these two reasons, and that the focus is centered on distance-decay parameter estimates, the scalar summary measures were not pursued here.

null model (i.e., gravity model) that does not include any spatial structure component. The results show that in all three scenarios, the CD model and SLX models always produce coefficients for their respective spatial structure components, δ and psi, that are near-zero and therefore do not indicate spatial structure effects when none are present[29]. In contrast, the SAR model produces non-zero spatial structure effects in all three scenarios and also produces distance-decay estimates (i.e., β) that are severely under-estimated in magnitude. These results provide evidence that the SAR specification is prone to overfitting to the data even when there is no spatial structure effect, and is in line with previous observations (LeSage and Pace, 2008; de la Mata and Llano, 2013; Kerkman *et al.*, 2017) that the spatial autoregressive component competes with the distance variable to explain the same variation. Therefore, though the SAR specification may be able to better recover some locational parameters in certain instances according to LeSage and Thomas-Agnan (2015), it does so at the cost of the interpretability of the distance-decay, which is an important behavioral indicator.

Third, it is instructive to analyze the results summarized in figure 54, which denote the outcome of calibrating a basic gravity model on data from the DGP's that account for spatial structure to gauge the effect of ignoring spatial structure. For data obtained via the CD DGP, the gravity model distance-decay parameter β is underestimated in magnitude while the destination parameter estimate $\hat{\alpha}$ varies around the true value. In the uniform scenario, the bias is minimal, since there is little-to-no variation amongst the distributions of distances between each location and all other locations. However, in the random and clustered scenarios, the bias and

[29]These estimates also resulted in very small t-values that would prompt one to conclude that these estimates are not statistically significant.

variation increase for both parameters and in the clustered scenario becomes strong enough that the destination population parameter estimate can take on a negative sign and the distance-decay can take on a positive sign! This is the circumstance outlined by Fotheringham (1981) where strong spatial structure has the effect of biasing distance-decay enough to yield unintuitive interpretations that suggest more trips can be expected over longer distances. In comparison to data derived from the CD DGP, the data derived from the SLX DGP does not produce any strong biases in any of the scenarios when parameters are estimated with a gravity model. At first this seems unintuitive, but if it is recalled that the destination populations are generated from a random uniform distribution, then these results may be explained. Since each location has the same number of neighbors, and since the values of these neighbor's destination populations is a sample from a random uniform distribution, then the average of each location's neighbors will be about the same. Hence, there is no variation in the spatial structure component of the SLX model, unless there is also spatial clustering of the locational attribute values (i.e., spatial autocorrelation). This confirms that the spatial structure effect captured by the CD model is entirely separate from those caused by spatial autocorrelation, though it also recognizes that the two phenomena are linked. Finally, for the data generated by the SAR DGP, calibrating the gravity model results in unbiased origin and destination population parameter estimates; however, the distance-decay parameter estimates are severely over-estimated in magnitude. Moreover, this trend is amplified as spatial structure is increased from the uniform scenario to the random scenario and then to the clustered scenario. This is likely due to the fact that the variation in the spatial autoregressive component strongly resembles the variation in distance and so when only distance is

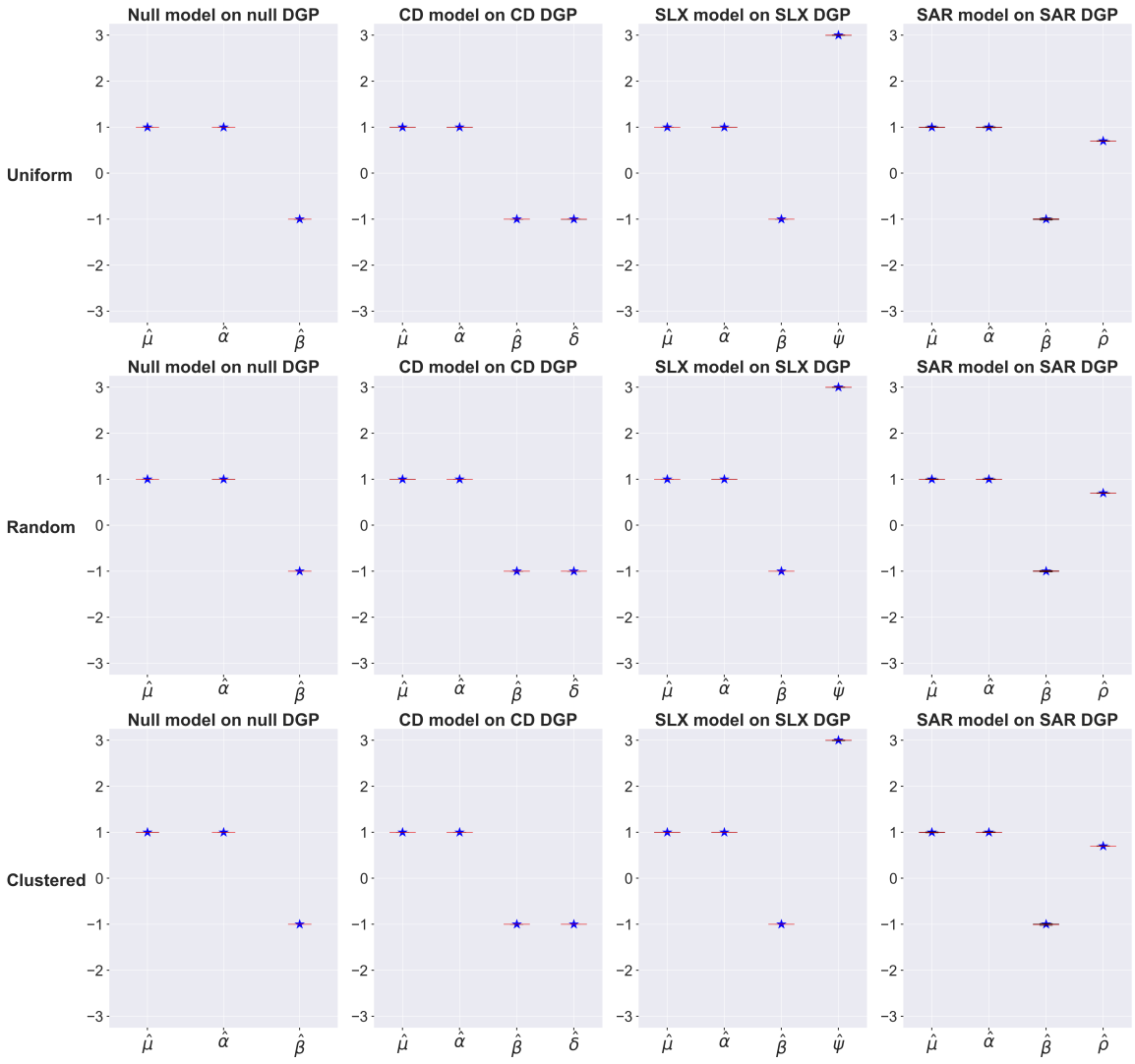


Figure 52: Results for models calibrated on datasets from their associated data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

available, the model tries to compensate by also assigning the effects associated with the autoregressive component to distance.

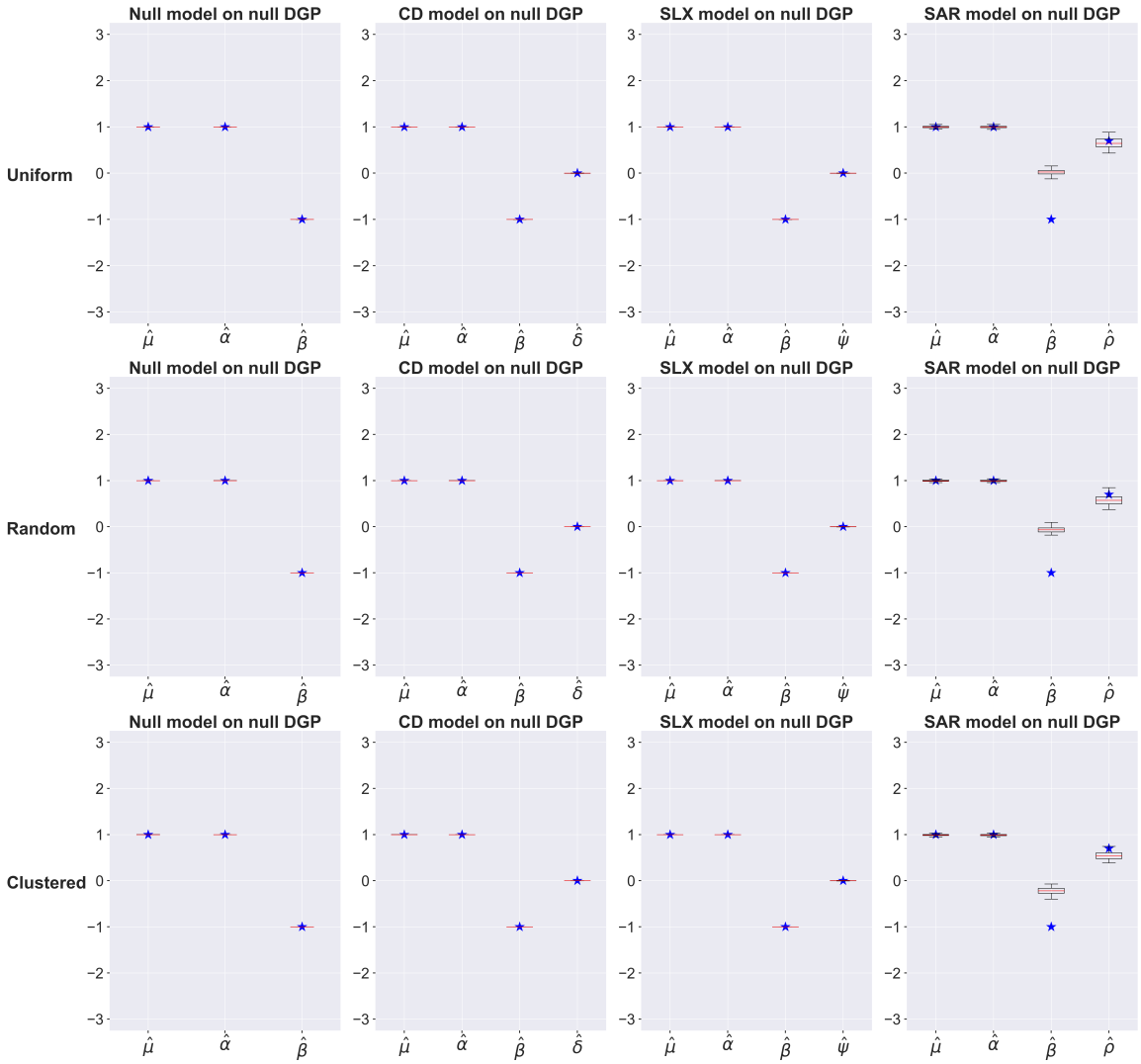


Figure 53: Results for models calibrated on datasets from the null gravity model data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

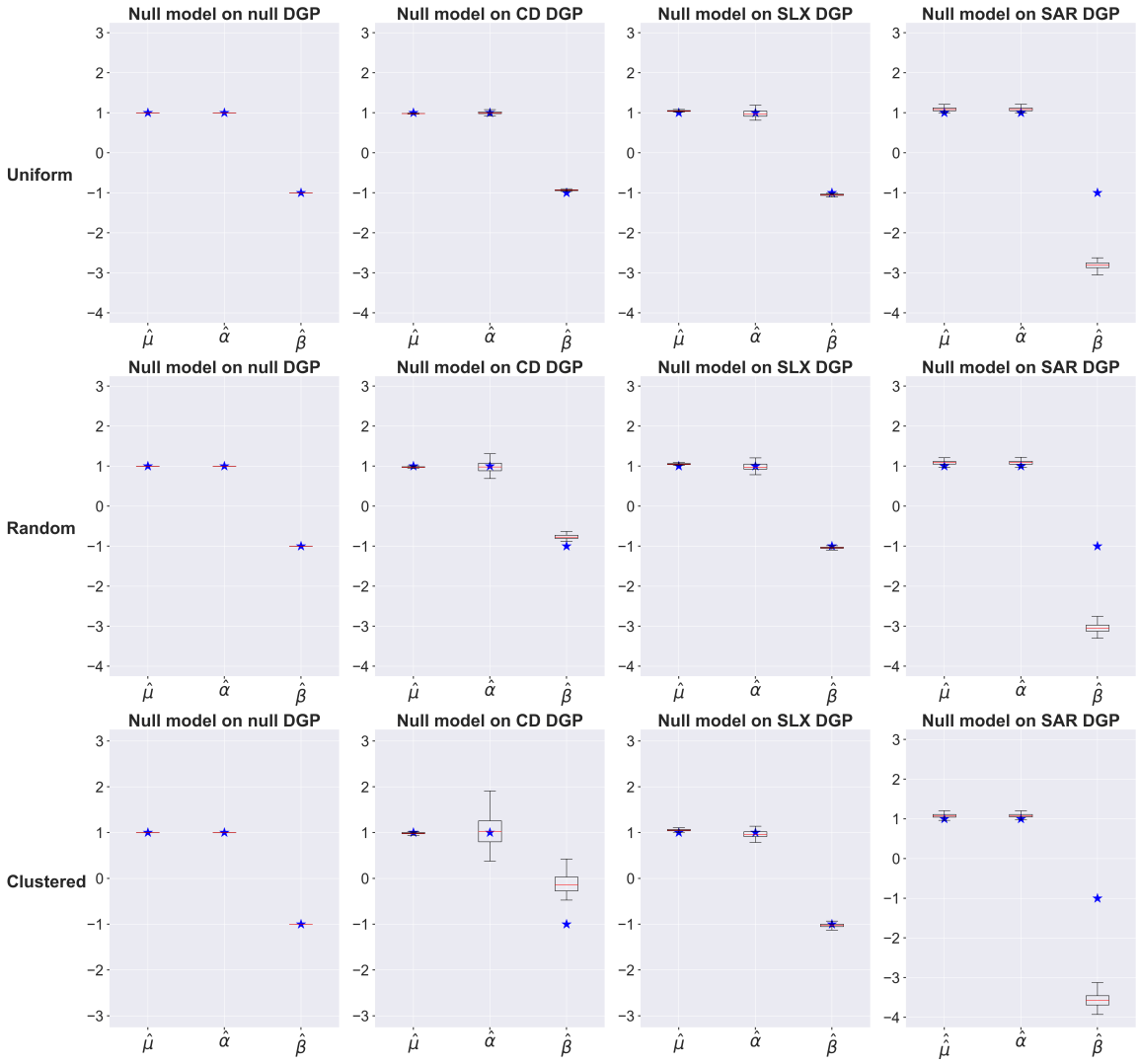


Figure 54: Results for the null model calibrated on datasets from each of the alternative data-generating processes. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

5.3.2 A cross-comparison of spatial structure specifications

The remaining results pertain to the parameters obtained from cross-calibrating the CD, SLX and SAR models on data derived from each of the three associated data-generating perspectives. These results are provided in figures 55 - 57 and facilitate an analysis of the consequences of calibrating each model with an incorrect underlying DGP. For the data derived from a CD DGP (figure 55), the outcome of calibrating an SLX model is similar to that of the null gravity model in that biases in the destination population parameter estimate, $\hat{\alpha}$, and the distance-decay parameter estimate, $\hat{\beta}$, are clearly seen. In addition, the SLX spatial structure parameter estimate, $\hat{\psi}$ often takes non-zero values[30], that increase as spatial structure effects increase. This, consequently, has the effect of inducing bias in the parameter estimates. Similarly, calibrating the SAR model on data from the CD DGP produces biased parameter estimates where the value of the SAR spatial structure parameter estimate, $\hat{\rho}$, becomes more extreme as spatial structure effects increase but the bias associated with $\hat{\beta}$ is not reduced. It can therefore be concluded that neither the SLX model nor the SAR model can account for the spatial structure effects contained in the CD DGP. It is important to also note that the SAR model still introduces bias into the distance-decay parameter estimate similar to when the data are generated from the null gravity DGP. In fact, this is also the case when the data are generated from the SLX DGP (figure 56) and demonstrates that the SAR model will potentially produce a non-zero spatial autoregressive component regardless of the underlying DGP.

For the SLX DGP (figure 56), the results are less interesting due to the fact that

[30]Out of the 300 realizations associated with these results, only 4 produced small t-values that would lead us to reject $\hat{\psi}$ as statistically insignificant.

previous results showed that calibrating the null gravity model instead of an SLX model does not result in biased parameters when there is no spatial clustering amongst the locational attributes. These conclusions are similar for the situation where the CD model is incorrectly calibrated on data from the SLX DGP; though it is possible to get a non-zero parameter estimate for δ in the uniform scenario, this issue disappears when there is spatial structure present in the locations (i.e., random or clustered). These non-zero values may be caused by collinearity between the CD spatial structure component and destination population, which also becomes slightly biased. Therefore, the risk of obtaining a false positive for the δ parameter will be low in since it only occurs in the circumstance where there is no spatial clustering amongst locations and no spatial autocorrelation in the locational attributes and this unlikely in reality.

Lastly, for the SAR DGP, the outcomes of incorrectly calibrating a CD model or a SLX model have some similarities and some differences. They are similar in that they both over-estimate the magnitudes of the distance-decay parameter estimates, similar to when the null gravity model is calibrated on data from the SAR DGP. However, they differ in that as the spatial structure in the points increases, the SLX spatial structure parameter estimate, $\hat{\psi}$, becomes increasingly biased, while the CD model produces a spatial structure parameter, δ, that is appropriately zero when there is spatial structure. This provides further evidence that there is also a low risk of obtaining a false positive for the δ parameter in the CD model in the circumstance of spatial clustering amongst locations and no spatial autocorrelation in the locational attributes.

Overall, these results support the conclusion that these three spatial interaction model specifications account for different facets of spatial structure in the locations of a spatial interaction system. More specifically, the CD model seems to be the most

robust, while the SAR model and SLX model frequently produce false positives for their respective spatial structure parameters.

5.3.3 Aggregation effects

An analysis of the results for the datasets aggregated to the grids of areal units can be greatly simplified for two reasons. First, the results for the two different grids are similar, with the 12 by 12 grid resulting in more pronounced patterns than the 24 by 24 grid. Therefore, only results for the 12 by 12 grid are presented here and the full results are made available in appendix C. Second, the results pertaining to the uniform points are very similar when they are aggregated because the points may be perfectly aggregated to the two different grids and therefore minimally bias the inter-location distances and definition of neighbors. For the 24 by 24 grid, an 8 nearest-neighbors definition was used to operationalize the SLX and SAR models since the checkerboard pattern means the areal units containing locations do not have any contiguous neighbors (i.e., that share a common border) that also contain locations, and this is the same definition of neighbors used to generate the data between the points. Similarly, when the points are aggregated to the 12 by 12 grid, each areal unit has 8 neighbors using a queen definition of contiguity, which again matches the spatial weight used to generate the data. Therefore, aggregation effects for the uniform scenario are minimal and results will only be presented for the random and clustered scenarios.

In general, data aggregation causes there to be more variation in the sets of

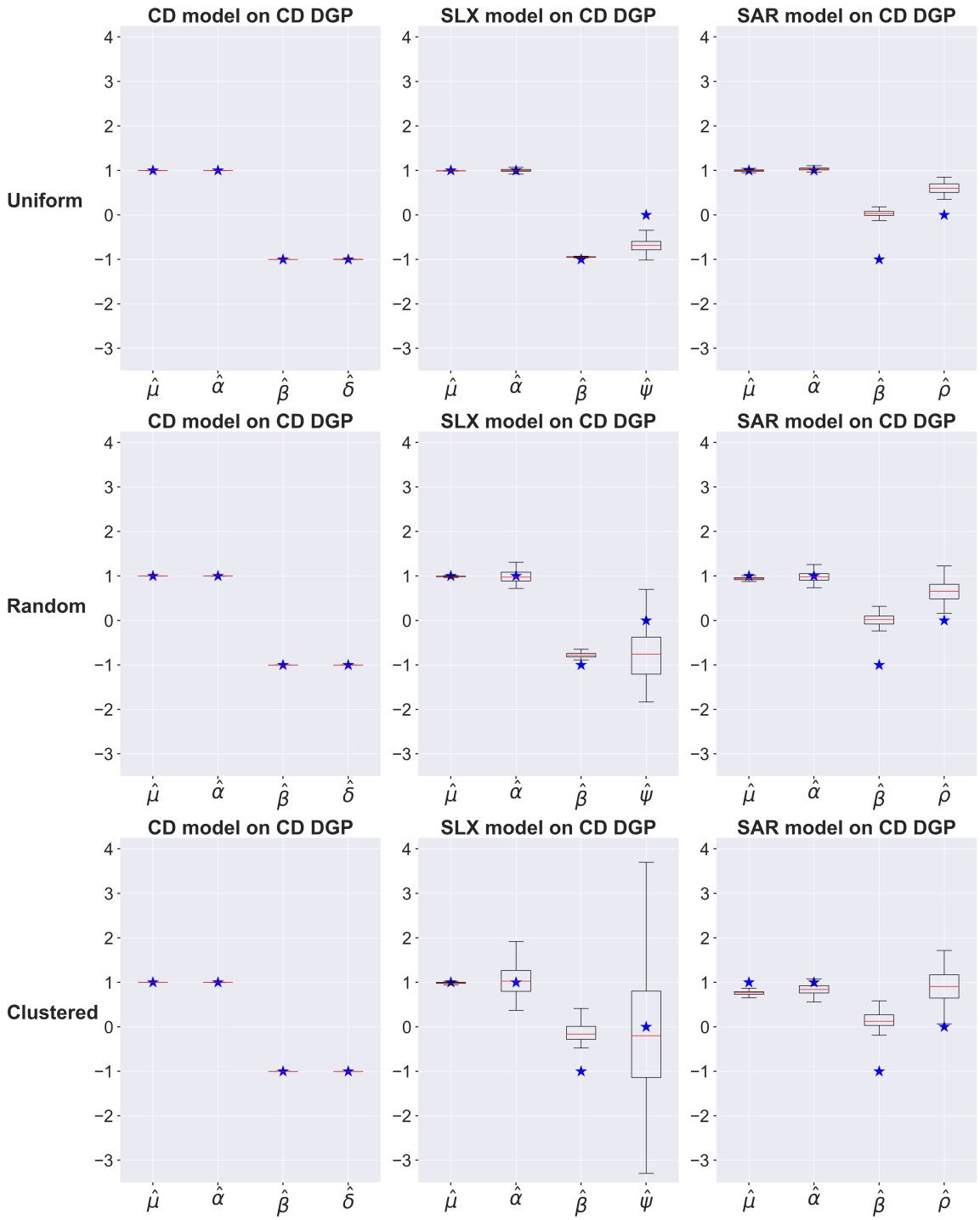


Figure 55: Results for CD, SLX, and SAR models calibrated on datasets from the CD data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

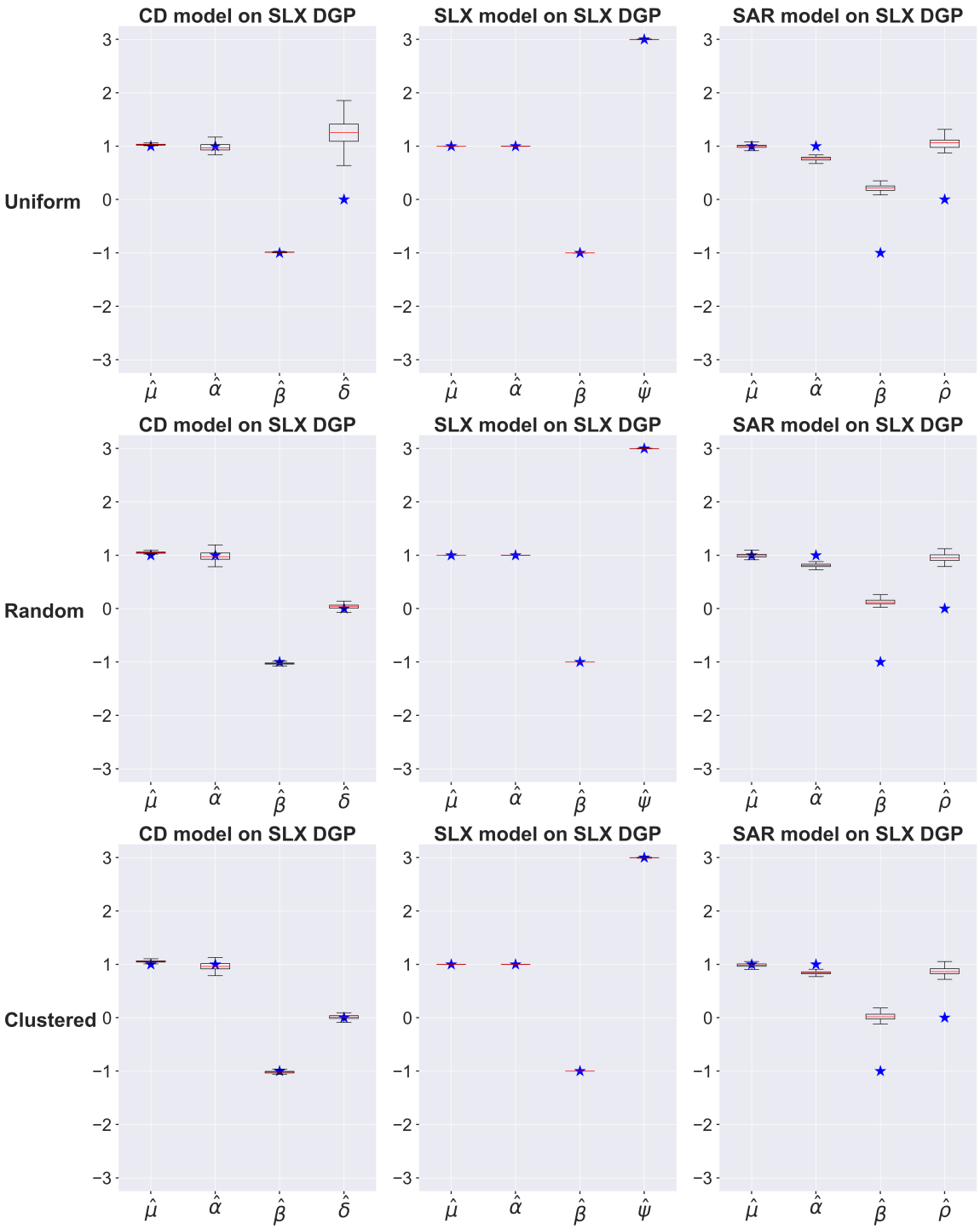


Figure 56: Results for CD, SLX, and SAR models calibrated on datasets from the SLX data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

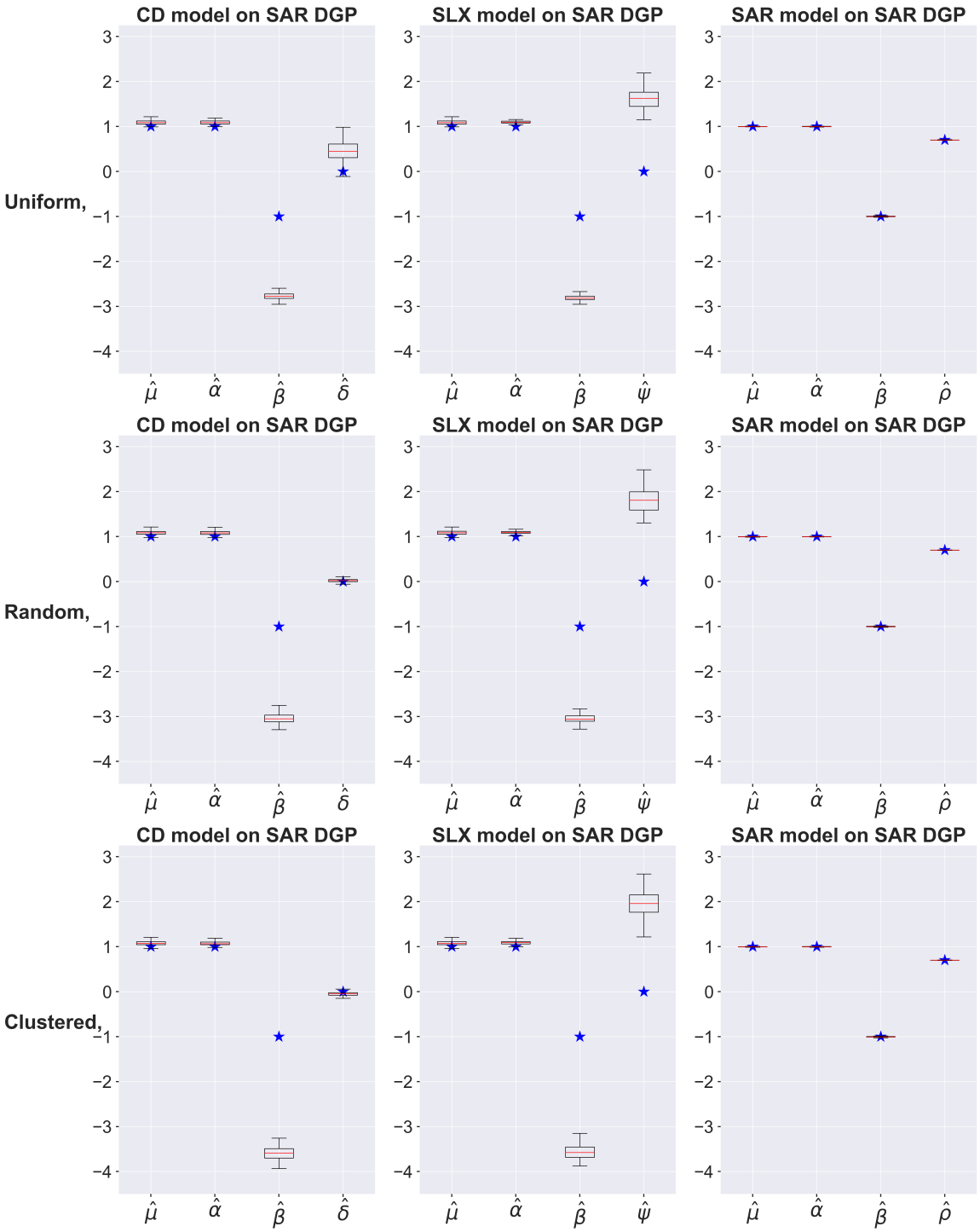


Figure 57: Results for CD, SLX, and SAR models calibrated on datasets from the SAR data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

calibrated parameter estimates; however, the nature of the variation is different depending on the model specification and the DGP from which the data are derived. Here, the discussion returns to the circumstances of framework validation, false positives, and the outcomes of ignoring spatial structure when it is present. Any discrepancies between these additional results and the baseline results can be attributed to the aggregation process.

In terms of framework validation for the aggregate data, it can be seen in figure 58 that all of the frameworks no longer produce parameter estimates without bias and/or variation. Each framework is discussed in order from the least affected by aggregation to the most seriously affected by aggregation. Though the null gravity model typically replicates the known parameter values reasonably accurately, there is some bias in the distance-decay estimates for the clustered scenario. For the CD model there is some variation and the destination population parameter estimate is underestimated, while the spatial structure parameter estimate ($\hat{\delta}$) is overestimated when there is intense spatial structure. Moreover, for neither the random or clustered scenario do the results produce parameter estimates with the incorrect sign that would imply unintuitive interpretations. In contrast, the combined effect of aggregation and spatial structure cause the spatial structure parameter estimate ($\hat{\psi}$) for the SLX model to be under-estimated, which is likely due to the contiguity-based definitions of neighbors at the aggregate level not being representative of true 8 nearest neighbor definition. Finally, the SAR models demonstrate erratic behavior. In the random scenario, the parameters are all over-estimated in magnitude, though they still take on the correct signs. In the clustered scenario, the origin and destination population parameter estimates are strongly negative and the distance-decay parameter estimate is severely over-estimated, reaching up to 25 times larger of an effect than the true

parameter would indicate. These results suggest that the cumulative distance-weighted accessibility term of the CD model is the most robust specification to account for spatial structure when the data are aggregated.

Results pertaining to false positives for the aggregated data are similar to those from the disaggregate data (figure 59). The CD model and SLX model still perform well and generally produce a near-zero parameter estimate associated with their respective spatial structure components when the data are produced from a null DGP that does not include spatial structure. Furthermore, the SAR model still competes with the distance variable and produces a non-zero, albeit small, parameter estimate for ρ even when data are produced from a null DGP.

The consequences of neglecting spatial structure for the aggregate data (figure 60 are also similar to the results from the disaggregate data for the CD DGP and the SLX DGP. For the former, the consequence is biased parameter estimates that increase with more clustered locations and potentially produce incorrect signs. For the latter, there is little to no effect, which is expected since it has been established that the lack of clustering amongst the location attribute value means the SLX DGP is not effectively incorporating any spatial structure component. Again, the results associated with the SAR framework are an exception and are erratic compared to the disaggregate data. In the random scenario, the origin and destination parameters are slightly over-estimated in magnitude, and the distance-decay parameter is strongly over estimated. However, in the clustered scenario, the origin and destination populations parameter estimates are strongly negative and the distance-decay parameter estimate is severely over-estimated, reaching up to almost 50 times larger of an effect than the true parameter would suggest. The additional insight that these results provide is that when data generated from the SAR DGP are aggregated, calibrating the null gravity

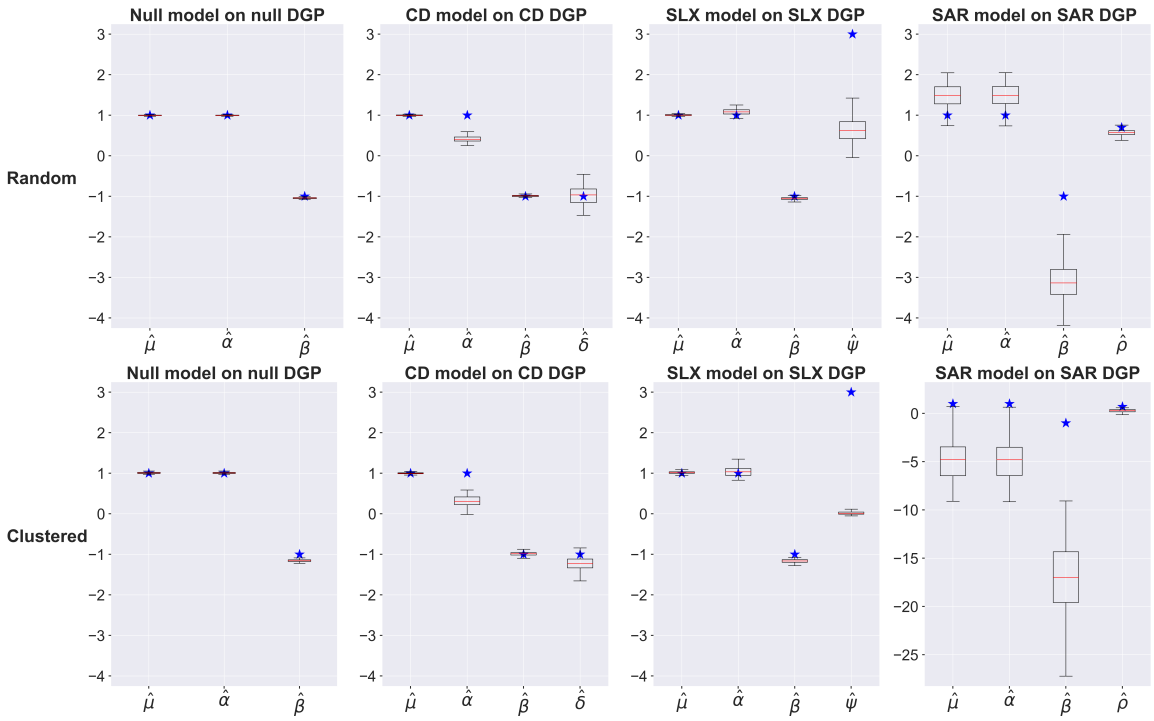


Figure 58: Aggregation results for models calibrated on datasets from their associated data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

model may produce extremely skewed results that provoke one to instead use the SAR model. However, the extremely skewed results may be due more to aggregation than to having the incorrect model as it was shown in figure 58 that results from a SAR model on the aggregated SAR DGP data are still very skewed.

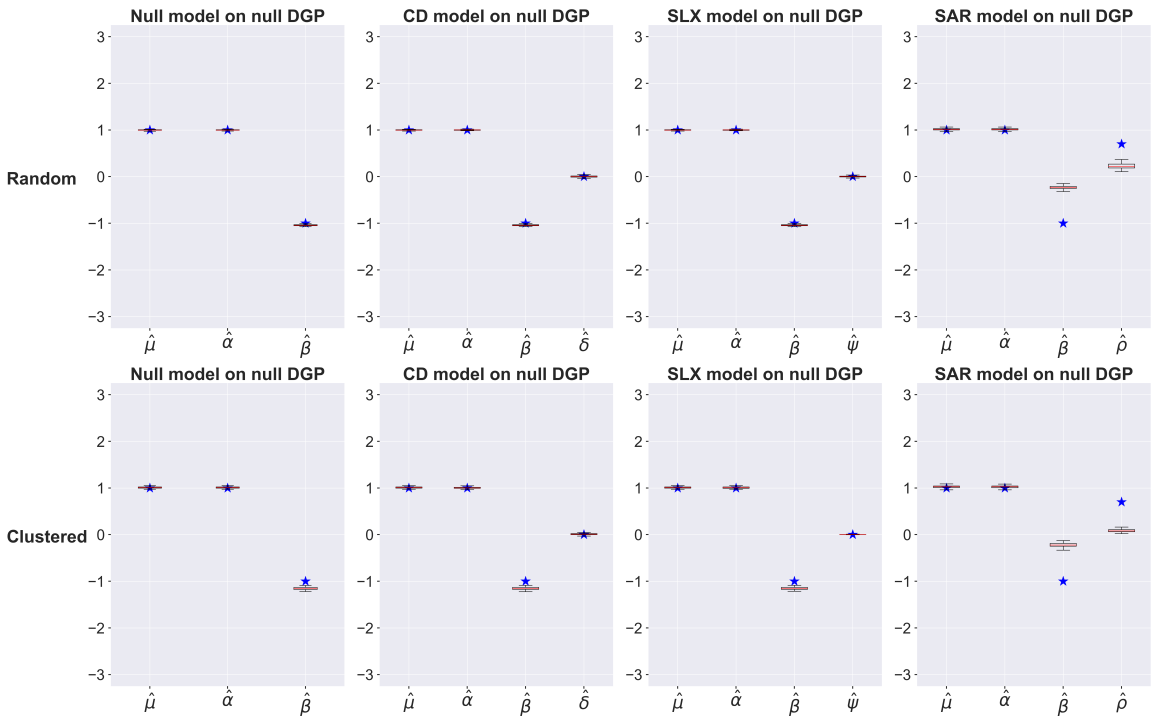


Figure 59: Aggregation results for models calibrated on datasets from the null gravity model data-generating process. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

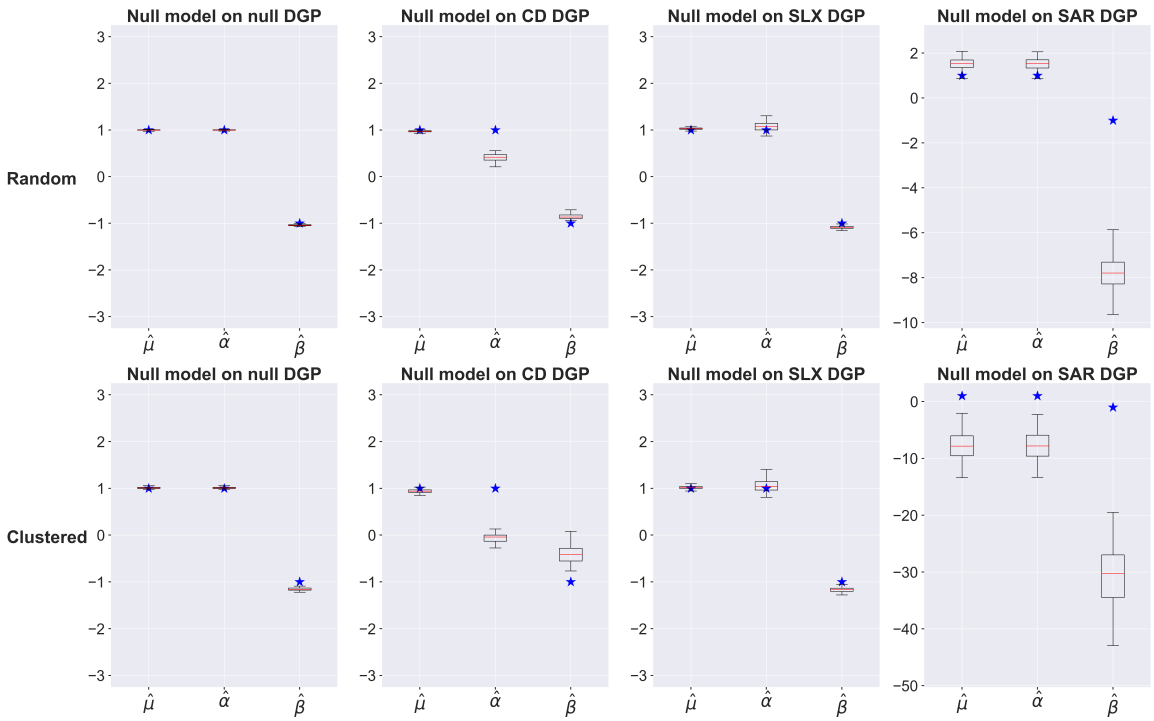


Figure 60: Aggregation results for null models calibrated on datasets from each of the data-generating processes. Top row provides results for the uniform scenario, middle row provides results for the random scenario, and the bottom row provides results for the clustered scenario.

5.4 Moving forward

In this chapter, an extensive simulated-based experiment was carried out to explore the effects of spatial structure, which was defined as the spatial clustering of the locations in a spatial interaction system. In particular, three spatial interaction specifications that have been proposed to deal explicitly with the problem of spatial structure were investigated. The ramifications of using aggregated spatial interaction data for model calibration were also explored.

The results produced several conclusions that can help guide empirical spatial interaction modeling. First, the SAR model is not a sufficient specification simply based on the model generating a non-zero autoregressive parameter estimate when no effect is present. It is shown that this can occur for any of the DGP's tested here, including the null gravity model. In fact, it was shown that the SAR model competes with the destination variable and produces biased distance-decay estimates. Therefore, the SAR spatial interaction model should be avoided when distance-decay is of substantive interest. Second, the spatial structure effect resulting from clustered locations is a distinct effect that is independent of the locational attribute values. It was shown that the CD model with a cumulative distance-weighted accessibility term was the most robust means for accounting for potential spatial structure effects. Lastly, aggregation can cause increased bias and variation, which can produce less reliable parameter estimates regardless of whether or not the correct model specification is employed. As a result, it is prudent to avoid aggregation bias by using the most accurate measure of distance and spatial neighbor associations that can be obtained.

These lessons will be employed in the subsequent chapters that focus on empirical spatial interaction models of bike and taxi trips in New York city.

Chapter 6

LOCAL MODELS OF LOCATION CHOICE IN NEW YORK CITY

6.1 Introduction

The previous chapters reviewed the theory and concepts underlying spatial interaction models, demonstrated the shortcomings of some exploratory techniques and investigated the effects of spatial structure on spatial interaction. The following two chapters build on these themes by calibrating various models of spatial interaction using both census commute-to-work data and the bike and taxi trips described in chapter 4. The former represents a traditional spatial interaction dataset while the latter two datasets are examples of newer, bigger data sources. In this chapter, the bike and taxi trips are evaluated for potential and pitfalls compared to the traditional census spatial interaction data and an assessment of the utility of these new data sources for capturing commute-to-work processes is provided. Since people who travel during morning rush hour by bike or taxi are unlikely to represent the average commuter, it is expected that model calibration results from each of these data sets will differ from the census data and from each other. Thus, it is interesting to determine how bike and taxi commuters behave differently than the average commuter. It is also valuable to identify whether or not the conclusions from models for the taxi and bike data overlap with the census data, since they can be collected more frequently than the census data and they are easier to collect because they can be harvested automatically by sensors.

Contemporary spatial interaction modeling literature draws largely upon the

wider statistical and econometric literature rather than on the geographical literature and consequently the focus is typically on global models that provide monolithic interpretations, rather than on local models. Therefore, local spatial interaction models are relatively absent from the recent literature, despite being popular several decades ago (see for example (Fotheringham, 1981) and others cited within). Local models, such as origin- or destination-specific models, produce a set of coefficients for each location that can be mapped and inspected for spatial variation in the processes that generated the data. The increased volume (i.e., spatial coverage) of big data provides a larger set of spatial interaction locations and therefore a higher resolution surface of local parameter estimates can be obtained, which may provide more nuanced and refined interpretations. Therefore, local models will also be employed in this chapter to determine how commuting behavior may vary across space.

The first section of this chapter establishes a baseline model of commute-to-work behavior using census spatial interaction data and considering both traditional and non-traditional location variables. Next, the bike and taxi data are subset using commuting semantics, such as the time of the day and the day of the week, and then the established baseline model is calibrated on them. By calibrating the same model across multiple datasets, it is possible to directly compare model results across the three distinct data sets. Throughout the chapter, perspectives on the potential and pitfalls of newer data sources in spatial interaction models of commuting are offered.

6.2 A baseline commute-to-work model

This initial section is devoted to establishing a basic spatial interaction model of commuting behavior using census commute-to-work data. Commuting has a long

history of being studied via spatial interaction models (Thorsen and Gitlesen, 1998; Gitlesen and Thorsen, 2000; Griffith, 2009a; Gitlesen *et al.*, 2010; Farmer, 2011). In this context, the number of commuting flows between locations often varies with the distance that needs to be traveled, population, income, the number of employed individuals and the number of job opportunities, amongst other factors. These well-known variables may also be extracted from the census and are therefore examined first.

The census commute-to-work data from the Census Transportation Planning Products (CTPP) unit of the Census Bureau, which is based on the 2006-2010 American Community Survey (ACS) contains 3,331,059 journeys between census tracts in New York City that occur between 138,524 origin-destination pairs where the origins are 2,115 census tracts containing individuals' residences and the destinations are 2149 census tracts containing individuals' workplaces. Removing intra-zonal flows reduces the dataset to 3,110,101 trips between 136,569 origin-destination pairs where the origins are still 2,115 census tracts but the destinations are now 2,136 census tracts. The reduction in the number of destinations was due to the fact there were 13 tracts that only had intra-zonal observations. After adding zero flow observations for the origin-destination pairs that did not report any commutes, the total number of observations becomes 4,517,640. This means that according to the census there are approximately 4,379,116 zero flows, which implies that the majority of the origin-destination pairs do not have any commuters traveling between them. The number of observations was further reduced to 4,483,520 after removing origin-destination pairs that do not contain a valid locational attribute, such as census tracts that do not have any observation or where a value of zero was recorded. The former is not valid input while the latter is an issue since the Poisson log-linear modeling framework requires

the logarithm of the explanatory variables. The final number of origins was then 2100 tracts and the final number of destinations was still 2136 tracts.

Commuting is perhaps most obviously thought of as the process of traveling from one's home residence to one's place of work. However, commuting routes are jointly determined by the locations of the residence and the place of work and individuals do not typically choose their place of work given that they live in a certain location. Rather, a more likely scenario is that given a place of work, an individual may choose to live somewhere that satisfies their preferences and minimizes the use of resources to get to work each day. Therefore, the commuting data were modeled using an attraction-constrained (destination-constrained) spatial interaction model that seeks to allocate the individuals arriving at each destination to the set of origin locations based on distances and origin locational attributes and is given in multiplicative form with a power function of distance-decay as,

$$T_{ij} = B_j V_i^{\mu} D_j d_{ij}^{\beta} \tag{6.1a}$$

$$B_j = \Big(\sum_i V_i^{\mu} d_{ij}^{\beta} \Big)^{-1} \tag{6.1b}$$

where T_{ij} is the number of flows between location i and destination j, D_j is the number of trips that terminate at destination j, V_i is a vector of origin location attributes, d_{ij} is the distance between i and j, and B_j is a balancing factor that ensures the total number of inflows D_j is replicated in the predicted flows. In this context, the primary interest is in describing the nature of the origins as if they were a set of destinations that an individual may choose from. This is accomplished by estimating[31] the origin

[31] All model calibrations in this chapter were carried out using custom software that leverages the sparse design matrices implied by constrained spatial interaction models for efficient computation

location parameter vector μ and the distance-decay parameter β and interpreting them in the context of commuting in New York City.

6.2.1 Census locational variables

Combinations of the variables introduced in chapter 4 were tested and a final set (table 2) of explanatory variables was chosen by using the (pseudo) adjusted R^2 and the AIC as indicators of goodness-of-fit[32]. The SSI and SRMSE were also available; however, they are both highly sensitive to the number of zero flows, which makes it difficult to use them to assess model fit more generally. While the SSI indicates little-to-no model fit in this scenario, the SRMSE has no upper bound and it becomes difficult to distinguish whether a large SRMSE is due to a larger number of flows or a lack of underlying associations or both.

Several variables were initially removed due to collinearity. First, population, number of housing units, and the number of employed individuals were all strongly correlated (Pearson's r correlation coefficient > 0.95) and therefore only the number of employed individuals (emp) was included in the model, since it is most closely related to the phenomenon of commuting. Second, population density and housing density were also strongly correlated and in this case, housing density (hd) was chosen because housing is directly related to the task of choosing a residential location to commute from. Thirdly, average income and percentage of people living in poverty were similarly strongly correlated, though inversely related, and in this case, it is not clear that

(Oshan, 2016). The code is part of the SpInt module of the Python spatial analysis library (PySAL) and is available at *https://github.com/pysal/spint*

[32]Higher R^2 and lower AIC indicate better model fit. In all cases an increase in the R^2 also resulted in a decrease in AIC making these two commensurate as a model fit criterion.

one is theoretically preferable over the other. It is equally conceivable that people would like to live in communities where the average income is higher (i.e., low poverty) and that people want to avoid areas of high poverty when possible. As a result, average income (inc) was selected over poverty based on the fact that average income results in a slightly better model fit and the coefficients on the other variables in the model are similar regardless of whether average income or poverty is included. Finally, Euclidian distance (dist) was computed between origin and destination centroids and entered into the model because no other form of separation/cost is available from the census. Manhattan distances were also computed and were highly correlated with the Euclidian distances (Pearson's r correlation coefficient > 0.98). Since the power function of distance is employed, which is not sensitive to the scale of the distances, either type of distance can be entered into the model with very similar results and in this case Euclidian was used.

The coefficient estimates and model fit for this baseline gravity-type attraction-constrained spatial interaction model (Grav) using housing density, employment, and average income from the census are provided in the top left column of table 2. The general interpretation of this model is as follows: larger commuting flows, *ceteris paribus*, are associated with origins that have: less dense housing; larger employed populations; lower average income; and that are closer to potential destinations. Since residential choice is being used here as a proxy for commuting behavior, these interpretations can also be stated in terms of the effect that each explanatory variable has on an individual to choose an origin, given that they will commute to a specific destination. In this case, individuals are more likely to choose an origin that is less densely populated (i.e., housing availability), has a high number of employed individuals, has a lower average income, and is near to their workplace.

Though there is a modest model fit ($R^2 = 0.41$), there are several issues with these interpretations. Most noticeably, it is counterintuitive that individuals would choose, on average, to live in neighborhoods with lower average income. The effect is likely to be a result of a strong relationship between house prices and average income - New York City has several highly affluent neighborhoods that the average worker cannot afford to live in. However, it is expected that most workers would choose to live in more affluent neighborhoods whenever possible. Another issue with these interpretations is that it is not clear that workers would always prefer a residence in a neighborhood with lower population density, especially in the context of a large city where it is known that many people choose to pay a premium to live in some the most densely populated neighborhoods. While these two contradictory interpretations are problematic, they can be clarified by calibrating destination-specific models and examining maps of the local parameter estimates.

Figures 61 - 64 display the local estimates for the parameters associated with the baseline gravity-type model (Grav) described above. In all the presented surfaces, some outliers (black) have been removed in order to make visual patterns more salient while those local estimates that are statistically insignificant at the 95% confidence interval and using a Bonferroni correction for multiple testing are indicated in grey. The surfaces for the number of employed persons (figure 62) demonstrates a predominantly positive relationship and the surface for distance-decay (figure 64) demonstrates a predominantly negative relationship, which are both as expected. This is, however, not the case for the housing density and average income parameter estimate surfaces. For these two surfaces (figures 61 and 63), it is shown that the outer boroughs and uptown Manhattan tend to have negative relationships while midtown and downtown Manhattan tend to have positive relationships. Three insights can be

Table 2: Parameter estimates and model fit for the census commute-to-work data. The top row refers to results for the entire dataset while the bottom row refers to results for a subset of the data pertaining to tracts in and around lower Manhattan. Grav refers to the gravity-type attraction-constrained spatial interaction model and CD1 and CD2 are competing destination models using different accessibility terms.

	Grav		CD1(hd)		CD2(poi's)	
All data	Estimate	SE	Estimate	SE	Estimate	SE
hd	-0.2815	0.0008	0.0492	0.0012	0.02469	0.0010
emp	1.0667	0.0010	0.8701	0.0014	0.9591	0.0013
inc	-0.1609	0.0011	-0.1747	0.0012	0.1664	0.0013
dist	-0.9833	0.0006	-1.042	0.0006	-1.1203	0.0007
cd	-	-	-1.040	0.0027	-0.8535	0.0016
adj. R^2	0.4100		0.4163		0.4222	
AIC	14165374		14014186		13873578	
	Grav		CD1(hd)		CD2(poi's)	
Subset	Estimate	SE	Estimate	SE	Estimate	SE
hd	0.1659	0.0040	0.2568	0.0043	0.2192	0.0041
emp	0.9484	0.0047	0.9282	0.0047	0.9215	0.0047
inc	0.1515	0.0030	0.1690	0.0030	0.2067	0.0031
dist	-0.5700	0.0021	-0.5759	0.0022	-0.6389	0.0023
cd	-	-	-0.6169	0.0027	-0.3589	0.0050
adj. R^2	0.5598		0.5621		0.5630	
AIC	744317		740396		739012	

gained from analyzing these surfaces in relation to their respective global parameter estimates. First, there are spatial non-stationarities in the relationships associated with commuting. Second, the spatial variation of the parameter estimates manifest in a pattern that roughly pertains to a divide between Manhattan and the outer boroughs, with this pattern being more pronounced in the housing density surface than in the average income surface. Third, there are more tracts in the outer boroughs, which are associated with negative relationships for these two variables. Therefore, it makes sense that the global parameters, which represent a weighted average of these

local estimates, would be negative even if there are logical interpretations for both positive and negative relationships.

Two conclusions can be made based on these insights. First, the interpretations of the parameter estimates need to be augmented to consider the underlying non-stationarity. For example, workers commuting to midtown and downtown Manhattan tend to choose residential neighborhoods with higher housing density and higher average income while workers commuting to the outer boroughs tend to choose residential neighborhoods with lower housing density and lower average income. This behavior is commensurate with individuals in high-paying employment in central Manhattan selecting high density and high price apartments nearby and individuals working in the outer boroughs having a preference for lower density and lower price living spaces.

The second conclusion that can made based on these local parameter estimate surfaces is that it may be useful to conduct additional analysis on a subset of the census tracts that demonstrate geographically homogeneous parameter estimates, such as tracts that are in and around lower Manhattan. Consequently, after subsetting the data for only commutes that begin and end in census tracts that also have bike stations[33] (midtown and downtown Manhattan and portions of Queens and Brooklyn that are adjacent to Manhattan), the same model was calibrated. The estimated parameters are shown in table 2 (bottom left) and now indicate a positive relationship between flows and housing density and between flows and average income. Also noticeable, is that the distance-decay is much weaker. Examining the local parameter estimate surface for the entire study area (figure 64) shows that even though there is

[33]Besides roughly correlating to the patterns in the local parameter estimate, this subset was chosen in order to make comparisons to the bike data in a later section.

a negative relationship throughout the study area, there is a less negative relationship in lower Manhattan so that this result is not surprising. The drop in distance-decay for the subset of data is likely due to the very high subway and bus accessibility in this region, both of which increase mobility.

After establishing a baseline gravity-type model, the next step is to consider spatial effects. In the previous chapters, it was shown that the eigenvector spatial filtering method yields ambiguous results while the spatial autoregressive (i.e., spatial lag) approach can obfuscate the coefficient estimates, especially in regards to the interpretation of distance-decay. The competing destinations approach with a cumulative distance-weighted accessibility term was shown to be more robust for capturing spatial structure (i.e., spatial clustering amongst locations) than a contiguity-based spatial lag of an explanatory variable. Therefore, the competing destinations methodology was used in subsequent analysis aimed at capturing spatial structure effects.

If only census data were available, which has been the case for most commuting research until relatively recently, then the accessibility term to account for spatial structure would be based on data from the census, such as population. Since commuting is modeled here from the perspective of residential choice, the appropriate accessibility term should be defined to capture spatial structure effects that measure the competition an origin faces from competing origins. An origin-based accessibility term is given as

$$A_{ij} = \sum_{\substack{k=1 \\ (k \neq j)}}^{n} \frac{W_k}{d_{ik}^{\sigma}} \tag{6.2}$$

where W_k is the attraction of each alternative origin other than i, and, σ, the second order distance-decay used for distance-weighting is set to -1. Though σ is sometimes

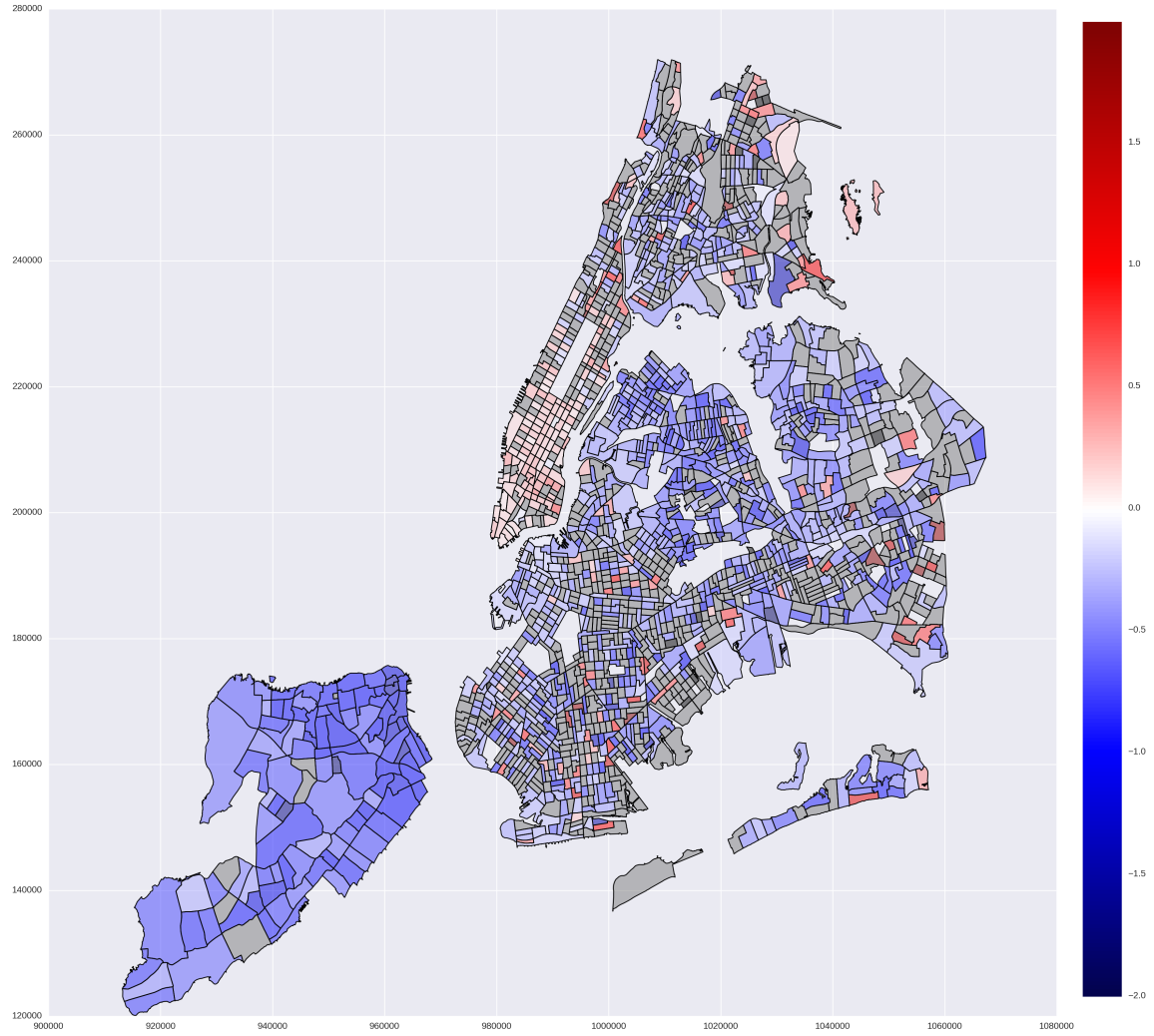


Figure 61: Local parameter estimates for the housing density (hd) variable in the baseline gravity-type attraction-constrained model using census commute-to-work data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

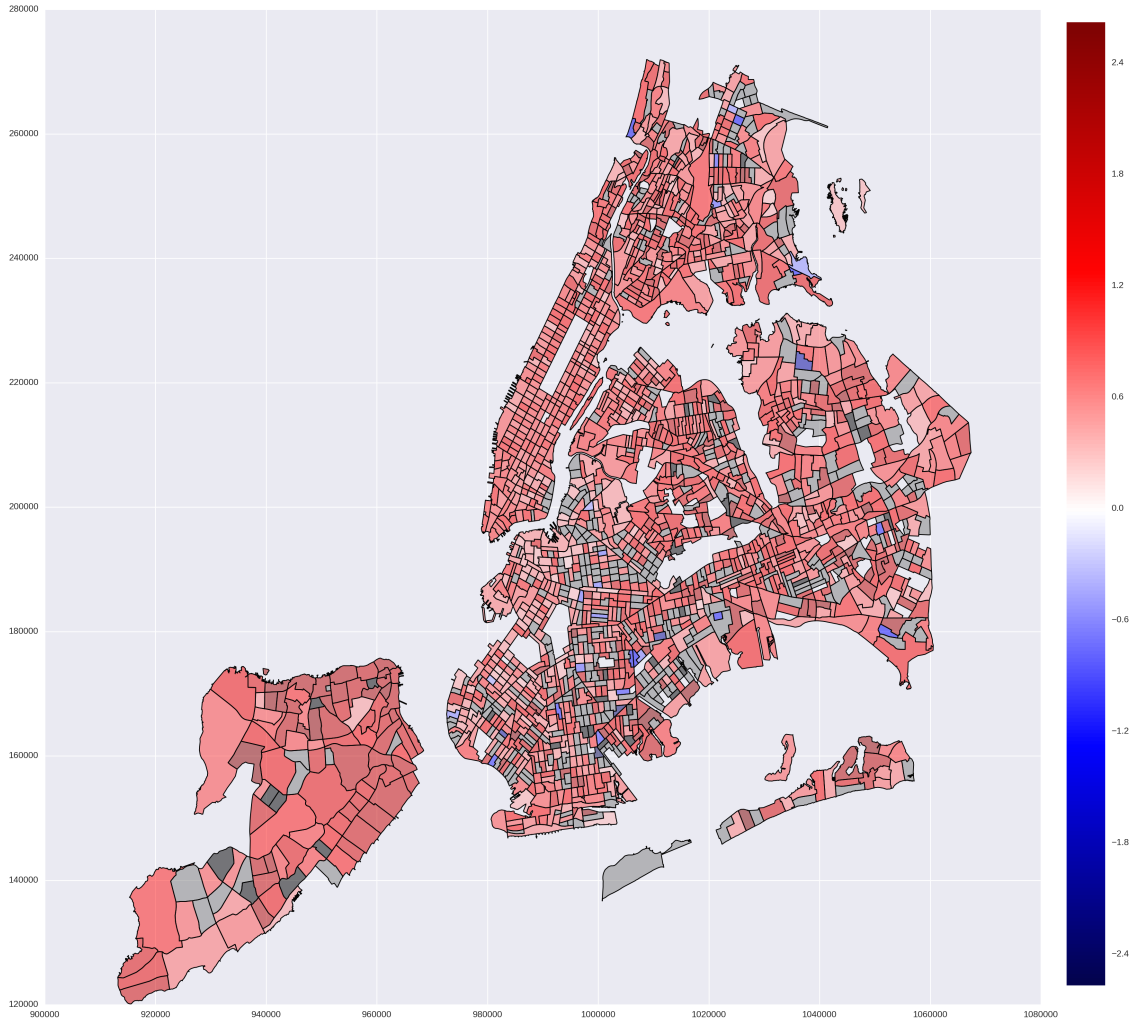


Figure 62: Local parameter estimates for the number of people employed (emp) in the baseline gravity-type attraction-constrained model using census commute-to-work data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

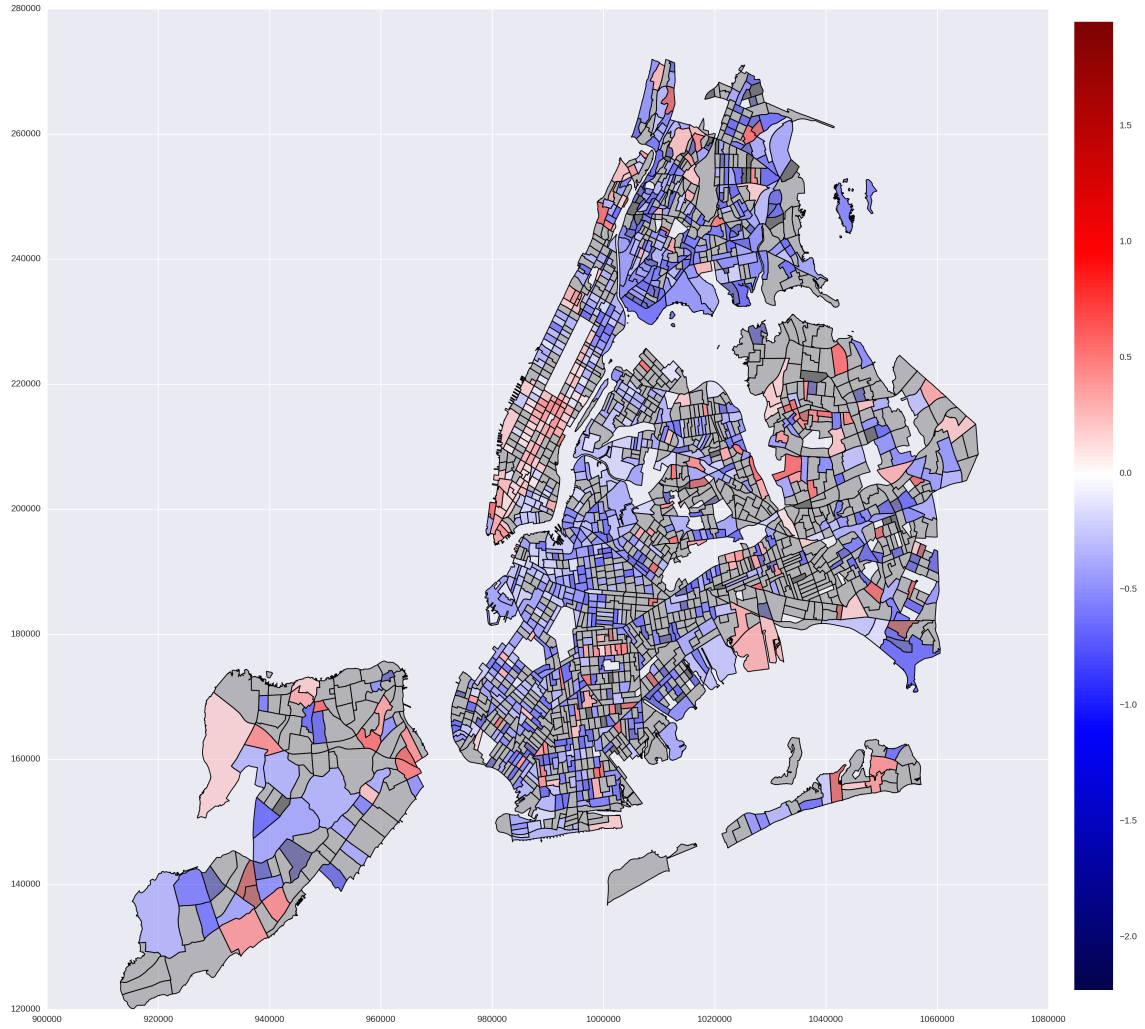


Figure 63: Local parameter estimates for the average income (inc) variable in the baseline gravity-type attraction-constrained model using census commute-to-work data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

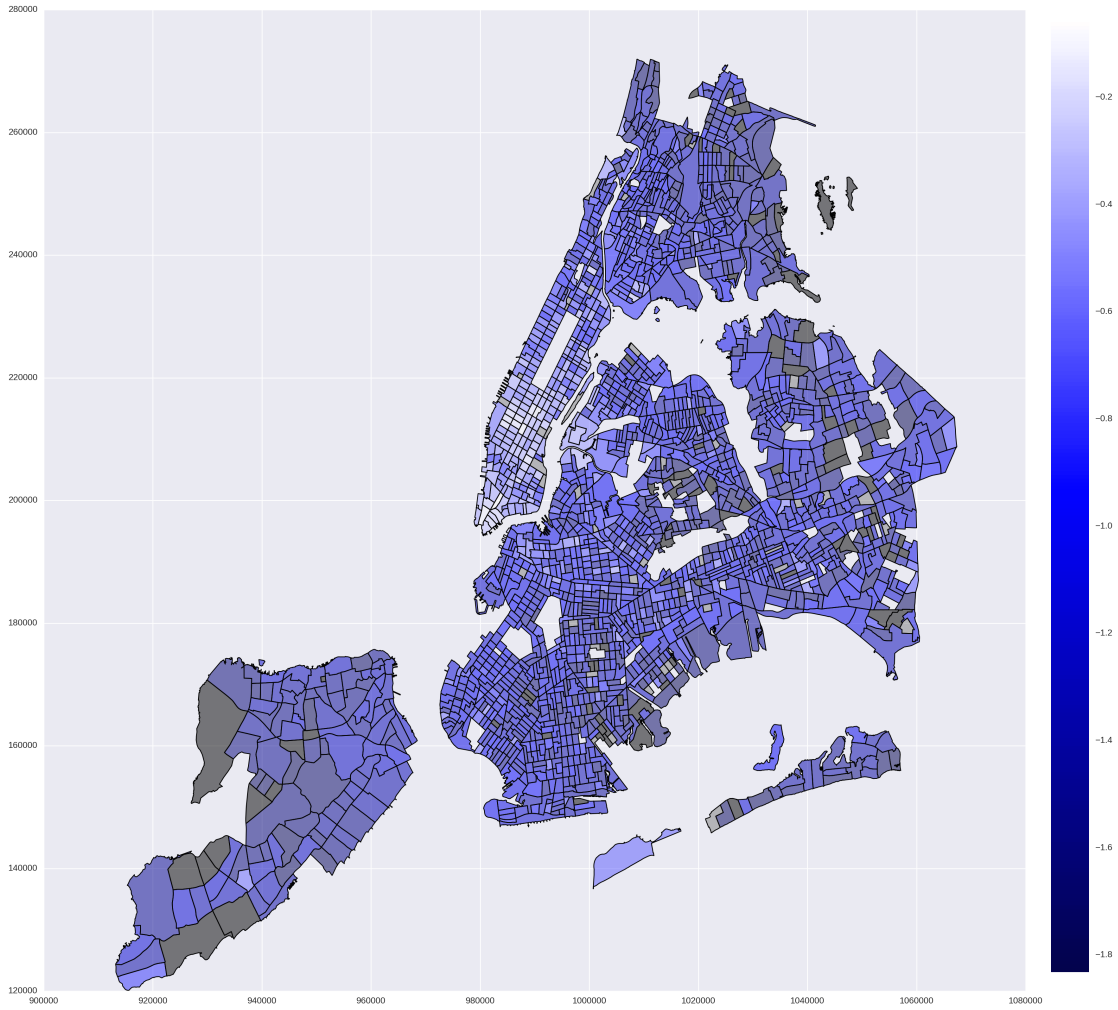


Figure 64: Local parameter estimates for distance-decay (dist) in the baseline gravity-type attraction-constrained model using census commute-to-work data. Lighter values indicate a weaker relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

derived iteratively with the first order distance-decay, β, until they converge, this was not done here since β was already very close to -1 and small fluctuations to σ produced similar results. Each of the three locational explanatory variables included in the basic gravity-type model were used to compute a potential accessibility term and then they were individually included in the model. Ultimately the model using an accessibility term based on housing density was selected, since it provided the largest increase in model fit, though all of the potential accessibility terms offered only marginal increases in model fit. An accessibility term based on housing density was also chosen over the other accessibility terms since it is more likely that an individual perceives clusters of locations in an urban context based on the built environment (i.e., housing density) rather than on their direct knowledge of employment rates or average income.

Results from a competing destinations model including an accessibility term defined using housing density (CD1) are provided in table 2 (top of middle column). Accounting for spatial structure produces an accessibility parameter estimate that is negative, indicating competition effects amongst residences across the majority of the study area (figure 65). Though the distance-decay and average income parameter estimates become slightly larger in absolute magnitude and the employment parameter estimate becomes smaller, their local parameter estimate surfaces do not change much qualitatively, and are therefore not shown here. In contrast, there is a big shift in the housing density coefficient, which is now positive instead of negative, though the magnitude is very small. Investigating the housing density parameter estimate surface for the Grav model (figure 61) and the CD1 model (figure 66), the reason for the shift in the global parameter estimate is not that the overall relationship has changed. Instead, Manhattan has remained positive, while several portions of the

outer boroughs have either positive or statistically insignificant relationships. That is, the outcome of accounting for spatial structure implies that if individuals first consider macro-level neighborhoods and then consider locations within these neighborhoods in their residential choice process, it becomes more likely that they will choose locations with higher housing densities than if they were only considering individual locations.

6.2.2 Additional locational variables

In the era of big data, many new forms of data are available that were previously either too costly to collect or the necessary technology for collection and dissemination were not available. This includes data collected automatically by sensors, data released by municipalities using data portals, and crowd-sourced data. Despite the availability of such diverse data sources, examples of their usage as explanatory variables within spatial interaction models is sparse. Therefore, in chapter 4, several variables, such as points-of-interest (POI's) from OpenStreetMap, subway usage, and building square footage were collected and presented as potential alternative variables within spatial interaction models. However, there are two pitfalls in using these new data forms: they either provide more noise than signal or are not uniformly sampled throughout the study area. The former issue was observed upon adding the building square footage variable to the previously presented gravity-type model (Grav) where the result was a decrease in model fit. The latter issue, uneven sampling, means that values of zero are recorded for many locations, which is problematic for Poisson log-linear regression that requires the logarithm of the explanatory variables. Zero or null observations occur either because sampling is biased, such as the reporting rate of crowd-sourced data

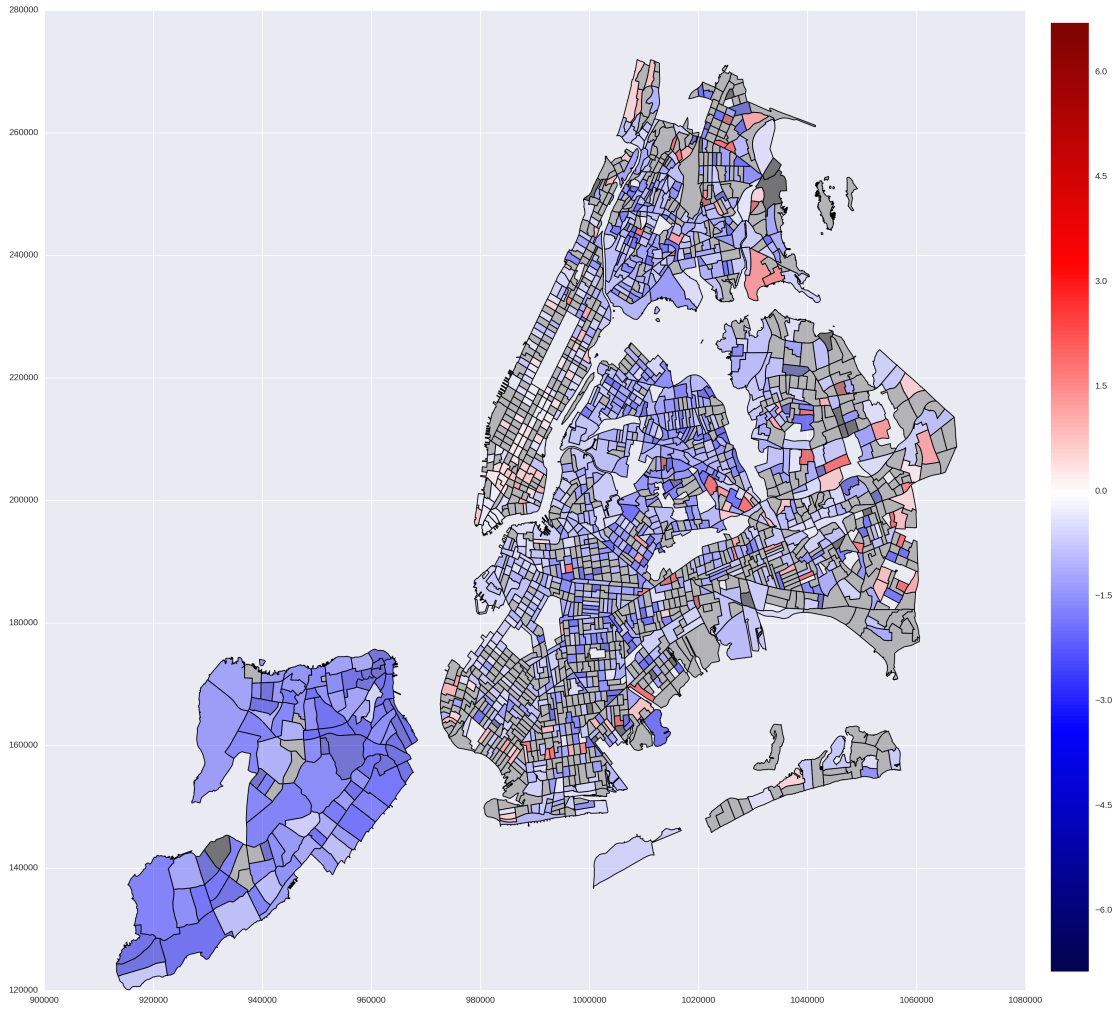


Figure 65: Local parameter estimates for the accessibility term (cd) defined using housing density in the attraction-constrained competing destination model (CD1). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

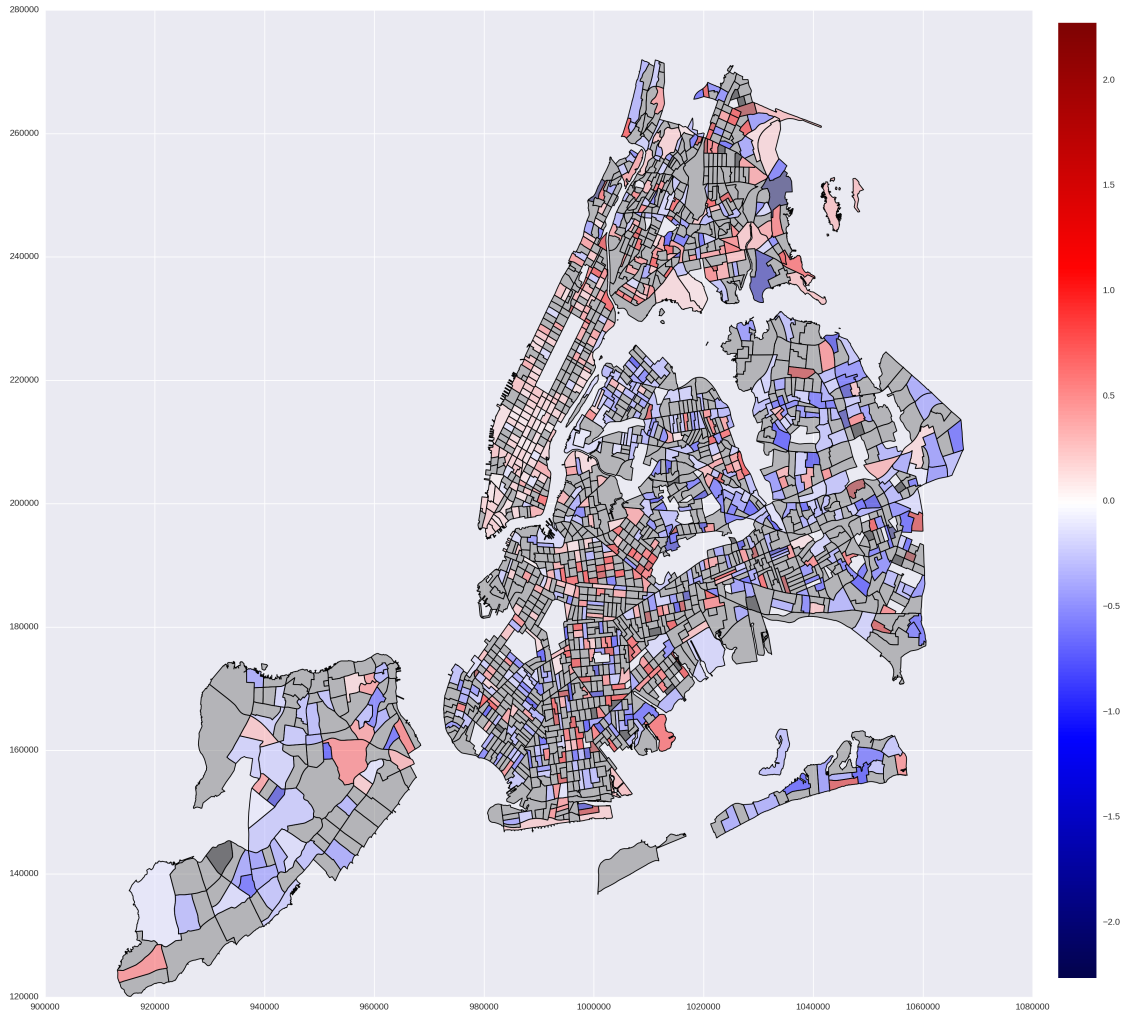


Figure 66: Local parameter estimates for the housing density (hd) variable in the attraction-constrained competing destination model of census commute-to-work data using an accessibility term defined using housing density (CD1). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

(i.e., OpenStreetMap POI's), or because an event is limited to a particular spatial domain (i.e., subway usage at subway stations).

One way around the issue of zero observations is to borrow data from neighboring locations, which is commonly done by computing a sum of distance-weighted neighboring observations for each location of interest, as is done in geographically weighted regression (Fotheringham *et al.*, 2002). Interestingly, this data-borrowing technique when applied to only the explanatory variable(s), is equivalent to computing the accessibility term in a competing destinations model. In this case, it was problematic to include additional distance-weighted variables beyond an initial competing destinations accessibility term within the model because all of the distance-weighted variables were highly collinear to each other. This is due to the fact that the variables were transformed using the same function and intensity (i.e., same value of σ in equation 6.2) and is sometimes referred to as concurvity (Ramsay *et al.*, 2003). While it may be possible to computationally explore variable-specific transformations, akin to those employed in multi-scale geographically-weighted regression (Fotheringham *et al.*, 2017; Wolf *et al.*, 2017), this is not taken up here. Instead, alternative accessibility terms computed using the subway usage and OpenStreetMap POI's were considered in lieu of including an accessibility term defined using the census variables. Equation 6.2 was still used with σ set to -1 so that the question of interest is whether or not similar results are obtained when these non-traditional sources of data are used to capture spatial structure.

After entering each alternative accessibility term into the model, the term computed using OpenStreetMap POI density (i.e., dividing a count of POI's by the area of a tract) resulted in the highest model fit, which is recorded in table 2 (top right) as CD2 along with the associated parameter estimates. Interestingly, this new accessibility

term resulted in a model fit only marginally higher than that achieved by model CD1 but also produced different parameter estimates. Though the parameter estimate associated with accessibility is still negative throughout the majority of the study area (figure 67) and therefore indicates competition effects, it is smaller than in CD1; however, the distance-decay parameter estimate is now even more negative than in either the Grav model or the CD1 model. Moreover, though the employment and housing density parameter estimates are similar to those estimated in CD1, the average income parameter estimate in CD2 is now positive instead of negative, which indicates that, on average, people will choose a residence in a neighborhood with higher average income (i.e., less poverty). Examining the associated local parameter estimate surface (figure 68 indicates that there are more positive relationships in Brooklyn and Queens than were present in the Grav model (figure 63). Overall, the increased model fit and more theoretically sound interpretations point towards POI density being more representative of urban spatial structure than housing density. This is perhaps because the POI density variable is based on a variety of urban amenities such as shops, cafes, restaurants and bars, amongst others, that can more accurately represent the diverse factors that contribute to neighborhood definition and attractiveness, though this may only hold for very dense urban areas, such as New York City, and needs to be further explored using data from other study areas.

A comparison of the results for the three models (Grav, CD1 and CD2) is also interesting in the context of the larger debate about spatial structure effects. Chun (2008) also reported flipped parameter estimate signs after introducing eigenvectors into spatial interaction models. It is possible that that the linear combination of eigenvectors that are selected using stepwise regression may wholly or in part explain

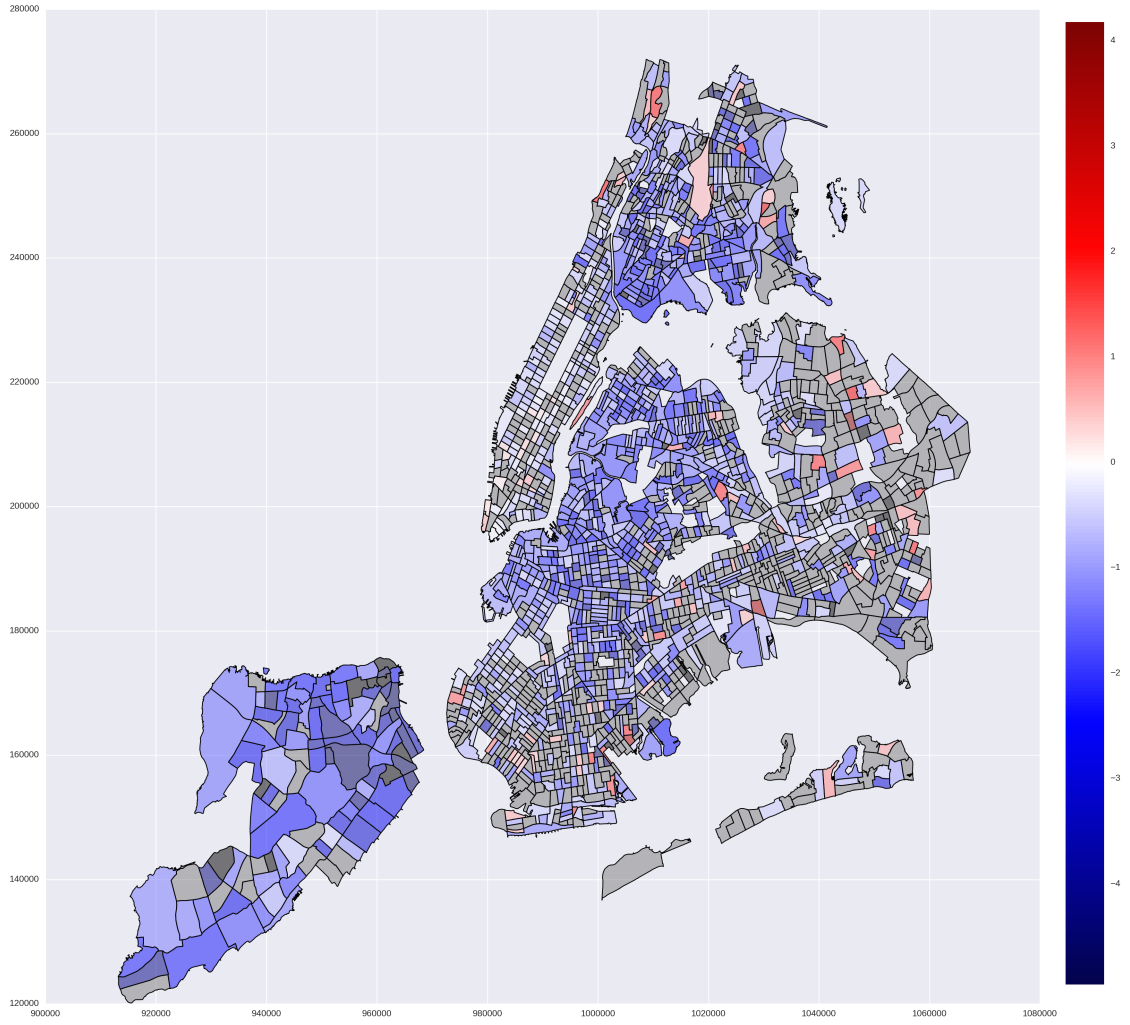


Figure 67: Local parameter estimates for the accessibility term (cd) defined using POI density in the attraction-constrained competing destination model of census commute-to-work data (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

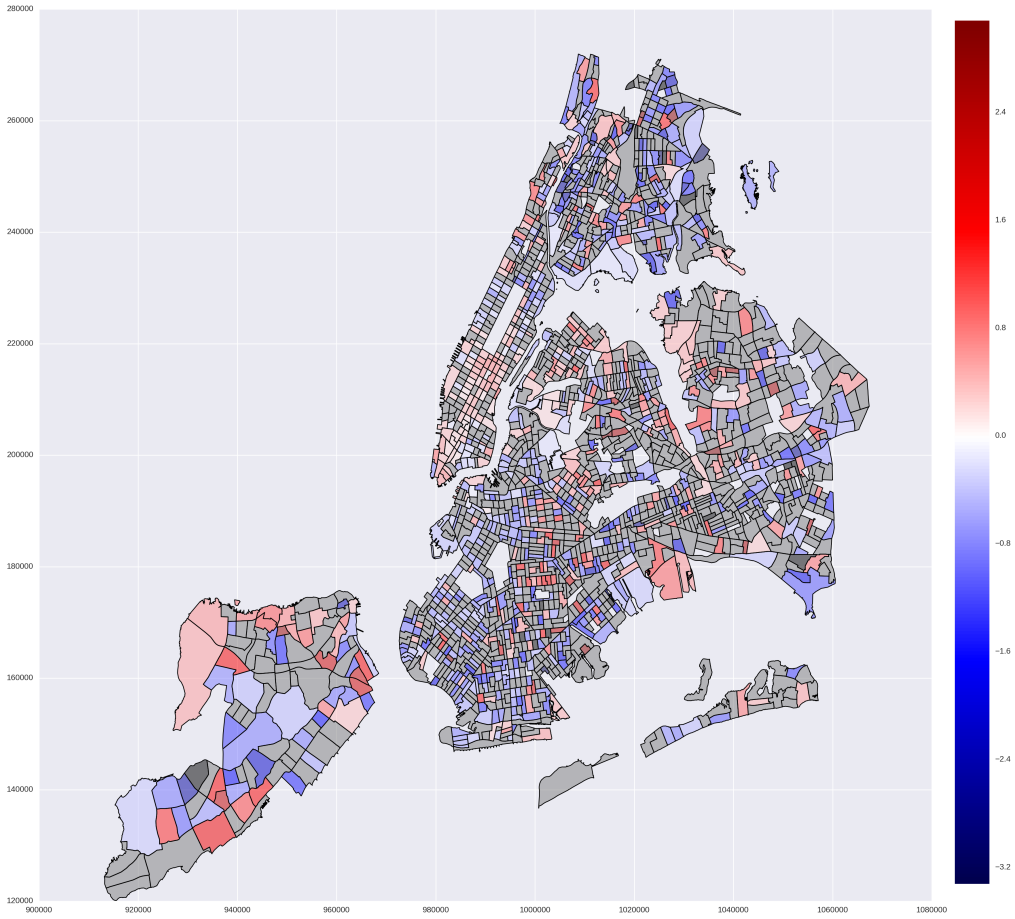


Figure 68: Local parameter estimates for the average income (inc) variable in the attraction-constrained competing destination model of census commute-to-work data using an accessibility term defined using POI density (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

the same variation that would be explained by an accessibility term. However, in the eigenvector spatial filtering methodology, there is no way to know what exactly is being accounted for or why doing so should change the model interpretations. Furthermore, the results presented here indicate that flipped signs may be due to omitted variable bias, but also due to non-stationarity. Since the global parameter estimates are a weighted average of the local parameter estimates, it may be more informative to redefine the study area based on non-stationary than to include mystery surrogate variables into the model. This notion was further supported by calibrating CD1 and CD2 on the subset of data previously defined that pertains to lower Manhattan and areas of Brooklyn and Queens that are close to lower Manhattan. Comparing parameter estimates for the three models on the subset data (table 2 bottom row) shows remarkably robust results. Though some competition effects are still present, none of the parameter estimates change drastically or have flipped signs. This provides further evidence that the flipped signs that were observed may be due relationships that are not constant across the study area rather than solely due to ommitted variable bias.

In the subsequent sections, it is determined whether or not similar model interpretations and conclusions can be made about commuting processes by using the bike and taxi datasets instead of the census commute-to-work data.

6.3 A commute-to-work model using bike trips

In this section, the same models that were calibrated on census commute-to-work data in the previous section are calibrated on the bike data to see whether or not similar interpretations can be gained. To make the bike data set commensurate

Table 3: Parameter estimates and model fit for the bike commute-to-work data. Grav refers to the gravity-type attraction-constrained spatial interaction model and CD1 and CD2 are competing destination models using different accessibility terms.

	Grav		CD1(hd)		CD2(poi's)	
Subset	Estimate	SE	Estimate	SE	Estimate	SE
hd	-0.2221	0.0011	-0.3963	0.0013	-0.3738	0.0012
emp	0.3210	0.0013	0.4412	0.0013	0.5465	0.0014
inc	0.1630	0.0010	0.1187	0.0011	0.0589	0.0011
dist	-1.367	0.0006	-1.308	0.0007	-1.2583	0.0007
cd	-	-	1.161	0.0035	0.8409	0.0015
adj. R^2	0.6633		0.6705		0.6810	
AIC	5289573		5175945		5011530	

with the census commute-to-work data, a subset of the bike data that pertains only to weekdays and that start between 5:00 and 10:00am was selected. This resulted in 3,661,877 bike trips aggregated to 32,333 origin-destination routes between 245 origins and 246 destinations. This implies that about half of the possible 60,270 origin-destination routes have non-zero trip count observations, which means there are far fewer zero flows in the bike data than in the census data. After further processing the bike data using the same methods used to prepare the census commuting data, there were 59,290 observations between 242 origins and 246 destinations.

The results for the three models (Grav, CD1, and CD2) calibrated on this subset of bike commute data are reported in table 3. It can be immediately seen that there are several differences between these results and those for the census commuting data for the same subset of locations (table 2 bottom row). It is useful to begin by first interpreting the parameter estimates in the gravity-type model (Grav) and comparing them to the equivalent values from the census commuting results.

First, the distance-decay parameter estimates in all of these bike data results

are more negative than in the census commuting results. It is not unreasonable to expect that cyclists perceive distance to be more of a deterrent to travel over longer distances than the average commuter represented in the census! In New York City, the primary method of transportation is via the subways and the buses and so it is likely that commuting patterns captured in the census reflect these more mobile forms of transportation.

Second, though origin employment numbers still have a positive influence on commuting flows, the relationship is less for the bike data than for the commuting flows. This may be due to the fact that the origins for the bike data are not necessarily indicative of residences as is the case in the census data. The bike data only records the station at which trips begin and terminate, and it is likely that the stations are located in communally accessible areas that require most people to walk several blocks from home before checking out a bike to begin the cycling segment of their commute. Therefore, it makes sense that in many census tracts employment is less important in determining the number flows.

Third, housing density is now negative instead of positive, which may be similarly explained. Since individuals in this data set are not choosing a residence, but rather an origin station, it is logical that they would often start their trips in census tracts with more bike docks, which are likely not the areas with higher housing density but are instead areas of mixed land use. Finally, the relationship between flows and average income in still positive. This indicates that bike commuters are more likely to originate in areas of higher average income.

Examining the local parameter estimate surfaces for these four variables (figure 69) demonstrates that there are few large clusters of spatial variation. One exception is that the distance-decay becomes stronger in the westernmost tracts of Brooklyn and

Queens where the density of stations is relatively low and there are more potential long distance trip that can be made, many of which would require crossing a bridge.

Overall, these bike data results furnish higher model fits than the census commuting data, which is likely to be due, in part, to the much lower level of zero flow observations. As with the census commuting data, the model fit increases marginally when including an accessibility term (CD1 and CD2); however, the parameter estimate associated with accessibility is now positive instead of negative, indicating agglomeration effects with bike station choice for early morning weekday trips. This is unsurprising for two reasons. First, urban planners working with the CITI bike-sharing program have likely chosen the locations of the bike stations at the most accessible locations so as to optimize bike usage. It follows then that origin stations that are more accessible will expect more trips than non-accessible stations purely because they are more accessible. Second, individuals likely choose to start their cycling trips from larger clusters of bike docks, which avoids the issue no bikes being available. Although the agglomeration effects are supported by these reasons, the interpretation does not hold for the entire study area. In fact, observing the local parameter estimate surfaces for the accessibility terms in CD1 and CD2 (figure 70 left and right, respectively) it can be seen that the agglomeration forces are strong in Manhattan, though in the outer boroughs competition forces are dominant. This dichotomy may arise because of the relative sparsity of stations and bike availability across Brooklyn and Queens compared to Manhattan. One discrepancy between the two surfaces of accessibility parameters is that the surface for the accessibility term defined using housing density indicates Alphabet City as having the strongest agglomeration effects whereas the surfaces for the accessibility term defined using POI density indicates midtown Manhattan as having the strongest agglomeration effects. Without further investigation it is not

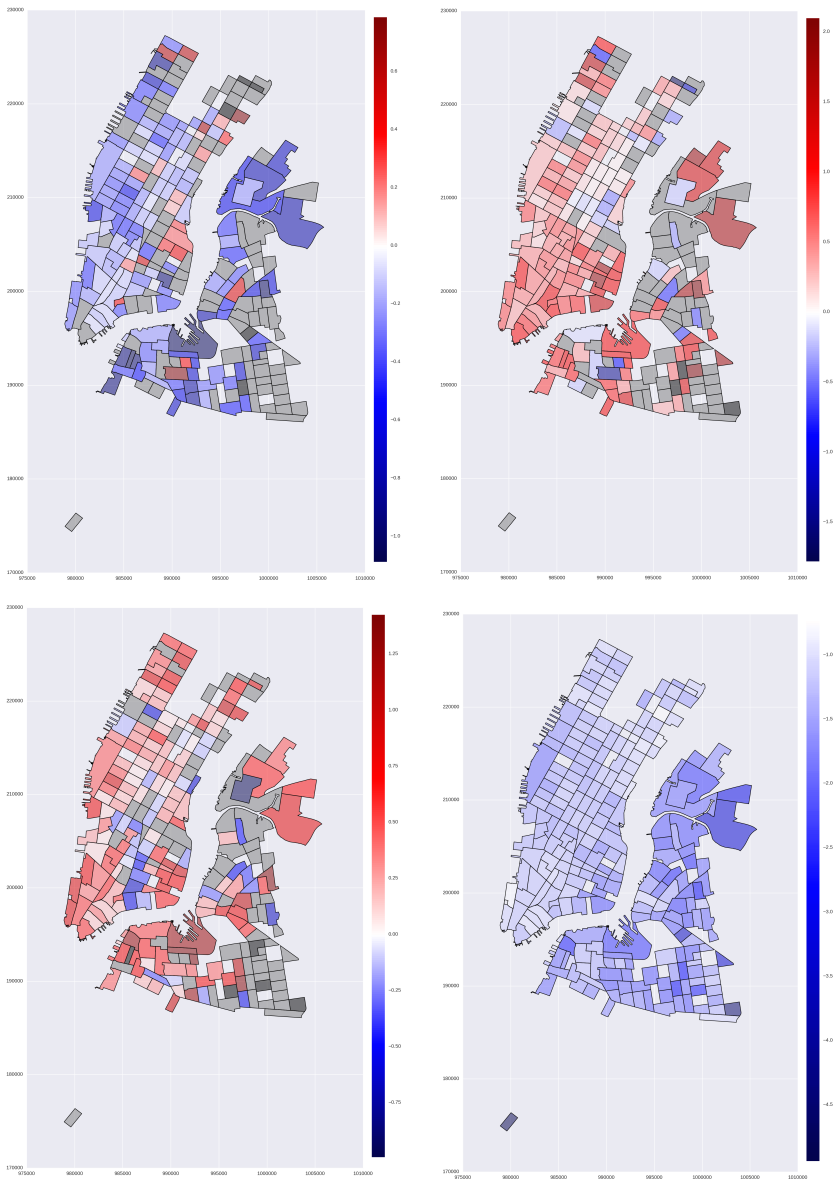


Figure 69: Local parameter estimate surfaces for the gravity-type attraction-constrained spatial interaction model (Grav) of the bike data. The top left surface corresponds to housing density (hd), the top right surface corresponds to employment (emp), the bottom left surface corresponds to average income (inc), and the bottom right surface corresponds to distance-decay (dist). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

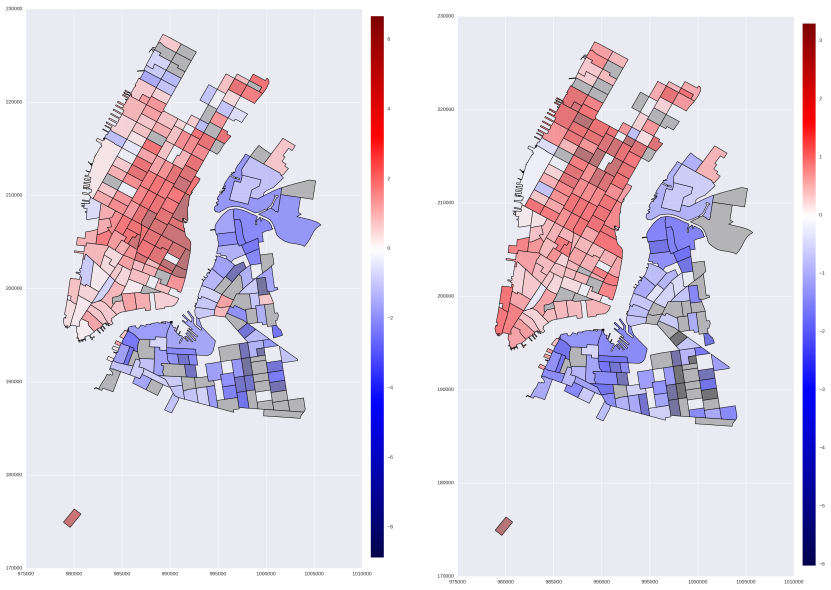


Figure 70: Local parameter estimates for the accessibility terms (cd) defined using housing density (left) and POI density (right) in the attraction-constrained competing destination models. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

possible to determine if one pattern is more theoretically appropriate than the other, and it is possible that both patterns may be important.

Of the two models with accessibility terms, CD1 and CD2, the one with an accessibility term based on the POI density provides a higher model fit than the one based on housing density. Furthermore, the outcomes of including the accessibility term are that the negative relationship with housing density becomes stronger, the positive relationship with employment becomes stronger, and the positive relationship with average income becomes greatly diminished. Removing average income from the model results in virtually no reduction of the model fit ($< .003$) and does not cause any large deviations in the other parameter estimates. It is therefore likely

that the significant non-zero effect detected in the gravity-type model was due to the model overcompensating for the omission of the accessibility variable. Of the remaining variables, noticeable deviations from those displayed in figure 69 were only apparent for the distance-decay parameter estimates where there were now several tracts that have positive distance-decay (figure 71). For the two tracts in the center of downtown Manhattan this may be because the individuals who commute to these destinations only commute from stations that are further afield, since many of the tracts surrounding this area are not very residential. For example, this area contains several large hotels, and famous attractions such as Rockefeller Center and Times Square. Most individuals living and working in this area would probably choose to walk since the commute distance is so short and cycling would require more effort than it would be worth. For the tract in Brooklyn (southeast corner) this tract is an outlier and further investigation is necessary to explain its presence.

The interpretations discussed here pertaining to the bike data are sufficiently different in nature from the census commute-to-work data to conclude that the bike data are not an suitable substitution for traditional survey data to study commuting behavior. This is probably primarily because the data are not as representative of residences and workplaces as are the commuting data, although it may also be because the methodology used here to capture commuting bike trips is only a crude approximation to the true set of bike trips associated with commuting. More advanced classification methods might be helpful in isolating a stronger signal that corresponds to the commuting processes represented in the census commute-to-work data.

It is also possible to expand the analysis of the bike data using variables that were not available for the census commute-to-work data. Most notably, a measure of bike capacity (i.e., number of bike docks) is available for each station and can be used

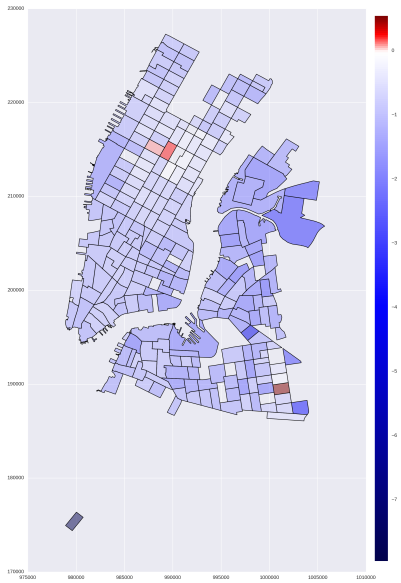


Figure 71: Local parameter estimates for distance-decay in the attraction-constrained competing destination model of the bike data using an accessibility term defined using POI density (CD2).

to compute an accessibility term that is more intuitive for bike trips. In addition, since the exact coordinates of the bike stations are available, it is possible to compute an accessibility term directly between individual stations instead of between tracks, which is can then be aggregated to tracts to be included in the model. An accessibility term defined using station capacity was calculated and entered in to the model at the tracts level (CD3) and the station level (CD4) and the results are presented in the top row of table 4. The results for the model using accessibility defined using station capacity at the tract level (CD3) produces very similar parameter estimates to those obtained from model CD1 (table 3) that defines accessibility using housing density, which means all of the previously discussed interpretations for model CD1 hold for model CD3. This is further supported by the local parameter estimate surface for accessibility from CD3 (figure 72 top left) that is similar to the surface produced from CD1. The other surfaces produced were also very similar to those from CD1

and Grav (table 3 and figure 69) and were therefore not shown here. In contrast, the results for the model using inter-station distance to define accessibility (CD4) resulted in very different global parameter estimates (table 4 top row). First, the parameter estimates for housing density, employment, and average income are all much smaller in magnitude and housing density and average income have flipped signs. Examining the local parameter estimate surfaces for these three variables (figure 73) it can be seen that the surfaces much noisier than those produced by the Grav model (or CD1 and CD3 models) that are portrayed in figure 69. In fact, removing all three variables from the model causes the pseudo R^2 to decrease less than 0.002 and suggests that when accessibility is measured accurately, it is unnecessary to account for these variables. Second, the accessibility parameter for model CD4 is much smaller in magnitude than for model CD3. Figure 72 (top right) illustrates the local parameter estimate surfaces for accessibility in model CD4 where many of the estimates that had negative values in model CD3 (top left) are now positive and overall the local parameter estimates take on a much smaller range of values. These results show that the definition of competing locations and accurate measurement of their proximity to each other is important and that errors in measuring these attributes can lead to different results.

Another enhancement that is possible for the bike data is to use distances that are computed by a routing algorithm that accounts for traffic patterns and gives preferences to roads that accomodate cyclers. These routed distances potentially yield more realistic measures of the effort needed to cycle between locations and may therefore produce different parameter estimates. The distance produced using the Mapzen *Matrix* routing service (chapter 4) was entered into models CD3 and CD4

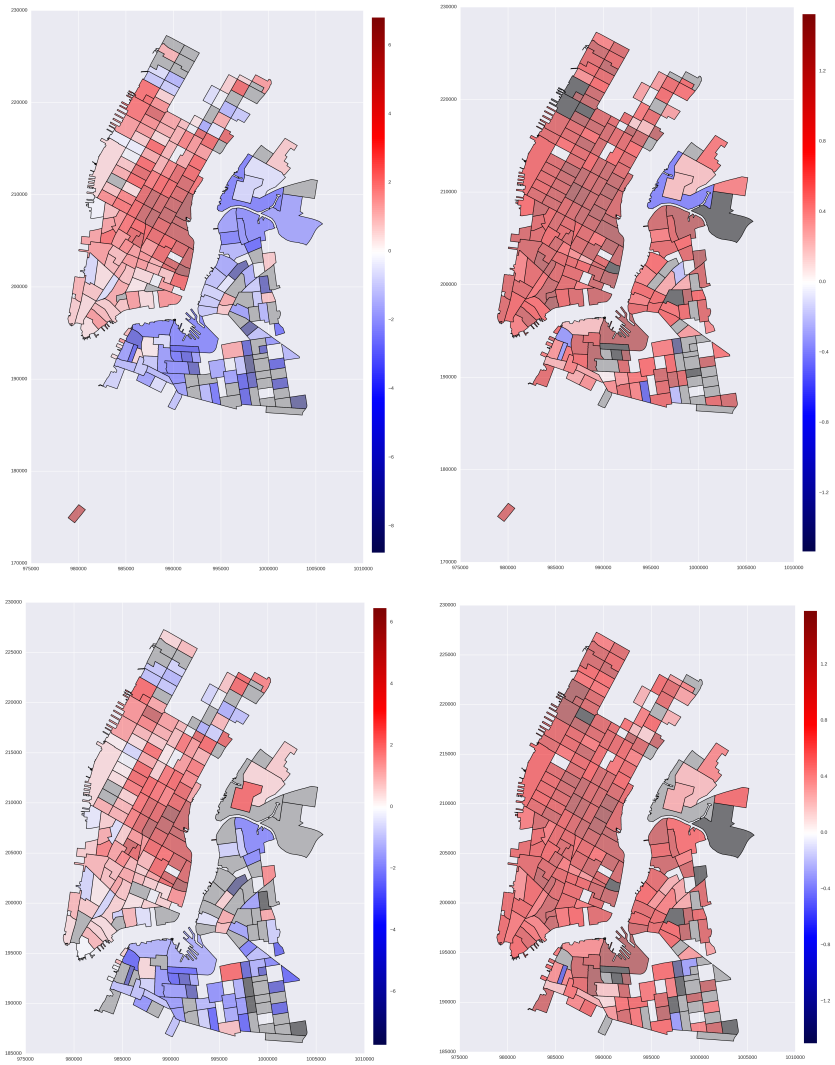


Figure 72: Local parameter estimates for accessibility parameter in attraction-constrained competing destination models of the bike data using an accessibility term defined using station capacity - top left pertains to CD3, top right pertains to CD4, bottom left pertains to CD5, and bottom right pertains to CD6. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

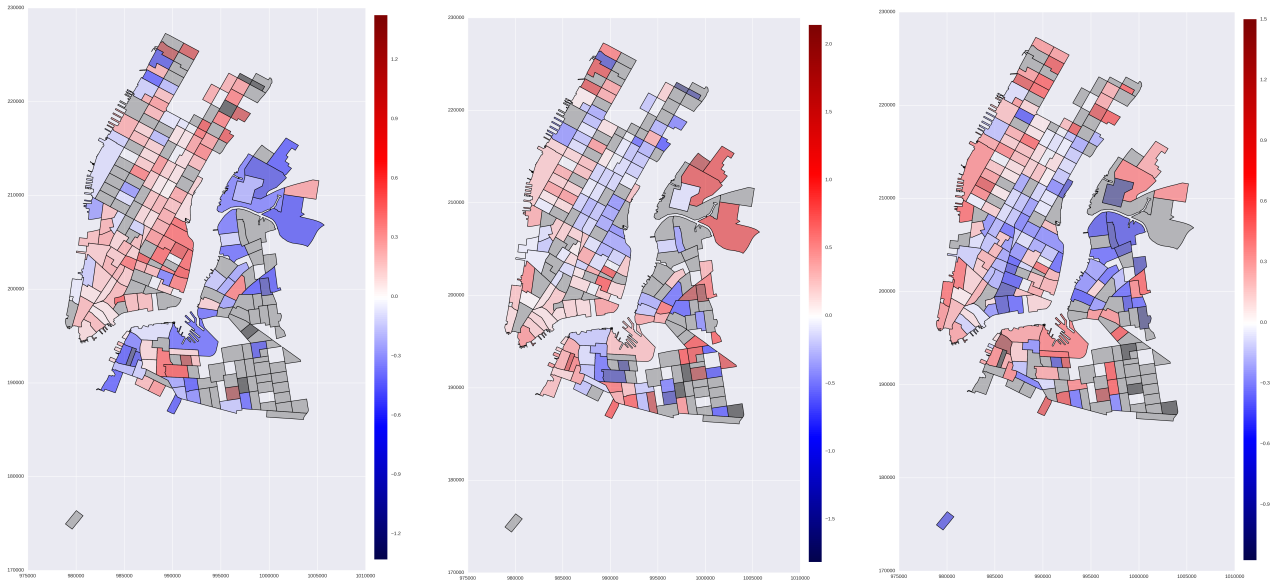


Figure 73: Local parameter estimates for housing density (left), employment (middle), and average income (right) in the attraction-constrained competing destination models of the bike data using an accessibility term defined using station capacity (CD4). Similar surfaces were obtained for CD6, which is the same model but uses distance derived using a routing algorithm. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

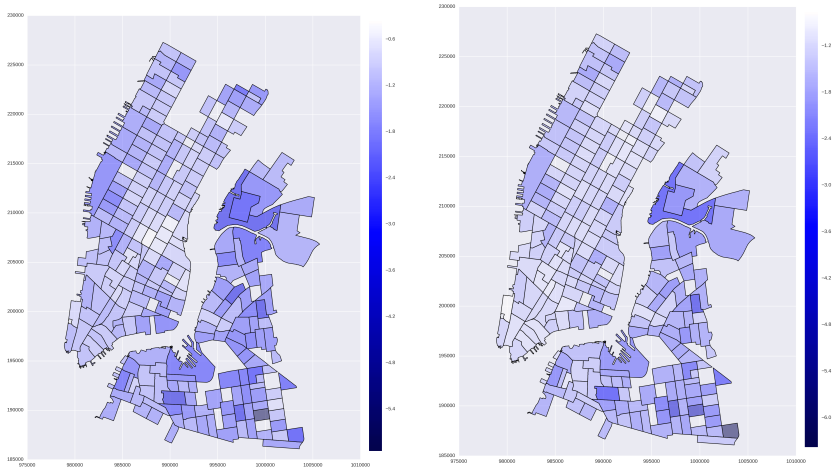


Figure 74: Local parameter estimates for distance-decay in attraction-constrained competing destination models of the bike data using an accessibility term defined using station capacity - left pertains to CD5, and right pertains to CD6.

Table 4: Parameter estimates and model fit for the extended models using the bike commute-to-work data. CD3 refers to a competing destination model using tract-level station capacity to define accessibility and CD4 refers to a competing destination model using point-level station capacity to define accessibility.

	CD3		CD4	
Euclidian	Estimate	SE	Estimate	SE
hd	-0.3875	0.0012	0.06332	0.0013
emp	0.4765	0.0013	0.02572	0.0014
inc	0.1042	0.0011	-0.0523	0.0011
dist	-1.262	0.0007	-1.240	0.0007
cd	1.457	0.0036	0.9887	0.0010
adj. R^2	0.6744		0.7242	
AIC	5115511		4332910	
	CD5		CD6	
Routed	Estimate	SE	Estimate	SE
hd	-0.4048	0.0013	0.0009	0.0013
emp	0.4788	0.0014	-0.0768	0.0014
inc	0.0594	0.0010	-0.1082	0.0011
dist	-1.506	0.0008	-1.499	0.0008
cd	1.223	0.0036	0.9714	0.0011
adj. R^2	0.6978		0.7491	
AIC	4677772		3882747	

instead of Euclidian distance to produce results for models CD5 and CD6 (bottom row of table 4). The global parameter estimates for housing density, employment and average income in models CD5 and CD6 are similar to those from models CD3 and CD4, though there are two exceptions. The first exception is that employment takes a negative value in CD6 compared to a positive value in CD5. However, like CD5, all of the estimates in CD6 are small, produce noisy local parameter surface estimates and dropping them from the model barely reduces the fit of the model. The

other exception is that using the routed distances produces distance-decay estimates that are higher in magnitude and accessibility estimates that are lower in magnitude. This indicates that more realistically measuring distance over the routes that cyclists are likely to use suggests that cyclists become more deterred by longer-distance trips. Moreover, this increase in the distance-decay parameter estimate does not create a local parameter estimate surface with patterns different from the surface for Grav model (or CD1, CD3 and CD4 models) that is portrayed in figure 69 (bottom right) nor does it produce local parameter estimate surfaces for accessibility that differ strongly from the surfaces produced from model CD3 and CD4 (figure 72 bottom row). Since model CD6 results in the highest model fit using the fewest variables (after dropping housing density, employment and average income) it provides the most parsimonious model. Two further conclusions can also be made. First, accurately capturing distance and accessibility, such as using inter-station distance and routed distances, is important for modeling the bike data. Second, these results provide evidence that the accessibility of the stations probably already reflects many other attraction variables that have been considered by city planners that optimized the system.

Results for the taxi data are presented in the next section and are compared to both the results from census commuting data and these results pertaining to the bike data.

6.4 A commute-to-work model using taxi trips

Following the previous section, the same models that were calibrated on census commute-to-work data are calibrated on the taxi data to see if similar interpretations

can be obtained. To make the taxi data commensurate with the census commute-to-work data, a subset of the taxi trips that pertains only to weekdays and that that start between 5:00 and 10:00am was extracted. This resulted in 39,702,911 taxi trips aggregated to 518,705 origin-destination routes between 2138 origins and 2163 destinations. This implies that a majority of origin-destination routes have zero trip count observations, however there are still fewer zero flows than are in the census commute-to-work data. After further processing the taxi data using the same methods used to prepare the census commuting data, there were 4,382,900 observations between 2,050 origins and 2139 destinations.

The same three models that were calibrated on the census commuting data (Grav, CD1, and CD2) were then calibrated on the taxi commuting data, and the results reported in table 5. One similarity between the results for the census and taxi commuting data is that including an accessibility term defined using POI density produces the highest model fit and both sets of data produce a very similar parameter estimate for distance-decay (approximately -1.12). However, many differences exist between the taxi data results and the census commute-to-work data. Once again, it is constructive to discuss each parameter estimate in the these models and describe any discrepancies between the two data sets.

First, in the gravity-type models (Grav) the distance decay derived from the taxi data is stronger than the distance-decay derived from the census commuting data. However, this may be due to omitted variable bias since including an accessibility term moderates the distance-decay to be less negative for the taxi data (table 5) and more negative for the census data (table 2), which leads to similar distance-decay, as was previously discussed.

Second, the relationship between the number of trips and average income at the

Table 5: Parameter estimates and model fit for the taxi commute-to-work data. Grav refers to the gravity-type attraction-constrained spatial interaction model and CD1 and CD2 are competing destination models using different accessibility terms.

	Grav		CD1(hd)		CD2(poi's)	
All data	Estimate	SE	Estimate	SE	Estimate	SE
hd	0.2082	0.0004	-0.3835	0.0004	-0.1143	0.004
emp	0.2357	0.0004	0.6123	0.0004	0.5842	0.0004
inc	1.377	0.0003	1.179	0.0003	0.9505	0.0004
dist	-1.365	0.0002	-1.281	0.0002	-1.121	0.0002
cd	-	-	3.700	0.0017	1.405	0.0004
adj. R^2	0.8532		0.8675		0.8827	
AIC	57142354		51559657		45819222	

trip origin is always positive and is much stronger in the taxi data. This is probably related to the relatively high costs of commuting by taxi compared to other forms of transportation such as subway or bus, such that individuals who can afford to take a taxi to work are more likely have the economic means to afford a residence in an area with a higher average income.

Third, the parameter estimate for the number of employed individuals is positive for both data sets, though it is smaller in magnitude for the taxi data. This could be related to the crude classification method used to extract commuting flows, which may also include many trips that are not commutes. Therefore, some trips begin in commercial areas rather than residential areas and may not be related to the associated employed population.

Fourth, like the parameter estimates for the accessibility terms in the competing destination models (CD1 and CD2) for the bike data, those for the taxi data are positive and therefore indicate agglomeration effects in residence choice (origin locations) rather than the competition effects (i.e., negative accessibility parameter estimates) observed for the census commute-to-work data. Though the origin of taxi trips are not restricted

to stations like the bike trips, it is less common for taxis to search for clients in more residential neighborhoods. Instead, they often choose to look for potential clients in areas of higher and more diverse activities, such as around major transportation hubs, parks, tourist attractions, shopping areas, and main avenues (typically running north and south) over more quiet streets (typically running east to west). This minimizes the time needed to find clients and is also useful for clients looking for a taxi who know to walk to the nearest intersection of an avenue or to a transportation hub to increase the likelihood of finding a taxi quickly. It also further implies, that, as with, the bike data, a pitfall of the taxi data is that they are not as representative of residence locations as is the census commute-to-work data. Related to this, the final difference is that the parameter estimate for housing density takes the opposite sign in the taxi results than it does in the census commuting results. While the census gravity-type model produced a negative estimate and the CD2 model produced a positive estimate, the taxi gravity-type model produced a positive estimate and the CD2 model produced a negative estimate. This may be due to individuals frequently walking from their residence to less residential areas with lower housing density to more quickly begin a taxi trip. Examining the local parameter estimate surfaces provides evidence that these flipped signs may be due to shifts in the spatial heterogeneity of the processes and is described in more detail below.

Figures 75 - 78 pertain to the local parameter estimate surfaces for the gravity-type model (Grav) of the taxi data and figures 79 - 82 pertain to the local parameter estimate surfaces for the competing destination model based on POI density for the taxi data (CD2). For housing density, the gravity-type model surface (figure 75) indicates that many census tracts have parameter estimates that are not statistically significant. Those that are significant tend to show a high level of heterogeneity with

several clusters of positive and negative parameter estimates throughout the study area. However, upon adding the accessibility term to the model in CD2, many of the positive parameter estimates become negative or statistically insignificant (figure 79) and demonstrates why the global parameter estimate flipped from positive to negative. Two particularly strong trends stand out in figure 79. First, there is a large cluster of negative parameter estimates in the eastern portion of Queens, which roughly corresponds to the area around Jamaica station, which is a major transportation hub that connects Manhattan, Queens, and Brooklyn, as well as Long Island to the east and JFK international airport to the south. The second cluster is a small enclave of strongly positive parameter estimates in the southeast of Brooklyn. This area is located between Coney Island to the south and a large park to the northeast. Furthermore, these two areas consistently display particularly strong opposing local relationships (figures 76 - 78 and figures 80 - 82) and demonstrate the usefulness of local models for detecting clusters with geographic context. Further research into these clusters would likely lead to more diverse theoretical interpretations.

A final trend, also noticed in the bike data, is that the competing destination model using accessibility defined with POI density (CD2) produces a local distance-decay parameter estimate surface that has very low distance-decay in central midtown Manhattan with one census tract producing a positive distance-decay estimate. It is

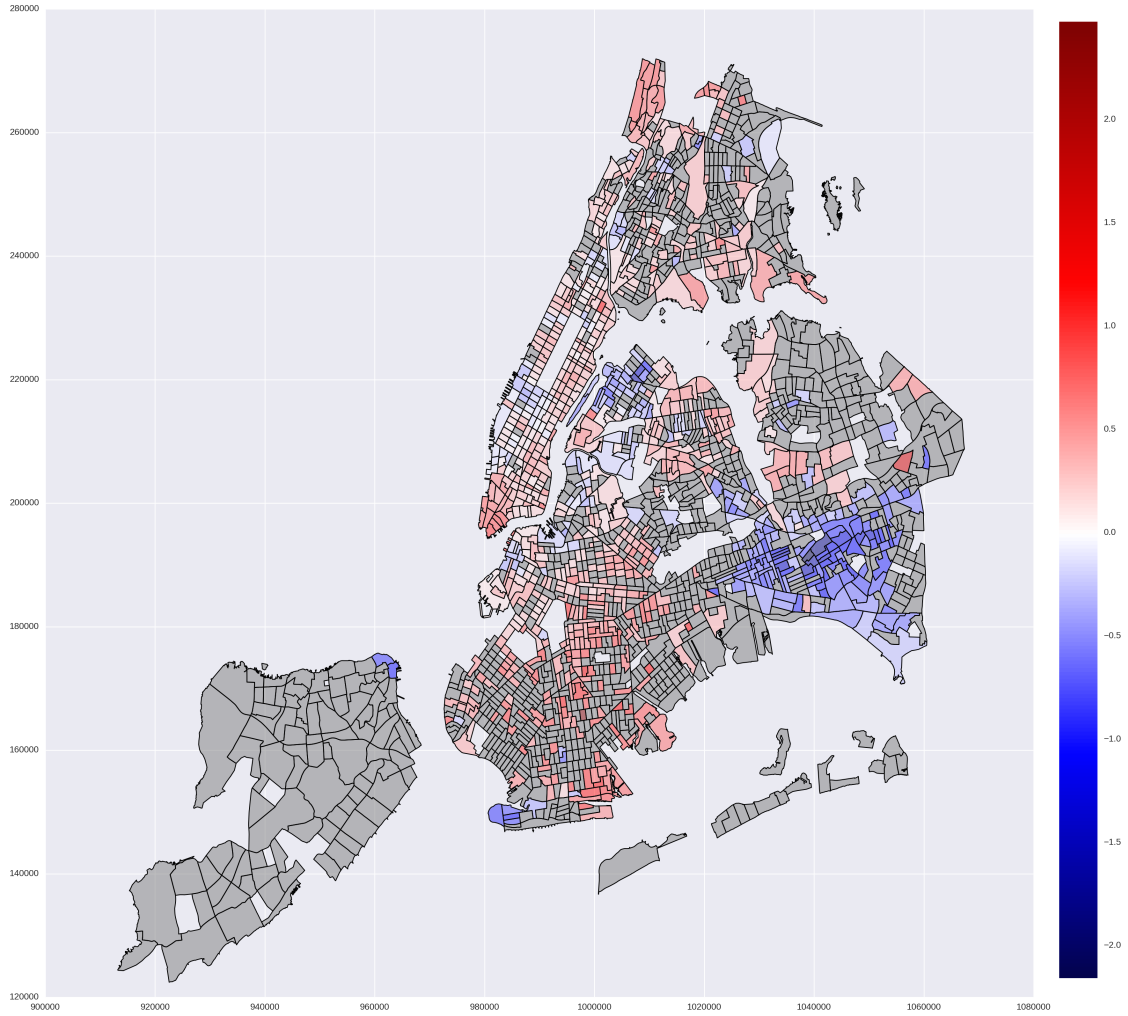


Figure 75: Local parameter estimates for the housing density (hd) variable in the gravity-type attraction-constrained model using taxi data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

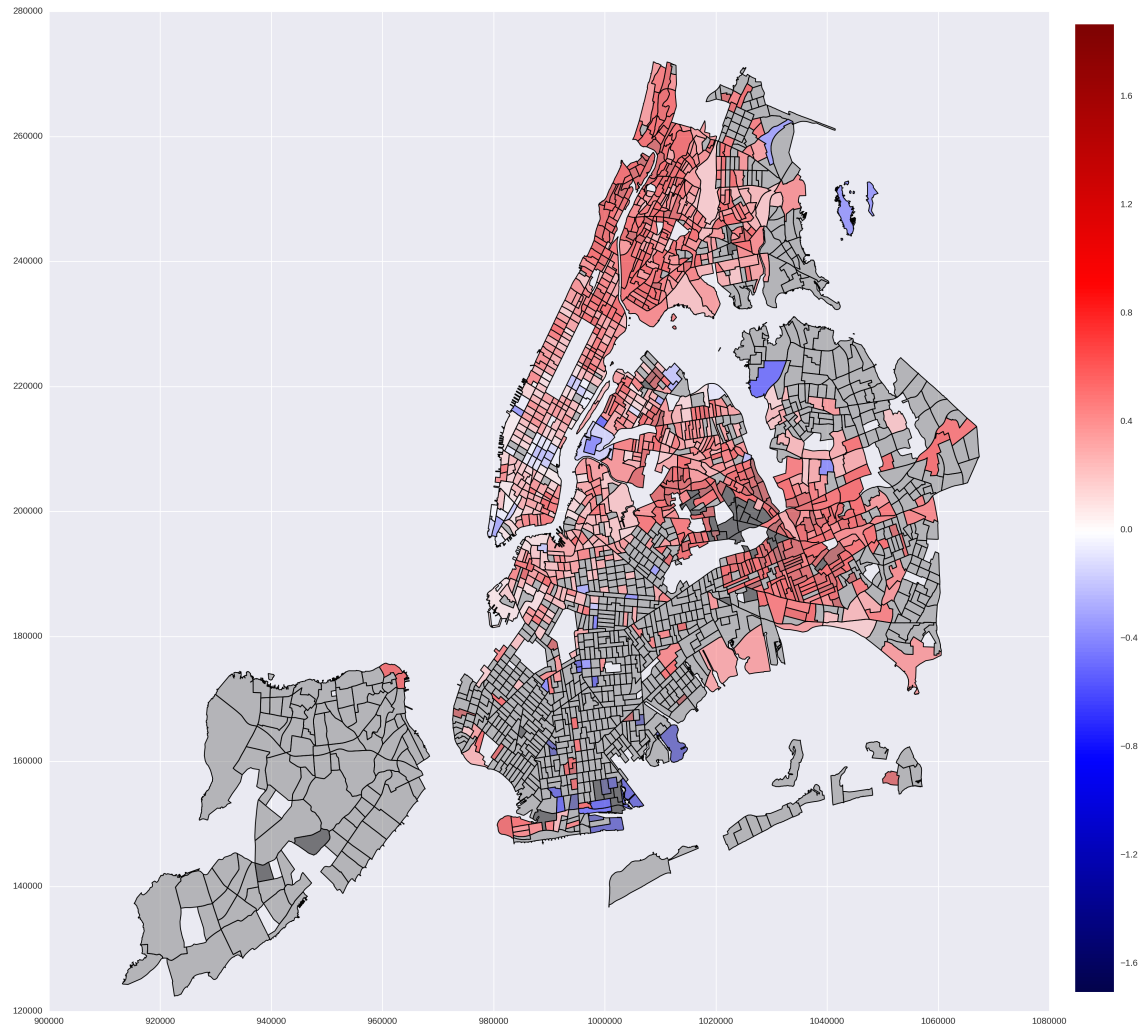


Figure 76: Local parameter estimates for the number of people employed (emp) in the gravity-type attraction-constrained model using taxi data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

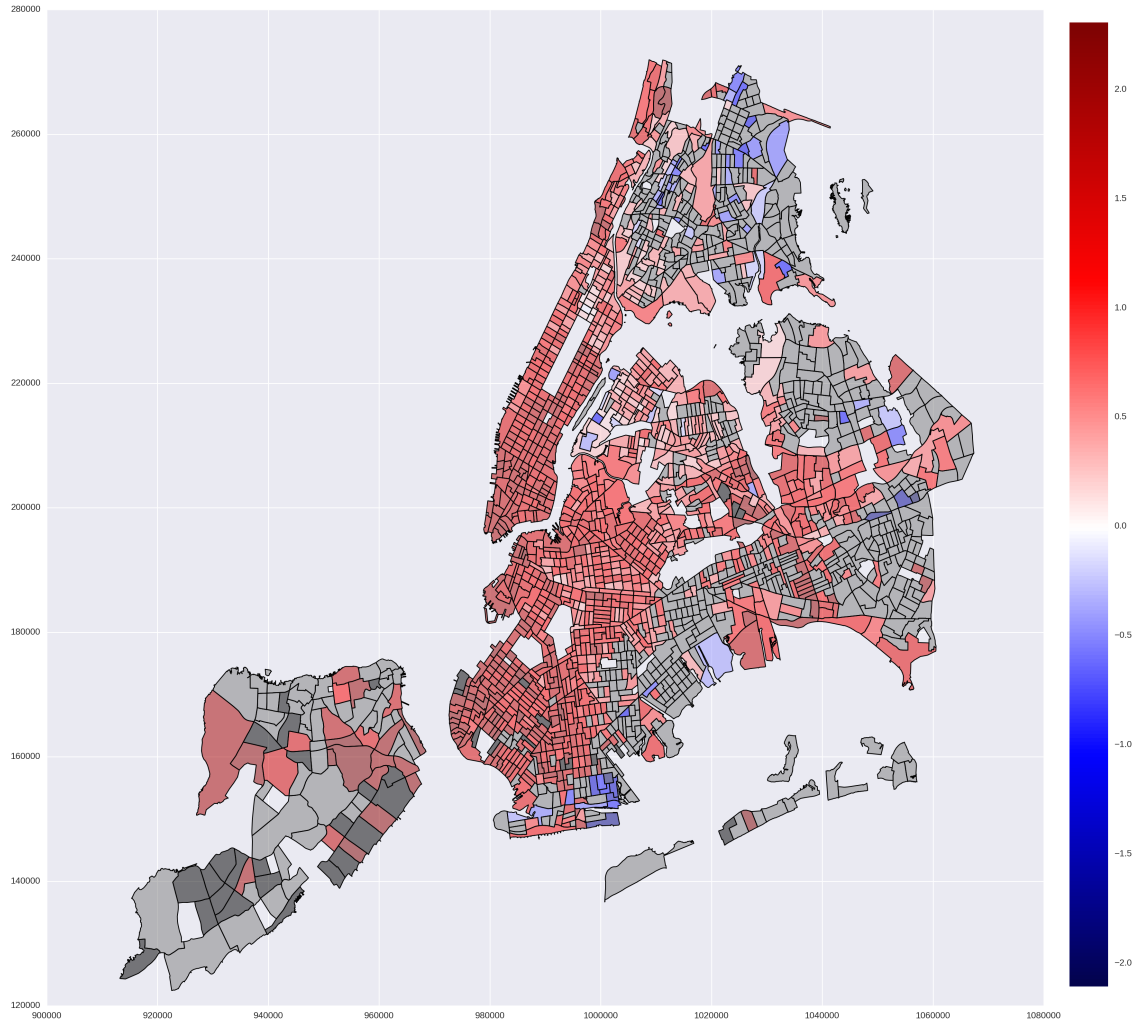


Figure 77: Local parameter estimates for the average income (inc) variable in the gravity-type attraction-constrained model using taxi data. Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

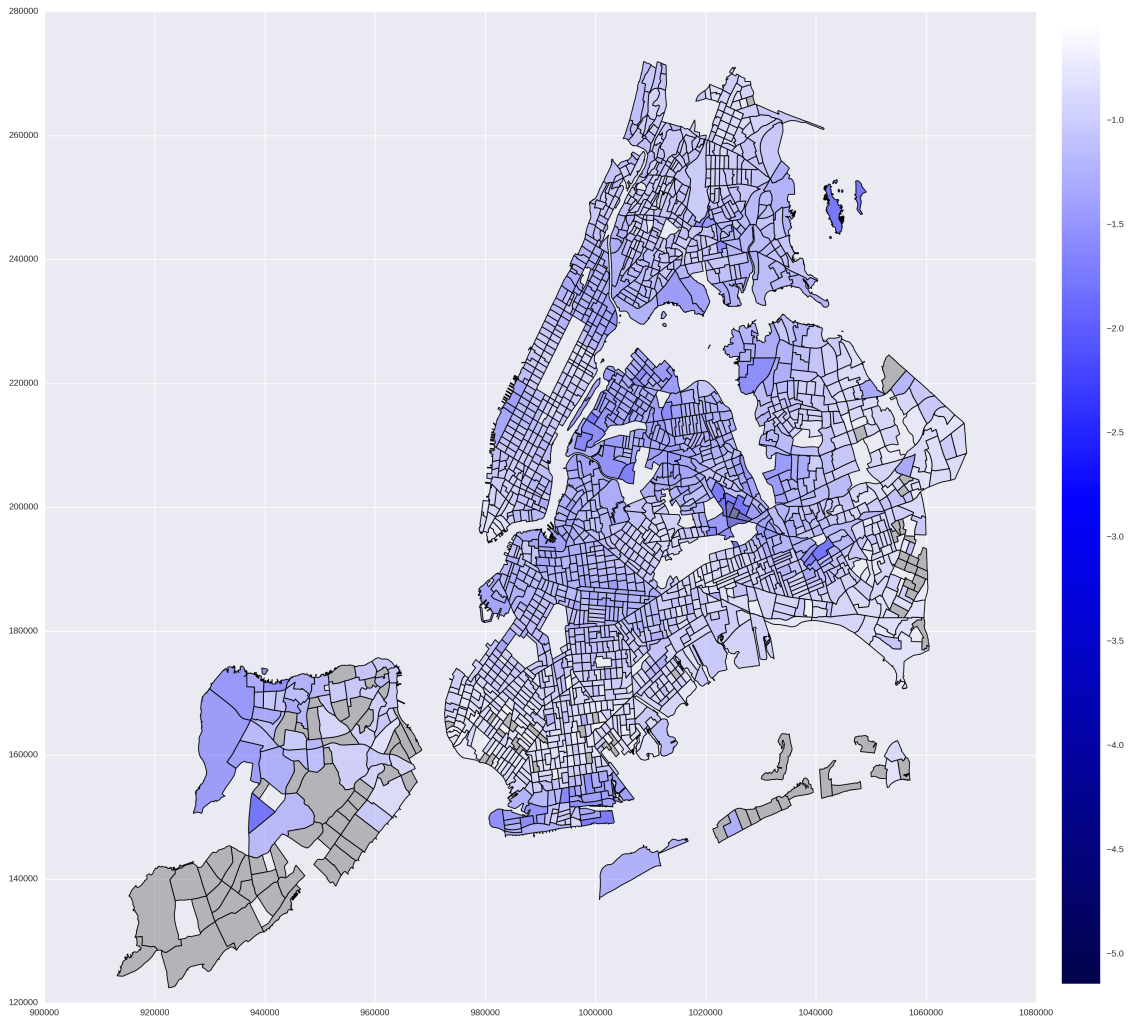


Figure 78: Local parameter estimates for distance-decay (dist) in the gravity-type attraction-constrained model using taxi data. Lighter values indicate a weaker relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

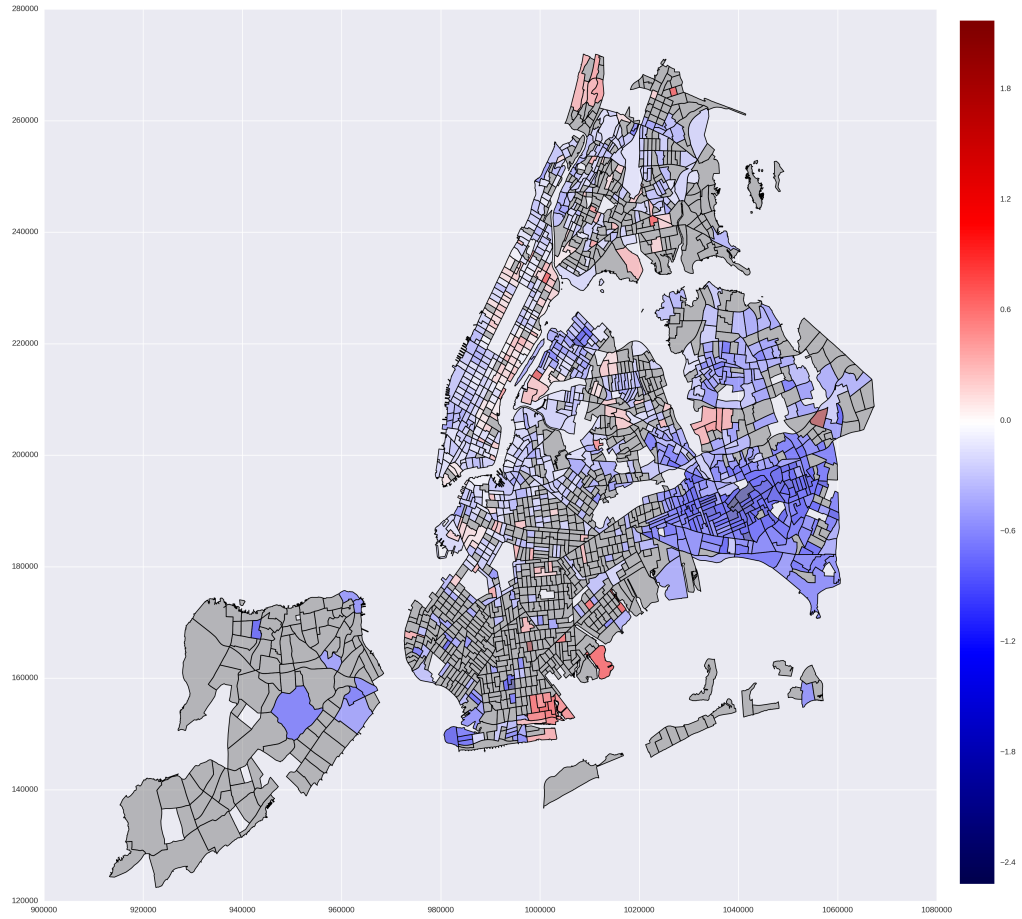


Figure 79: Local parameter estimates for the housing density (hd) variable in the attraction-constrained competing destination model of the taxi data using an accessibility term defined using POI density (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

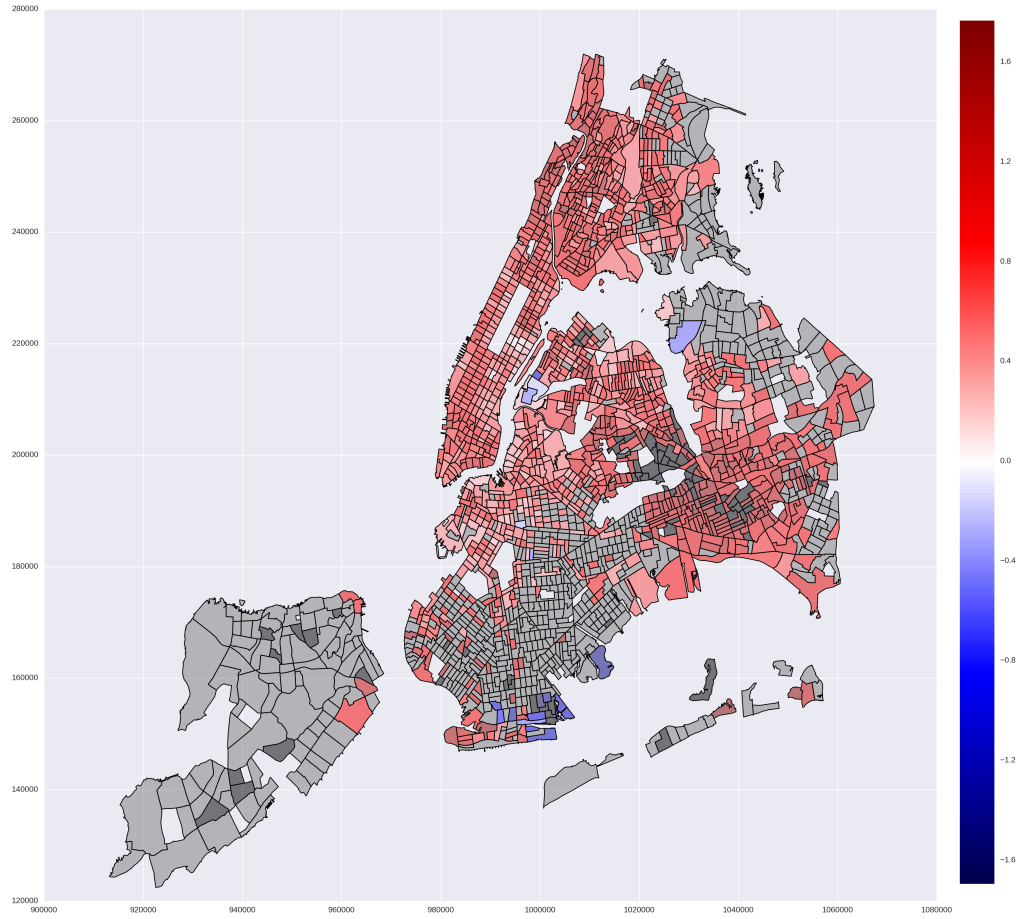


Figure 80: Local parameter estimates for the number of people employed (emp) in the attraction-constrained competing destination model of the taxi data using an accessibility term defined using POI density (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

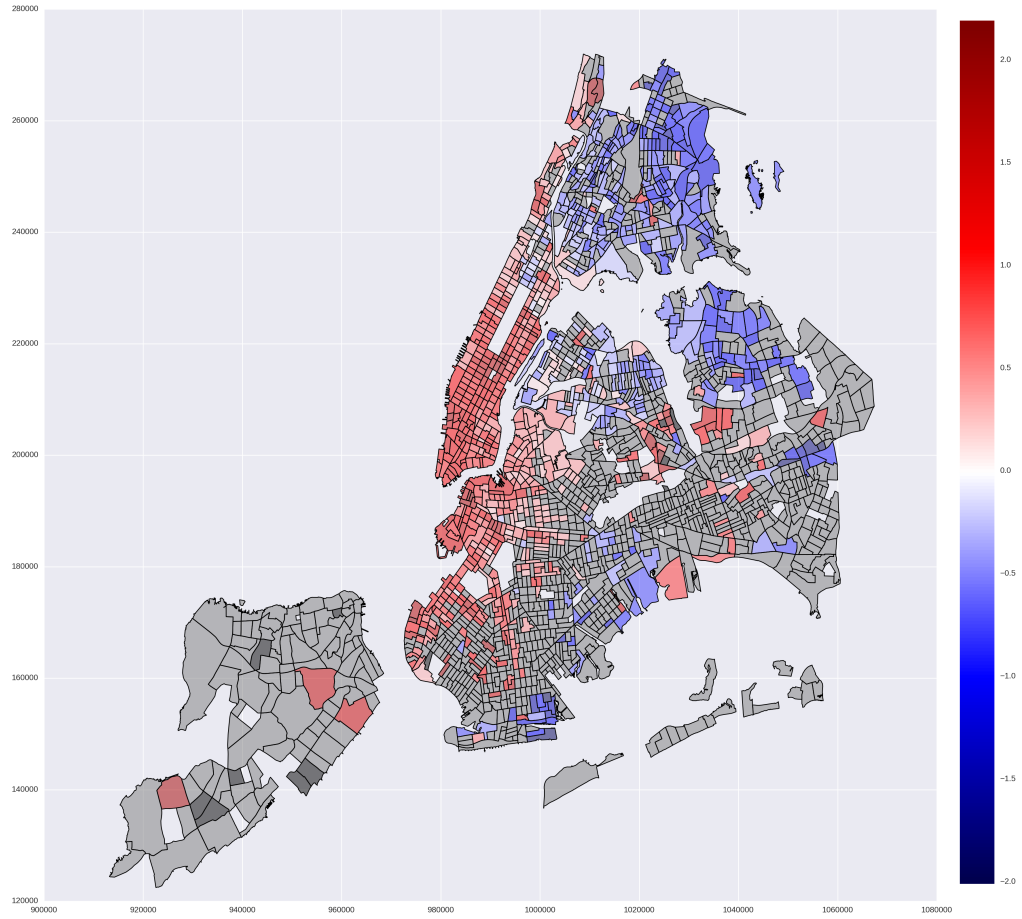


Figure 81: Local parameter estimates for the average income (inc) variable in the attraction-constrained competing destination model of the taxi data using an accessibility term defined using POI density (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

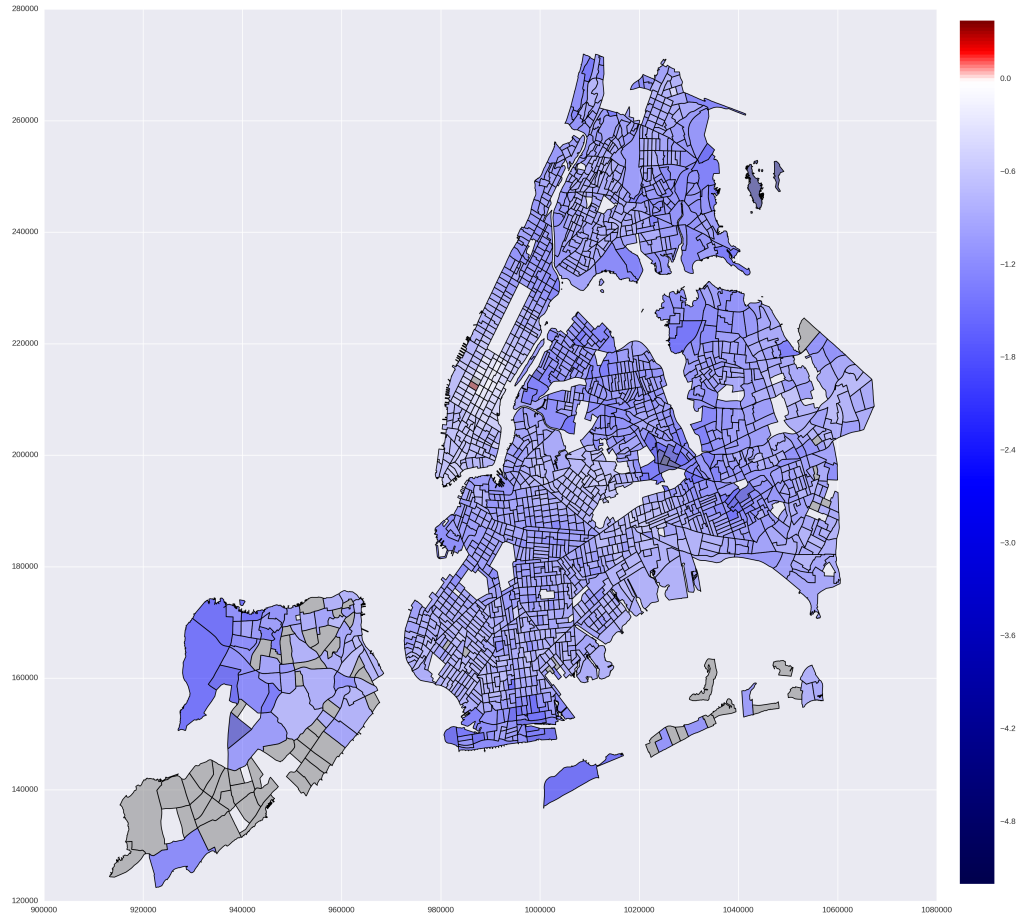


Figure 82: Local parameter estimates for distance-decay (dist) in the attraction-constrained competing destination model of the taxi data using an accessibility term defined using POI density (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

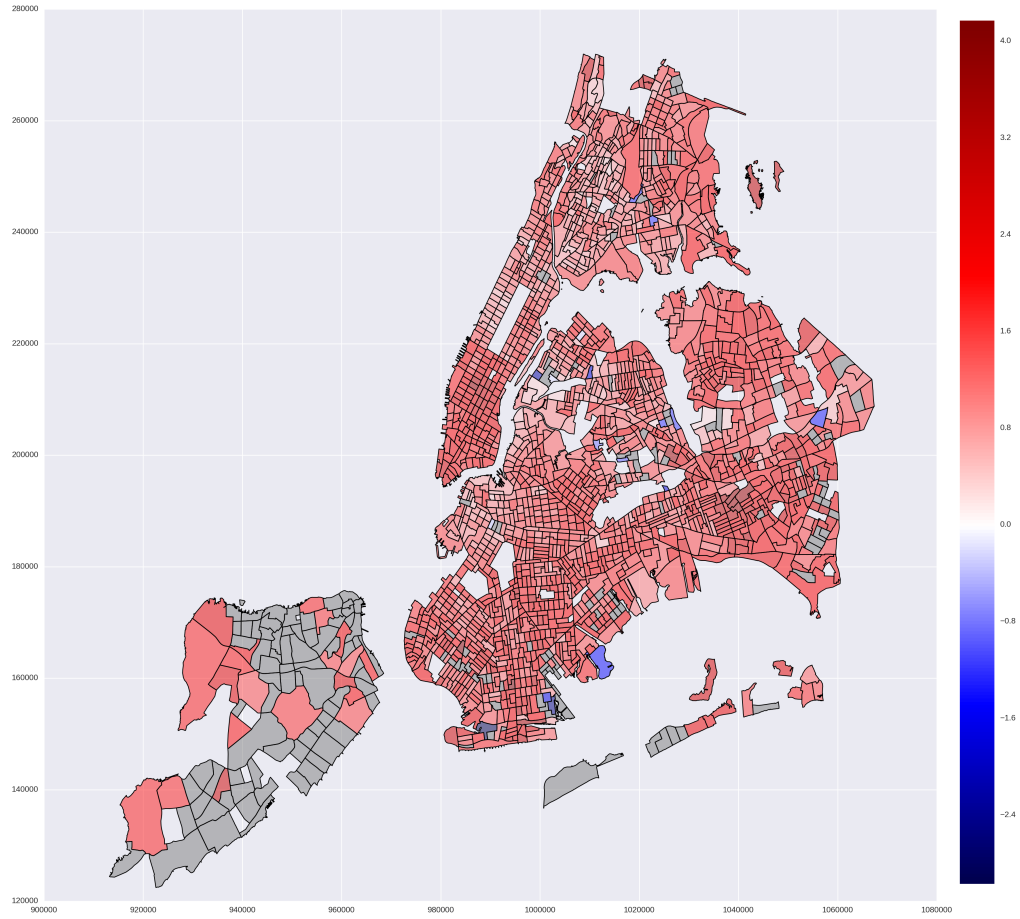


Figure 83: Local parameter estimates for accessibility term (cd) defined using POI density in the attraction-constrained competing destination model of the taxi data (CD2). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

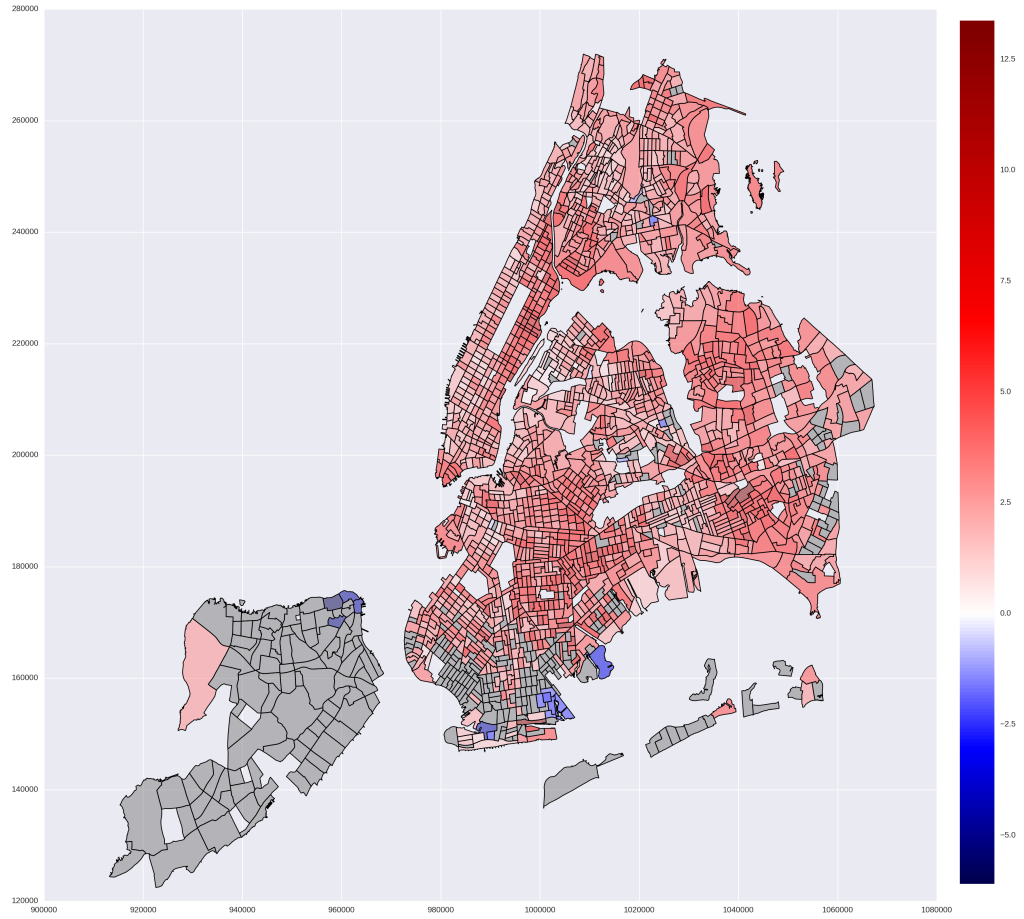


Figure 84: Local parameter estimates for accessibility term (cd) defined using housing density in the attraction-constrained competing destination model of the taxi data (CD1). Red values indicate a positive relationship and blue values indicate a negative relationship. Black tracts have been removed as outliers to aid in visualizing trends while grey tracts are deemed statistically insignificant at the 95% CI using a Bonferroni correction for multiple testing.

intuitive that this area, which has already been designated as a very attractive area, might draw individuals regardless of their potential journey length and therefore results in very weak distance-decay. More specifically, the positive anomaly corresponds very closely to Pennsylvania station, which is one of largest transportation hubs in Manhattan. It is reasonable to assume that many of the trips recorded to this destination are either not commuting trips or only represent one segment of a longer commuting trip. Since the station connects midtown Manhattan to the entire study area and to the greater metropolitan region, it would not be surprising if it attracted many long distance trips. Visualizing the local parameter estimate surface for the accessibility term defined using POI density (figure 83) supports this theory, as it can be seen that midtown Manhattan has the strongest agglomeration forces in Manhattan and means there is a strong relationship between flows and accessibility. However, it is also important to note that, like the bike data, an accessibility term defined instead using housing density (figure 84) resulted in a similar surface, but indicated that the strongest agglomeration forces are on the upper west side, rather than in midtown. Again, it is not possible to determine if one surface is more appropriate and perhaps some combinations of two variables should be explored in future research.

As with the bike data, the taxi data portrayed similar pitfalls in terms of modeling commute-to-work behavior when compared to the census commuting data. Overall, the conclusions reached are dissimilar possibly for three reasons. First, taxi commuters represent a small fraction of the overall commuting population and therefore the data set is not representative of the average commuter who primarily use the public transport system and this is apparent when comparing the parameter estimates. Second, the taxi data are not necessarily representative of individuals' residences

and workplaces. Thirdly, a more accurate classification method is needed to isolate commuting trips from non-commuting trips.

6.5 Moving forward

Through the use of increased computing capabilities it was possible to efficiently calibrate global and local models with over 4,000,000 origin-destination observations and over 2,000 binary indicator variables needed to singly-constrained spatial interaction models.This facilitated the visualization of higher resolution local parameter estimate surfaces, which were used to explore spatial non-stationarity, identify clusters and assign behavioral interpretations to them. Consequently, this work demonstrates the usefulness of the higher spatial resolution of newer, bigger data sources. In this case, census data were also available at the same resolution; however, it would be possible to aggregate the bike or taxi data to an arbitrary grid with even higher resolution that would allow for the exploration of finer spatial variation in the parameter estimates, albeit at the potential cost of increased zero-inflation.

By building models using the census commute-to-work data and then applying them to the bike and taxi data, it was possible to determine that these newer datasets are not sufficient substitutes for traditional survey data sources in that they capture different sets of individuals. People who travel by taxi or by bike at rush hour are not necessarily representative of commuters in general. Consequently, the results from the taxi and bike flows generate information on different sets of people moving across the urban landscape during the morning rush hours and are useful for studying different facets of urban transportation.

Finally, non-traditional variables were used to define accessibility terms in com-

peting destinations models that resulted in different interpretations compared to accessibility terms defined using traditional data sources. Consequently, this chapter identified both potential and pitfalls of some big data sources in the context of spatial interaction modeling of commuting behavior and transportation usage. In the next chapter, the increased temporal resolution of these data sources will be evaluated.

Chapter 7

TEMPORAL SUBSET MODELS OF LOCATION CHOICE IN NEW YORK CITY

7.1 Introduction

The previous chapter evaluated the potential and pitfalls associated with using 'big data' (bike and taxi datasets in New York City) for spatial interaction modeling. Specifically, local models were used to investigate geographic variation in commuting processes, such as residential choice. It was determined that although the bike and taxi data are not representative of the average commuter, they may still be used to study behavior associated with specific modes of transportation. It was also demonstrated that in both the bike and taxi datasets, agglomeration forces tend to be present across origins and, consequently, it is important to include an accessibility measure to capture spatial structure effects. Examining surfaces of local parameter estimates allowed the nature of the spatial structure effects (i.e., agglomeration effects across origins) to be further illuminated and it was shown that different accessibility terms can be computed using either traditional census data or newer non-traditional sources of data.

Spatial interaction models are typically calibrated on data that are aggregated over several months or years and are rarely calibrated on data disaggregated to smaller temporal units. This is primarily because human movement data are usually available only in course temporal and spatial aggregations; an example being migration and commuting data obtained from the US census. Such data can only provide a broad average representation of movement behavior and it has to be taken on faith that

this behavior is constant across the timer period for which the data are collected. However, movement data with a very high number of observations and increasingly finer temporal resolutions, such as the bike and taxi in New York City, are now becoming available and provide an exciting means to explore potential dynamics in movement behavior by calibrating spatial interaction models on temporal subsets of the data and compare parameter estimates across these subsets. Therefore, in this chapter spatial interaction models are calibrated for monthly, weekly, daily, and even hourly subsets of the bike and taxi data and the parameter estimates are presented as time series in order to investigate potential temporal patterns.

The overall goals of this chapter are threefold. First is to establish baseline models of destination choice using bike and taxi data that are aggregated from June 2014 through May 2016. Second is to investigate whether or not parameter estimates can be reliably obtained for increasingly finer temporal subsets of the bike and taxi data. If they can, the third goal is to determine how behavior associated with the bike and taxi trips varies over time and across different temporal resolutions. In particular, attention is paid to a period of blizzard activity that occurred in late January 2016 to explore how extreme weather affects the relationships between the urban environment and mobility. As a result, this chapter provides further evaluation of the potential and pitfalls of the bike and taxi data for understanding urban mobility and dynamic human behavior. The overall modeling framework used to carry out these goals is introduced below.

7.2 Modeling framework

Attraction-constrained models were employed in chapter 6, since the main goal was to model the choice of residence location (i.e., origins) associated with commuting trips. However, this framework is not adopted here. It was already discussed that the bike and taxi data may not be as representative of a single process as the census data because they can contain many types of trips. Though the diversity of processes represented within these new sources of data can be considered a weakness of the data, it may also be considered a strength, since it provides the opportunity to study more aspects of mobility. For instance, the census commute-to-work data would not be useful to study any of the movements in New York City that are not associated with commuting! The bike and taxi data also represent trips associated with different types of shopping, leisure, and social interaction, which are a strong component of urban life. Unlike the commuting data that describes average behavior over the long-term, each observation in the bike and taxi data represent choices made in real-time. As a result, it is likely that individuals represented in the bike and taxi data would be making short-term decisions about potential destinations rather than origins so the focus in this chapter is on destination choice and production-constrained models are employed, which are discussed in detail in chapter 2, but are reviewed here for convenience. Production-constrained models are useful for allocating known numbers of individuals from each origin to a set of destinations and the model is formally given in multiplicative form with a power function of distance-decay as

$$T_{ij} = A_i O_i W_j^{\alpha} d_{ij}^{\beta} \tag{7.1a}$$

$$A_i = \Big(\sum_j W_j^{\alpha} d_{ij}^{\beta} \Big)^{-1} \tag{7.1b}$$

where T_{ij} is the number of flows between location i and destination j, O_i is the number of trips that begin at origin i, W_j is a vector of destination location attributes, d_{ij} is the distance between i and j, and A_i is a balancing factor that ensures the total number of inflows O_i is replicated in the predicted flows. In this context, the primary interest is in describing the nature of the destinations that individuals choose from and the effect of distance, which is accomplished by estimating the origin location parameter vector α and the distance-decay parameter β and interpreting them in the context of urban activity within New York City. Consequently, a production constrained model is first calibrated for the bike and taxi data aggregated across the entire study period. Since the interest in this chapter is on describing the nature of the destinations and the effect of distance over time, models are calibrated[34] for monthly, weekly, daily, and hourly subsets of the data. For the monthly and weekly calibrations, the entire dataset is used since the number of calibrations remains reasonable at approximately 24 and 104 temporal units, respectively. However, the daily and hourly subsets would result in hundreds and thousands of model calibrations for the entire study period and therefore only a portion of the study period was used in order to limit the number of calibrations. More specifically, daily calibrations were limited to one summer month (June 2015) and one winter month (January 2016) and hourly calibrations were limited to approximately one week within the summer month (June 17-24) and approximately one week within the winter month (January 20-26). Results are first presented for the bike data and then for the taxi data.

[34] All model calibrations in this chapter were carried out using custom software that leverages the sparse design matrices implied by constrained spatial interaction models for efficient computation (Oshan, 2016). The code is part of the SpInt module of the Python spatial analysis library (PySAL) and is available at *https://github.com/pysal/spint*

7.3 Bike data results

7.3.1 Baseline model

Before calibrating the temporal subset models on the bike data, an initial specification must be proposed and validated on the entire dataset. A simple model was selected based on the results from chapter 6 where the most parsimonious model for the bike data was obtained using an accessibility term defined using station capacity and inter-station distances. However, an important difference is that inter-station trips were used here instead of aggregating trips between census tracts. In chapter 6, a primary goal was to compare the models from the bike data to those from the census commute-to-work data and therefore it was useful for the sake of comparison to maintain the same spatial unit of analysis across the different data sets. This restriction is not necessary in this chapter and it is more intuitive to calibrate models of bike movements using the most spatially disaggregate data possible, which are the flows between the individual bike stations. The flow models to be calibrated include three explanatory variables: station capacity (cap), accessibility defined using station capacity (cd), and inter-station distance computed using the Mapzen *Matrix* routing algorithm that prefers roads favored by cyclists (dist). These are used to understand the determinants of approximately 20 million bike trips aggregated to 242,557 origin-destination routes between 497 origin stations and 497 destination stations [35]. Parameter estimates and model fit for these data are presented in table 6 where it can be seen that a moderate-to-high pseudo R^2 of 0.6925 is obtained.

[35]Intra-station trips are excluded, as well trips to stations that did not have a reliable measure of station capacity. About 82,626 of the routes contain zero flows, indicating about two-thirds of the routes have non-zero flow magnitudes.

Table 6: Parameter estimates and model fit for the bike data. Grav refers to the gravity-type production-constrained spatial interaction model and CD refers to the competing destination model.

	Grav		CD(cap)	
	Estimate	SE	Estimate	SE
cap	0.8466	0.0008	0.6565	0.0008
dist	-1.356	0.0003	-1.298	0.0003
cd	-	-	1.198	0.0015
adj. R^2	0.6818		0.6925	
AIC	20024094		19349287	

The parameter estimates in table 6 provide the interpretation that, *ceteris paribus*, larger flows of bike trips are associated with destinations that: have larger station capacity; are closer to other bike stations; and are nearby to where the trip started. It is intuitive that individuals would want to minimize the distance traveled and that they would be more likely to ride to stations with a higher capacity of bike docks to maximize the probability that there will be an empty dock into which the user can park their bike at the end of their trip. Accessibility in this case is representative of the station capacity in areas around a potential destination station and the positive parameter estimate obtained here denotes agglomeration effects suggesting that stations within easy reach of alternative stations are favored by bike renters, presumably to minimize inconvenience should parking slots not be available in the intended destination station. Finally, it can be seen that adding the accessibility term (CD) slightly improves the model fit and has the effect of reducing the magnitude of the measured distance-decay and the attraction of larger bike stations; again the latter makes sense as bike renters will be less concerned about a station's capacity if there are alternative stations nearby.

7.3.2 Temporal subsets

Having established a baseline model for the whole time period of the data, models were then calibrated separately for monthly, weekly, daily, and hourly subsets of the bike data using the same model specification. The goals of this experiment are to (1) assess if robust parameter estimates can be obtained for these subsets of bike data and (2) determine whether the calibrations allow us to examine if and how movement behavior changes over time. Figures 85 through 90 provide parameter estimates, model fit, and the number of destination stations for each of the monthly, weekly, daily, and hourly subsets of the bike data, with separate figures for the daily and hourly results pertaining to warmer and colder periods in the study area. The blue dotted horizontal line in each time series indicates the value of the corresponding baseline parameter estimate.

Examining figures 85 through 90, it can be seen that there is an expected trend where the larger temporal units, such as months (figure 85) and weeks (figure 86) yield parameter estimates with very low uncertainty, and that this uncertainty increases as the data are subset to the smaller temporal units of days and hours. Uncertainty is evaluated here by the 95% confidence interval that is plotted for each parameter estimate time series. The dashed green line represents the upper limit of the confidence interval and the dashed red line represents the lower limit of the confidence intervals. It may be seen in figure 85 and figure 86 that the confidence intervals are tightly centered on the solid blue line that represents the parameter estimates. For the daily subset models, the confidence intervals increase in size (figure 87 and 88), but never incorporate zero. The results from the hourly subset models (figures 89 and 90) indicate that parameter estimates with acceptably low uncertainty are only obtained during

times of high usage, which is typically during the morning and afternoon commutes. During times of low usage, and especially late at night, the confidence intervals are far too large for any sensible conclusions to be drawn from the results. These results hold for daily and hourly subsets pertaining to periods of both warmer and colder weather. One exception to this trend is immediately after the bike-share system was re-activated after being deactivated during the blizzard activity. The period of deactivation is captured by breaks in figures 88 and 90 and figure 90 demonstrates the exception where very high parameter estimates are obtained with low uncertainty. However, the high positive value of distance-decay observed cannot be logically interpreted and this exception should likely be treated as an anomaly since this parameter estimate was obtained for a period that contained few trips between a small number of stations.

Several interesting temporal trends can be seen in the parameter estimates. For distance-decay, three trends are apparent. First, distance-decay varies seasonally (figures 85 and 86) with a stronger effect in the colder winter months when fewer cyclists would want to make longer trips and a weaker effect in the warmer summer months when weather is less of a deterrent. Second, distance-decay tends to be stronger during the weekends and on holidays when individuals do not typically need to commute to work and might choose to stay closer to home (figure 87 {6th-7th, 13th-14th, 20th-21st, 27th-28th} and figure 88 {1st-3rd, 9th-10th, 16th-17th}). Third, distance-decay remains between -1.0 and -1.5 when it can be reliably estimated throughout the day, with lower distance-decay during morning and afternoon commuting hours. This trend is visible in figure 89, though it is not visible in figure 90 because the previously mentioned anomaly makes the range of the series larger, resulting in a less focused plot. Though not shown here, the series in figure 90 is also examined by manually decreasing the range of the y-axis and it was found that the parameter estimates

are similar to those for the warm hourly subsets (figure 89). Therefore, subsequent analysis of the parameter estimates focuses on only the warm hourly subset results. These three trends demonstrate that the estimated distance-decay varies over time and that the nature of this variation is dependent upon which temporal frequency is being used to subset the data. Each of the temporal frequencies employed in this experiment is associated with cycles of human activity or human responses to physical activity (i.e., weather). Since the variations in the measured distance-decay are linked to these human behaviors, this provides evidence that the distance-decay parameter estimates obtained can reasonably be interpreted as reflections of human behavior.

Furthermore, it can be seen in the monthly and weekly results that the baseline distance-decay parameter estimate tends to be an approximate average of the series values since the blue dotted line divides the series with similar portions occurring above and below the baseline. This is not the case for the daily and hourly results. For the daily distance-decay estimates pertaining to warmer weather, the series tends to run above the baseline while the opposite trend is true for the daily distance-decay estimates pertaining to colder weather, which has a series that runs below the baseline. A similar, though less pronounced pattern occurs in the hourly results. This pattern exists because two different seasonal trends are reflected in the warm and cold results and is therefore not problematic. However, this pattern also suggests that when data are aggregated over a time period that only represents part of a temporal cycle (i.e., a season) that the estimated distance-decay may only be accurate for that subset of time.

The parameter estimates for the station capacity and accessibility variables also contain temporal variation that seems to depend on the temporal disaggregation of the data so that the following results are described for each temporal unit separately.

For the monthly subsets (figure 85), the series of parameter estimates for station capacity and accessibility (i.e., nearby capacity) are inversely related. Station capacity is more important in the warmer months and less important in the cooler months. This may perhaps be due to cyclists making trips regardless of station capacity in the cold months when demand is generally lower and the availability of docking places is not an issue. In addition, the variation in station capacity is less intense for the second half of the series probably because of the increase in the number of stations in July 2016 which seems to have the outcome of stabilized the impact of station capacity. Accessibility parameter estimates exhibit the opposite trend where stations that may be perceived as part of a larger cluster of stations are more attractive during the colder months. This may be that during the colder months, bike trips represent a larger share of commuters and a lower share of tourists and commuters are more likely to have destinations where there are many bike stations. Interestingly, monthly accessibility parameter estimates are always below the baseline measure indicating that the baseline measure is not an average of the monthly measures and this is revisited later.

Moving to the weekly subset results (figure 86), it can be seen that both capacity and accessibility parameter estimates maintain a similar trend to that found for the monthly subset results. In contrast, a much different pattern is produced for the capacity and accessibility parameter estimates in the daily results. For both the warmer period (figure 87) and the colder period (figure 88) station capacity has a similar range of estimates to that observed in the monthly and weekly results but it can now be seen that station capacity is less important on the weekends ({6th-7th, 13th-14th, 20th-21st, 27th-28th} for the warm period and {2nd-3rd, 9th-10th, 16th-17th, 23rd-24th} for the cold period). During the weekends, cyclists are less likely to travel

to areas of high employment that might be associated with stations of high capacity. A similar cyclical pattern exists in the daily accessibility parameter estimates, except now the series for the warmer period spans both positive and negative values with the negative values occurring on the weekends. Those who cycle during weekend and in warm weather are likely less constrained by weather or time and may choose to cycle to areas with relatively fewer bike docks so that there are fewer trips than would be expected to areas where bike stations are located relatively close to each other. In addition, the warm weather accessibility parameter estimates have values well below the baseline estimate but the cold weather estimates are much closer to the baseline value.

Even more detail can be observed in the hourly subset results pertaining to the capacity and accessibility parameter estimates (figures 89 and 90). As previously discussed, the parameter estimates contain less uncertainty during the day for high usage periods, though different hourly patterns emerge for capacity and accessibility. For the competition/agglomeration variable, the parameter estimates are only reliable (i.e., when the confidence interval does not incorporate zero) for periods related to the morning and afternoon commuting rush. In contrast, station capacity parameter estimates are their highest and least uncertain in the morning and then decline throughout the day before becoming statistically insignificant for the later evening until the next morning commute. The trends for both parameter estimates are related to commuting behavior, and indicate that nearby station capacity is important for both the morning and afternoon commute. Station capacity becomes less important throughout the day because people often go out for dinner or go shopping before heading home and places of leisure and residential destinations may not have stations with capacity as high as stations in areas of high employment. A final feature that

can be observed in the hourly data is that similar to distance-decay, there is an anomaly immediately after the bike-share system commences following the blizzard activity. In fact, a positive accessibility parameter estimate of approximately 6.5 is obtained, though it can not be seen in figure 90 which has been zoomed to focus on the finer trends. It is perhaps possible that this extreme outlier is pulling the baseline accessibility parameter estimate upwards. It was previously mentioned that data during this period consists of very few trips between a small number of stations and excluding this data from future analyses might result in a baseline parameter estimate that is smaller in magnitude.

It appears that the bike data are not particularly helpful for capturing dynamic behavior related to blizzard activity and this is due to two reasons. First, for the coarser temporal units, the data are highly aggregated and the blizzard only pertains to a tiny portion of the data. In figures 85 and 86 the dashed red line denotes the time period associated with blizzard activity and no strong differences in the parameter estimates can be observed. The second reason is due to the fact that the bike-share system was temporarily closed down during and after the blizzard, which means no data were recorded for those days and hours that are most closely associated with the blizzard. Parameter estimates obtained after the gap (figures 88 and 90) are anomalous for a single time period and then seem to return back to typical behavior and this is due to relatively few trips being recorded when there was snow and ice present and cycling is unattractive.

Despite the bike data being insufficient to capture behavior associated with blizzard activity, the results indicate that the temporal subsets of the bike data are highly useful for extracting more consistent temporal patterns that occur cyclically. Parameter estimates vary seasonally, between weekdays and weekends, and over the course of a

day and these cycles are associated with several different types of behavior, including commuting and leisure. Furthermore, parameter estimates are very robust when aggregated monthly, weekly, and daily. Only when the data are aggregated hourly are some statistically insignificant (i.e., the confidence interval includes zero) parameter estimates produced, and these tended to be more numerous during the later hours of the evening and earlier hours of the morning when the fewest bike trips occur. This suggests that perhaps as the bike-share system continues to expand and becomes more popular that parameter estimates for hourly subsets may be obtained with less uncertainty. It might also be possible to capture more diverse behavior by classifying destinations based on the different types of attractions that are available, such as employment, dining, residences, or green space and then estimating a separate series for each type of destination. In the next section, a similar experiment and evaluation is provided for the taxi data using a baseline model that includes more destination explanatory variables.

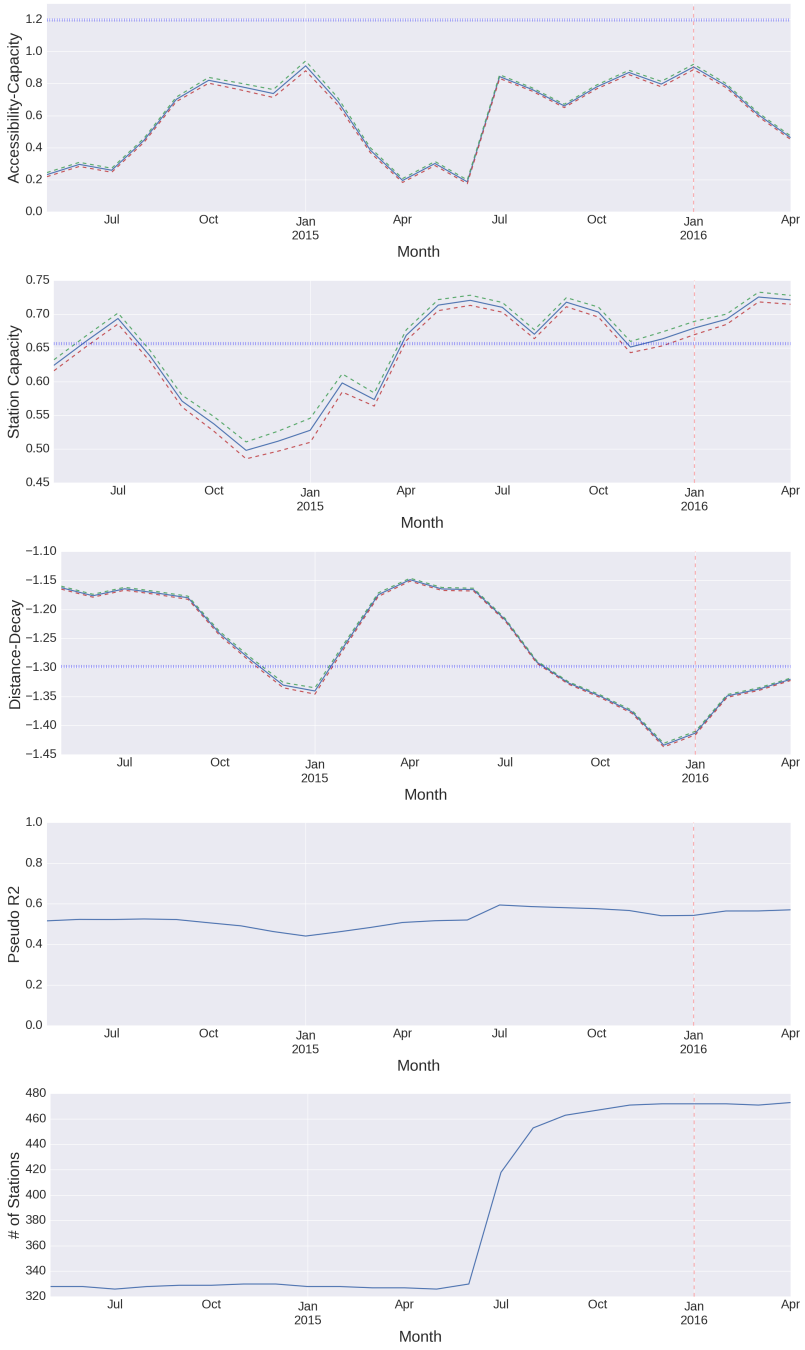


Figure 85: Monthly sampling results for the bike data, including parameter estimates, model fit, and number of origin spatial units. Dashed red line denotes blizzard activity.

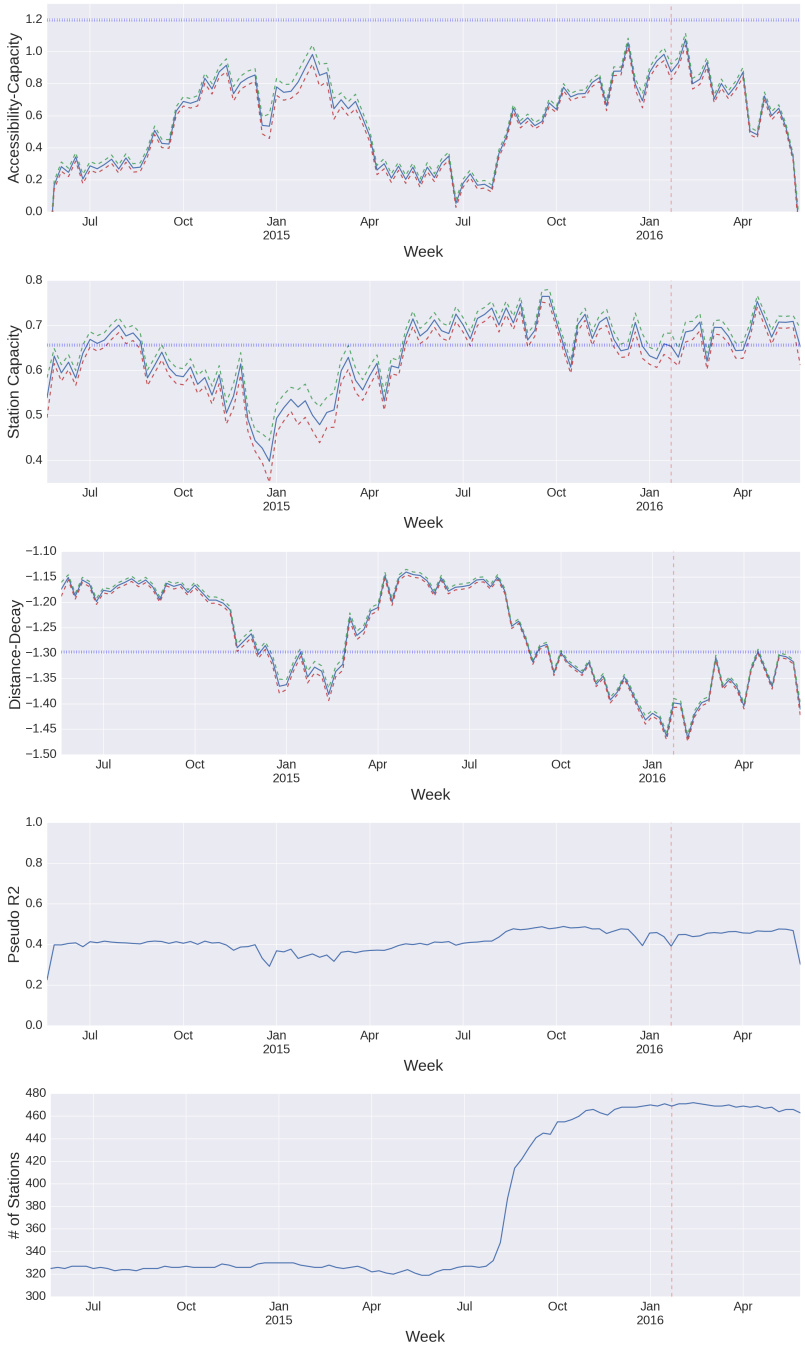


Figure 86: Weekly sampling results for the bike data, including parameter estimates, model fit, and number of origin spatial units. Dashed red line denotes blizzard activity.

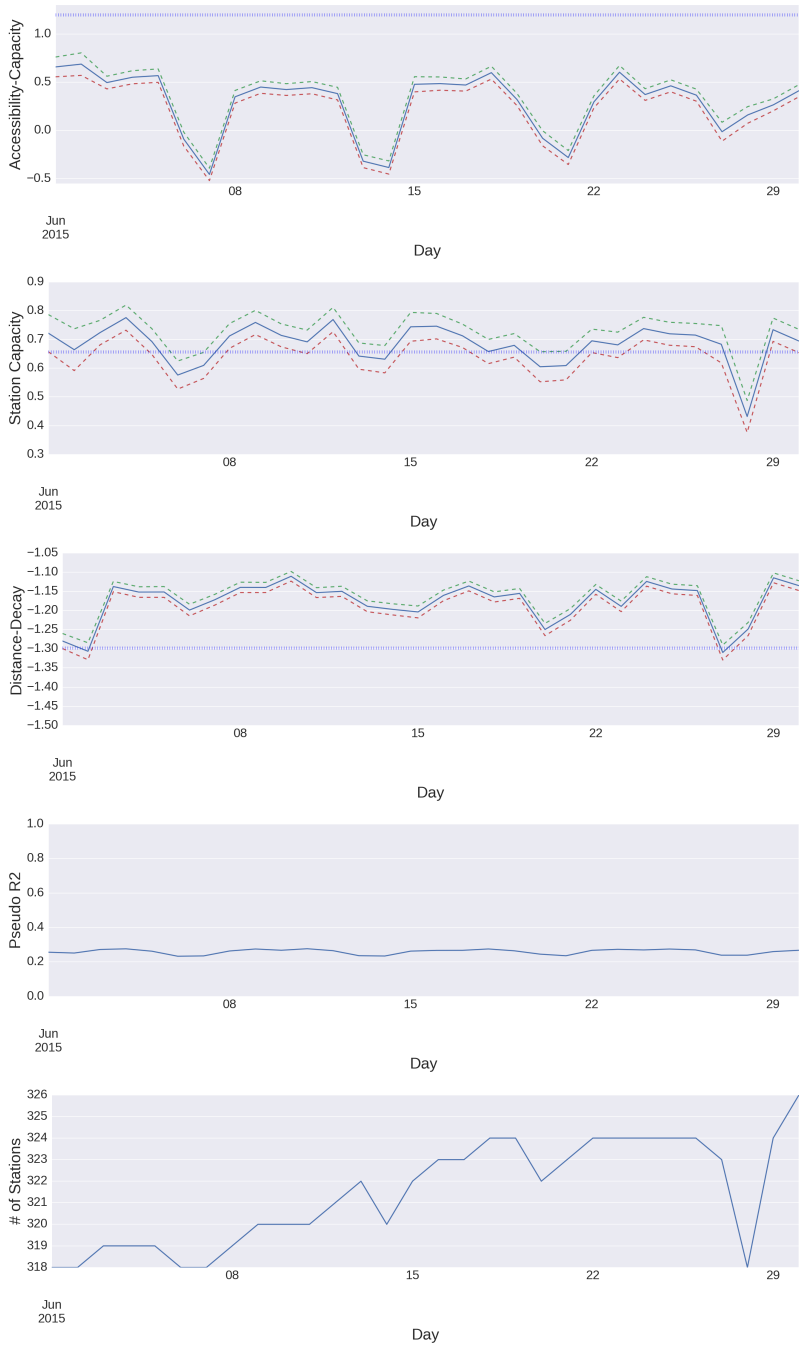


Figure 87: Daily sampling results for the bike data using a single month representative of warmer weather (June 2015), including parameter estimates, model fit, and number of origin spatial units.

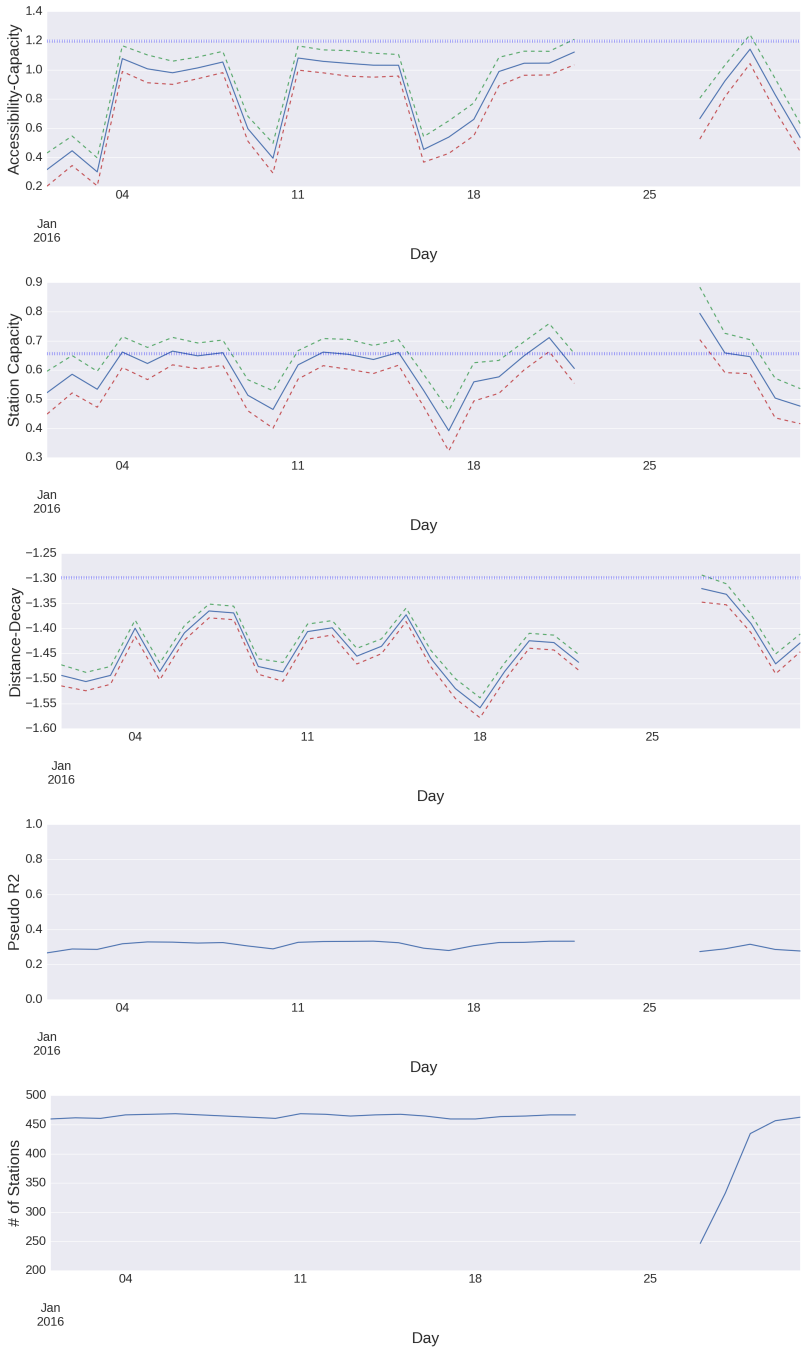


Figure 88: Daily sampling results for the bike data using a single month representative of colder weather (January 2016), including parameter estimates, model fit, and number of origin spatial units. The gaps in the trends denote blizzard activity when the stations were temporarily closed down.

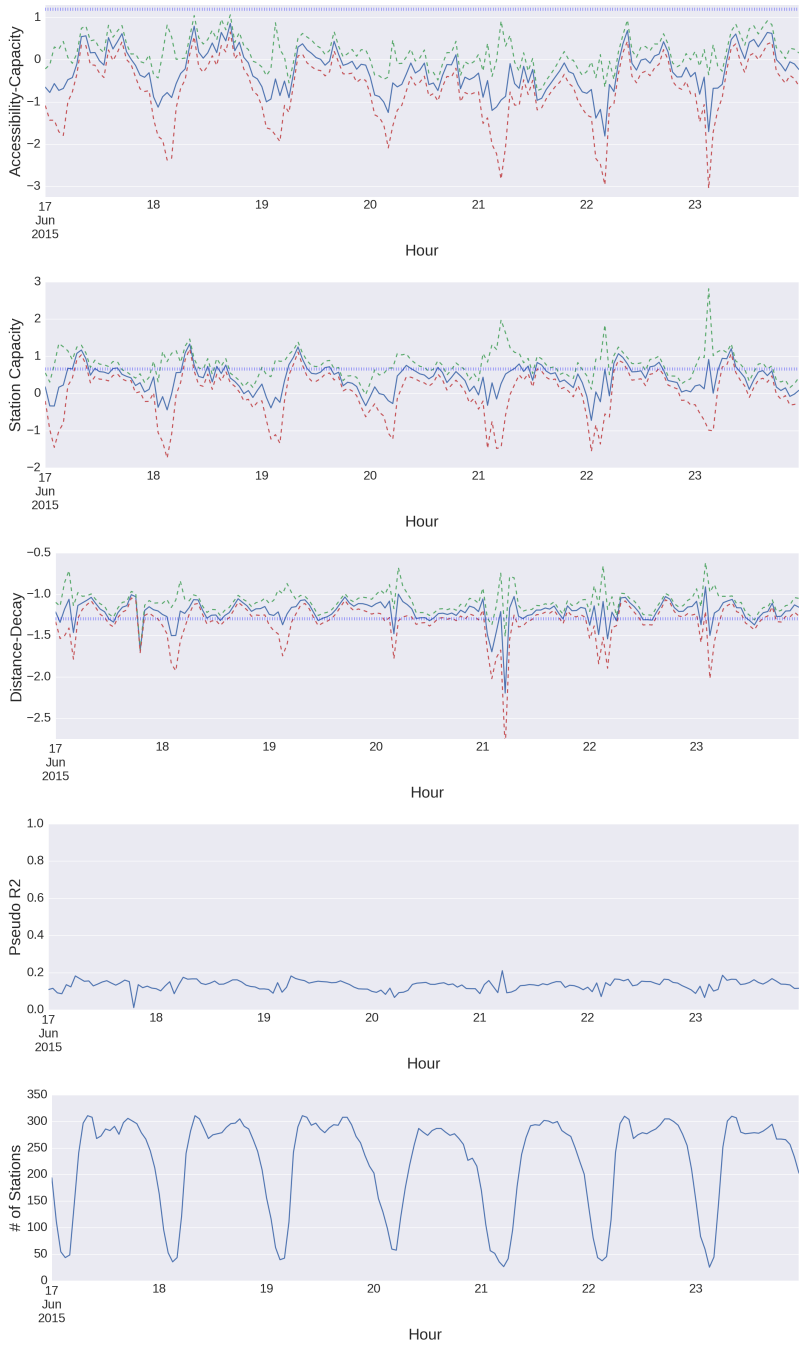


Figure 89: Hourly sampling results for the bike data using a single week representative of warmer weather (June 17-24 2016), including parameter estimates, model fit, and number of origin spatial units.

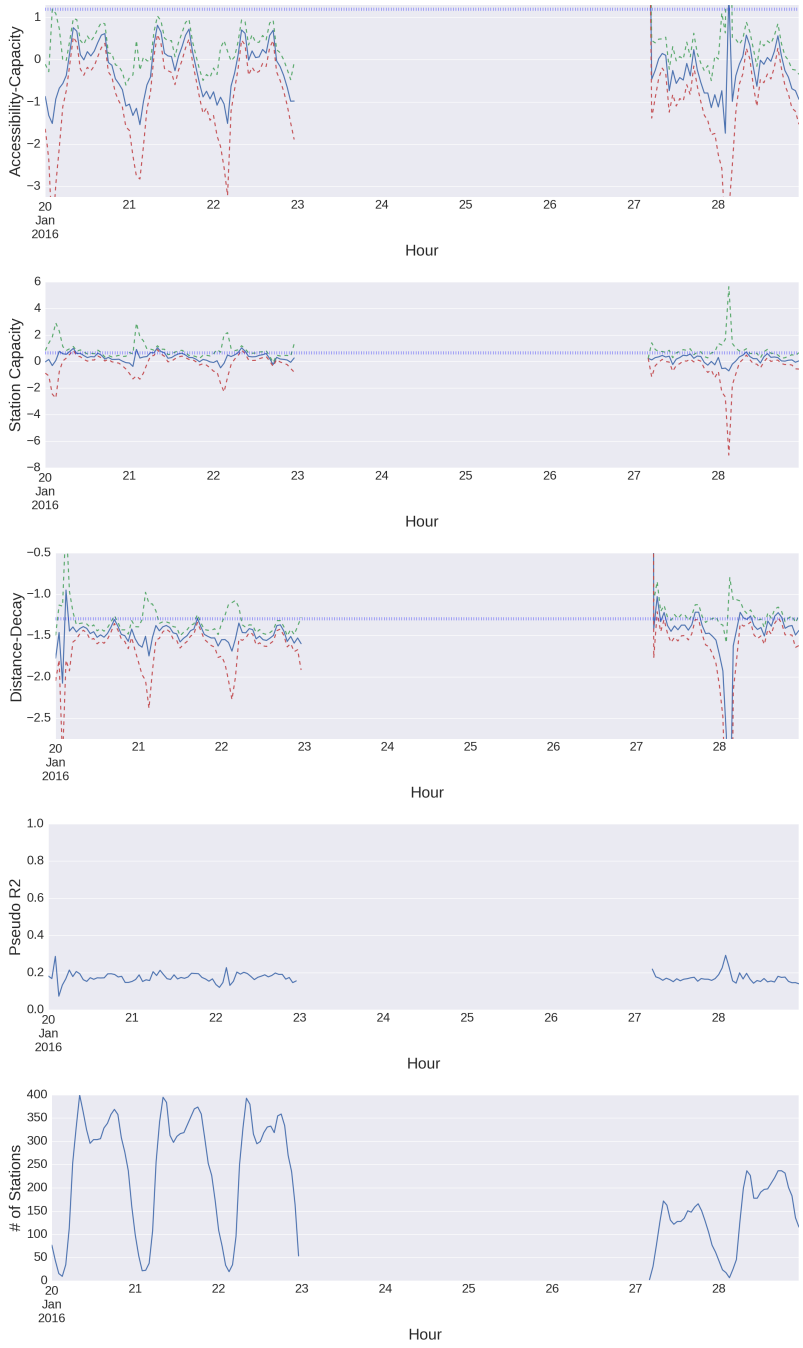


Figure 90: Hourly sampling results for the bike data using a single week representative of colder weather (January 20-29 2016), including parameter estimates, model fit, and number of origin spatial units. The gaps in the trends denote blizzard activity when the stations were temporarily closed down.

7.4 Taxi data results

7.4.1 Baseline model

As with the bike data, a baseline model is needed for the taxi flows to assist with the interpretation of the results from the temporal subset models. Here, the analysis is carried out at the tract-level, since no point-level aggregation is available for the taxi trips. The chosen specification includes five explanatory variables of destination choice – population (pop), average income (inc), the number of jobs (jobs), accessibility defined using point-of-interest density (cd), and inter-tract Euclidian distance (dist) – that are used to explain approximately 300 million taxi trips aggregated to 4,442,703 origin-destination routes between 2140 origin tracts and 2077 destination tracts [36]. Parameter estimates and model fit for a gravity model (i.e., without an accessibility variable) and a competing destinations (i.e., with accessibility variable) are presented in table 7 where it can be seen that a high pseudo R^2 of 0.9152 is achieved.

The interpretations provided by the parameter estimates in table 7 are that, *ceteris paribus*, larger flows of taxi trips are associated with destinations that: have higher populations; have higher average income; have a higher number of jobs; are more accessible to points of interest and are closer. Again, it is intuitive that individuals would want to minimize distance traveled, since longer taxi trips will cost more in terms of both time and money. The model also indicates that taxi riders are more likely to travel to destinations with higher population, that are wealthier, and have a

[36] Intra-tract trips are excluded, as well trips to tracts that did not have a reliable measure of one or more destination attributes. About 3,260,081 of the routes contain zero flows, indicating about a quarter of the routes have non-zero flow magnitudes

Table 7: Parameter estimates and model fit for the taxi data. Grav refers to the gravity-type production-constrained spatial interaction model and CD refers to the competing destination model.

	Grav		CD(poi's)	
	Estimate	SE	Estimate	SE
pop	0.3054	0.00008	0.3533	0.0008
inc	0.5602	0.00010	0.4318	0.00010
jobs	0.4492	0.00004	0.2850	0.00050
dist	-1.244	0.00006	-1.130	0.00007
cd	-	-	0.8745	0.0002
adj. R^2	0.9044		0.9152	
AIC	265154781		235233752	

higher number of jobs. Residential areas with higher populations are more attractive destinations for individuals returning from work, a shopping trip, or leisure activities. Areas with higher numbers of jobs represent commercial regions, where people travel for employment (i.e., commuting). Areas of higher income are more likely to contain individuals for whom taxi fares are less of a constraint on usage and areas with many points of interest nearby are more likely to be destinations associated with leisure and tourism. Adding this latter term (CD) slightly increases the model fit and reduces the measured effects of distance-decay, average income and the number of jobs while increasing the measured attraction of population.

7.4.2 Temporal subsets

Similarly to the experiment carried out for the bike data, models were calibrated for monthly, weekly, daily, and hourly subsets of the taxi data using the specification described in the previous section. The goals of this experiment are the same as for the

bike data, which are: (1) to assess if robust parameter estimates can be obtained for various temporal subsets of taxi data and to (2) to determine whether the resulting parameter estimates are useful indicators of temporal variations in travel behavior. Figures 91 through 96 provide parameter estimates, model fit[37], and the number of destination locations for monthly, weekly, daily, and hourly subsets of the taxi data, with separate results for daily and hourly results pertaining to warmer and cooler samples of the study period. The blue dotted horizontal line in each series indicates the associated baseline parameter estimate.

Very few parameter estimates have any noticeable uncertainty and those few that do can typically be deemed statistically insignificant because the respective confidence intervals incorporate zero. Such uncertainty occurs in two circumstances. The first circumstance is for the hourly parameter estimates during the period of blizzard activity, which can be seen in figure 96. No dashed red line is present to indicate the blizzard activity because it would obfuscate the pattern amongst the confidence intervals for parameter estimates from the 23rd to the 24th. This period relates to the end of the snowfall of the blizzard and before the city was able to effectively clear the roads. It is clear that during this time the confidence intervals generally become very large for all of the parameter estimates, most of which include both positive and negative values. The second circumstance is for the average income and accessibility parameter estimates that result from data that are sampled hourly from approximately 3:00am-5:00am, which is when there are relatively few taxi trips (figure 95). Hence, it can be seen that results for the taxi data produce exceptionally robust

[37]It is possible for the pseudo R^2 to take on values of zero if any portion of the data are ill-conditioned and this may be observed for the hourly subset results.

parameter estimates for almost all of the temporal subsets and this means that this data set is ideal for studying dynamic urban behavior.

Starting with the monthly subset results (figure 91), it can be seen that the parameter estimates for the population and jobs variables are essentially static and have very little variation compared to the baseline parameter estimate. In a large city, the population and employment opportunities would be expected to be robust over time so these results are not surprising. In contrast, distance-decay has the most variation and displays a seasonal trend where distance-decay is stronger than the baseline parameter estimate during the colder months of the year and less intense than the baseline parameter estimate during the warmer months. This trend is interesting because it might be expected that individuals prefer to travel longer distances by taxi in the winter when it is less convenient to do so by bike or on foot. However, stronger distance-decay in the colder months may be capturing the general habit of individuals to travel shorter distances when it is cold and windy outside and it could also be that in colder conditions, individuals forego walking or biking short distances in favor of taxis. Unlike the bike data, January 2016 for the taxi data produced one of the strongest distance-decay estimates, which is associated with blizzard activity and is denoted by the dashed red line. The parameter estimates associated with average income and accessibility also portray some variation, though it is difficult to explain what might cause this. In the case of average income, there is a decrease in the size of the effect for the two months of July included in the study period. In the case of accessibility, there is a slight decrease in the effect over time and this also requires further investigation. For all of the monthly results, the series are centered on the baseline parameter estimate, which indicates the baseline provides an estimate that is representative of an average effect.

Results for the monthly subsets (figure 92) are similar to the patterns described here, though they contain more short-term variation. This may be due to regular weather patterns, large-scale social events and festivals, or random fluctuations. It could also be interesting to compare these series either to weather data to see if the fluctuations are correlated to levels of precipitation, air quality or UV index or to social media data to see if they are associated with trending topics.

A bimodal pattern is present in the parameter estimates resulting from data that are subset daily (figures 93 and 94), which roughly correlates to weekdays and weekends. For the results pertaining to warmer weather (figure 93), it can be observed that on the weekends ({6th-7th, 13th-14th, 20th-21st, 27th-28th}) individuals are more likely to take a taxi to a destination that is associated with higher POI accessibility, a lower number of jobs, a lower average income, and higher population. The trends regarding POI accessibility and number of jobs indicate that more leisure trips take place on the weekends. Though this is not a groundbreaking insight, it does suggest that this methodology provides quantitative evidence and can be used to make time-specific hypothesis about different types of behavior. In contrast, average income and population indicate that during the weekends more trips are likely to terminate in residential neighborhoods and are less likely to terminate in neighborhoods with higher income. These are less obvious trends that may not have been otherwise observed. Interestingly, there seems to be an anomaly in the effects of all four destination attributes around June 16th. After a search of current events occurring in New York City on that date, three were discovered that could have produced anomalous behavior. First, Donald Trump announced his presidential candidacy in New York City that afternoon, which was quickly met with protests [38]. Second, there were two subway

[38] $https://en.wikipedia.org/wiki/Timeline_of_protests_against_Donald_Trump$

stations closures due to activity on the tracks [39]. Third, New York City bus drivers were purposely taking long pauses at crosswalks to protest a proposed new law, which created unusual traffic [40]. Curiously, the anomaly does not seem to exist when the taxi data are sampled hourly (figure 95) and it is difficult to determine which of these three events, if any, contributed towards the anomaly in the daily results. One possible way to further investigate the anomaly could be to calibrate origin-specific (i.e., local) production-constrained models for each hourly or daily subset and to map the surfaces of parameter estimates to see if the anomaly manifests spatially.

Results pertaining to daily subsets for colder weather (figure 94) are similar to those described above for the four destination attributes and support the pattern that behavior is different during the weekend days ({1st-3rd, 9th-10th, 16th-17th, 23rd-24th}) than during the weekdays. Though there are no anomalies, it is evident from figure 94 that blizzard activity causes some shift in behavior. The number of destinations at which trips terminate is at a minimum during the blizzard (i.e., just before the dashed red line) and the parameter estimate for average income is almost zero. This may be due to people in neighborhoods of all income levels wanting to get home more quickly before and during the storm.

Behavioral trends are the most pronounced in the hourly subset results for the warmer weather (figure 95) where strong hourly temporal trends are present for all of the relationships in the model and a rich narrative emerges. While the number of jobs has a larger effect on the number of trips in the morning and decreases throughout the day, population has a small effect in the morning and tends in increase throughout the

[39] $http://www.nydailynews.com/new-york/leap-front-subway-trains-queens-manhattan-article-1.2259417$

[40] $http://nypost.com/2015/06/16/nyc-bus-drivers-de-blasio-protest-brings-traffic-to-a-crawl/$

day and into the evening. This trend nicely captures morning and evening commuting behavior. Distance-decay also tends to be weakest during the morning and evening hours when people need to travel over longer distance to commute to and from work. The parameter estimate for average income is typically between 0.6 and 0.4 during the day; however, the first few hours after midnight tends to produce negative parameter estimates of approximately -0.2. This could be representative of people who work relatively lower-wage service jobs at restaurants and bars traveling home later in the evening. In addition, the positive effect of average income tapers off in the afternoon and evening during the weekend (20th and 21st), which suggests that weekend leisure trips are less constrained by financial considerations. That is, free time is valuable and it is worth the monetary costs associated with taxi transportation to get around more quickly, regardless of whether one lives in a wealthier neighborhood or not. In addition, the subways run less frequently on the weekends, which makes them a less attractive transportation option compared to taxis. Finally, POI accessibility tends to spike in the morning during the commuting rush, and then fall off briefly before increasing throughout the day when people get off work and are more likely to stop at bars, restaurants, and shops before heading home. This trend is the most intense during Friday and Saturday evening (19th and 20th) and the associated parameter estimates are likely pulling the baseline parameter estimate upward since the majority of the series occurs below the baseline parameter estimate. With only a few explanatory variables it is possible to capture a diverse array of behavior.

Though the hourly results pertaining to colder weather (figure 96) can be used to detect periods of blizzard activity (around the 24th), they cannot be used to reliably investigate associated changes in behavior. This is because during the blizzard activity the uncertainty becomes very high and most of the parameter estimates are

statistically insignificant. The cause of the uncertainty is likely related to the low number of trips during the blizzard and that the trips that do occur are associated with intricate behavior that may require an alternative baseline modeling framework to accurately describe them. It is however possible to observe that the spatial interaction system quickly returns to the same behavior that existed prior to the storm. Thus, it is possible to determine whether an external shock has a strong impact on a spatial interaction system and how long it takes for the system to return to its previous equilibrium. This knowledge could be very useful for emergency planning and disaster management.

These results for the taxi data are remarkably robust in that parameter estimates are obtained with low uncertainty for almost all of the subsets of the data. It is clear this methodology allows for a more intricate analysis of the socio-economic processes underlying taxi trips than was previously possible. Furthermore, several temporal trends in the model relationships were detected that could not have otherwise been discovered and supported with quantitative evidence. Finally, the series of parameter estimates at the daily sampling frequency proved useful for capturing changes in behavior associated with blizzard activity and for detecting unknown anomalies.

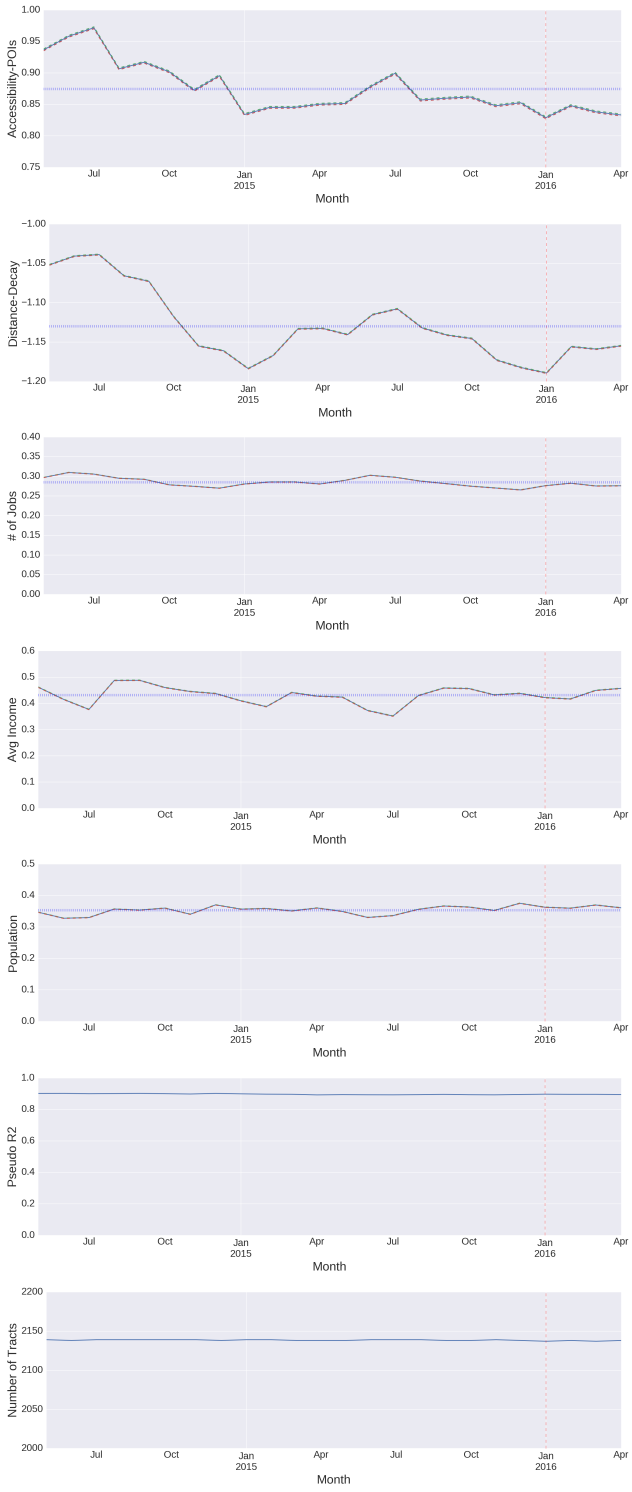


Figure 91: Monthly sampling results for the taxi data, including parameter estimates, model fit, and number of origin spatial units. Dashed red line denotes blizzard activity.

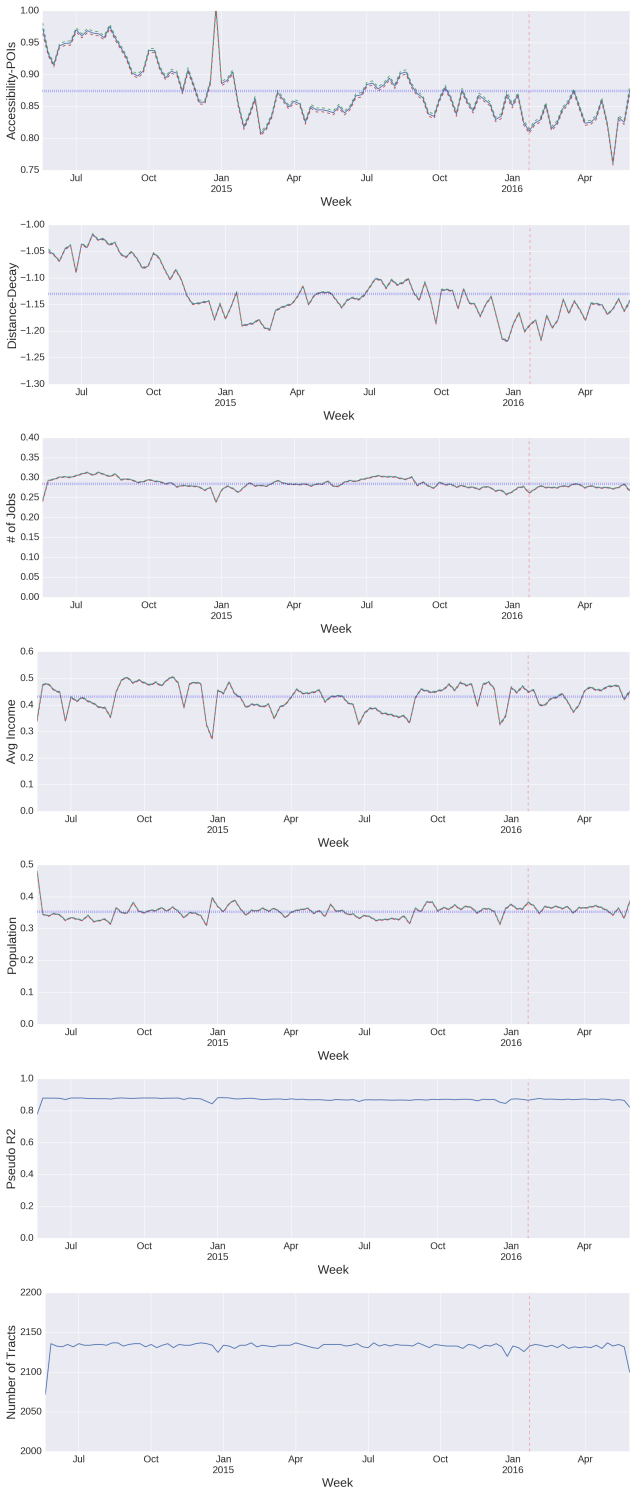


Figure 92: Weekly sampling results for the taxi data, including parameter estimates, model fit, and number of origin spatial units. Dashed red line denotes blizzard activity.

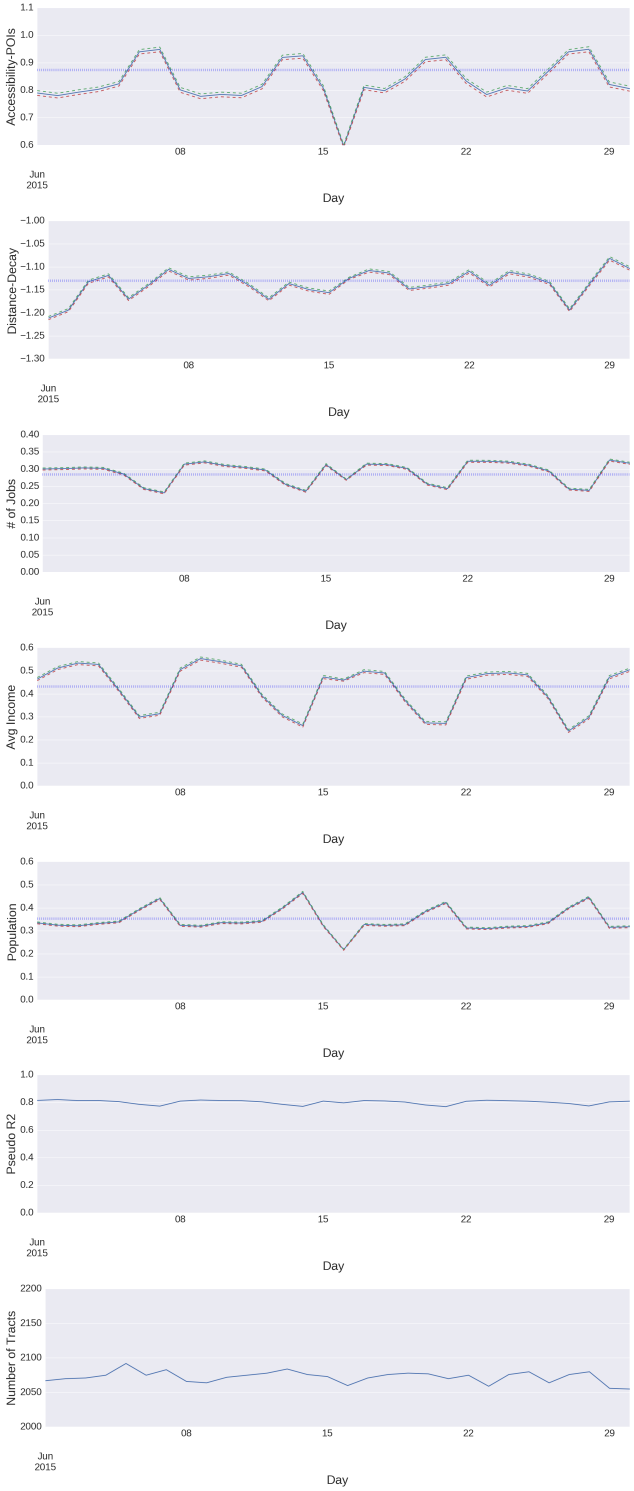


Figure 93: Daily sampling results for the taxi data using a single month representative of warmer weather (June 2015), including parameter estimates, model fit, and number of origin spatial units

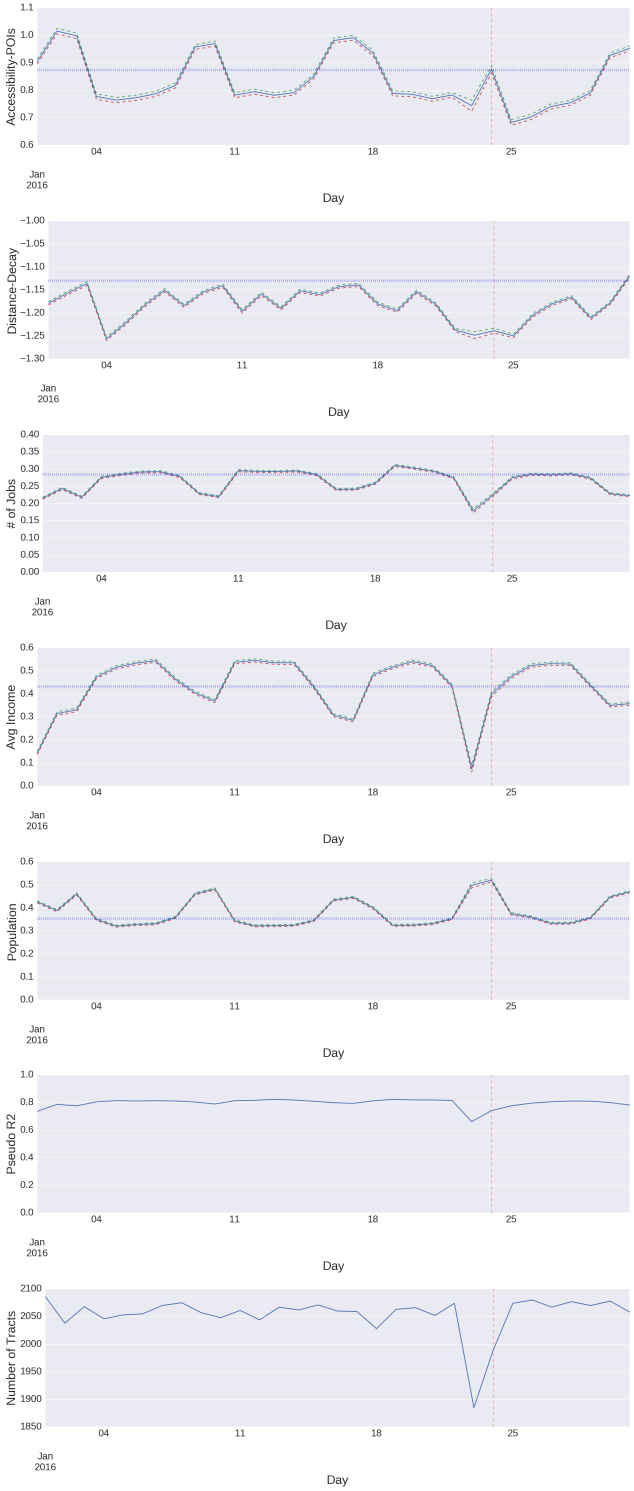


Figure 94: Daily sampling results for the taxi data using a single month representative of colder weather (January 2016), including parameter estimates, model fit, and number of origin spatial units. Dashed red line denote blizzard activity.

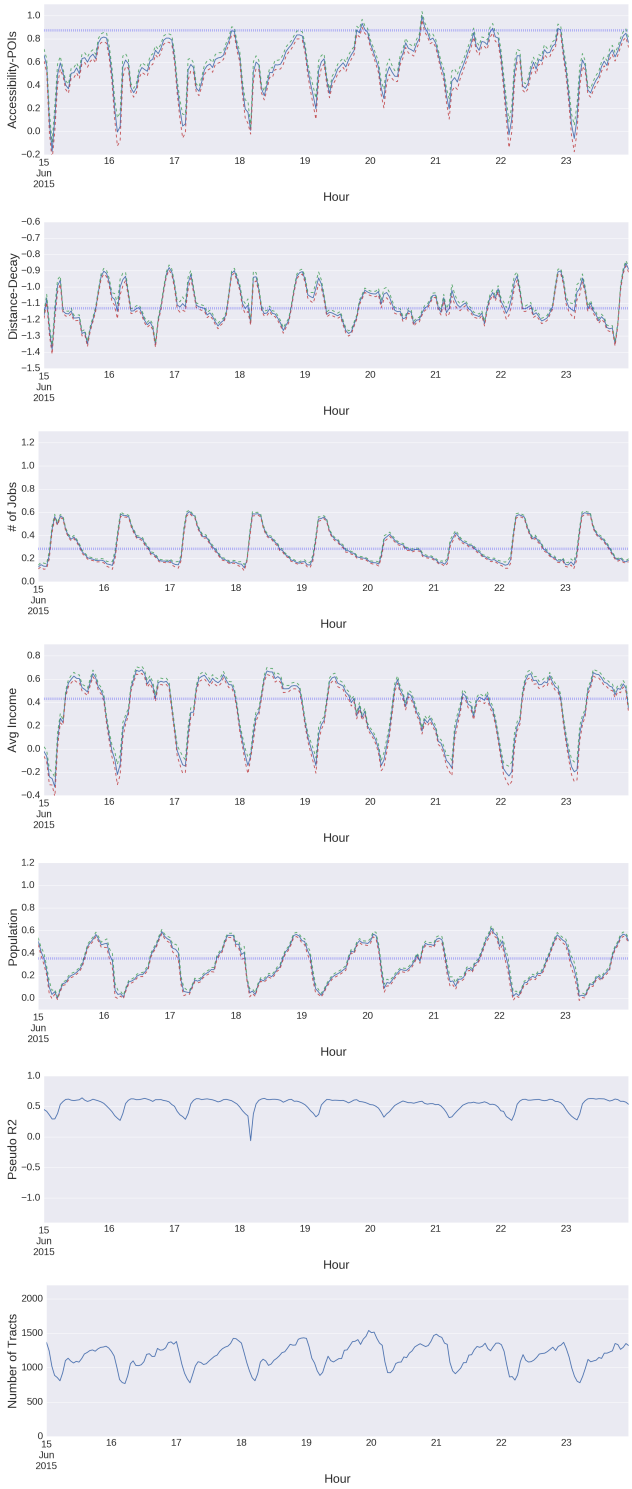


Figure 95: Hourly sampling results for the taxi data using a single week representative of warmer weather (June 15-24 2016), including parameter estimates, model fit, and number of origin spatial units.

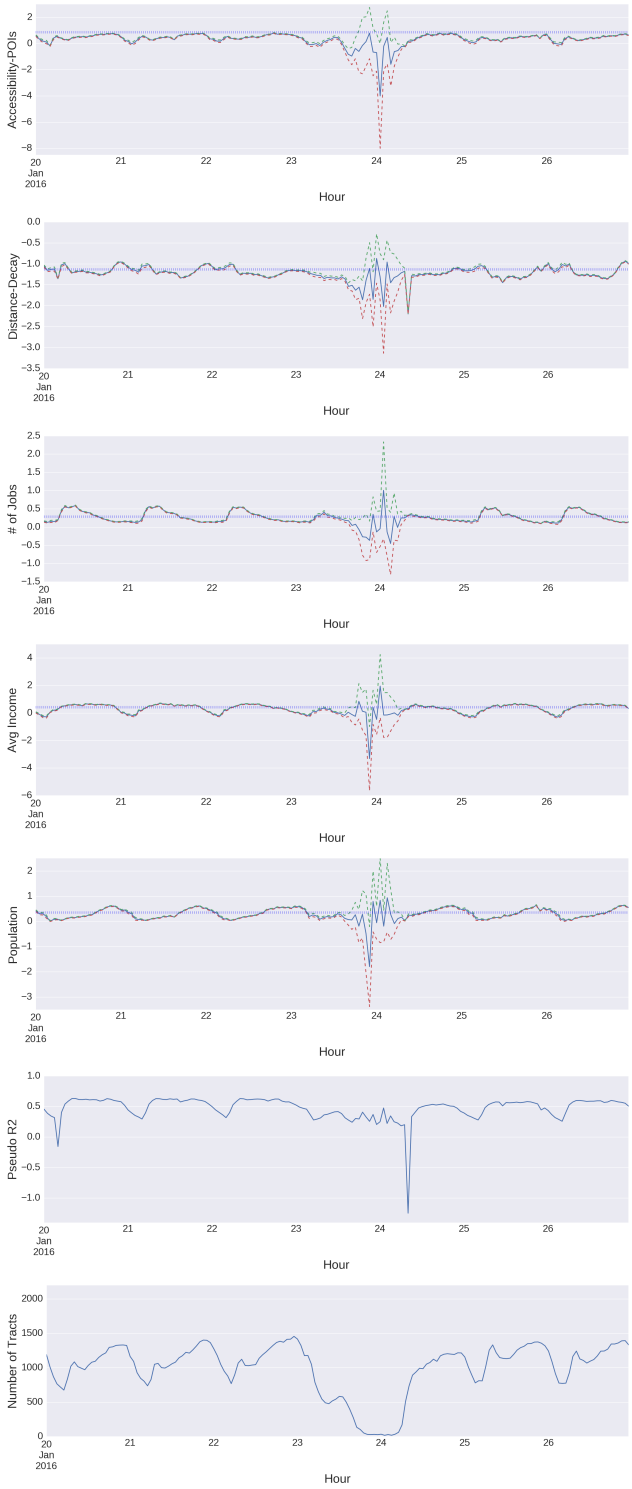


Figure 96: Hourly sampling results for the taxi data using a single week representative of colder weather (January 20-27 2016), including parameter estimates, model fit, and number of origin spatial units.

7.5 Moving Forward

Newer spatial interaction data sources, such as bike and taxi data analyzed here, allow for much finer temporal analysis of movement behavior than has previously been possible. In this chapter, temporal subset models of spatial interaction were calibrated in New York City to capture the dynamic nature of the behavioral processes underlying bike and taxi trips. It will be important to test the methodology on other data sets, time periods, and study areas. If the results obtained here can be corroborated and robust parameter estimates and dynamic behaviors can be extracted from similar data sets in other cities it may perhaps mark the dawn of a new era of spatial interaction modeling and usage.

The results presented here demonstrate that temporal subset spatial interaction models provide a new tool to analyze urban processes and illuminates how the processes determining movement patterns across a city vary over time. For instance, a global model produces one set of parameter estimates and may be thought of as representing static relationships. Some methods have been proposed that claim to provide an indication of how behavior changes when a spatial interaction system enters a new steady-state equilibrium (LeSage and Pace, 2008; LeSage and Fischer, 2014; LeSage and Thomas-Agnan, 2015) or whether a spatial interaction system is resilient to external shocks (Griffith and Chun, 2015). However, these methods were demonstrated to have several flaws and may be unnecessary. When temporal subsets of the data are available, it is possible to estimates a series of parameter estimates and directly observe how behavior changes over time and whether or not the spatial interaction system return to an equilibrium after an external shock. More specifically, the monthly and weekly parameter estimates represent the slow dynamics in how

various attributes affect travel decisions and the daily and hourly parameter estimates represent the fast dynamics within these relationships. Therefore, temporal subset models eliminate the need to make tenuous temporal assumptions in order to draw conclusions about spatial interaction dynamics.

Furthermore, for this type of analysis, there seems to be a general tradeoff between parameter estimate uncertainty and detail with which dynamic processes can be captured. However, this tradeoff is less important for the taxi trips, which occur more frequently and more evenly over the course of a typical day. Therefore, the methodology is promising for studying urban behavior and will become more valuable as more data such as those for taxi trips becomes available. It will also be interesting to link these results with weather data and social media data, as well as combine them with other methodologies to further classify and disaggregate flow by different trip purposes.

In the previous chapter, the potential and pitfalls associated with the increased volume and variety of spatial interaction data were evaluated. This theme was continued here by evaluating the potential and pitfalls associated with the increased velocity of spatial interaction data, and it was determined that the increased velocity provides immense opportunity to increase our knowledge about human behavior. In the subsequent and final chapter of this research, the potential and pitfalls discovered in each chapter are further discussed and some suggestions for further research are proposed.

Chapter 8

DISCUSSION AND CONCLUSION

8.1 Introduction

Numerous research objectives were addressed in this dissertation. First, spatial interaction model specifications that account for spatial structure effects were compared and contrasted. Second, empirical spatial interaction models were developed using traditional and non-traditional data sources within New York City and the potential and pitfalls associated with newer data sources were evaluated. Third, the spatial and temporal dynamics of spatial interaction processes were explored using local models and temporal subset models. The findings from the research will hopefully improve the way that spatial interaction data and the urban behavior represented within them are modeled. To summarize the merit and potential impact of these findings, the outcomes from each section of the research are highlighted, some suggestions for future directions are provided, and several general conclusions are offered.

8.2 Research findings

8.2.1 Spatial structure effects

In chapter 5, several simulation experiments were undertaken in order to compare three spatial interaction specifications that were reviewed in chapter 2 and to evaluate their ability to account for spatial structure effects. This was achieved by simulating

flows between sets of points that contain different degrees of spatial clustering using the data-generating processes associated with each of the three specifications. Then each specification was used to calibrate a model on each of the simulated data sets. This facilitated the investigation of how each model specification performed when it was calibrated on data containing: (i) the expected spatial structure effects; (ii) no spatial structure effects; and (iii) unexpected spatial structure effects. Several important results were obtained.

First, it was demonstrated that the effects due to the spatial clustering of locations and those associated with the concept of spatial autocorrelation are different effects, even if they may potentially be inter-related. Since none of the location-based explanatory variables were spatially autocorrelated, it was possible to evaluate the ability of each specification to capture effects associated solely with spatial clustering. The competing destinations models proved to be the most effective in this regard. This model also had the attractive feature that it did not indicate the presence of spatial effects when none were present.

Second, the spatial autoregressive spatial interaction model produced results that indicate it is highly susceptible to falsely identifying the presence of spatial effects. It indicated the presence of spatial effects even when it was calibrated on data generated from a null gravity model where there were no spatial structure effects. It was surmised that this is because the autoregressive component seems to be competing for the same variation accounted for by the distance variable in the model and a further finding was that this cannot be remedied using the scalar summary measures proposed by LeSage and Thomas-Agnan (2015), which only apply to the parameter estimates associated with the origin and destination explanatory variables. Therefore using the spatial

autoregressive specification complicates the interpretability of the distance-decay parameter estimate.

An additional experiment was carried out that aggregated the flows simulated between points to flows between areal units. The purpose this experiment was to assess the ability of the specifications to capture spatial structure effects when the data are aggregated. It provided evidence that aggregated data may be less reliable for capturing spatial structure effects and should therefore be avoided whenever possible.

8.2.2 Local models of spatial interaction

Local spatial interactions models were calibrated using both traditional and non-traditional datasets in chapter 6 with three primary goals in mind. First was to investigate any spatial heterogeneities (i.e., spatial non-stationarity) within the model relationships that represent movement processes. The results indicate that there is spatial heterogeneity and that it may be responsible for unstable global parameter estimates that sometimes change between positive and negative values (i.e., flipped sign) when additional explanatory variables are added or removed from the model. By observing surfaces of local parameter estimates it was discovered that only certain regions would actually change or become statistically insignificant. Since the global parameter estimates are a weighted average of the local parameter estimates, it is possible for changes in only a portion of study area to severely affect the global parameter estimate. Thus, when flipped signs are observed for global parameter estimates, it may indicate that a relationship changed for only a portion of the study area, rather than a change for the entire study area. When increasingly finer spatial units are employed for local spatial interaction models, a more detailed parameter

estimate surface is obtained and can be used to investigate spatial non-stationarity. Therefore, a potential of non-traditional data sets is that:

- increased spatial resolution permits data to be aggregated to increasingly finer spatial units and allows for more detailed surfaces of local parameter estimates to be estimated.

The second purpose of the research carried out in chapter 6 was to determine whether or not bike and taxi trips in New York City are practical for studying commuting behavior in comparison to traditional data from the US census. This was done by establishing a baseline model using the census commute-to-work data and then calibrating the same model on the bike and taxi data. Many of the parameter estimates from the models for the bike and taxi data were different from those obtained for the model using the census data. A likely reason for this is that:

- bike and taxi trips origins and destinations may not be representative of the residences and workplaces of commuting trips.

Another potential cause of the different parameter estimates observed between census commuting data and the bike and taxi data is that:

- automatically recorded data may not be representative of a single type of trip and do not provide sufficient information to distinguish between different types of trips.

These two pitfalls suggest that the bike and taxi data are not ideal for studying the behavior of the general commuting population. However, once can reserve this criticism because:

- bike and taxi data are each representative of a single mode of transportation, which makes them more useful for studying these modes than some traditional sources of data.

Finally, the third objective of chapter 6 was to explore the usefulness of new data sources as explanatory variables in spatial interaction models. Accessibility was defined using point-of-interest density and was shown to capture different spatial structure effects than traditional variables. Furthermore, accessibility was measured with higher precision using disaggregated inter-location distances. Thus:

- the variety of newer data sources and their increased spatial resolution allow a more diverse and more precise set of explanatory variables to be defined.

8.2.3 Temporal subset models of spatial interaction

Some additional potential and pitfalls were defined in chapter 7 using increasingly finer temporal units to define subsets of the bike and taxi data. Remarkably, calibrating models on the subsets resulted in parameter estimates with low uncertainty for monthly, weekly, daily, and even many hourly subsets. Only in the case of the hourly subsets with a low number of trips (i.e., late night periods) and during a blizzard did the confidence intervals for the parameter estimates become very large and include zero. Therefore, a pitfall of newer data sets is that:

- disaggregating the data to increasingly finer temporal units for use in spatial interaction models may be limited if sample size becomes sparse.

However, sparse sample size was rarely problematic in this research and the parameter estimates from calibrating spatial interaction models on the temporal subsets provided

an effective tool to study the dynamic behavior of cyclists and taxi patrons across New York City. It was possible to develop a rich narrative about the spatial decisions of individuals and how they vary annually, throughout the week, and over the course of each day. The series of parameter estimates provided quantitative evidence that confirmed several expected spatial-temporal relationships and also facilitated the discovery of new relationships. As a result, it was illustrated that a major potential of bigger spatial interaction data sets is that:

- finer temporal resolution in new spatial interaction data sets permits a more nuanced analysis of behavior than has previously been possible giving great potential to uncover much more detailed information of the dynamics regarding human movement behavior.

Furthermore, the series of parameter estimates obtained from calibrating spatial interaction models on the temporal subsets were useful for establishing a set of baseline equilibrium behavior associated with urban activity in New York City. It was then possible to detect periods where behavior deviated from the expected equilibrium. For example, the series of parameter estimates were used to identify the period of blizzard activity, as well as how quickly the system returned to equilibrium after this extreme weather event. This indicates that a further potential of newer data sources is that:

- finer temporal resolution enables the detection of anomalous behavior and the impacts of external shocks to the spatial interaction system.

8.3 Future directions

8.3.1 Spatial structure effects

The simulations used in chapter 5 are useful for generating new insights into spatial structure effects in spatial interaction modeling, a problem that has been around for almost 40 years. However, some additional experiments will further aid the process of selecting an appropriate spatial interaction specification. For example, it would be interesting to carry out a similar experiment to those in chapter 5 using only the uniformly distributed locations (i.e., no spatial clustering), but with varying levels of spatial autocorrelation for the origin and destination variables. This would isolate the effects of spatial autocorrelation and could more clearly demonstrate the differences between the effects associated with spatial clustering and spatial autocorrelation. It could also be insightful to examine different types of autoregressive effects. Here, only a cross-product of origin and destination effects was included and it is possible to instead include only a destination effect or multiple effects defined separately for origins and destinations. It is perhaps possible that an alternative conceptualization of the autoregressive components may be more effective; however, there is little-to-no theoretical basis for these different conceptualizations. Finally, the simulation experiment could also be extended by employing irregular areal units instead of uniform grids to perform the aggregation of the data. This would be more representative of reality and could help substantiate the findings found here.

8.3.2 Local models of spatial interaction

Local models were used in chapter 6 to explore the nature of non-stationarity in parameter estimates from spatial interactions models. This was carried out using destination-specific attraction-constrained models that each use only trips that begin at a single location and terminate at all other locations, which provides a natural way to create mutually exclusive subsets of the data. It could also be possible to employ a geographically weighted framework for calibrating local models at any number of locations in the study area (Kordi and Fotheringham, 2016). The advantage of this framework is that the locations used for creating local samples of the data are not dependent upon pre-defined areal units and can occur with higher density in areas where it is suspected that there may be finer resolution spatial heterogeneity in the processes.

Multicollinearity was encountered when trying to include several distance-weighted variables in some of the models from chapter 6. This was problematic because the competing destination accessibility term and some novel point-level explanatory variables both leverage the distance-weighting technique in order to be used in the models. Furthermore, results from experimenting with accessibility terms defined using different explanatory variables indicated that it might be possible to capture different aspects of urban spatial structure using different explanatory variables. Therefore, a multi-scale distance-weighting algorithm, such as the one used within multi-scale geographically weighted regression (Fotheringham *et al.*, 2017) could be useful for creating a compound accessibility term where each explanatory variable included in the model is distance-weighted using a uniquely defined transformation. This could

allow for an accessibility term that captures diverse spatial structure and that may reduce multicollinearity when distance-weighting several explanatory variables.

Though many new variables were introduced for use in spatial interaction models in this research, there are still many more that can be explored. It is useful to continuously harvest new data sets, creatively define novel explanatory variables, and assess their usefulness in spatial interaction models. The outcome of this process will undoubtedly produce models that capture more specific relationships and ultimately provide new insights.

8.3.3 Temporal subset models of spatial interaction

Several extensions for the temporal subset methodology presented in chapter 7 are possible. First, the robustness of the parameter estimates obtained from spatial interaction models calibrated on monthly, weekly, daily, and hourly subsets of bike and taxi data are very encouraging and suggest that it may be possible to calibrate models on even finer temporal subsets. In particular, the results for the taxi data were extraordinarily robust and it would be interesting to investigate if it is possible to reliably obtain parameter estimates for half-hour or fifteen-minute intervals during periods of peak usage. This would produce an even more detailed series of parameter estimates for characterizing dynamic urban behavior.

Another useful extension would be to link the series of parameter estimates with other temporal data, such as weather data and social media data. It might be possible to use these auxiliary data sources to define relationships amongst some of the irregular variation in the parameter estimate series that is not apparent in the annual, weekly, and daily temporal cycles. For instance, irregular spikes and valleys in the parameter

estimate series might be related to high levels of precipitation or could be associated with large-scale cultural events such as 'restaurant week', where many restaurants offer discounts and show off their flagship menu items. The former could be expected to deter leisure trips while the latter might promote them.

It is perhaps most important to extend the research presented here by applying the temporal subset methodology to other types of spatial interaction data and for different study areas. If similar conclusions can be made to those presented here then this will substantiate the usefulness of the temporal subset methodology. Moreover, this would help construct a more general understanding of dynamic urban behavior that includes more places and types of data. It could also be beneficial to document further instances of extreme weather and other significant events that can be considered external shocks to urban systems. By utilizing the temporal subset method over periods that include these events it would be possible to catalogue whether or not each type of event causes disequilibrium (i.e., changes in model relationships), and how long it takes the respective spatial interaction system to return to equilibrium. Such knowledge would be extremely valuable to municipal agencies involved in both long-term planning and emergency response efforts.

8.4 Conclusions

Within this dissertation, a comparison of several spatial interaction model specifications were made and the potential and pitfalls of new sources of 'big' spatial interaction data were investigated. Based on the outcomes discussed above, several general conclusions can be made.

Concerning spatial interaction specifications, future spatial interaction research

needs to be more specific about *which* spatial effects are being addressed, including how each effect arises, the consequences of failing to account for them, and the outcome and interpretation that is expected when a particular method is used to account for them. In this work, it was shown that the effects that arise from spatial clustering and spatial autocorrelation are not identical. In addition, the spatial interaction literature is full of instances where a method designed for areal or point data is applied to spatial interaction data (i.e., eigenvector spatial filtering methods or spatial autoregressive methods) rather than considering the unique characteristics of spatial interaction data and the processes they represent. Therefore, it is not enough to simply indicate that some spatial effects exist and that an advanced method was used to account for them. To truly advance the current state of knowledge, it is of the utmost importance to be specific about why a method is used and how using it provides a better understanding of spatial processes.

Both potential and pitfalls were identified for the various newer data sources employed in this research. Though there are some pitfalls, there is a tremendous amount of potential to diversify our understanding of human behavior using these new data sets within a spatial interaction modeling framework. The focus here was on relationships between movement data and other external factors, which is unlike much of the recent scholarship on new 'big' movement data that tends to be descriptive. It was shown that spatially local and temporal subset models are effective tools for identifying the relationships that are representative of human behavior. Furthermore, in the case of the bike data, it was demonstrated that new forms of spatial interaction data should be modeled based on the processes that they are most closely associated with (i.e., cycling between stations), since they may not be representative of the processes that are classically studied via spatial interaction models. It is tempting to

use new sources of spatial interaction data within a model formulated for traditional spatial interaction data (i.e., commuting between census units), though the results of such efforts may not be representative of reality.

Finally, new data sources allow us to detect useful temporal dynamics in spatial interaction data by employing temporal subset spatial interaction models. These dynamics capture the diversity of human activities that are present within urban environments and cannot be detected using traditional spatial interaction data sources. Furthermore, the temporal dynamics identified in this research were sensitive to the various temporal resolutions that were utilized. Different dynamics emerge for different resolutions, which provides a rich opportunity to learn what drives location choice preferences during different periods of time.

8.5 Closing remarks

In this final chapter, the main findings were highlighted, some future avenues of research were discussed, and several conclusions were made. It is evident from this body of work that new sources of spatial interaction data are very valuable and that the current increase in their availability provides the occasion to establish new standards for knowledge production. Throughout this dissertation several efforts have been made toward this cause within the context of spatial interaction modeling. Various novel insights were generated and some methodological innovations were proposed and validated. Although there is still much work to do, the research presented here provides the foundation for a new era of spatial interaction modeling.

NOTES

REFERENCES

Abdmoulah, W., "Arab Trade Integration: Evidence from Zero-Inflated Negative Binomial Model", Journal of Economic Cooperation & Development **32**, 2, 39–65, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/1114068484?pq-origsite=summon (2011).

Afandizadeh, S. and S. M. Y. Hamedani, "A fuzzy intervening opportunity model to predict home-based shopping trips", Canadian Journal of Civil Engineering **39**, 2, 203–222, URL http://www.nrcresearchpress.com/doi/abs/10.1139/l11-097 (2012).

Akaike, H., "A new look at the statistical model identification", Automatic Control, IEEE Transactions on **19**, 6, 716–723, URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1100705 (1974).

Akwawua, S. and J. A. Pooler, "The development of an intervening opportunities model with spatial dominance effects", Journal of Geographical Systems **3**, 1, 69–86, URL http://link.springer.com.ezproxy1.lib.asu.edu/article/10.1007/PL00011468 (2001).

Anas, A., "Discrete choice theory, information-theory and the multinomial logit and gravity models", Transportation Research Part B: Methodological **17**, 1, 13–23 (1983).

Anselin, L., "Spatial Econometrics", in "Palgrave Handbook of Econometrics", pp. 901–969 (Palgrave Macmillan, 2006).

Anselin, L. and S. J. Rey, *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL* (GeoDa Press LLC, Chicago, IL, 2014).

Arribas-Bel, D., "Accidental, open and everywhere: Emerging data sources for the understanding of cities", Applied Geography **49**, 45–53, URL http://www.sciencedirect.com/science/article/pii/S0143622813002178 (2014).

Arvis, J.-F. and B. Shepherd, "The Poisson quasi-maximum likelihood estimator: a solution to the "adding up" problem in gravity models", Applied Economics Letters **20**, 6, 515–519, URL http://www.tandfonline.com/doi/abs/10.1080/13504851.2012.718052 (2013).

Bachand-Marleau, J., B. H. Y. Lee and A. M. El-Geneidy, "Better Understanding of Factors Influencing Likelihood of Using Shared Bicycle Systems and Frequency of Use", Transportation Research Record: Journal of the Transportation Research Board **2314**, -1, 66–71, URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2314-09 (2012).

Barabasi, A.-L. and R. Albert, "Emergence of scaling in random networks", Science **286**, 5439, 509–512, URL http://arxiv.org/abs/cond-mat/9910332, arXiv: cond-mat/9910332 (1999).

Barchiesi, D., H. S. Moat, C. Alis, S. Bishop and T. Preis, "Quantifying International Travel Flows Using Flickr", PLoS ONE **10**, 7, e0128470, URL http://dx.doi.org/10.1371/journal.pone.0128470 (2015).

Batty, M., "Resilient cities, networks, and disruption", Environment and Planning B: Planning and Design **40**, 4, 571–573, URL http://www.envplan.com/abstract.cgi?id=b4004ed (2013).

Batty, M., "Data About Cities: Redefining Big, Recasting Small", Working Paper 203, UCL, London (2015).

Baxter, M., "Model Misspecification and Spatial Structure in Spatial-Interaction Models", Environemnt and Planning A **15**, 3, 319–327 (1983).

Baxter, M., "Misspecification in spatial interaction models: further results", Environment and Planning A **17**, 5, 673–678 (1985).

Bazzani, A., B. Giorgini, S. Rambaldi, R. Gallotti and L. Giovannini, "Statistical laws in urban mobility from microscopic GPS data in the area of Florence", Journal of Statistical Mechanics: Theory and Experiment **2010**, 05, P05001, URL http://iopscience.iop.org/1742-5468/2010/05/P05001 (2010).

Beecham, R. and J. Wood, "Exploring gendered cycling behaviours within a large-scale behavioural data-set", Transportation Planning and Technology **37**, 1, 83–97, URL http://dx.doi.org/10.1080/03081060.2013.844903 (2013).

Beecham, R., J. Wood and A. Bowerman, "Studying commuting behaviours using collaborative visual analytics", Computers, Environment and Urban Systems **47**, 5–15, URL http://www.sciencedirect.com/science/article/pii/S0198971513001014 (2014).

Berglund, S. and A. Karlström, "Identifying local spatial association in flow data", Journal of Geographical Systems **1**, 3, 219–236, URL http://link.springer.com/article/10.1007/s101090050013 (1999).

Bernardin, V. L., F. Koppelman and D. Boyce, "Enhanced Destination Choice Models Incorporating Agglomeration Related to Trip Chaining While Controlling for Spatial Competition", Transportation Research Record: Journal of the Transportation Research Board **2132**, -1, 143–151, URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2132-16 (2009).

Bernasco, W., "Modeling Micro-Level Crime Location Choice: Application of the Discrete Choice Framework to Crime at Places", Journal of Quantitative Criminology **26**, 1, 113–138, URL http://link.springer.com.ezproxy1.lib.asu.edu/article/10.1007/s10940-009-9086-6 (2010).

Birkin, M., G. Clarke and M. Clarke, "Refining and Operationalizing Entropy-Maximizing Models for Business Applications", Geographical Analysis **42**, 4, 422–445, URL http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=54473795&site=ehost-live (2010).

Birkin, M. and A. Heppenstall, "Extending Spatial Interaction Models with Agents for Understanding Relationships in a Dynamic Retail Market", Urban Studies Research **2011**, 1–12, URL http://www.hindawi.com/journals/usr/2011/403969/ (2011).

Bivand, R., J. Hauke and T. Kossowski, "Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods: Computing the Jacobian in Spatial Autoregressive Models", Geographical Analysis **45**, 2, 150–179, URL http://doi.wiley.com/10.1111/gean.12008 (2013).

Black, F., "The Trouble with Econometric Models", Financial Analusts Journal **38**, 2, 29–37 (1982).

Black, W. R., "Network autocorrelation in transport network and flow systems", Geographical Analysis **24**, 3, 207–222, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1992.tb00262.x/abstract (1992).

Black, W. R., "Spatial interaction modeling using artificial neural networks", Journal of Transport Geography **3**, 3, 159–166, URL http://www.sciencedirect.com/science/article/pii/096669239500013S (1995).

Black, W. R. and I. M. Thomas, "Accidents on Belgium's Motorways: A Network Autocorrelation Analysis", Journal of Transport Geography **6**, 1, 23–31 (1998).

Blanchet, F. G., P. Legendre and D. Borcard, "Modelling directional spatial processes in ecological data", Ecological Modelling **215**, 4, 325–336, URL http://linkinghub.elsevier.com/retrieve/pii/S0304380008001798 (2008).

Bolduc, D., M. J. GAUDRY and M. G. DAGENAI, "Spatially Autocorrelated Errors In Origin-Destination Models: A New Specification Applied To Aggregate Mode Choice", Transportation Research Part B: Methodological **23**, 5, 361–372, URL http://www.academia.edu/download/41726058/SPATIALLY_AUTOCORRELATED_ERRORS_IN_ORIGI20160129-1725-1yhvnq.pdf (1989).

Bolduc, D., R. Laferriere and G. Santarossa, "Spatial autoregressive error components in travel flow models: An application to aggregate mode choice", in "New Directions in Spatial Econometrics", pp. 96–108 (Springer, 1995).

Bolduc, D., R. Laferrière and G. Santarossa, "Spatial autoregressive error components in travel flow models", Regional Science and Urban Economics **22**, 3, 371–385, URL http://www.sciencedirect.com/science/article/pii/016604629290035Y (1992).

Boots, B., "A variance-stabilizing coding scheme for spatial link matrices^", Environment and Planning A **31**, 165–180, URL http://www.envplan.com/epa/fulltext/a31/a310165.pdf (1999).

Boots, B., N. and P. S. Kanaroglou, "Incorporating the effects of spatial structure in discrete choice models of migration", Journal of Regional Science **28**, 4, 495–509 (1988).

Borcard, D. and P. Legendre, "All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices", Ecological Modelling **153**, 1, 51–68, URL http://www.sciencedirect.com/science/article/pii/S0304380001005014 (2002).

Borcard, D., P. Legendre, C. Avois-Jacquet and H. Tuomisto, "Dissecting the spatial structure of ecological data at multiple scales", Ecology **85**, 7, 1826–1832, URL http://www.esajournals.org/doi/abs/10.1890/03-3111 (2004).

Borgnat, P., E. Fleury, C. Robardet, A. Scherrer and others, "Spatial analysis of dynamic movements of Vélo'v, Lyon's shared bicycle program", in "European Conference on Complex Systems 2009", (2009), URL http://hal.inria.fr/ensl-00408150/.

Borgnat, P., C. Robardet, J.-B. Rouquier, P. Abry, E. Fleury and P. Flandrin, "Shared Bicycles In A City: A Signal Processing and Data Analysis Perspective", Advances in Complex Systems **10**, 0 (2010).

Boyandin, I., E. Bertini, P. Bak and D. Lalanne, "Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data", Computer Graphics Forum **30**, 3, 971–980, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1467-8659.2011.01946.x/abstract (2011).

Brandsma, A. and R. H. Ketellapper, "A bivariate approach to spatial autocorrelation", Environemnt and Planning A **11**, 1, 51–58 (1979).

Brockmann, D., L. Hufnagel and T. Geisel, "The scaling laws of human travel", Nature **439**, 7075, 462–465, URL http://www.nature.com/nature/journal/v439/n7075/full/nature04292.html (2006).

Burger, M., F. Van Oort and G.-J. Linders, "On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation", Spatial Economic Analysis **4**, 2, 167–190, URL http://www.tandfonline.com/doi/abs/10.1080/17421770902834327 (2009).

Caroll, R. J., D. Ruppert, L. A. Stefanski and C. M. Crainiceanu, *Measurement Error in Nonlinear Models*, Monographs on Statistics and Probability 105 (Chapman & Hall/CRC, New York, 2006), 2nd edition edn.

Cascetta, E., F. Pagliara and A. Papola, "Alternative approaches to trip distribution modelling: A retrospective review and suggestions for combining different approaches", Papers in Regional Science **86**, 4, 597–620, URL http://doi.wiley.com/10.1111/j.1435-5957.2007.00135.x (2007).

Charles-Edwards, E., T. Wilson and N. Sander, "Visualizing Australian internal and international migration flows", Regional Studies, Regional Science **2**, 1, 431–433, URL http://dx.doi.org/10.1080/21681376.2015.1066267 (2015).

Chu, D., D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng and G. Chen, "Visualizing Hidden Themes of Taxi Movement with Semantic Transformation", in "Pacific Visualization Symposium (PacificVis), 2014 IEEE", pp. 137–144 (IEEE, 2014), URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6787160.

Chun, Y., "Modeling network autocorrelation within migration flows by eigenvector spatial filtering", Journal of Geographical Systems **10**, 4, 317–344, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/289806945?pq-origsite=summon (2008).

Chun, Y. and D. A. Griffith, "Modeling Network Autocorrelation in Space–Time Migration Flow Data: An Eigenvector Spatial Filtering Approach", Annals of the Association of American Geographers **101**, 3, 523–536, URL http://www.tandfonline.com/doi/abs/10.1080/00045608.2011.561070 (2011).

Chun, Y., D. A. Griffith, M. Lee and P. Sinha, "Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters", Journal of Geographical Systems **18**, 1, 67–85, URL http://link.springer.com/10.1007/s10109-015-0225-3 (2016).

Chun, Y., H. Kim and C. Kim, "Modeling interregional commodity flows with incorporating network autocorrelation in spatial interaction models: An application of the US interstate commodity flows", Computers, Environment and Urban Systems **36**, 6, 583–591, URL http://linkinghub.elsevier.com/retrieve/pii/S0198971512000373 (2012).

Claeson, C.-F., "Zone Preference in Intraregional Population Movement: Sounding into a Migrant Population on a Coordinate Basis", Geografiska Annaler. Series B, Human Geography **50**, 2, 133, URL http://www.jstor.org/stable/491013?origin=crossref (1968).

Claeson, C.-F., "A Two-Stage Model of In-Migration to Urban Centres: Deductive Development of a Variant of the Gravity Formulation", Geografiska Annaler. Series B, Human Geography **51**, 2, 127, URL http://www.jstor.org/stable/490539?origin=crossref (1969).

Clarke, G., R. Langley and W. Cardwell, "Empirical applications of dynamic spatial interaction models", Computers, Environment and Urban Systems **22**, 2, 157–184, URL http://www.sciencedirect.com/science/article/pii/S0198971598000210 (1998).

Cliff, A., R. Martin and J. Ord, "Evaluating the friction of distance parameter in gravity models", Regional Studies **8**, 281–286 (1974).

Cliff, A., R. Martin and J. Ord, "Map pattern and friction of distance parameters: reply to comments by R. J. Johnston, and by L. Curry, D. A. Griffith and E. S. Sheppard", Regional Studies **9**, 3, 285–288, URL http://www.tandfonline.com/doi/abs/10.1080/09595237500185301 (1975).

Cliff, A., R. Martin and J. Ord, "A reply to the final comment", Regional Studies **10**, 3, 341–342, URL http://www.tandfonline.com/doi/abs/10.1080/09595237600185351 (1976).

Corrado, L. and B. Fingleton, "Where Is The Economics In Spatial Econometrics?", Journal of Regional Science **52**, 2, 210–239, URL http://doi.wiley.com/10.1111/j.1467-9787.2011.00726.x (2012).

Curry, L., "A spatial analysis of gravity flows", Regional Studies **6**, 2, 131–147, URL http://dx.doi.org/10.1080/09595237200185141 (1972).

Curry, L., D. A. Griffith and E. S. Sheppard, "Those gravity parameters again", Regional Studies **9**, 3, 289–296, URL http://www.tandfonline.com/doi/abs/10.1080/09595237500185311 (1975).

Curtis, A. and A. S. Fotheringham, "Large-scale information surfaces: an analysis of city-name recalls in the United States", Geoforum **26**, 1, 75–87, URL http://www.sciencedirect.com/science/article/pii/001671859500014C (1995).

Daley, M. and C. Rissel, "Perspectives and images of cycling as a barrier or facilitator of cycling", Transport Policy **18**, 1, 211–216, URL http://linkinghub.elsevier.com/retrieve/pii/S0967070X10000995 (2011).

de la Mata, T. and C. Llano, "Social networks and trade of services: modelling interregional flows with spatial and network autocorrelation effects", Journal of Geographical Systems **15**, 3, 319–367, URL http://link.springer.com/10.1007/s10109-013-0183-6 (2013).

Diebold, F. X., "On the Origin (s) and Development of the Term'Big Data"', PIER Working Paper PIER Working Paper 12-037, University of Pennsylvania, Philadelphia, URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2152421 (2012).

Dison, D., W. and C. Hale, W., "Gravity versus Intervening Opportunity Models in Explanation of Spatial Trade Flows", Growth and Change **8**, 4, 15–22 (1977).

Do, T. M. T. and D. Gatica-Perez, "Where and what: Using smartphones to predict next locations and applications in daily life", Pervasive and Mobile Computing **12**, 79–91, URL http://www.sciencedirect.com/science/article/pii/S1574119213000576 (2014).

Dray, S., P. Legendre and P. R. Peres-Neto, "Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM)", Ecological Modelling **196**, 3-4, 483–493, URL http://linkinghub.elsevier.com/retrieve/pii/S0304380006000925 (2006).

Elffers, H., D. Reynald, M. Averdijk, W. Bernasco and R. Block, "Modelling Crime Flow between Neighbourhoods in Terms of Distance and of Intervening Opportunities", Crime Prevention and Community Safety: An International Journal **10**, 2, 85–96, URL http://www.palgrave-journals.com/doifinder/10.1057/palgrave.cpcs.8150062 (2008).

Faghih-Imani, A., N. Eluru, A. M. El-Geneidy, M. Rabbat and U. Haq, "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal", Journal of Transport Geography **41**, 306–314, URL http://linkinghub.elsevier.com/retrieve/pii/S0966692314000234 (2014).

Farmer, C. and T. Oshan, "Spatial Interaction", The Geographic Information Science & Technology Body of Knowledge **(4th Quarter 2017 Edition)** (2017).

Farmer, C. J. and A. Pozdnoukhov, "Building streaming GIScience from context, theory, and intelligence", in "Proceedings of the Workshop on GIScience in the Big Data Age. Columbus, Ohio", (2012), URL http://ncg.nuim.ie/ncg/GWR/content/staff/staff/downloads/apozdnoukhov/GIScience2012.pdf.

Farmer, C. J. Q., *Commuting flows & local labour markets: Spatial interaction modelling of travel-to-work*, phd, National University of Ireland Maynooth, URL http://eprints.nuim.ie/2857/ (2011).

Ferreira, N., J. Poco, H. Vo, J. Freire and C. Silva, "Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips", IEEE Transactions on Visualization and Computer Graphics **19**, 12, 2149–2158 (2013).

Fik, T. J., R. G. Amey and G. F. Mulligan, "Labor migration amongst hierarchically competing and intervening origins and destinations", Environment and Planning A **24**, 9, 1271–1290, URL http://www.envplan.com/epa/fulltext/a24/a241271.pdf (1992).

Fik, T. J. and G. F. Mulligan, "Spatial flows and competing central places: towards a general theory of hierarchical interaction", Environment and Planning A **22**, 4, 527–549, URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1217752 (1990).

Fischer, M. and S. Hlavackova, "Spatial interaction modelling: Neural network methods and global optimiation", in "Evolving cities: geocomputation in territorial planning", pp. 45–61 (Routledge, 2004).

Fischer, M. M., "Learning in neural spatial interaction models: a statistical perspective", Journal of Geographical Systems **4**, 3, 287–299, URL http://link.springer.com/article/10.1007/s101090200090 (2002).

Fischer, M. M., "Neural Networks: A General Framework for Non-Linear Function Approximation", Transactions in GIS **10**, 4, 521–533, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1467-9671.2006.01010.x/abstract (2006).

Fischer, M. M., "Neural Spatial Interaction Models: Network Training, Model Complexity and Generalization Performance", in "Computational Science and Its Applications–ICCSA 2013", pp. 1–16 (Springer, 2013), URL http://link.springer.com/chapter/10.1007/978-3-642-39649-6_1.

Fischer, M. M. and S. Gopal, "Artificial Neural Networks: A New Approach to Modeling Interregional Telecommunication Flows", Journal of Regional Science **34**, 4, 503–527, URL http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0349776&site=ehost-live (1994).

Fischer, M. M. and D. A. Griffith, "Modeling Spatial Autocorrelation In Spatial Interaction Data: An Application To Patent Citation Data In The European Union", Journal of Regional Science **48**, 5, 969–989, URL http://doi.wiley.com/10.1111/j.1467-9787.2008.00572.x (2008).

Fischer, M. M. and M. Reismann, "Evaluating neural spatial interaction modelling by bootstrapping", Networks and Spatial Economics **2**, 3, 255–268, URL http://link.springer.com/article/10.1023/A:1019923727752 (2002).

Fischer, M. M., M. Reismann and K. Hlavackova–Schindler, "Neural network modeling of constrained spatial interaction flows: Design, estimation, and performance issues", Journal of Regional Science **43**, 1, 35–61, URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9787.00288/abstract (2003).

Fishman, E., "Bikeshare: A Review of Recent Literature", Transport Reviews **36**, 1, 92–113, URL http://www.tandfonline.com/doi/full/10.1080/01441647.2015.1033036 (2016a).

Fishman, E., "Cycling as transport", Transport Reviews **36**, 1, 1–8 (2016b).

Flowerdew, R. and M. Aitkin, "A Method of Fitting the Gravity Model Based on the Poisson Distribution", Journal of Regional Science **22**, 2, 191–202, URL http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0132624&site=ehost-live (1982).

Flowerdew, R. and A. Lovett, "Fitting Constrained Poisson Regression Models to Interurban Migration Flows", Geographical Analysis **20**, 4, 297–307, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1988.tb00184.x/abstract (1988).

Forsyth, A. and K. Krizek, "Urban Design: Is there a Distinctive View from the Bicycle?", Journal of Urban Design **16**, 4, 531–549, URL http://dx.doi.org/10.1080/13574809.2011.586239 (2011).

Fotheringham, A. S., "Spatial Structure and Distance-Decay Parameters", Annals of the Association of American Geographers **71**, 3, 425–436, URL http://www.jstor.org/stable/2562901 (1981).

Fotheringham, A. S., "A new set of spatial-interaction models: the theory of competing destinations", Environment and Planning A **15**, 1, 15–36, URL https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=5650f59a2a&view=att&th=148ae8cddef115b1&attid=0.6&disp=safe&realattid=f_i0ikg6ao5&zw&saduie=AG9B_P8fvyYIEEg5Fz7mRvOkWv25&sadet=1411859676375&sads=dTj63ZP8-fpLYUwn9Moc9RdFzoo (1983a).

Fotheringham, A. S., "Some theoretical aspects of destination choice and their relevance to production-constrained gravity models", Environment and Planning A **15**, 8, 1121–1132, URL https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=5650f59a2a&view=att&th=148ae8cddef115b1&attid=0.1&disp=safe&realattid=f_i0ikft7c0&zw&sadnir=3&saduie=AG9B_P8fvyYIEEg5Fz7mRvOkWv25&sadet=1411858733276&sads=_yX3zmJcD6bu58bA66Uo2N84zT8 (1983b).

Fotheringham, A. S., "Spatial flows and spatial patterns", Environment and Planning A **16**, 4, 529–543, URL https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=5650f59a2a&view=att&th=148ae8cddef115b1&attid=0.2&disp=safe&realattid=f_i0ikfvjs1&zw&saduie=AG9B_P8fvyYIEEg5Fz7mRvOkWv25&sadet=1411858966277&sads=D9vDC2uC4GFdNsOzLiafGnJji1E (1984).

Fotheringham, A. S., "Spatial competition and agglomeration in urban modelling", Environment and Planning A **17**, 2, 213–230, URL https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=5650f59a2a&view=att&th=148ae8cddef115b1&attid=0.3&disp=safe&realattid=f_i0ikfxuf2&zw&saduie=AG9B_P8fvyYIEEg5Fz7mRvOkWv25&sadet=1411860138933&sads=lcO3TTHYFq0DgwWtRB-gAp5KzVk (1985).

Fotheringham, A. S., "Modelling hierarchical destination choice", Environment and Planning A **18**, 3, 401–418, URL https://mail-attachment.googleusercontent.com/attachment/u/1/?ui=2&ik=5650f59a2a&view=att&th=148ae8cddef115b1&attid=0.5&disp=safe&realattid=f_i0ikg3if4&zw&saduie=AG9B_P8fvyYIEEg5Fz7mRvOkWv25&sadet=1411859853440&sads=92jwfesGJ9i_ZwtI0Hxf8WlEiOM (1986).

Fotheringham, A. S., "Consumer Store Choice and Choice Set Definition", Marketing Science **7**, 3, 299, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/212226570/207368958/1?accountid=4485 (1988).

Fotheringham, A. S., C. Brunsdon and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships* (John Wiley & Sons, 2002).

Fotheringham, A. S., T. Champion, C. Wymer and M. Coombes, "Measuring destination attractivity: a migration example", International Journal of Population Geography **6**, 6, 391–421, URL http://onlinelibrary.wiley.com/doi/10.1002/1099-1220(200011/12)6:6<391::AID-IJPG200>3.0.CO;2-5/abstract (2000).

Fotheringham, A. S. and A. Curtis, "Regularities in Spatial Information Processing: Implications for Modeling Destination Choice", The Professional Geographer **51**, 2, 227–239, URL http://onlinelibrary.wiley.com/doi/10.1111/0033-0124.00159/abstract (1999).

Fotheringham, A. S. and D. C. Knudsen, "Modeling Discontinuous Change in Retailing Systems: Extensions of the Harris-Wilson Framework With Results From a Simulated Urban Retailing System", Geographical Analysis **18**, 4, 295–312, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1986.tb00103.x/abstract (1986).

Fotheringham, A. S., T. Nakaya, K. Yano, S. Openshaw and Y. Ishikawa, "Hierarchical destination choice and spatial interaction modelling: a simulation experiment", Environment and Planning A **33**, 5, 901–920, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=a33136 (2001).

Fotheringham, A. S. and M. E. O'Kelly, *Spatial Interaction Models:Formulations and Applications* (Kluwer Academic Publishers, London, 1989), URL

http://www.springer.com/earth+sciences+and+geography/geography/book/978-0-7923-0021-2.

Fotheringham, A. S., P. Rees, T. Champion, S. Kalogirou and A. R. Tremayne, "The development of a migration model for England and Wales: overview and modelling out-migration", Environment and Planning A **36**, 9, 1633–1672, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=a36136 (2004).

Fotheringham, A. S. and M. J. Webber, "Spatial Structure and the Parameters of Spatial Interaction Models", Geographical Analysis **12**, 1, 33–46, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1980.tb00016.x/abstract (1980).

Fotheringham, A. S., W. Yang and W. Kang, "Multi-scale geographically weighted regression", Annals of the American Association of Geographers (2017).

Fotheringham, S., "Distance-decay parameters: A reply", Annals of the Association of American Geographers **72**, 3, 551–553 (1982).

Froehlich, J., J. Neumann and N. Oliver, "Measuring the pulse of the city through shared bicycle programs", Proc. of UrbanSense08 pp. 16–20, URL http://sensorlab.cs.dartmouth.edu/urbansensing/papers/urbansense08_proceedings.pdf#page=22 (2008).

Froehlich, J., J. Neumann and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling.", in "IJCAI", pp. 1420–1426 (2009), URL http://www.aaai.org/ocs/index.php/IJCAI/IJCAI-09/paper/download/578/910.

Gardiner, C. and R. Hill, "Cycling on the Journey to Work: Analysis of Socioeconomic Variables from the UK 1991 Population Census Samples of Anonymised Records", Planning Practice and Research **12**, 3, 251–261, URL http://dx.doi.org/10.1080/02697459716491 (1997).

Gargiulo, F., M. Lenormand, S. Huet and O. Baqueiro Espinosa, "Commuting Network Models: Getting the Essentials", Journal of Artificial Societies and Social Simulation **15**, 2, 6 (2011).

Gatersleben, B. and K. M. Appleton, "Contemplating cycling to work: Attitudes and perceptions in different stages of change", Transportation Research Part A: Policy and Practice **41**, 4, 302–312, URL http://linkinghub.elsevier.com/retrieve/pii/S0965856406001091 (2007).

Ghoniem, M., J. Fekete and P. Castagliola, "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations", in "IEEE Symposium on Information Visualization, 2004. INFOVIS 2004", pp. 17–24 (2004).

Gibbons, S. and H. G. Overman, "Mostly Pointless Spatial Econometrics?", Journal of Regional Science **52**, 2, 172–191, URL http://doi.wiley.com/10.1111/j.1467-9787.2012.00760.x (2012).

Gitlesen, J. P., G. Kleppe, I. Thorsen and J. Ubøe, "An Empirically Based Implementation and Evaluation of a Hierarchical Model for Commuting Flows", Geographical Analysis **42**, 3, 267–287, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1538-4632.2010.00793.x/abstract (2010).

Gitlesen, J. P. and I. Thorsen, "A competing destinations approach to modeling commuting flows: a theoretical interpretation and an empirical application of the model", Environment and Planning A **32**, 11, 2057–2074, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=a3329 (2000).

Glejser, H., "A gravity model of interdependent equations to estimate flow creation and diversion", Journal of regional science **9**, 3, 439–449 (1969).

Goh, S., K. Lee, J. S. Park and M. Y. Choi, "Modification of the gravity model and application to the metropolitan Seoul subway system", Physical Review E **86**, 2, 026102, URL http://link.aps.org/doi/10.1103/PhysRevE.86.026102 (2012).

Gong, L., X. Liu, L. Wu and Y. Liu, "Inferring trip purposes and uncovering travel patterns from taxi trajectory data", Cartography and Geographic Information Science **0**, 0, 1–12, URL http://dx.doi.org/10.1080/15230406.2015.1014424 (2015).

González, M. C., C. A. Hidalgo and A.-L. Barabási, "Understanding individual human mobility patterns", Nature **453**, 7196, 779–782, URL http://www.nature.com/nature/journal/v453/n7196/full/nature06958.html (2008).

Gordon, I. R., "Economic explanations of spatial variation in distance detterence", Environemnt and Planning A **17**, 1, 59–72 (1985).

Griffith, D. and Y. Chun, "Evaluating Eigenvector Spatial Filter Corrections for Omitted Georeferenced Variables", Econometrics **4**, 2, 29, URL http://www.mdpi.com/2225-1146/4/2/29 (2016).

Griffith, D. A., "Spatial structure and spatial interaction: A review", Environment and Planning A **8**, 7, 731–740 (1976).

Griffith, D. A., "Spatial Autocorrelation and Eigenfunctions Of The Geographic Weights Matrix Accompanying Geo-Referenced Data", Canadian Geographer / Le Géographe canadien **40**, 4, 351–367, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0064.1996.tb00462.x/abstract (1996).

Griffith, D. A., "A linear regression solution to the spatial autocorrelation problem", Journal of Geographical Systems **2**, 2, 141–156, URL http://link.springer.com.ezproxy1.lib.asu.edu/article/10.1007/PL00011451 (2000).

Griffith, D. A., "A spatial filtering specification for the auto-Poisson model", Statistics & Probability Letters **58**, 3, 245–251, URL http://www.sciencedirect.com/science/article/pii/S0167715202000998 (2002).

Griffith, D. A., "A spatial filtering specification for the autologistic model", Environment and Planning A **36**, 10, 1791–1811, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=a36247 (2004).

Griffith, D. A., "Spatial Structure and Spatial Interaction 25 years later", The review of regional Studies **37**, 1, 28–38 (2007).

Griffith, D. A., "Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 Germany journey-to-work flows", Journal of Geographical Systems **11**, 2, 117–140, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/294690116/illustrataImage/$N/1/OB-301-0006979525/812706222/docView?accountid=4485 (2009a).

Griffith, D. A., "Spatial Autocorrelation in Spatial Interaction", in "Complexity and Spatial Networks", edited by A. Reggiani and P. Nijkamp, pp. 221–237 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009b), URL http://link.springer.com/10.1007/978-3-642-01554-0_16.

Griffith, D. A., "Visualizing analytical spatial autocorrelation components latent in spatial interaction data: An eigenvector spatial filter approach", Computers, Environment and Urban Systems **35**, 2, 140–149, URL http://linkinghub.elsevier.com/retrieve/pii/S0198971510000827 (2011).

Griffith, D. A. and Y. Chun, "Spatial Autocorrelation in Spatial Interactions Models: Geographic Scale and Resolution Implications for Network Resilience and Vulnerability", Networks and Spatial Economics **15**, 2, 337–365, URL http://link.springer.com/10.1007/s11067-014-9256-4 (2015).

Griffith, D. A. and M. M. Fischer, "Constrained variants of the gravity model and spatial dependence: model specification and estimation issues", Journal of Geographical Systems **15**, 3, 291–317, URL http://link.springer.com/10.1007/s10109-013-0182-7 (2013).

Griffith, D. A., M. M. Fischer and J. LeSage, "The spatial autocorrelation problem in spatial interaction modelling: a comparison of two common solutions", Letters in Spatial and Resource Sciences URL http://link.springer.com/10.1007/s12076-016-0172-8 (2016).

Griffith, D. A. and K. G. Jones, "Explorations into the relationship between spatial structure and spatial interaction", Environment and Planning A **12**, 187–201 (1980).

Griffith, D. A. and P. R. Peres-Neto, "Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses", Ecology **87**, 10, 2603–2613, URL http://www.esajournals.org/doi/abs/10.1890/0012-9658(2006)87%5B2603:SMIETF%5D2.0.CO%3B2 (2006).

Guinn, J. M. and P. Stangl, "Pedestrian and bicyclist motivation: an assessment of influences on pedestrians' and bicyclists' mode choice in Mt. Pleasant, Vancouver", Urban, Planning and Transport Research **2**, 1, 105–125, URL http://dx.doi.org/10.1080/21650020.2014.906907 (2014).

Guldmann, J.-M., "Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows*", Papers in Regional Science **78**, 2, 179–194, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1435-5597.1999.tb00739.x/abstract (1999).

Guo, D., "Visual analytics of spatial interaction patterns for pandemic decision support", International Journal of Geographical Information Science **21**, 8, 859–877, URL http://www.tandfonline.com/doi/abs/10.1080/13658810701349037 (2007).

Guo, D., "Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data", IEEE Transactions on Visualization and Computer Graphics **15**, 6, 1041–1048 (2009).

Guo, D., S. Liu and H. Jin, "A graph-based approach to vehicle trajectory analysis", Journal of Location Based Services **4**, 3-4, 183–199, URL http://www.tandfonline.com/doi/abs/10.1080/17489725.2010.537449 (2010).

Guo, D. and X. Zhu, "Origin-Destination Flow Data Smoothing and Mapping", IEEE Transactions on Visualization and Computer Graphics **20**, 12, 2043–2052, URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6875983 (2014).

Guo, D., X. Zhu, H. Jin, P. Gao and C. Andris, "Discovering Spatial Patterns in Origin-Destination Mobility Data: Discovering Spatial Patterns in Origin-Destination Mobility Data", Transactions in GIS **16**, 3, 411–429, URL http://doi.wiley.com/10.1111/j.1467-9671.2012.01344.x (2012).

Guy, C. M., "Recent advances in spatial interaction modelling: an application to the forecasting of shopping travel", Environment and Planning A **19**, 2, 173–186, URL http://journals.sagepub.com/doi/abs/10.1068/a190173 (1987).

Halleck Vega, S. and J. P. Elhorst, "The SLX Model", Journal of Regional Science **55**, 3, 339–363, URL http://doi.wiley.com/10.1111/jors.12188 (2015).

Hampshire, R. C. and L. Marla, "An Analysis of Bike Sharing Usage: Explaining Trip Generation and Attraction from Observed Demand", in "Transportation Research Board 91st Annual Meeting", (2012), URL http://trid.trb.org/view.aspx?id=1129620.

Han, X., Q. Hao, B. Wang and T. Zhou, "Origin of the Scaling Law in Human Mobility: Hierarchical Organization of Traffic Systems", arXiv:0908.1221 [physics] URL http://arxiv.org/abs/0908.1221, arXiv: 0908.1221 (2009).

Handy, S., B. van Wee and M. Kroesen, "Promoting Cycling for Transport: Research Needs and Challenges", Transport Reviews **34**, 1, 4–24, URL http://dx.doi.org/10.1080/01441647.2013.860204 (2014).

Handy, S. L. and Y. Xing, "Factors Correlated with Bicycle Commuting: A Study in Six Small U.S. Cities", International Journal of Sustainable Transportation **5**, 2, 91–110, URL http://dx.doi.org/10.1080/15568310903514789 (2011).

Harris, B. and A. G. Wilson, "Equilibrium values and dynamics of attractiveness terms in production-constrained spatial-interaction models", Environment and planning A **10**, 4, 371–388, URL http://envplan.com/epa/fulltext/a10/a100371.pdf (1978).

Haynes, K. E. and A. S. Fotheringham, *Gravity and spatial interaction models*, vol. 2 (Sage publications, Beverly Hills, 1984), URL http://www.web.pdx.edu/~stipakb/download/PA557/ReadingsPA557sec1-2.pdf.

Haynes, K. E., D. L. Poston and P. Schnirring, "Intermetropolitan Migration in High and Low Opportunity Areas: Indirect Tests of the Distance and Intervening Opportunities Hypotheses", Economic Geography **49**, 1, 68, URL http://www.jstor.org/stable/142746?origin=crossref (1973).

Heide, H. T., "Migration Models and Their Significance for Population Forecasts", The Milbank Memorial Fund Quarterly **41**, 1, 56, URL http://www.jstor.org/stable/3348680?origin=crossref (1963).

Heinen, E., K. Maat and B. v. Wee, "The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances", Transportation Research Part D: Transport and Environment **16**, 2, 102–109, URL http://www.sciencedirect.com/science/article/pii/S1361920910001306 (2011).

Helbich, M. and D. A. Griffith, "Spatially varying coefficient models in real estate: Eigenvector spatial filtering and alternative approaches", Computers, Environment and Urban Systems **57**, 1–11, URL http://linkinghub.elsevier.com/retrieve/pii/S0198971515300387 (2016).

Hendry, D. F., "Econometrics-Alchemy or Science?", Economica **47**, 188, 387, URL http://www.jstor.org/stable/2553385?origin=crossref (1980).

Hirtle, S. C. and J. Jonides, "Evidence of hierarchies in cognitive maps", Memory & Cognition **13**, 3, 208–217, URL http://link.springer.com.ezproxy1.lib.asu.edu/article/10.3758/BF03197683 (1985).

Hodges, J. S. and B. J. Reich, "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love", The American Statistician **64**, 4, 325–334, URL http://www.tandfonline.com/doi/abs/10.1198/tast.2010.10052 (2010).

Holten, D. and J. J. Van Wijk, "Force-Directed Edge Bundling for Graph Visualization", in "Computer Graphics Forum", vol. 28, pp. 983–990 (Wiley Online Library, 2009), URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8659.2009.01450.x/full.

Hu, P. and J. Pooler, "An empirical test of the competing destinations model", Journal of geographical systems **4**, 3, 301–323, URL http://link.springer.com/article/10.1007/s101090200088 (2002).

Huff, D. L., "A Probabilistic Analysis of Shopping Center Trade Areas", Land Economics **39**, 1, 81–90, URL http://www.jstor.org/stable/3144521 (1963).

Huff, D. L., "Defining and Estimating a Trading Area", Journal of Marketing **28**, 3, 34–38, URL http://www.jstor.org/stable/1249154 (1964).

Huque, M. H., H. D. Bondell, R. J. Carroll and L. M. Ryan, "Spatial regression with covariate measurement error: A semiparametric approach: On the Impact of Covariate Measurement Error", Biometrics **72**, 3, 678–686, URL http://doi.wiley.com/10.1111/biom.12474 (2016).

Huque, M. H., H. D. Bondell and L. Ryan, "On the impact of covariate measurement error on spatial regression modelling: COVARIATE MEASUREMENT ERROR", Environmetrics **25**, 8, 560–570, URL http://doi.wiley.com/10.1002/env.2305 (2014).

Ishikawa, Y., "An empirical study of the competing destinations model using Japanese interaction data", Environment and Planning A **19**, 10 (1987).

Ishikawa, Y., "Explorations into the two-stage destination choice", Geographical Review of Japan B **62**, 75–85 (1990).

Johnston, R., "Map pattern and friction of distance parameters: a comment", Regional Studies **9**, 3, 281–283, URL http://www.tandfonline.com/doi/abs/10.1080/09595237500185291 (1975).

Johnston, R., "On Regression Coefficients in Comparitive Studies of the Friction of Distance", Journal of economic and social geography **67**, 15–28 (1976).

Johnston, R. J., "On Frictions of Distance and Regression Coefficients", Area **5**, 3, 187–191, URL http://www.jstor.org/stable/20000751 (1973).

Jäppinen, S., T. Toivonen and M. Salonen, "Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach", Applied Geography **43**, 13–24, URL http://www.sciencedirect.com/science/article/pii/S014362281300132X (2013).

Kalogirou, S., "Destination Choice of Athenians: An Application of Geographically Weighted Versions of Standard and Zero Inflated Poisson Spatial Interaction Models: Destination Choice of Athenians", Geographical Analysis pp. n/a–n/a, URL http://doi.wiley.com/10.1111/gean.12092 (2015).

Kaltenbach, K., D., "Application of Gravity and Intervening Opportunities Models to Recreational Travel in Kentucky", Tech. Rep. 336, Department of Highways Division of Research, Kentucky (1972).

Kang, C., Y. Liu, D. Guo and K. Qin, "A Generalized Radiation Model for Human Mobility: Spatial Scale, Searching Direction and Trip Constraint", PLoS ONE **10**, 11, e0143500, URL http://dx.doi.org/10.1371/journal.pone.0143500 (2015).

Kerkman, K., K. Martens and H. Meurs, "A multilevel spatial interaction model of transit flows incorporating spatial and network autocorrelation", Journal of Transport Geography **60**, 155–166, URL http://linkinghub.elsevier.com/retrieve/pii/S0966692316302058 (2017).

Kim, K. and J. E. Cohen, "Determinants of International Migration Flows to and from Industrialized Countries: A Panel Data Approach Beyond Gravity1", International Migration Review **44**, 4, 899–932, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1747-7379.2010.00830.x/abstract (2010).

Kitchin, R., "Big Data, new epistemologies and paradigm shifts", Big Data & Society **1**, 1, 2053951714528481, URL http://bds.sagepub.com/content/1/1/2053951714528481.short (2014a).

Kitchin, R., "The real-time city? Big data and smart urbanism", GeoJournal **79**, 1, 1–14, URL http://link.springer.com/article/10.1007/S10708-013-9516-8 (2014b).

Knudsen, D. and A. Fotheringham, "Matrix comparison, Goodness-of-fit, and spatial interaction modeling", International Regional Science Review **10**, 127–147 (1986).

Kordi, M. and A. S. Fotheringham, "Spatially Weighted Interaction Models (SWIM)", Annals of the American Association of Geographers **106**, 5, 990–1012, URL http://www.tandfonline.com/doi/full/10.1080/24694452.2016.1191990 (2016).

Kordi, M., C. Kaiser and A. S. Fotheringham, "A possible solution for the centroid-to-centroid and intra-zonal trip length problems", in "International Conference on Geographic Information Science, Avignon", (2012), URL http:

//www.agile-online.orgwww.agile-online.org/Conference_Paper/CDs/agile_2012/proceedings/papers/Paper_Kordi_A_possible_solution_for_the_centroid-to-centroid_and_intra-zonal_trip_length_problems_2012.pdf.

Koylu, C. and D. Guo, "Smoothing locational measures in spatial interaction networks", Computers, Environment and Urban Systems **41**, 12–25, URL http://www.sciencedirect.com/science/article/pii/S0198971513000215 (2013).

Kruger, R., D. Thom and T. Ertl, "Semantic Enrichment of Movement Behavior with Foursquare - A Visual Analytics Approach", IEEE Transactions on Visualization and Computer Graphics **PP**, 99, 1–1 (2014).

Laney, D., "3d Data Management: Controlling Data Volume, Velocity, and Variety", Tech. rep., The Meta Group, Stamford, URL http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. (2001).

Le Gallo, J. and B. Fingleton, "Measurement errors in a spatial context", Regional Science and Urban Economics **42**, 1-2, 114–125, URL http://linkinghub.elsevier.com/retrieve/pii/S0166046211000986 (2012).

Leamer, E., E., "Let's Take the Con Out of Econometrics", The American Economic Review **73**, 1, 31–43, URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.333.8024 (1983).

Lee, M.-L. and R. K. Pace, "Spatial distribution of retail sales", The journal of real estate finance and economics **31**, 1, 53–69, URL http://www.springerlink.com/index/X344437QLJ132262.pdf (2005).

Legendre, P., M. R. Dale, M.-J. Fortin, J. Gurevitch, M. Hohn and D. Myers, "The consequences of spatial structure for the design and analysis of ecological field surveys", Ecography **25**, 5, 601–615, URL http://onlinelibrary.wiley.com/doi/10.1034/j.1600-0587.2002.250508.x/pdf (2002).

Lenormand, M., A. Bassolas and J. J. Ramasco, "Systematic comparison of trip distribution laws and models", Journal of Transport Geography **51**, 158–169, URL http://linkinghub.elsevier.com/retrieve/pii/S0966692315002422 (2016).

Lenormand, M., S. Huet and F. Gargiulo, "Generating French virtual commuting network at municipality level", arXiv:1109.6759 [math, stat] URL http://arxiv.org/abs/1109.6759, arXiv: 1109.6759 (2011).

Lenormand, M., S. Huet, F. Gargiulo and G. Deffuant, "A Universal Model of Commuting Networks", PLoS ONE **7**, 10, e45985, URL http://dx.doi.org/10.1371/journal.pone.0045985 (2012).

LeSage, J. and R. Pace, "The Biggest Myth in Spatial Econometrics", Econometrics **2**, 4, 217–249, URL http://www.mdpi.com/2225-1146/2/4/217/ (2014).

LeSage, J. P. and M. M. Fischer, "Spatial regression-based model specifications for exogenous and endogenous spatial interaction", Available at SSRN 2420746 URL http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2420746 (2014).

LeSage, J. P., M. M. Fischer and T. Scherngell, "Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects*", Papers in Regional Science **86**, 3, 393–421, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1435-5957.2007.00125.x/abstract (2007).

LeSage, J. P. and C. Llano, "A spatial interaction model with spatially structured origin and destination effects", Journal of Geographical Systems **15**, 3, 265–289, URL http://link.springer.com/10.1007/s10109-013-0181-8 (2013).

LeSage, J. P. and R. K. Pace, "Spatial Econometric Modeling of Origin-Destination Flows*", Journal of Regional Science **48**, 5, 941–967, URL http://doi.wiley.com/10.1111/j.1467-9787.2008.00573.x (2008).

LeSage, J. P. and W. Polasek, "Incorporating Transportation Network Structure in Spatial Econometric Models of Commodity Flows", Spatial Economic Analysis **3**, 2, 225–245 (2008).

LeSage, J. P. and E. Satici, "A Bayesian Spatial Interaction Model Variant of the Poisson Pseudo-Maximum Likelihood Estimator", SSRN (2013).

LeSage, J. P. and C. Thomas-Agnan, "Interpreting Spatial Econometric Origin-Destination Flow Models", Journal of Regional Science **55**, 2, 188–208, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/jors.12114/abstract (2015).

Li, L., L. Yang, H. Zhu and R. Dai, "Explorative Analysis of Wuhan Intra-Urban Human Mobility Using Social Media Check-In Data", PLoS ONE **10**, 8, e0135286, URL http://dx.doi.org/10.1371/journal.pone.0135286 (2015).

Li, Y., H. Tang and X. Lin, "Spatial linear mixed models with covariate measurement errors", Statistica Sinica **19**, 3, 1077, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2695401/ (2009).

Liang, X., J. Zhao, L. Dong and K. Xu, "Unraveling the origin of exponential law in intra-urban human mobility", Scientific Reports **3**, URL http://www.nature.com/srep/2013/131018/srep02983/full/srep02983.html (2013).

Liang, X., X. Zheng, W. Lv, T. Zhu and K. Xu, "The scaling of human mobility by taxis is exponential", Physica A: Statistical Mechanics and its Applications **391**, 5, 2135–2144, URL http://www.sciencedirect.com/science/article/pii/S0378437111008703 (2012).

Lin, M., W.-J. Hsu and Z. Q. Lee, "Modeling High Predictability and Scaling Laws of Human Mobility", in "2013 IEEE 14th International Conference on Mobile Data Management (MDM)", vol. 2, pp. 125–130 (2013).

Linders, G.-J. M. and H. L. De Groot, "Estimation of the gravity equation in the presence of zero flows", Tech. rep., Tinbergen Institute Discussion Paper, URL http://www.econstor.eu/handle/10419/86589 (2006).

Liu, X., L. Gong, Y. Gong and Y. Liu, "Revealing travel patterns and city structure with taxi trip data", Journal of Transport Geography **43**, 78–90, URL http://www.sciencedirect.com/science/article/pii/S0966692315000253 (2015).

Liu, Y., C. Kang, S. Gao, Y. Xiao and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data", Journal of Geographical Systems **14**, 4, 463–483, URL http://link.springer.com/10.1007/s10109-012-0166-z (2012).

Liu, Y., D. Tong and X. Liu, "Measuring Spatial Autocorrelation of Vectors: Measuring Spatial Autocorrelation of Vectors", Geographical Analysis pp. n/a–n/a, URL http://doi.wiley.com/10.1111/gean.12069 (2014).

Llorente, A., M. Garcia-Herranz, M. Cebrian and E. Moro, "Social Media Fingerprints of Unemployment", PLoS ONE **10**, 5, e0128692, URL http://dx.doi.org/10.1371/journal.pone.0128692 (2015).

Lo, L., "Spatial structure and spatial interaction: a simulation approach", Environment and planning A **23**, 9, 1279–1300, URL http://journals.sagepub.com/doi/abs/10.1068/a231279 (1991a).

Lo, L., "Substitutability, Spatial Structure, and Spatial Interaction", Geographical Analysis **23**, 2, 132–146, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1991.tb00229.x/abstract (1991b).

Lo, L., "Destination interdependence and the competing-destinations model", Environment and Planning A **24**, 8, 1191–1204, URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1217744 (1992).

Long, W. H. and R. B. Uris, "Distance, intervening opportunities, city hierarchy, and air travel", Annals of Regional Science **5**, 152–161 (1971).

Lovelace, R., M. Birkin, P. Cross and M. Clarke, "From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows: From Big Noise to Big Data", Geographical Analysis pp. n/a–n/a, URL http://doi.wiley.com/10.1111/gean.12081 (2015).

Lovelace, R., N. Malleson, K. Harland and M. Birkin, "Geotagged tweets to inform a spatial interaction model: a case study of museums", arXiv:1403.5118 [cs, stat] URL http://arxiv.org/abs/1403.5118, arXiv: 1403.5118 (2014).

Lu, Y. and J.-C. Thill, "Assessing the cluster correspondence between paired point locations", Geographical Analysis **35**, 4, 290–309, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.2003.tb01116.x/abstract (2003).

Lu, Y. and J.-C. Thill, "Cross-scale analysis of cluster correspondence using different operational neighborhoods", Journal of Geographical Systems **10**, 3, 241–261, URL http://link.springer.com/10.1007/s10109-008-0069-1 (2008).

Margaretic, P., C. Thomas-Agnan and R. Doucet, "Spatial dependence in (origin-destination) air passenger flows: Spatial dependence in air passenger flows", Papers in Regional Science **96**, 2, 357–380, URL http://doi.wiley.com/10.1111/pirs.12189 (2017).

Martens, K., "The bicycle as a feedering mode: experiences from three European countries", Transportation Research Part D: Transport and Environment **9**, 4, 281–294, URL http://www.sciencedirect.com/science/article/pii/S1361920904000100 (2004).

Martínez, L. M. and J. M. Viegas, "A new approach to modelling distance-decay functions for accessibility assessment in transport studies", Journal of Transport Geography **26**, 87–96, URL http://www.sciencedirect.com/science/article/pii/S096669231200230X (2013).

Masucci, A. P., J. Serras, A. Johansson and M. Batty, "Gravity vs radiation model: on the importance of scale and heterogeneity in commuting flows", arXiv:1206.5735 [physics] URL http://arxiv.org/abs/1206.5735, arXiv: 1206.5735 (2012).

McFadden, D., "Conditional logit analysis of qualitative choice behavior", in "Frontiers in Econometrics", pp. 105–142 (Academic Press, New York, 1974).

McFadden, D., "Modelling the choice of residential location", Tech. rep., URL http://cowles.econ.yale.edu/P/cd/d04b/d0477.pdf (1977).

McMillen, D. P., "Spatial Autocorrelation Or Model Misspecification?", International Regional Science Review **26**, 2, 208–217, URL http://irx.sagepub.com/cgi/doi/10.1177/0160017602250977 (2003).

McMillen, D. P., "Perspectives On Spatial Econometrics: Linear Smoothing With Structured Models", Journal of Regional Science **52**, 2, 192–209, URL http://doi.wiley.com/10.1111/j.1467-9787.2011.00746.x (2012).

McNamara, T. P., "Mental representations of spatial relations", Cognitive Psychology **18**, 1, 87–121, URL http://www.sciencedirect.com/science/article/pii/0010028586900162 (1986).

Metulini, R., "Spatial gravity models for international trade: a panel analysis among OECD countries", in "ERSA conference papers", (2013), URL http://www.academia.edu/download/30646583/application_on_OECD_countries.pdf.

Miller, H. J., "Geocomputation", in "The SAGE Handbook of Spatial Analysis", pp. 398–418 (SAGE Publications, Ltd, London, 2009), URL http://sk.sagepub.com/reference/hdbk_spatialanalysis/n21.xml.

Mooney, P., P. Corcoran and A. Winstanley, "Preliminary Results of a Spatial Analysis of Dublin City's Bike Rental Scheme", in "GIS Research UK 18th Annual Conference", pp. 325–330 (London, 2010), URL http://eprints.maynoothuniversity.ie/4919/.

Mozolin, M., J. C. Thill and E. Lynn Usery, "Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation", Transportation Research Part B: Methodological **34**, 1, 53–73, URL http://www.sciencedirect.com/science/article/pii/S0191261599000144 (2000).

Murakami, D. and D. A. Griffith, "Random effects specifications in eigenvector spatial filtering: a simulation study", Journal of Geographical Systems **17**, 4, 311–331, URL http://link.springer.com/10.1007/s10109-015-0213-7 (2015).

Murray, A. T., Y. Liu, S. J. Rey and L. Anselin, "Exploring movement object patterns", The Annals of Regional Science **49**, 2, 471–484, URL http://link.springer.com/10.1007/s00168-011-0459-z (2012).

Nakaya, T., "Local spatial interaction modelling based on the geographically weighted regression approach", GeoJournal **53**, 4, 347–358, URL http://link.springer.com/article/10.1023/A%3A1020149315435 (2001).

Nazem, M., M. Trépanier and C. Morency, "Revisiting the destination ranking procedure in development of an Intervening Opportunities Model for public transit trip distribution", Journal of Geographical Systems **17**, 1, 61–81, URL http://link.springer.com/10.1007/s10109-014-0203-1 (2015).

Newing, A., G. P. Clarke and M. Clarke, "Developing and Applying a Disaggregated Retail Location Model with Extended Retail Demand Estimations: Disaggregated Retail Location Model", Geographical Analysis **47**, 3, 219–239, URL http://doi.wiley.com/10.1111/gean.12052 (2015).

Newman, M. E. J., "Power laws, Pareto distributions and Zipf's law", Cities **30**, 59–67, URL http://arxiv.org/abs/cond-mat/0412004, arXiv: cond-mat/0412004 (2013).

Nijkamp, P. and A. Reggiani, "Dynamic spatial interaction models: new directions", Environment and Planning A **20**, 11, 1449–1460, URL http://www.envplan.com/epa/fulltext/a20/a201449.pdf (1988).

Nissi, E. and A. Sarra, "Detecting Local Variations in Spatial Interaction Models by Means of Geographically Weighted Regression", Journal of Applied Sciences **11**, 4, 630–638, URL http://www.scialert.net/abstract/?doi=jas.2011.630.638 (2011).

Noulas, A., S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo, "A Tale of Many Cities: Universal Patterns in Human Urban Mobility", PLoS ONE **7**, 5, e37027, URL http://dx.doi.org/10.1371/journal.pone.0037027 (2012).

NYC, "NYC Open Data", URL https://opendata.cityofnewyork.us/ (2017).

Okabe, A., "A theoretical comparison of the opportunity and gravity models", Regional Science and Urban Economics **6**, 381–397 (1976).

Okabe, A. and N. Tagashira, "Spatial aggregation bias in a regression model containing a distance variable", Geographical Systems **3**, 77–99 (1996).

Olsson, G., "Central places systems, spatial interaction, and stochastic processes", Papers and proceedings of the regional sciences association **18**, 13–45 (1967).

Openshaw, S., "Modelling spatial interaction using a neural net", in "Geographic Information Systems, Spatial Modelling and Policy Evaluation", edited by M. M. Fischer and P. Nijkamp, pp. 147–164 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1993), URL http://www.springerlink.com/index/10.1007/978-3-642-77500-0_10.

Oshan, T., C. Farmer and O. Eoin, "Spatial Interaction Simulation Methods for Ancient Settlement Distributions in Central Italy", in "CAA 2014: 21st Century Archaeology: Concepts, methods and tools. Proceedings of the 42nd Annual Conference on Computer Applications and Quantitative Methods in Archaeology", pp. 631–640 (Oxbow books, Paris, France, 2014).

Oshan, T. and A. Fotheringham, Stewart, "A Closer Examination of Spatial-Filter-Based Local Models", International Conference on GIScience Short Paper Proceedings **1**, URL http://escholarship.org/uc/item/04t0t6ds (2016).

Oshan, T. M., "A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL)", REGION **3**, 2, 11, URL http://openjournals.wu.ac.at/ojs/index.php/region/article/view/175 (2016).

Oshan, T. M. and A. S. Fotheringham, "A Comparison of Spatially Varying Regression Coefficient Estimates Using Geographically Weighted and Spatial-Filter-Based Techniques: A Comparison of Spatially Varying Regression", Geographical Analysis URL http://doi.wiley.com/10.1111/gean.12133 (2017).

O'Brien, O., J. Cheshire and M. Batty, "Mining bicycle sharing data for generating insights into sustainable transport systems", Journal of Transport Geography **34**, 262–273, URL http://www.sciencedirect.com/science/article/pii/S0966692313001178 (2014).

Pace, R. K., J. P. Lesage and S. Zhu, "Interpretation and Computation of Estimates from Regression Models using Spatial Filtering", Spatial Economic Analysis **8**, 3, 352–369, URL http://www.tandfonline.com/doi/abs/10.1080/17421772.2013.807355 (2013).

Paciorek, C. J., "The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators", Statistical Science **25**, 1, 107–125, URL http://projecteuclid.org/euclid.ss/1280841736 (2010).

Padgham, M., "Human Movement Is Both Diffusive and Directed", PLoS ONE **7**, 5, e37754, URL http://dx.doi.org/10.1371/journal.pone.0037754 (2012).

Partridge, M. D., M. Boarnet, S. Brakman and G. Ottaviano, "INTRODUCTION: WHITHER SPATIAL ECONOMETRICS?", Journal of Regional Science **52**, 2, 167–171, URL http://doi.wiley.com/10.1111/j.1467-9787.2012.00767.x (2012).

Patuelli, R., G.-J. Linders, R. Metulini, D. A. Griffith and others, "The Space of Gravity: Spatially Filtered Estimation of a Gravity Model for Bilateral Trade", in "Spatial Econometric Interaction Modelling", Advances in Spatial Science, pp. 145–169 (Springer, 2015), URL http://amsacta.unibo.it/4332/1/WP1022.pdf.

Patuelli, R., M. Mussoni and G. Candela, "The effects of World Heritage Sites on domestic tourism: a spatial interaction model for Italy", Journal of Geographical Systems **15**, 3, 369–402, URL http://link.springer.com/10.1007/s10109-013-0184-5 (2013).

Pedersen, P. O., "Innovation diffusion within and between national urban systems", Geographical Analysis **2**, 3, 203–254, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1970.tb00858.x/abstract (1970).

Pellegrini, P. A. and A. S. Fotheringham, "Intermetropolitan migration and hierarchical destination choice: a disaggregate analysis from the US Public Use Microdata Samples", Environment and Planning A **31**, 6, 1093–1118, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=a311093 (1999).

Pellegrini, P. A. and A. S. Fotheringham, "Modelling spatial choice: a review and synthesis in a migration context", Progress in Human Geography **26**, 4, 487–510, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/230727749/abstract?accountid=4485 (2002).

Pellegrini, P. A., A. S. Fotheringham and G. Lin, "An Empirical Evaluation of Parameter Sensitivity to Choice Set Definition in Shopping Destination Choice Models", Papers in Regional Science **76**, 2, 257–284, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/j.1435-5597.1997.tb00691.x/abstract (1997).

Peng, C., X. Jin, K.-C. Wong, M. Shi and P. Liò, "Collective Human Mobility Pattern from Taxi Trips in Urban Area", PLoS ONE **7**, 4, e34487, URL http://dx.doi.org/10.1371/journal.pone.0034487 (2012).

Philippidis, G., H. Resano-Ezcaray and A. I. Sanjuán-López, "Capturing zero-trade values in gravity equations of trade: an analysis of protectionism in agro-food sectors", Agricultural Economics **44**, 2, 141–159, URL http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1111/agec.12000/abstract (2013).

Pinkse, J. and M. E. Slade, "The Future Of Spatial Econometrics", Journal of Regional Science **50**, 1, 103–117, URL http://doi.wiley.com/10.1111/j.1467-9787.2009.00645.x (2010).

Pirozmand, P., G. Wu, B. Jedari and F. Xia, "Human mobility in opportunistic networks: Characteristics, models and prediction methods", Journal of Network and Computer Applications **42**, 45–58, URL http://www.sciencedirect.com/science/article/pii/S1084804514000587 (2014).

Pooler, J., "Competition among destinations in spatial interaction models: a new point of view", Chinese Geographical Science **8**, 3, 212–224, URL http://link.springer.com/article/10.1007/s11769-997-0014-0 (1998).

Porojan, A., "Trade Flows and Spatial Effects: The Gravity Model Revisited", Open Economies Review **12**, 3, 265, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/206484331/abstract?accountid=4485 (2001).

Pucher, J., R. Buehler and M. Seinen, "Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies", Transportation Research Part A: Policy and Practice **45**, 6, 451–475, URL http://www.sciencedirect.com/science/article/pii/S0965856411000474 (2011).

Rae, A., "From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census", Computers, Environment and Urban Systems **33**, 3, 161–178, URL http://linkinghub.elsevier.com/retrieve/pii/S019897150900009X (2009).

Rae, A., "Flow-Data Analysis with Geographical Information Systems: A Visual Approach", Environment and Planning B: Planning and Design **38**, 5, 776–794, URL http://epb.sagepub.com.ezproxy1.lib.asu.edu/content/38/5/776 (2011).

Ramsay, T. O., R. T. Burnett and D. Krewski, "The effect of concurvity in generalized additive models linking mortality to ambient particulate matter", Epidemiology **14**, 1, 18–23, URL http://journals.lww.com/epidem/Abstract/2003/01000/The_Effect_of_Concurvity_in_Generalized_Additive.9.aspx (2003).

Ren, Y., M. Ercsey-Ravasz, P. Wang, M. C. González and Z. Toroczkai, "Predicting commuter flows in spatial networks using a radiation model based on temporal ranges", Nature Communications **5**, 5347, URL http://www.nature.com/doifinder/10.1038/ncomms6347 (2014).

Rietveld, P. and V. Daniel, "Determinants of bicycle use: do municipal policies matter?", Transportation Research Part A: Policy and Practice **38**, 7, 531–550, URL http://www.sciencedirect.com/science/article/pii/S0965856404000382 (2004).

Rogerson, P. A., "Parameter estimation in the intervening opportunities model", Geographical Analysis **18**, 4, 357–360, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1986.tb00107.x/full (1986).

Romanillos, G., M. Zaltz Austwick, D. Ettema and J. De Kruijf, "Big Data and Cycling", Transport Reviews **36**, 1, 114–133, URL http://www.tandfonline.com/doi/full/10.1080/01441647.2015.1084067 (2016).

Roth, C., S. M. Kang, M. Batty and M. Barthélemy, "Structure of Urban Movements: Polycentric Activity and Entangled Hierarchical Flows", PLoS ONE **6**, 1, e15923, URL http://dx.doi.org/10.1371/journal.pone.0015923 (2011).

Roy, J. R., *Spatial Interaction Modelling: A Regional Science Context*, Advances in Spatial Science (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004), URL http://link.springer.com/10.1007/978-3-540-24807-1.

Sander, N., G. J. Abel, R. Bauer and J. Schmidt, "Visualising migration flow data with circular plots", Tech. rep., Vienna Institute of Demography Working Papers, URL http://www.econstor.eu/handle/10419/97018 (2014).

Santos Silva, J. M. C. and S. Tenreyro, "The Log of Gravity", The Review of Econoomics and Statsitics **88**, 4, 641–658 (2006).

Savage, T. and H. Vo, "Yellow cabs as red corpuscles", in "2013 IEEE International Conference on Big Data", pp. 22–28 (2013).

Sayer, R. A., "Gravity models and spatial autocorrelation, or atrophy in urban and regional modelling", Area pp. 183–189, URL http://www.jstor.org/stable/20001229 (1977).

Scherngell, T. and R. Lata, "Towards an integrated European Research Area? Findings from Eigenvector spatially filtered spatial interaction models using European Framework Programme data*: Towards an integrated European Research Area?", Papers in Regional Science pp. no–no, URL http://doi.wiley.com/10.1111/j.1435-5957.2012.00419.x (2012).

Schmitt, R. R. and D. L. Greene, "An alternative derivation of the intervening opportunities model", Geographical Analysis **10**, 1, 73–77, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1978.tb00646.x/abstract (1978).

Schneider, T. W., "Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance", URL http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/ (2015).

Schuijbroek, J., R. Hampshire and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems", Working Paper URL http://repository.cmu.edu/tepper/1491/?utm_source=repository.cmu.edu/tepper/1491&utm_medium=PDF&utm_campaign=PDFCoverPages (2013).

Sellner, R., M. M. Fischer and M. Koch, "A Spatial Autoregressive Poisson Gravity Model: A SAR Poisson Gravity Model", Geographical Analysis **45**, 2, 180–201, URL http://doi.wiley.com/10.1111/gean.12007 (2013).

Sen, A. and T. Smith, *Gravity Models of Spatial Interaction Behavior*, Advances in Spatial and Network Economics (Springer, 1995).

Seya, H., D. Murakami, M. Tsutsumi and Y. Yamagata, "Application of LASSO to the Eigenvector Selection Problem in Eigenvector-based Spatial Filtering: Application of LASSO", Geographical Analysis **47**, 3, 284–299, URL http://doi.wiley.com/10.1111/gean.12054 (2015).

Shaheen, S. A., S. Guzman and H. Zhang, "Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future", Transportation Research Record: Journal of the Transportation Research Board **2143**, -1, 159–167, URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2143-20 (2010).

Shekhar, R., "Geospatial Operations at Scale with Dask and Geopandas", URL https://medium.com/towards-data-science/geospatial-operations-at-scale-with-dask-and-geopandas-4d92d00eb7e8 (2017).

Sheppard, E., "Theoretical Underpinnings of the Gravity Hypothesis", Geogaphical Analysis **10**, 4, 386–402 (1979a).

Sheppard, E., "Distance-decay parameters: A comment", Annals of the Association of American Geographers **72**, 4, 549–550 (1982).

Sheppard, E., "The Distance-decay Gravity Model Debate", in "Spatial Statistics and models", pp. 367–384 (Dreidel, Boston, 1984).

Sheppard, E. S., "Gravity parameter estimation", Geographical Analysis **11**, 2, 120–132, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1979.tb00681.x/abstract (1979b).

Sheppard, E. S., "Notes On Spatial Interaction", The Professional Geographer **31**, 1, 8–15, URL http://www.tandfonline.com/doi/abs/10.1111/j.0033-0124.1979.00008.x (1979c).

Sheppard, E. S., D. A. Griffith and L. Curry, "A final comment on mis-specification and autocorrelation in those gravity parameters", Regional Studies **10**, 3, 337–339, URL http://www.tandfonline.com/doi/abs/10.1080/09595237600185341 (1976).

Shu, J., M. C. Chou, Q. Liu, C.-P. Teo and I.-L. Wang, "Models for Effective Deployment and Redistribution of Bicycles Within Public Bicycle-Sharing Systems", Operations Research **61**, 6, 1346–1359, URL http://pubsonline.informs.org/doi/abs/10.1287/opre.2013.1215 (2013).

Sila-Nowicka, K., *Using GPS to further understanding of spatial behavior*, Ph.D. thesis, University of St Andrews, St. Andrews (2016).

Simini, F., M. C. González, A. Maritan and A.-L. Barabási, "A universal model for mobility and migration patterns", Nature **484**, 7392, 96–100, URL http://www.nature.com/nature/journal/v484/n7392/full/nature10856.html (2012).

Siła-Nowicka, K., J. Vandrol, T. Oshan, J. A. Long, U. Demšar and A. S. Fotheringham, "Analysis of human mobility patterns from GPS trajectories and contextual information", International Journal of Geographical Information Science **30**, 5, 881–906, URL http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1100731 (2016).

Smith, S., L.J., "Intervening opportunities and travel to urban recreational centers", Journal of Leisure Research **12**, 4, 296–308 (1980).

Song, C., T. Koren, P. Wang and A.-L. Barabási, "Modelling the scaling properties of human mobility", Nature Physics **6**, 10, 818–823, URL http://www.nature.com/nphys/journal/v6/n10/full/nphys1760.html (2010a).

Song, C., Z. Qu, N. Blumm and A.-L. Barabási, "Limits of Predictability in Human Mobility", Science **327**, 5968, 1018–1021, URL http://www.sciencemag.org/content/327/5968/1018 (2010b).

Stouffer, S., A., "Intervening opportunities and competing migrants", The Journal of Regional Science **2**, 1, 1–26 (1960).

Stouffer, S. A., "Intervening Opportunities: A Theory Relating Mobility and Distance", American Sociological Review **5**, 6, 845–867, URL http://www.jstor.org/stable/2084520 (1940).

Tagashira, N. and A. Okabe, "The Modifiable Areal Unit Problem in a Repression Model Whose Independent Variable Is a Distance from a Predetermined Point", Geographical Analysis **34**, 1, 1–20, URL http://doi.wiley.com/10.1353/geo.2002.0006 (2002).

Tao, R. and J.-C. Thill, "Spatial Cluster Detection in Spatial Flow Data: Spatial Cluster Detection", Geographical Analysis URL http://doi.wiley.com/10.1111/gean.12100 (2016).

Thill, J.-C., "Choice set formation for destination choice modelling", Progress in Human Geography **16**, 3, 361–382, URL http://search.proquest.com.ezproxy1.lib.asu.edu/docview/810628368?pq-origsite=summon (1992).

Thill, J.-C. and J. L. Horowitz, "Travel-Time Constraints on Destination-Choice Sets", Geographical Analysis **29**, 2, 108–123, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1997.tb00951.x/abstract (1997).

Thorsen, I. and J. P. Gitlesen, "Empirical Evaluation of Alternative Model Specifications to Predict Commuting Flows", Journal of Regional Science **38**, 2, 273–292, URL http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0472443&site=ehost-live (1998).

Tiefelsdorf, M., "The Saddlepoint Approximation of Moran's I's and Local Moran's I i's Reference Distributions and Their Numerical Evaluation", Geographical Analysis **34**, 3, 187–206, URL http://doi.wiley.com/10.1353/geo.2002.0018 (2002).

Tiefelsdorf, M., "Misspecifications in interaction model distance decay relations: A spatial structure effect", Journal of Geographical Systems **5**, 1, 25–50, URL http://link.springer.com/article/10.1007/s101090300102 (2003).

Tiefelsdorf, M. and B. Boots, "The specification of constrained interaction models using the SPSS loglinear procedure", Geographical Systems **2**, 21–38 (1995).

Tiefelsdorf, M. and D. A. Griffith, "Semiparametric filtering of spatial autocorrelation: the eigenvector approach", Environment and Planning A **39**, 5, 1193–1221, URL http://www.envplan.com/abstract.cgi?id=a37378 (2007).

TLC, N., "TLC Trip Record Data", URL http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml (2017).

Tran, N., N. Wilson and D. Hite, "Choosing the best model in the presence of zero trade: a fish product analysis", Nontariff Measures with Market Imperfections: Trade and Welfare Implications (Frontiers of Economics and Globalization, Volume 12), edited by JC Beghin pp. 127–48, URL http://books.google.com/books?hl=en&lr=&id=vHHNZ4hxyroC&oi=fnd&pg=PA127&dq=%22the+consistency+of+estimates+is%22+%22performs+well+if+one+can+find+true+excluded+variables.+However,+Liu+(2009)+argues+that+since%22+%22applied+economic+research+has+explored+alternative+specifications+to+address%22+&ots=moc4cPpk_-&sig=Qw8NUsDgug5L__oRy4rIi7pXjLw (2013).

Tsutsumi, M. and K. Tamesue, "Intraregional flow problem in spatial econometric model for origin–destination flows", Environment and Planning B: Planning and Design **39**, 6, 1006–1015, URL http://www.envplan.com/abstract.cgi?id=b38029 (2012).

Ulyssea-Neto, I., "The development of a new gravity-opportunity model for trip distribution", Environment and Planning A **25**, 817–826, URL http://www.envplan.com/epa/fulltext/a25/a250817.pdf (1993).

United Nations and Department of Economic and Social Affairs, *World urbanization prospects, the 2014 revision: highlights* (2014), URL http://proxy.uqtr.ca/login.cgi?action=login&u=uqtr&db=ebsco&ezurl=http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=857993, oCLC: 973117582.

Vandenbulcke, G., C. Dujardin, I. Thomas, B. d. Geus, B. Degraeuwe, R. Meeusen and L. I. Panis, "Cycle commuting in Belgium: Spatial determinants and 're-cycling' strategies", Transportation Research Part A: Policy and Practice **45**, 2, 118–137, URL http://linkinghub.elsevier.com/retrieve/pii/S0965856410001588 (2011).

Vaona, A., "Spatial autocorrelation and the sensitivity of RESET: a simulation study", Journal of Geographical Systems **12**, 1, 89–103, URL http://link.springer.com/10.1007/s10109-009-0093-9 (2010).

Vogel, P., T. Greiser and D. C. Mattfeld, "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns", Procedia - Social and Behavioral Sciences **20**, 514–523, URL http://www.sciencedirect.com/science/article/pii/S1877042811014388 (2011).

Vries, J. J. D., P. Nijkamp and P. Rietveld, "Exponential or power distance-decay for commuting? An alternative specification", Environment and Planning A **41**, 2, 461–480, URL http://www.envplan.com/abstract.cgi?id=a39369 (2009).

Wang, J., Y. Mao, J. Li, Z. Xiong and W.-X. Wang, "Predictability of Road Traffic and Congestion in Urban Areas", PLoS ONE **10**, 4, e0121825, URL http://dx.doi.org/10.1371/journal.pone.0121825 (2015).

Wang, X.-W., X.-P. Han and B.-H. Wang, "Correlations and Scaling Laws in Human Mobility", PLoS ONE **9**, 1, e84954, URL http://dx.doi.org/10.1371/journal.pone.0084954 (2014).

Ward, J. S. and A. Barker, "Undefined By Data: A Survey of Big Data Definitions", arXiv:1309.5821 [cs] URL http://arxiv.org/abs/1309.5821, arXiv: 1309.5821 (2013).

Wardman, M., M. Tight and M. Page, "Factors influencing the propensity to cycle to work", Transportation Research Part A: Policy and Practice **41**, 4, 339–350, URL http://www.sciencedirect.com/science/article/pii/S0965856406001212 (2007).

Webber, M. J., "A Theoretical Analysis of Aggregation in Spatial Interaction Models", Geographical Analysis **12**, 2, 129–141, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1980.tb00023.x/abstract (1980).

Whong, C., "FOILing NYC's *Boro* Taxi Trip Data", URL http://chriswhong.com/open-data/foiling-nycs-boro-taxi-trip-data/ (2014).

Wills, M., "A flexible gravity-opportunity model for trip distribution", Transportation Research **20B**, 2, 89–111 (1986).

Wilson, A., "Boltzmann, Lotka and Volterra and spatial structural evolution: an integrated methodology for some dynamical systems", Journal of The Royal Society Interface **5**, 25, 865–871, URL http://rsif.royalsocietypublishing.org/content/5/25/865 (2008).

Wilson, A., "Entropy in Urban and Regional Modelling: Retrospect and Prospect", Geographical Analysis **42**, 4, 364–394, URL http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.2010.00799.x/full (2010a).

Wilson, A., "Urban and regional dynamics from the global to the local: hierarchies, 'DNA', and 'genetic' planning", Environment and Planning B: Planning and Design **37**, 5, 823–837, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=b36141 (2010b).

Wilson, A. G., "A statistical theory of spatial distribution models", Transportation Research **1**, 253–269, URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1221809 (1967).

Wilson, A. G., "Use of Entropy Maximizing Models im theory of trip distribution, mode splot, and route split.", Journal of Transport Economics **3**, 1, 108–126, URL http://www.bath.ac.uk/e-journals/jtep/pdf/Volume_111_No_1_108-126.pdf (1969).

Wilson, A. G., "Advances and problems in distribution modelling", Transportation Research **4**, 1–18, URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1221818 (1970).

Wilson, A. G., "A family of spatial interaction models, and associated developments", Environment and Planning A **3**, 1–32, URL https://illiad.lib.asu.edu/illiad/illiad.dll?Action=10&Form=75&Value=1221819 (1971).

Wilson, A. G., "Further developments of entropy maximising transport models", Transportation Planning and Technology **1**, 3, 183–193, URL http://www.tandfonline.com/doi/abs/10.1080/03081067308717045 (1973).

Wolf, L. J., T. M. Oshan and A. S. Fotheringham, "Single and Multiscale Models of Process Spatial Heterogeneity: Single and Multiscale Models", Geographical Analysis URL http://doi.wiley.com/10.1111/gean.12147 (2017).

Wood, J., J. Dykes and A. Slingsby, "Visualisation of Origins, Destinations and Flows with OD Maps", The Cartographic Journal **47**, 2, 117–129, URL http://www.maneyonline.com/doi/abs/10.1179/000870410X12658023467367 (2010).

Wood, J., A. Slingsby and J. Dykes, "Visualizing the dynamics of London's bicycle hire scheme", Cartographica **46**, 4, 239–251, URL zotero://attachment/526/ (2011).

Wu, L., Y. Zhi, Z. Sui and Y. Liu, "Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data", PLoS ONE **9**, 5, e97010, URL http://dx.doi.org/10.1371/journal.pone.0097010 (2014).

Xiao, N. and Y. Chun, "Visualizing Migration Flows Using Kriskograms", Cartography and Geographic Information Science **36**, 2, 183–191, URL http://www.tandfonline.com/doi/abs/10.1559/152304009788188763 (2009).

Yamada, I. and J.-C. Thill, "Local Indicators of Network-Constrained Clusters in Spatial Patterns Represented by a Link Attribute", Annals of the Association of American Geographers **100**, 2, 269–285, URL http://www.tandfonline.com/doi/abs/10.1080/00045600903550337 (2010).

Yan, J. and J.-C. Thill, "Visual data mining in spatial interaction analysis with self-organizing maps", Environment and Planning B: Planning and Design **36**, 3, 466–486, URL http://www.envplan.com.ezproxy1.lib.asu.edu/abstract.cgi?id=b34019 (2009).

Yan, X.-Y., C. Zhao, Y. Fan, Z. Di and W.-X. Wang, "Universal Predictability of Mobility Patterns in Cities", arXiv:1307.7502 [physics] URL http://arxiv.org/abs/1307.7502, arXiv: 1307.7502 (2013).

Yang, Y., C. Herrera, N. Eagle and M. C. González, "Limits of Predictability in Commuting Flows in the Absence of Data for Calibration", Scientific Reports **4**, URL http://www.nature.com/srep/2014/140711/srep05662/full/srep05662.html (2014).

Yano, K., T. Nakaya, A. S. Fotheringham, S. Openshaw and Y. Ishikawa, "A comparison of migration behaviour in Japan and Britain using spatial interaction models", International Journal of Population Geography **9**, 5, 419–431, URL http://doi.wiley.com/10.1002/ijpg.297 (2003).

Yoon, J. W., F. Pinelli and F. Calabrese, "Cityride: A Predictive Bike Sharing Journey Advisor", in "2012 IEEE 13th International Conference on Mobile Data Management", pp. 306–311 (IEEE, 2012), URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6341407.

Yue, Y., H.-d. Wang, B. Hu, Q.-q. Li, Y.-g. Li and A. G. O. Yeh, "Exploratory calibration of a spatial interaction model using taxi GPS trajectories", Computers, Environment and Urban Systems **36**, 2, 140–153, URL http://www.sciencedirect.com/science/article/pii/S0198971511000901 (2012).

Yufei, H. A. N., L. Oukhellou and E. Come, "Towards bicycle demand prediction of large-scale bicycle sharing system", 93rd Annual Meeting of the Transportation Research Board URL http://www.comeetie.fr/pdfrepos/TRBYUFEI.pdf (2014).

Zaltz Austwick, M., O. O'Brien, E. Strano and M. Viana, "The Structure of Spatial Networks and Communities in Bicycle Sharing Systems", PLoS ONE **8**, 9, e74685, URL http://dx.doi.org/10.1371/journal.pone.0074685 (2013).

Zheng, Z., S. Rasouli and H. Timmermans, "Two-regime Pattern in Human Mobility: Evidence from GPS Taxi Trajectory Data: Two-regime Pattern in Human Mobility", Geographical Analysis pp. n/a–n/a, URL http://doi.wiley.com/10.1111/gean.12087 (2015).

Zhong, C., S. M. Arisona, X. Huang, M. Batty and G. Schmitt, "Detecting the dynamics of urban structure through spatial network analysis", International Journal of Geographical Information Science **28**, 11, 2178–2199, URL http://www.tandfonline.com/doi/full/10.1080/13658816.2014.914521 (2014).

Zhu, X. and D. Guo, "Mapping Large Spatial Flow Data with Hierarchical Clustering: Mapping Large Spatial Flow Data with Hierarchical Clustering", Transactions in GIS **18**, 3, 421–435, URL http://doi.wiley.com/10.1111/tgis.12100 (2014).

Zhu, X. and D. Guo, "Urban event detection with big data of taxi OD trips: A time series decomposition approach: ZHU and GUO", Transactions in GIS **21**, 3, 560–574, URL http://doi.wiley.com/10.1111/tgis.12288 (2017).

APPENDIX A

EXPLORING THE BOX-COX METHODOLOGY

1 Replicating and Validating Analysis from Tiefelsdorf (2003)

Several analyses from Tiefelsdorf (2003) replicated here to further explore his unexpected results and test the boxcox method.

```
In [2]: #Load libraries
        import numpy as np
        import scipy.stats as stats
        from statsmodels.api import formula as smf
        from pysal.contrib.spint import gravity as grav
        import pandas as pd
        from statsmodels.api import families
        from patsy.contrasts import Sum
        import scipy.spatial.distance as spatial
        import geopandas as gp
        import matplotlib.pyplot as plt
        %pylab inline
        import scipy as sp
```

```
Populating the interactive namespace from numpy and matplotlib
```

1.1 First, the simulation results are replicated (p. 40) to validate model specification specification that uses origin-specific and destination-specific effects for boxcox transformed distance.

1.1.1 Recreate data from Tiefelsdorf's simulation using q=2 and q=0

```
In [68]: oi = np.array([10,10,10,10,10,10,
                        10,10,10,10,10,10,
                        10,10,10,10,10,10,
                        10,10,10,10,10,10,
                        10,10,10,10,10,10,
                        10,10,10,10,10,10,
                        10,10,10,10,10,10])

In [69]: dj = np.array([10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,
                        10,10,10,10,10,10,10])

In [70]: dij = np.array([1.0,2.0,3.0,4.0,5.0,6.0,
                         1.0,1.0,2.0,3.0,4.0,5.0,
                         2.0,1.0,1.0,2.0,3.0,4.0,
                         3.0,2.0,1.0,1.0,2.0,3.0,
                         4.0,3.0,2.0,1.0,1.0,2.0,
                         5.0,4.0,3.0,2.0,1.0,1.0,
                         6.0,5.0,4.0,3.0,2.0,1.0])
```

```
In [71]: origins = np.array([1,1,1,1,1,1,
                             2,2,2,2,2,2,
                             3,3,3,3,3,3,
                             4,4,4,4,4,4,
                             5,5,5,5,5,5,
                             6,6,6,6,6,6,
                             7,7,7,7,7,7])

In [72]: dests = np.array([2,3,4,5,6,7,
                           1,3,4,5,6,7,
                           1,2,4,5,6,7,
                           1,2,3,5,6,7,
                           1,2,3,4,6,7,
                           1,2,3,4,5,7,
                           1,2,3,4,5,6])

In [73]: flows = np.exp(np.log(oi) + \
                        np.log(dj) + -1.0*dij) + \
                        np.random.normal(0, .0001)

In [74]: data = pd.DataFrame({'Data':flows,
                              'Origin': origins,
                              'Oi': oi,
                              'Dj': dj,
                              'Dij':dij,
                              'Dest':dests})

In [75]: data['Origin'] = data['Origin'].astype(str)
         data['Dest'] = data['Dest'].astype(str)
         data['q2'] = (data['Dij'].astype(float)**2.0 -1.0)/2.0
         data['q0'] = np.log((data['Dij'].astype(float)))
```

1.1.2 Estimate Tiefelsdorf model using categorical interactions for local distance decay (deviation coding)

boxcox: q = 2

```
In [76]: q2 = smf.glm('Data~C(Origin, Sum):q2+C(Dest, Sum):q2+np.log(Oi)+np.log(Dj)',
                      data=data, family=families.Poisson()).fit()

         print 'Origin specific distance-decay estimates for q = 2'
         print ''
         print q2.params[1:8]
```

```
Origin specific distance-decay estimates for q = 2

C(Origin, Sum)[mean]:q2   -0.480763
C(Origin, Sum)[S.1]:q2     0.103351
C(Origin, Sum)[S.2]:q2     0.043551
C(Origin, Sum)[S.3]:q2    -0.076930
C(Origin, Sum)[S.4]:q2    -0.139945
C(Origin, Sum)[S.5]:q2    -0.076930
C(Origin, Sum)[S.6]:q2     0.043551
dtype: float64
```

boxcox: q = 0

```
In [77]: q0 = smf.glm('Data~C(Origin, Sum):q0+C(Dest, Sum):q0+np.log(Oi)+np.log(Dj)',
                      data=data, family=families.Poisson()).fit()

         print 'Origin specific distance-decay estimates for q = 0'
         print ''
         print q0.params[1:8]
```

```
Origin specific distance-decay estimates for q = 0

C(Origin, Sum)[mean]:q0   -1.855542
C(Origin, Sum)[S.1]:q0    -0.314918
C(Origin, Sum)[S.2]:q0    -0.168438
C(Origin, Sum)[S.3]:q0     0.260716
C(Origin, Sum)[S.4]:q0     0.445282
C(Origin, Sum)[S.5]:q0     0.260716
C(Origin, Sum)[S.6]:q0    -0.168438
dtype: float64
```

Comparing these results for q = 2 and q = 0 to those presented in Tiefelsdorf (2003) we can see that our results exactly match, so we know we have the correct model specification.

1.2 Now Tiefelsdorf's migration example is replicated. The same data is used, thought we do not have population weighted centroids. Nevertheless, it will be seen that the results are very similar

1.2.1 Prepare data

In [78]:
```
#read migration data
migration_90 = pd.read_csv('1990_migration.csv', index_col='state')

#remove data pertaining to Alaska and Hawaii
migration_90 = migration_90[migration_90.index!='Alaska']
migration_90 = migration_90[migration_90.index!='Hawaii']
states = migration_90.index.to_native_types()
states[7] = 'District of Columbia'
migration_90.drop('Alaska', axis=1, inplace=True)
migration_90.drop('Hawaii', axis=1, inplace=True)

#reformat migration data
migration_90 = migration_90.as_matrix().astype(float)
flows = migration_90.reshape((-1,))
```

In [79]:
```
#read population data
pop = pd.read_csv('1990_pop.csv', index_col='state')

#remove data pertaining to Alaska and Hawaii
pop = pop[pop.index!='Alaska']
pop = pop[pop.index!='Hawaii']

#read states shapefile, reproject, and sort by alphabetically
state_shp = gp.read_file('tl_2014_us_state.shp')
state_shp = state_shp[state_shp['NAME'].isin(states)]
state_shp = state_shp.to_crs(epsg='3857')
state_shp.sort_values('NAME', inplace=True)
```

In [80]:
```
# prep origin and destinatio variables for flow dyads

origins = np.repeat(states, 49)
Oi = np.repeat(pop.values, 49)

dests = np.tile(states, 49)
Dj = np.tile(pop.values.flatten(), 49).reshape((-1,))

origin_geom = np.repeat(state_shp.centroid, 49)
dest_geom = np.tile(state_shp.centroid, 49).flatten()
dest_geom = pd.DataFrame({'dests': dests, 'dest_geom': dest_geom})
dest_geom = gp.GeoDataFrame(dest_geom,
                            geometry='dest_geom',
```

```
                                    crs={'init': 'epsg:3857',
                                         'no_defs': True})
```

```
In [16]: #concatenate all the data into a dataframe

         mig_90 = pd.DataFrame({'Data': flows,
                                'Origin': origins,
                                'Destination': dests,
                                'Oi': Oi,
                                'Dj': Dj,
                                'origin_geom':origin_geom})

         mig_90 = gp.GeoDataFrame(mig_90,
                                  geometry='origin_geom',
                                  crs={'init': 'epsg:3857',
                                       'no_defs': True})

         mig_90.reset_index(drop=True, inplace=True)
```

```
In [17]: #compute distance
         mig_90['Dij'] = mig_90.distance(dest_geom)
```

```
In [18]: #remove intra-zonal flows
         mig_90 = mig_90[mig_90['Origin'] != mig_90['Destination']]
```

```
In [19]: #apply boxcox transform with q = -0.2 to distances
         mig_90['boxcox'] = stats.boxcox(mig_90['Dij'], -.2)
```

1.2.2 Lets first estimate his model that includes an origin-specific distance-decay estimate, a destination-specific distance-decay estimate, and a single estimate for origin and destination population.

```
In [20]: #Tiefelsdorf's model
         model = smf.glm('Data~C(Origin, Sum):boxcox+C(Destination,
         Sum):boxcox+np.log(Oi)+np.log(Dj)',
                         data=mig_90, family=families.Poisson()).fit()
         df = model.params.to_frame()

         with pd.option_context('display.max_rows', None, 'display.max_columns', 3):
             print(df)
```

```
                                                          0
Intercept                                        173.668230
C(Origin, Sum)[mean]:boxcox                      -17.956129
C(Origin, Sum)[S.Alabama]:boxcox                   0.160532
C(Origin, Sum)[S.Arizona]:boxcox                   0.323655
C(Origin, Sum)[S.Arkansas]:boxcox                 -0.228826
C(Origin, Sum)[S.California]:boxcox                1.901954
C(Origin, Sum)[S.Colorado]:boxcox                  0.174147
C(Origin, Sum)[S.Connecticut]:boxcox              -0.059304
C(Origin, Sum)[S.Delaware]:boxcox                 -1.308107
C(Origin, Sum)[S.District of Columbia]:boxcox     -1.579128
C(Origin, Sum)[S.Florida]:boxcox                   1.359368
C(Origin, Sum)[S.Georgia]:boxcox                   0.640888
C(Origin, Sum)[S.Idaho]:boxcox                    -0.778474
C(Origin, Sum)[S.Illinois]:boxcox                  0.946250
C(Origin, Sum)[S.Indiana]:boxcox                   0.379437
C(Origin, Sum)[S.Iowa]:boxcox                     -0.155303
C(Origin, Sum)[S.Kansas]:boxcox                   -0.154231
C(Origin, Sum)[S.Kentucky]:boxcox                  0.041982
C(Origin, Sum)[S.Louisiana]:boxcox                 0.138122
C(Origin, Sum)[S.Maine]:boxcox                    -0.668079
```

```
C(Origin, Sum)[S.Maryland]:boxcox                     0.249410
C(Origin, Sum)[S.Massachusetts]:boxcox                0.437842
C(Origin, Sum)[S.Michigan]:boxcox                     0.785607
C(Origin, Sum)[S.Minnesota]:boxcox                    0.273698
C(Origin, Sum)[S.Mississippi]:boxcox                 -0.230062
C(Origin, Sum)[S.Missouri]:boxcox                     0.372460
C(Origin, Sum)[S.Montana]:boxcox                     -0.989473
C(Origin, Sum)[S.Nebraska]:boxcox                    -0.546235
C(Origin, Sum)[S.Nevada]:boxcox                      -0.560289
C(Origin, Sum)[S.New Hampshire]:boxcox               -0.762817
C(Origin, Sum)[S.New Jersey]:boxcox                   0.645933
C(Origin, Sum)[S.New Mexico]:boxcox                  -0.499373
C(Origin, Sum)[S.New York]:boxcox                     1.260408
C(Origin, Sum)[S.North Carolina]:boxcox               0.630077
C(Origin, Sum)[S.North Dakota]:boxcox                -1.229284
C(Origin, Sum)[S.Ohio]:boxcox                         0.888105
C(Origin, Sum)[S.Oklahoma]:boxcox                    -0.011182
C(Origin, Sum)[S.Oregon]:boxcox                       0.064339
C(Origin, Sum)[S.Pennsylvania]:boxcox                 0.943050
C(Origin, Sum)[S.Rhode Island]:boxcox                -1.097635
C(Origin, Sum)[S.South Carolina]:boxcox               0.104833
C(Origin, Sum)[S.South Dakota]:boxcox                -1.173425
C(Origin, Sum)[S.Tennessee]:boxcox                    0.353824
C(Origin, Sum)[S.Texas]:boxcox                        1.380090
C(Origin, Sum)[S.Utah]:boxcox                        -0.445062
C(Origin, Sum)[S.Vermont]:boxcox                     -1.392153
C(Origin, Sum)[S.Virginia]:boxcox                     0.600860
C(Origin, Sum)[S.Washington]:boxcox                   0.546039
C(Origin, Sum)[S.West Virginia]:boxcox               -0.583558
C(Origin, Sum)[S.Wisconsin]:boxcox                    0.289591
C(Destination, Sum)[S.Alabama]:boxcox                 0.087461
C(Destination, Sum)[S.Arizona]:boxcox                 0.212637
C(Destination, Sum)[S.Arkansas]:boxcox               -0.243473
C(Destination, Sum)[S.California]:boxcox              1.741744
C(Destination, Sum)[S.Colorado]:boxcox                0.191695
C(Destination, Sum)[S.Connecticut]:boxcox             0.001977
C(Destination, Sum)[S.Delaware]:boxcox               -1.288698
C(Destination, Sum)[S.District of Columbia]:boxcox   -1.427668
C(Destination, Sum)[S.Florida]:boxcox                 1.125218
C(Destination, Sum)[S.Georgia]:boxcox                 0.460639
C(Destination, Sum)[S.Idaho]:boxcox                  -0.718299
C(Destination, Sum)[S.Illinois]:boxcox                0.959362
C(Destination, Sum)[S.Indiana]:boxcox                 0.351919
C(Destination, Sum)[S.Iowa]:boxcox                   -0.064706
C(Destination, Sum)[S.Kansas]:boxcox                 -0.125325
C(Destination, Sum)[S.Kentucky]:boxcox                0.030959
C(Destination, Sum)[S.Louisiana]:boxcox               0.259956
C(Destination, Sum)[S.Maine]:boxcox                  -0.672630
C(Destination, Sum)[S.Maryland]:boxcox                0.194508
C(Destination, Sum)[S.Massachusetts]:boxcox           0.462526
C(Destination, Sum)[S.Michigan]:boxcox                0.783468
C(Destination, Sum)[S.Minnesota]:boxcox               0.254635
C(Destination, Sum)[S.Mississippi]:boxcox            -0.209600
C(Destination, Sum)[S.Missouri]:boxcox                0.333352
C(Destination, Sum)[S.Montana]:boxcox                -0.835284
C(Destination, Sum)[S.Nebraska]:boxcox               -0.459640
C(Destination, Sum)[S.Nevada]:boxcox                 -0.701004
C(Destination, Sum)[S.New Hampshire]:boxcox          -0.769210
C(Destination, Sum)[S.New Jersey]:boxcox              0.669383
```

```
C(Destination, Sum)[S.New Mexico]:boxcox             -0.464286
C(Destination, Sum)[S.New York]:boxcox                1.326184
C(Destination, Sum)[S.North Carolina]:boxcox          0.461714
C(Destination, Sum)[S.North Dakota]:boxcox           -1.016626
C(Destination, Sum)[S.Ohio]:boxcox                    0.863405
C(Destination, Sum)[S.Oklahoma]:boxcox                0.064020
C(Destination, Sum)[S.Oregon]:boxcox                 -0.016879
C(Destination, Sum)[S.Pennsylvania]:boxcox            0.907927
C(Destination, Sum)[S.Rhode Island]:boxcox           -1.023301
C(Destination, Sum)[S.South Carolina]:boxcox         -0.017305
C(Destination, Sum)[S.South Dakota]:boxcox           -1.037689
C(Destination, Sum)[S.Tennessee]:boxcox               0.246767
C(Destination, Sum)[S.Texas]:boxcox                   1.330532
C(Destination, Sum)[S.Utah]:boxcox                   -0.399505
C(Destination, Sum)[S.Vermont]:boxcox                -1.337763
C(Destination, Sum)[S.Virginia]:boxcox                0.485051
C(Destination, Sum)[S.Washington]:boxcox              0.415370
C(Destination, Sum)[S.West Virginia]:boxcox          -0.470879
C(Destination, Sum)[S.Wisconsin]:boxcox               0.295156
np.log(Oi)                                           -2.840261
np.log(Dj)                                           -2.599338
```

While the population weighted centroids and trend surfaces that Tiefelsdorf uses are not available, the model on the basic population and distance data results in local distance coefficient estimates that are very similar in sign and magnitude. Even where there are differences in signs, it is important to note that these are local coefficients that are centered on a global mean coefficients and variations in the local coefficients are therefore very slight. More importantly, it can be seen above that the unexpected negative coefficients for the origin and destination are replicated.

1.3 Now lets estimate the same model but with only an origin-specific local distance-decay.

```
In [21]: model = smf.glm('Data~C(Origin, Sum):boxcox+np.log(Oi)+np.log(Dj)',
                         data=mig_90, family=families.Poisson()).fit()
         df = model.params.to_frame()

         with pd.option_context('display.max_rows', None, 'display.max_columns', 3):
             print(df)
```

```
                                                    0
Intercept                                     152.248487
C(Origin, Sum)[mean]:boxcox                   -16.211557
C(Origin, Sum)[S.Alabama]:boxcox                0.298973
C(Origin, Sum)[S.Arizona]:boxcox                0.432889
C(Origin, Sum)[S.Arkansas]:boxcox              -0.384872
C(Origin, Sum)[S.California]:boxcox             3.164634
C(Origin, Sum)[S.Colorado]:boxcox               0.221817
C(Origin, Sum)[S.Connecticut]:boxcox           -0.036357
C(Origin, Sum)[S.Delaware]:boxcox              -2.214747
C(Origin, Sum)[S.District of Columbia]:boxcox  -2.565113
C(Origin, Sum)[S.Florida]:boxcox                2.140246
C(Origin, Sum)[S.Georgia]:boxcox                1.038871
C(Origin, Sum)[S.Idaho]:boxcox                 -1.386434
C(Origin, Sum)[S.Illinois]:boxcox               1.666185
C(Origin, Sum)[S.Indiana]:boxcox                0.690968
```

```
C(Origin, Sum)[S.Iowa]:boxcox                    -0.220273
C(Origin, Sum)[S.Kansas]:boxcox                  -0.269598
C(Origin, Sum)[S.Kentucky]:boxcox                 0.119269
C(Origin, Sum)[S.Louisiana]:boxcox                0.302201
C(Origin, Sum)[S.Maine]:boxcox                   -1.205903
C(Origin, Sum)[S.Maryland]:boxcox                 0.480455
C(Origin, Sum)[S.Massachusetts]:boxcox            0.804996
C(Origin, Sum)[S.Michigan]:boxcox                 1.384898
C(Origin, Sum)[S.Minnesota]:boxcox                0.460126
C(Origin, Sum)[S.Mississippi]:boxcox             -0.343512
C(Origin, Sum)[S.Missouri]:boxcox                 0.652058
C(Origin, Sum)[S.Montana]:boxcox                 -1.730365
C(Origin, Sum)[S.Nebraska]:boxcox                -0.912358
C(Origin, Sum)[S.Nevada]:boxcox                  -1.062662
C(Origin, Sum)[S.New Hampshire]:boxcox           -1.358901
C(Origin, Sum)[S.New Jersey]:boxcox               1.140122
C(Origin, Sum)[S.New Mexico]:boxcox              -0.881860
C(Origin, Sum)[S.New York]:boxcox                 2.228159
C(Origin, Sum)[S.North Carolina]:boxcox           1.040452
C(Origin, Sum)[S.North Dakota]:boxcox            -2.108765
C(Origin, Sum)[S.Ohio]:boxcox                     1.573724
C(Origin, Sum)[S.Oklahoma]:boxcox                 0.009751
C(Origin, Sum)[S.Oregon]:boxcox                   0.042120
C(Origin, Sum)[S.Pennsylvania]:boxcox             1.690851
C(Origin, Sum)[S.Rhode Island]:boxcox            -1.738649
C(Origin, Sum)[S.South Carolina]:boxcox           0.145812
C(Origin, Sum)[S.South Dakota]:boxcox            -2.000920
C(Origin, Sum)[S.Tennessee]:boxcox                0.593037
C(Origin, Sum)[S.Texas]:boxcox                    2.326843
C(Origin, Sum)[S.Utah]:boxcox                    -0.760149
C(Origin, Sum)[S.Vermont]:boxcox                 -2.364460
C(Origin, Sum)[S.Virginia]:boxcox                 0.960692
C(Origin, Sum)[S.Washington]:boxcox               0.809132
C(Origin, Sum)[S.West Virginia]:boxcox           -0.916183
C(Origin, Sum)[S.Wisconsin]:boxcox                0.535825
np.log(Oi)                                       -5.482206
np.log(Dj)                                        0.930544
```

Since the destination-specific distance-decay is removed, the destination population coefficient estimate is no longer negative and is now 0.93, which would be a typical value that is expected. The origin population coefficient estimate is still negative, but origin variables are typically dropped when estimation an orihin-specific model like this.

1.4 We can also try removing the origin-specific distance-decay and only have a destination-specific distance-decay.

```
In [106]: model = smf.glm('Data~C(Destination, Sum):boxcox+np.log(Oi)+np.log(Dj)',
                          data=mig_90, family=families.Poisson()).fit()
          df = model.params.to_frame()

          with pd.option_context('display.max_rows', None, 'display.max_columns', 3):
              print(df)
```

```
                                                    0
Intercept                                  140.496952
C(Destination, Sum)[mean]:boxcox           -15.251498
```

```
C(Destination, Sum)[S.Alabama]:boxcox                 0.236691
C(Destination, Sum)[S.Arizona]:boxcox                 0.314741
C(Destination, Sum)[S.Arkansas]:boxcox               -0.386925
C(Destination, Sum)[S.California]:boxcox              2.881259
C(Destination, Sum)[S.Colorado]:boxcox                0.234018
C(Destination, Sum)[S.Connecticut]:boxcox            -0.005452
C(Destination, Sum)[S.Delaware]:boxcox               -2.107530
C(Destination, Sum)[S.District of Columbia]:boxcox   -2.249567
C(Destination, Sum)[S.Florida]:boxcox                 1.820241
C(Destination, Sum)[S.Georgia]:boxcox                 0.848451
C(Destination, Sum)[S.Idaho]:boxcox                  -1.248326
C(Destination, Sum)[S.Illinois]:boxcox                1.603619
C(Destination, Sum)[S.Indiana]:boxcox                 0.621115
C(Destination, Sum)[S.Iowa]:boxcox                   -0.133930
C(Destination, Sum)[S.Kansas]:boxcox                 -0.238806
C(Destination, Sum)[S.Kentucky]:boxcox                0.109520
C(Destination, Sum)[S.Louisiana]:boxcox               0.417939
C(Destination, Sum)[S.Maine]:boxcox                  -1.177481
C(Destination, Sum)[S.Maryland]:boxcox                0.394777
C(Destination, Sum)[S.Massachusetts]:boxcox           0.772196
C(Destination, Sum)[S.Michigan]:boxcox                1.307596
C(Destination, Sum)[S.Minnesota]:boxcox               0.406518
C(Destination, Sum)[S.Mississippi]:boxcox            -0.302365
C(Destination, Sum)[S.Missouri]:boxcox                0.574100
C(Destination, Sum)[S.Montana]:boxcox                -1.498614
C(Destination, Sum)[S.Nebraska]:boxcox               -0.797514
C(Destination, Sum)[S.Nevada]:boxcox                 -1.124508
C(Destination, Sum)[S.New Hampshire]:boxcox          -1.331216
C(Destination, Sum)[S.New Jersey]:boxcox              1.099371
C(Destination, Sum)[S.New Mexico]:boxcox             -0.803850
C(Destination, Sum)[S.New York]:boxcox                2.189629
C(Destination, Sum)[S.North Carolina]:boxcox          0.848611
C(Destination, Sum)[S.North Dakota]:boxcox           -1.815237
C(Destination, Sum)[S.Ohio]:boxcox                    1.473246
C(Destination, Sum)[S.Oklahoma]:boxcox                0.077994
C(Destination, Sum)[S.Oregon]:boxcox                 -0.015494
C(Destination, Sum)[S.Pennsylvania]:boxcox            1.561410
C(Destination, Sum)[S.Rhode Island]:boxcox           -1.623319
C(Destination, Sum)[S.South Carolina]:boxcox          0.057334
C(Destination, Sum)[S.South Dakota]:boxcox           -1.792218
C(Destination, Sum)[S.Tennessee]:boxcox               0.470417
C(Destination, Sum)[S.Texas]:boxcox                   2.180622
C(Destination, Sum)[S.Utah]:boxcox                   -0.666752
C(Destination, Sum)[S.Vermont]:boxcox                -2.231926
C(Destination, Sum)[S.Virginia]:boxcox                0.814587
C(Destination, Sum)[S.Washington]:boxcox              0.655937
C(Destination, Sum)[S.West Virginia]:boxcox          -0.759251
C(Destination, Sum)[S.Wisconsin]:boxcox               0.500290
np.log(Oi)                                            0.907872
np.log(Dj)                                           -4.967968
```

The results are now the opposite of the previous models where the origin population coefficient estimate is no longer negative and is takes on a typical value and now the distance population variable is negative, thought it should be dropped since this is a destination-specific model.

1.5 Finally, if neither an origin- or destination-specific model is used it can be seen that neither the origin or the destination population coefficient estimates is negative.

```
In [23]: model = smf.glm('Data~boxcox+np.log(Oi)+np.log(Dj)', data=mig_90,
                         family=families.Poisson()).fit()
         df = model.params.to_frame()

         with pd.option_context('display.max_rows', None, 'display.max_columns', 3):
             print(df)
```

```
                     0
Intercept    37.029388
boxcox      -11.698338
np.log(Oi)    0.875737
np.log(Dj)    0.896145
```

1.6 This next experiment demonstrates that mis-specified population functional form can cause the boxcox methodology to over-compensate on the transformation of the distance variable to try to correct for having the wrong functional form of the population variable. This means that Tiefelsdorf's method cannot isolate distance/spatial structure effects. The experiment involves boxcox transforming the destination population variable using q = {-4, -3, -2, -1, 0, 1, 2, 3, 4}. Then flows are generated using Tiefelsdorf's simulated data and a basic gravity model data-generating process. Here, the flows are generated using a logged populations and distance (q = 0), which implies power functions. Then, instead of directly using logged destination population, each of the boxcox transformed destination populations are used for the above values of q. For each one, an optimal boxcox transformation on the distance variable is produced using Tiefelsdorf's methodology of optimizing the log likelihood. If the method is robust then it should always return q = 0 for the distance variable indicating the correct functional form of a power function. However, it will be seen that this is not the case.

1.6.1 Prepare the data and all of the boxcox transforms

```
In [103]: oi = np.array([10,10,10,10,10,10,
                         10,10,10,10,10,10,
                         10,20,20,20,20,20,
                         20,30,30,30,30,30,
                         30,20,20,20,20,20,
                         20,10,10,10,10,10,
                         10,10,10,10,10,10])

          dj = np.array([10,20,30,20,10,10,
                         10,20,30,20,10,10,
                         10,10,30,20,10,10,
                         10,10,20,20,10,10,
                         10,10,20,30,10,10,
                         10,10,20,30,20,10,
                         10,10,20,30,20,10])

          dij = np.array([1.0,2.0,3.0,4.0,5.0,6.0,
                          1.0,1.0,2.0,3.0,4.0,5.0,
                          2.0,1.0,1.0,2.0,3.0,4.0,
                          3.0,2.0,1.0,1.0,2.0,3.0,
                          4.0,3.0,2.0,1.0,1.0,2.0,
                          5.0,4.0,3.0,2.0,1.0,1.0,
```

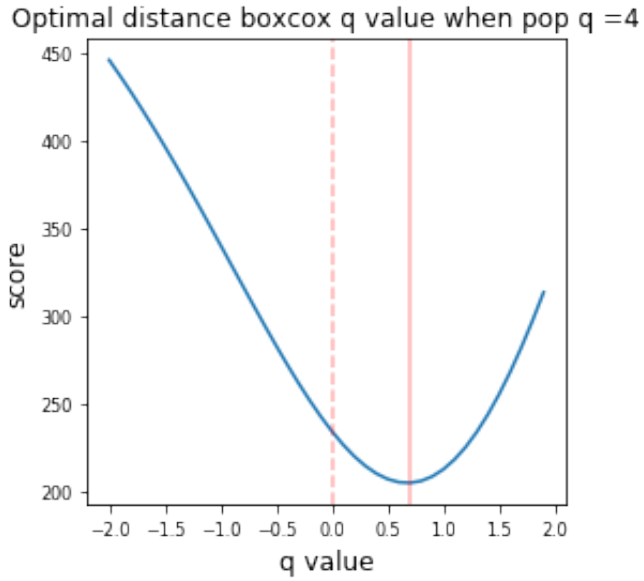
```
                    6.0,5.0,4.0,3.0,2.0,1.0])

origins = np.array([1,1,1,1,1,1,
                    2,2,2,2,2,2,
                    3,3,3,3,3,3,
                    4,4,4,4,4,4,
                    5,5,5,5,5,5,
                    6,6,6,6,6,6,
                    7,7,7,7,7,7])

dests = np.array([2,3,4,5,6,7,
                  1,3,4,5,6,7,
                  1,2,4,5,6,7,
                  1,2,3,5,6,7,
                  1,2,3,4,6,7,
                  1,2,3,4,5,7,
                  1,2,3,4,5,6])

#Set dj to have different transforms
q4 = stats.boxcox(dj, 4)
q3 = stats.boxcox(dj, 3)
q2 = stats.boxcox(dj, 2)
q1 = stats.boxcox(dj, 1)
q0 = stats.boxcox(dj, 0)
qn4 = stats.boxcox(dj, -4)
qn3 = stats.boxcox(dj, -3)
qn2 = stats.boxcox(dj, -2)
qn1 = stats.boxcox(dj, -1)



#Simulate data using power function of distance,
#boxcox transformed destination pop and natural (power) origin pop
flows = np.exp(np.log(oi) +\
               np.log(dj) + \
               -1.0*np.log(dij)) + \
               np.random.normal(0, .005)

data = pd.DataFrame({'Data':flows,
                     'Origin': origins,
                     'Oi': oi,
                     'Dj': dj,
                     'Dij':dij,
                     'Dest':dests,
                     'q2': q2,
                     'q0':q0})

data['Origin'] = data['Origin'].astype(str)
data['Dest'] = data['Dest'].astype(str)
```

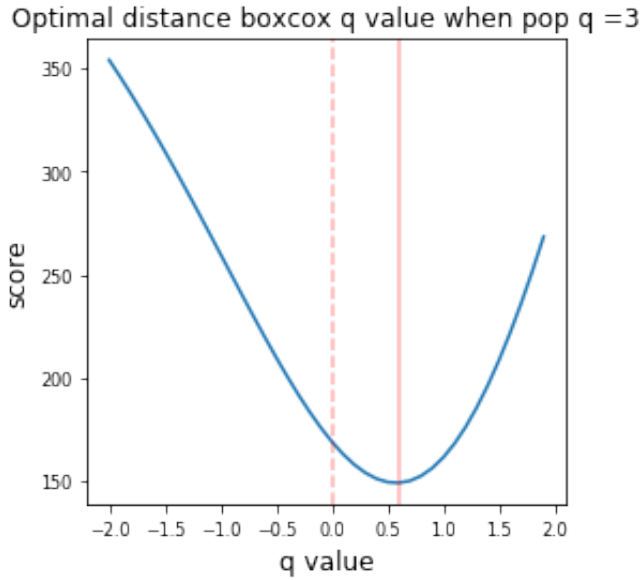
In [93]:
```
#helper function for running each optimization and collecting the results
def plot_scores(Q, i):
    scores = {}
    #Search q values for specification using local distance-decay
    #but global origin and destination population estimates
    for q in np.arange(-2,2, .1):
        data['boxcox'] = stats.boxcox(data['Dij'], q)
        model = smf.glm('Data~C(Origin, Sum):boxcox+np.log(Oi)+'+Q,
                        data=data, family=families.Poisson()).fit()
        scores[q] = model.deviance

    scores = sorted(scores.items(), key=lambda s: s[0])
    x,y = zip(*scores)

    fig, ax = plt.subplots(figsize=(4,4))
    plt.tick_params(labelsize=8)
    plt.plot(x,y)
    plt.title('Optimal distance boxcox q value when pop q =' + str(i))
```

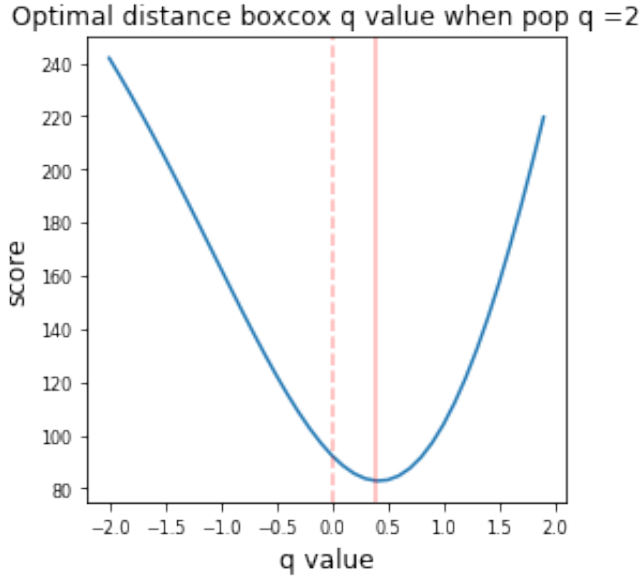
```
    plt.xlabel('q value', size=12)
    plt.ylabel('score', size = 12)


    min_x =  x[np.where(np.array(y) == min(y))[0][0]]
    ax.axvline(min_x, color='r', alpha=.3)
    ax.axvline(0, color='r', linestyle='--', alpha=.3)
    return min_x
```

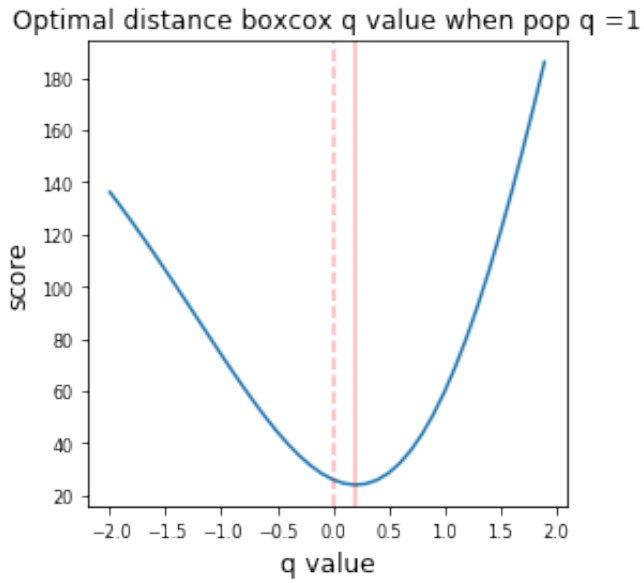
1.6.2 Now a plot is produced for each scenario where the destination population has a misspecified functional form due to the different q values that indicates the score function (blue), the optimal q value that was found (solid red line) and the actual q value that should have been found if the method was robust (dashed red line).

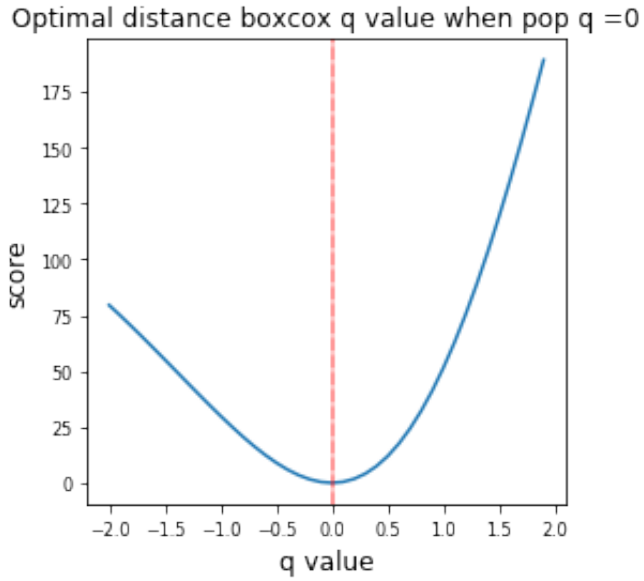
```
In [94]: qs = ['q4', 'q3', 'q2', 'q1', 'q0', 'qn1', 'qn2', 'qn3', 'qn4']
         q_num = [4, 3, 2, 1, 0 , -1, -2, -3, -4]
         mins = []
         for i, q in enumerate(qs):
             mins.append(plot_scores(q, q_num[i]))
```

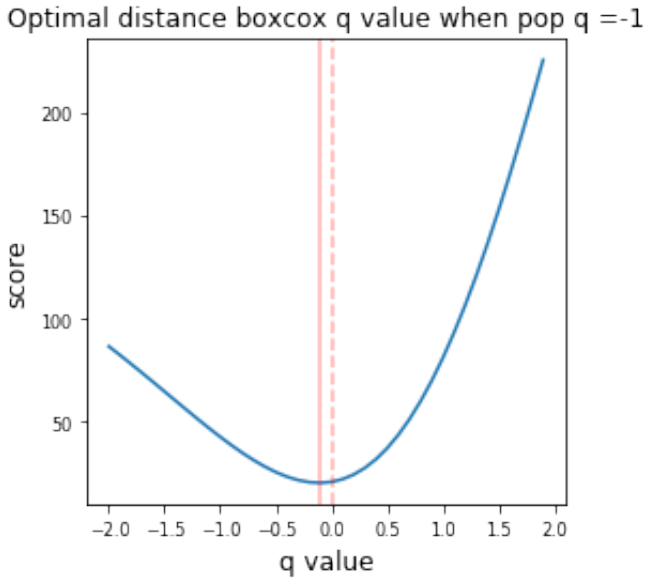


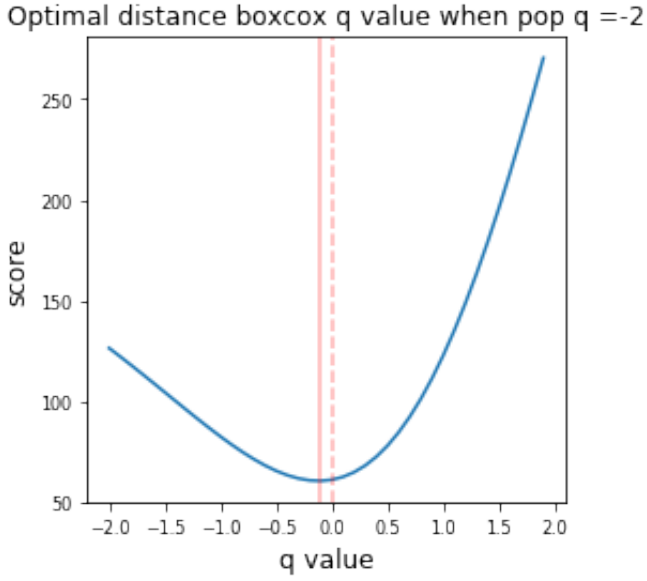

Optimal distance boxcox q value when pop q =3
score
q value
350
300
250
200
150
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

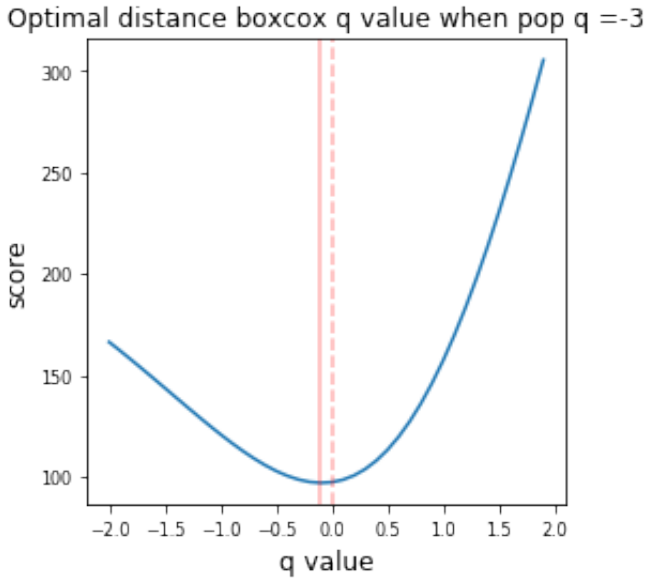

Optimal distance boxcox q value when pop q =2
score
q value
240
220
200
180
160
140
120
100
80
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

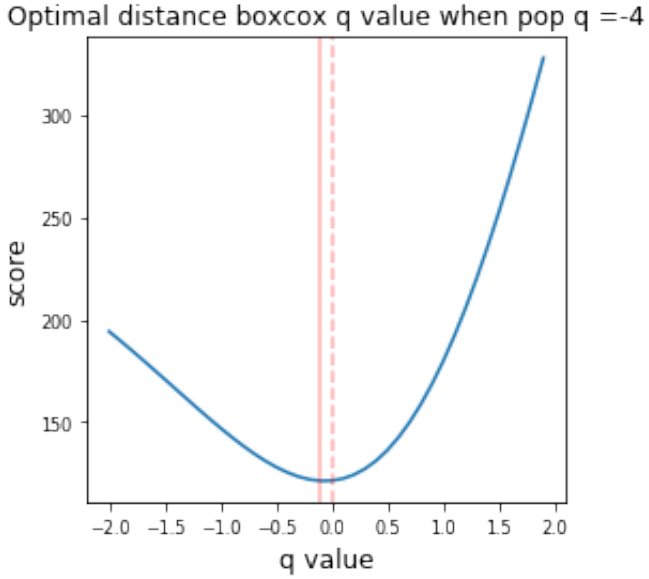

Optimal distance boxcox q value when pop q =1
score
q value
180
160
140
120
100
80
60
40
20
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Optimal distance boxcox q value when pop q =0
score
q value
175
150
125
100
75
50
25
0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Optimal distance boxcox q value when pop q =-1
score
q value
200
150
100
50
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Optimal distance boxcox q value when pop q =-2
score
q value
250
200
150
100
50
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

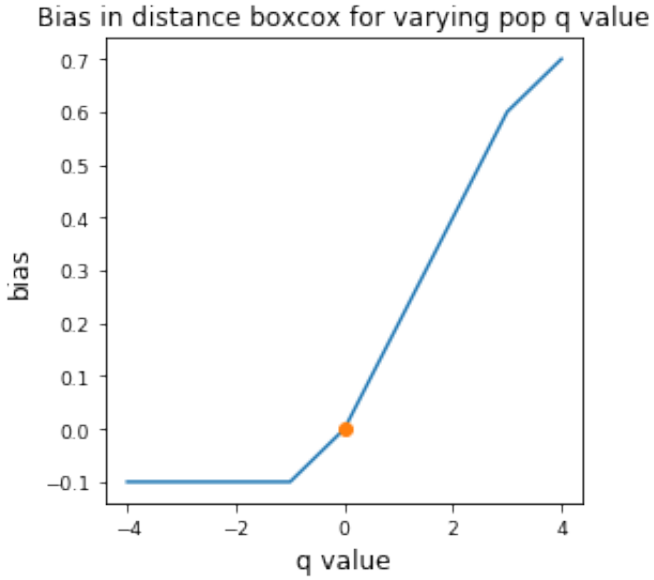

Optimal distance boxcox q value when pop q =-3
score
q value
300
250
200
150
100
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


Optimal distance boxcox q value when pop q =-4
score
q value
300
250
200
150
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

These plots show that as q is increased from 0 to 4 for the boxcox transform of the destination population variable, there is larger bias in the optimal q value that is selected for the distance variable. This means that as the functional form mispecification increases on one variable, it can cause the boxcox methodology to overcompensate on another variable. In addition, the plots also show that some bias can occur in the q value selected for distance when the q value for destination population decreases from 0 to -4. However, this bias is smaller and there seems to be a lower bound on the potential bias because the same amount of bias is incurred for q values of -1, -2, -3, and -4. A summar of the bias incurred for each q = {-4, 4} can be seen in the plot below where bias is defined as the difference between the selected q value for boxcox transformed distance and the true functional form used to generate the data (i.e. q =0).

```
In [95]: fig, ax = plt.subplots(figsize=(4,4))
         plt.plot([4, 3, 2, 1, 0, -1, -2, -3, -4], mins)
         plt.tick_params(labelsize=9)
         plt.title('Bias in distance boxcox for varying pop q value')
         plt.xlabel('q value', size=12)
         plt.ylabel('bias', size = 12)
         ax.plot(0, 0, marker='o' )
```

```
Out[95]: [<matplotlib.lines.Line2D at 0x121acf550>]
```



It seems the reason there is a lower bound on the bias that is incurred by negative q values is that by q = -1 the boxcox transformation has already altered the data to have essentially no variation. It can be seen below that for q = {-4, -1} the destination population barely varies for negative q values.

```
In [34]: print qn1
```

```
[ 0.9         0.95        0.96666667  0.95        0.9         0.9         0.9
  0.95        0.96666667  0.95        0.9         0.9         0.9         0.9
  0.96666667  0.95        0.9         0.9         0.9         0.9         0.95
  0.95        0.9         0.9         0.9         0.9         0.95
  0.96666667  0.9         0.9         0.9         0.9         0.95
  0.96666667  0.95        0.9         0.9         0.9         0.95
  0.96666667  0.95        0.9       ]
```

```
In [35]: print qn2
```

```
[ 0.495       0.49875     0.49944444  0.49875     0.495       0.495       0.495
  0.49875     0.49944444  0.49875     0.495       0.495       0.495       0.495
  0.49944444  0.49875     0.495       0.495       0.495       0.495
  0.49875     0.49875     0.495       0.495       0.495       0.495
  0.49875     0.49944444  0.495       0.495       0.495       0.495
  0.49875     0.49944444  0.49875     0.495       0.495       0.495
  0.49875     0.49944444  0.49875     0.495     ]
```

```
In [36]: print qn3
```

```
[ 0.333       0.33329167  0.33332099  0.33329167  0.333       0.333       0.333
  0.33329167  0.33332099  0.33329167  0.333       0.333       0.333       0.333
  0.33332099  0.33329167  0.333       0.333       0.333       0.333
  0.33329167  0.33329167  0.333       0.333       0.333       0.333
  0.33329167  0.33332099  0.333       0.333       0.333       0.333
  0.33329167  0.33332099  0.33329167  0.333       0.333       0.333
  0.33329167  0.33332099  0.33329167  0.333     ]
```

```
In [37]: print qn4
```

```
[ 0.249975    0.24999844  0.24999969  0.24999844  0.249975    0.249975
  0.249975    0.24999844  0.24999969  0.24999844  0.249975    0.249975
  0.249975    0.249975    0.24999969  0.24999844  0.249975    0.249975
  0.249975    0.249975    0.24999844  0.24999844  0.249975    0.249975
  0.249975    0.249975    0.24999844  0.24999969  0.249975    0.249975
  0.249975    0.249975    0.24999844  0.24999969  0.24999844  0.249975
  0.249975    0.249975    0.24999844  0.24999969  0.24999844  0.249975  ]
```

While the power functional forms of the locational attributes the and the power and exponetional function form of distance are rooted in the conceptual and technical history of spatial interaction models, other functional forms implied by the boxcox methodology have little-to-no justification. It appears that the boxcox transform removes (tempers) the variation (i.e., spatial structure) so that we cannot capture/understand it. For instance, using the boxcox method would make it impossible to spot outliers like more flows than expected from NY to FL, which can likely be accounted for by a measurable variable. Therefore, boxcox is not good for model building.

```
In [ ]:
```

APPENDIX B

EXPLORING THE VECTOR-BASED MORAN'S I TECHNIQUE

1 Replicating and Validating the Vector-based Moran's I technique

In this appendix, the vector-based spatial autocorrelation statistic based on Moran's I (VMI) of (Liu *et al.*, 2014) is explored. The VMI is a relatively new statistic and has only been applied once to taxi trips in China. Furthermore, to the knowledge of the author, no software is currently available for employing this statistic. Therefore, before it can be applied in this research, it needs to be replicated and validated. The statistic seeks to compare the direction and magnitude of a single vector to all others, which can be computed as

$$I = \frac{n}{\sum_i \sum_j M_{ij}} \frac{\sum_i \sum_j M_{ij}(u_i u_j + v_i v_j)}{\sum_i (u_{i^2} + v_i^2)} \tag{1}$$

where $u = (x^D - x^O) - (\bar{x}^D - \bar{x}^O)$, $v = (y^D - y^O) - (\bar{y}^D - \bar{y}^O)$, the O and D superscripts denote whether the planar coordinates x and y are associated with origins or destinations, and M_{ij} is a spatial weight matrix used to define spatial association amongst flows. Within the original conception of this vector-based Moran's I, distance-decay weights are utilized to define proximity between either vector origins or vector destinations, though the authors claim any theoretically justifiable weight could be used just as well.

1.1 Replicating the vector-based Moran's I autocorrelation statistic

Four scenarios are presented below in figure 97 that Liu *et al.* (2014) theorize to represent (a) positive autocorrelation from the perspective of vector origins, (b) positive autocorrelation from the perspective of vector destinations, (c) negative autocorrelation from the perspective of vector origins, and (d) no autocorrelation (random) from the perspective of either vector origins or destinations. After developing the code to carry out the VMI statistic, vectors that are representative of scenarios (a) and (b) were synthesized. Details regarding the vectors are summarized in table 8 where the vector coordinates as given represent scenario (a) and swapping the origin and destination points results in scenario (b). Then, the test statistic was computed from the perspective of origins, VMI_o, and the perspective of destinations, VMI_d, which resulted in $VMI_o = 0.65$ for and $VMI_d = -0.76$ for scenario (a) and $VMI_o = -0.76$ and $VMI_d = 0.65$ for scenario (b). This pattern very closely matches the pattern found in the results reported by Liu *et al.* (2014) that $VMI_o = 0.996$ for and $VMI_d = -0.194$ for scenario (a) and that $VMI_o = -0.194$ and $VMI_d = 0.996$ for scenario (b). Of course, the VMI values are not the same because the coordinates of the original vectors are not known and the statistic will be sensitive to any deviations in coordinates form the original vectors to those generated here. Nevertheless, the

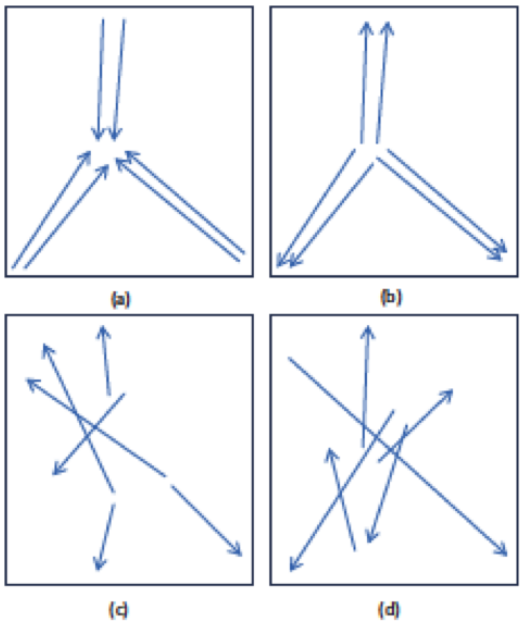


Figure 97: Four patterns of vectors. Pattern (a) results in positive vector autocorrelation from the perspective of origins; pattern (b) results in positive vector autocorrelation from the perspective of destinations; pattern (c) results in negative vector autocorrelation from the perspective of origins; and pattern (d) results in no vector autocorrelation. Image is reproduced from (Liu *et al.*, 2014).

fact that the individual signs and overall patterns match provides initial evidence that the statistic has been sufficiently replicated.

Further validation of the VMI statistic was achieved by simulating 100 random vectors in a given study area and computing the VMI, as is done by Liu *et al.* (2014). Figure 98 considers three scenarios:

- (a) vectors are random anywhere in the study area;
- (b) origins are constrained to a moderately sized nucleus in the center of the study area and destinations are random anywhere in the study area;

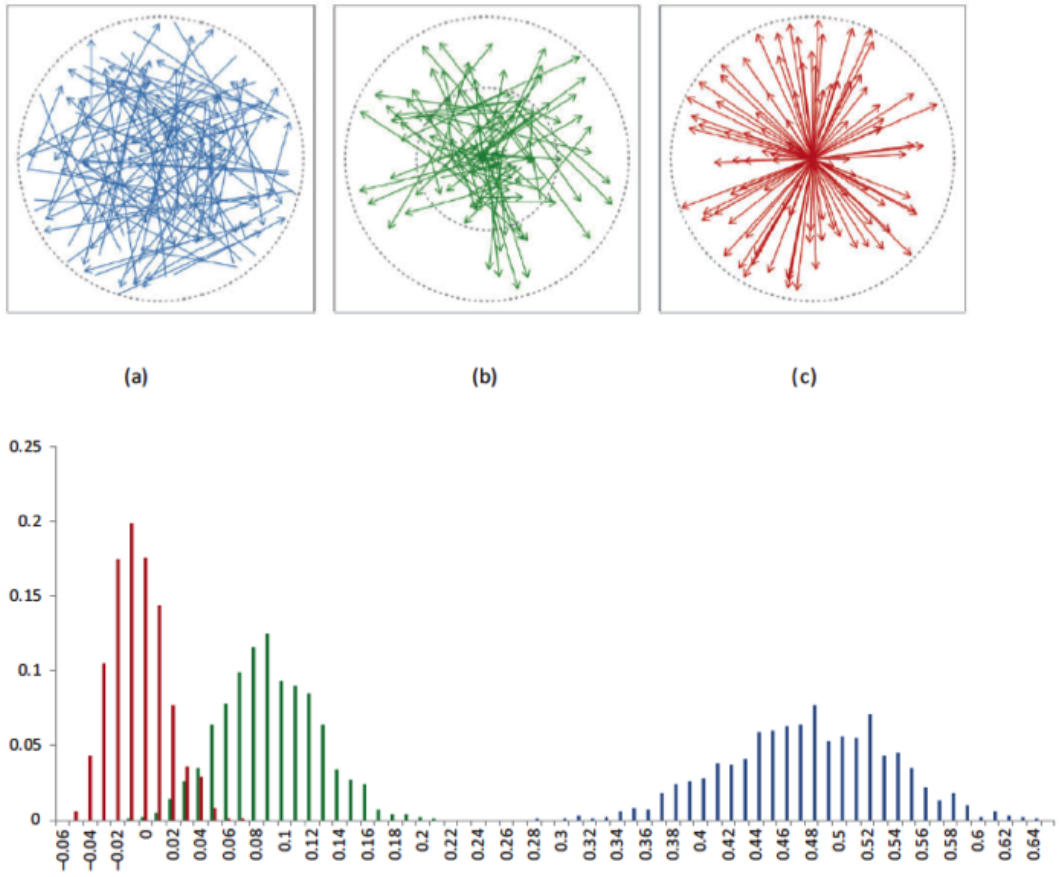


Figure 98: Three scenarios of 100 randomly generated vectors and their corresponding distribution of VMI for 1000 realizations. The top row illustrates the random vectors where scenario (a) is generated by selecting origins and destinations at random from the entire study area; scenario (b) is generated by selecting origins from inside a moderately sized nucleus in the center of the study area and selecting destinations at random from the entire study area; and scenario (c) is generated by selecting origins from inside a very small nucleus in the center of the study area and selecting destinations at random from the entire study area. The bottom row displays the distribution of 1000 VMI statistics computed from 1000 realizations of 100 vectors for each scenario. Image is reproduced from (Liu *et al.*, 2014).

Table 8: A set of six vectors and their corresponding origin and destination coordinates. These vectors are representative of the pattern in figure 97 (a). Swapping the origins and destinations of these vectors results in a set of vectors that are representative of the pattern in figure 97 (b).

Number	Origin-x	Origin-y	Destination-x	Destination-y
1	55	60	100	500
2	60	55	105	501
3	500	55	155	500
4	505	60	160	500
5	105	950	105	500
6	155	950	155	499

- (c) origins are constrained to a very small nucleus in the center of the study area and destinations are random anywhere in the study area.

The top of figure 98 provides a visual example of each scenario, while the bottom illustrates the distribution of VMI values computed on the sets of random vectors for 1000 simulations. It can be seen that only scenario (c) yields results centered on the expected null VMI value of -0.01 for 100 vectors. Scenarios (a) and (b) have distributions centered on values larger than the expected null VMI value with scenario (a) being centered on a much larger value (approximately 0.48). This experiment was replicated for scenario (a) and (c) using 50 vectors. The results are summarized below in figure 99 where the top row provides a visual that is very similar to the top row of figure 98 (left and right) from the original experiment. In addition, the observed mean of the VMI for scenario (c) is essentially equal to the expected null values of -0.0204 for 50 vectors (figure 98 bottom right) as was the case for the original experiment. Finally, the observed mean of the VMI for scenario (a) is higher than the expected null value (figure 98 bottom left), as was also the case for the original experiment. Two important conclusions may be drawn. First, replicating this

experiment further confirms that the VMI statistic was accurately replicated. Second, and more importantly, any application of this statistic should thoroughly discuss what the null pattern of random vectors is thought to be for a given process and acknowledge that low to mild autocorrelation as typically indicated by a Moran's I statistic may actually indicate what would intuitively be regarded as complete spatial randomness. A consequence is that significance testing of the statistic is exceedingly important.

Two methodologies for randomly permuting vectors are put forth by Liu *et al.* (2014) to carry out significance testing, since the underlying distribution of the statistic needed to develop an analytical test is unknown. Method (1) creates random permutations by assigning each vector origin to the origin of another vector while maintaining the same destination. This is effectively a geometric translation as the original vector origin and destination are shifted by the coordinates of the new origin. A consequence is that some of the randomized vectors may be outside the study area. Method (2) creates random permutations by assigning each vector origin to the destination of another vector, which ensures all vectors are inside the study area. Both permutation methods are tested by obtaining 1000 randomizations and computing pseudo p-values for a set of 100 vectors that indicate moderate autocorrelation (VMI = 0.436), which results in $p = 0.001$ for method (1) and $p = 0.023$ for method (2). As a result, it is concluded by Liu *et al.* (2014) that both methods provide sufficient theoretical random patterns with method (1) being more liberal than method (2) because it does not respect the limitations of the study area boundaries (i.e., edge effects). However, this reasoning is at odds with statistical intuition since each method preserves (or does not preserve) different features of the original set of vectors. Method (1) preserves the distribution of vector magnitudes, the distribution of vector

directions and the set of origin points, though it does not preserve the set of destination points. In contrast, method (2) preserves both the set of origin points and destination points, but it preserves neither the distribution of magnitudes nor the distribution of directions. Naturally, one would expect these two different methods to result in different statistical conclusions. Furthermore, results are only presented for a single realization of the experiment, which means they cannot indicate how this test will behave when sampling variation is considered. To fully explore the validity of these randomization techniques, an additional experiment is needed that produces many realizations of the hypothesis tests for each method and then analyzes the power and size of the tests.

1.2 Evaluating the power and size of significance tests

As indicated above, no analysis of the properties of the significance tests was provided by Liu *et al.* (2014). Two important properties regarding significance testing are the power and the size of a test. Test power, which ranges from 0 to 1, refers to the ability of the test to confirm an expected positive test (accept an alternative hypothesis rather than the null hypothesis). Higher power is desirable, denoting a high rate of true positives, a decrease in the tendency to yield false negatives, and that a test can accurately identify a trend from noise. In contrast, the test size, which also ranges from 0 to 1, refers to the ability of a test to reject an expected negative test (accept null hypothesis). Smaller test size is preferable with the expectation that some false positives may occur purely by random chance. Many tests are performed using a 95% confidence interval and a critical value of $\alpha = 0.05$. This implies that out of 100 realizations of a test, no more than 5% (five out of 100) of the tests are

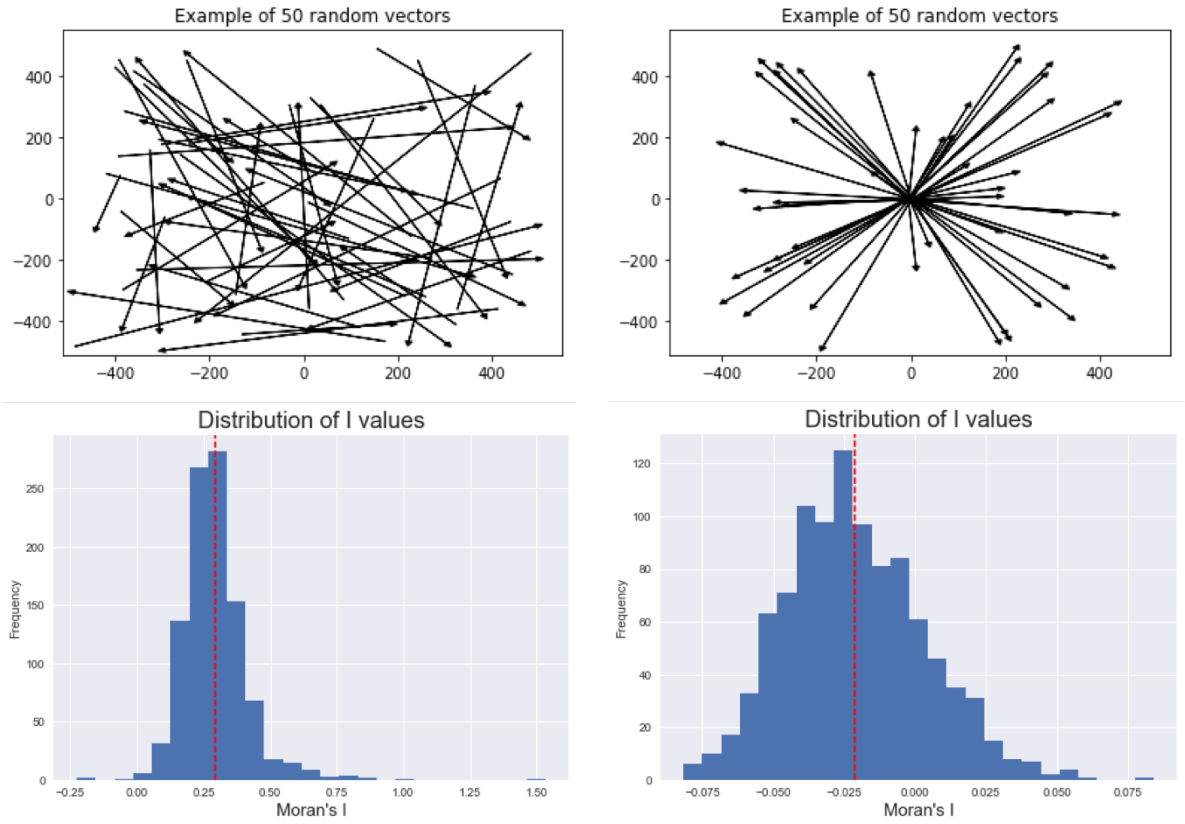


Figure 99: A partial replication of the results summarized in figure 98 that includes scenario (a) (left) and scenario (c) (right) using 50 vectors. The top row is illustrative of random vectors for each scenario while the bottom row is the distribution of the VMI statistics from 1000 realizations for each scenario. Here, the bottom right distribution pertains to the vectors generated from scenario (c) and the bottom left distribution pertains to the vectors generated from scenario (a).

expected to yield false positives. To summarize, test power measure the ability (i.e. probability) of a test to detect an effect when it is present and test size measures the ability of a test to accept the null hypothesis when there is no effect present. It is important to evaluate the power and size of a test to make sure that the test will actually help in making effective decisions regarding a particular statistic.

A simulation study was therefore designed to remedy the lack of analysis of the power and size of significance tests for both permutation methods. To explore the

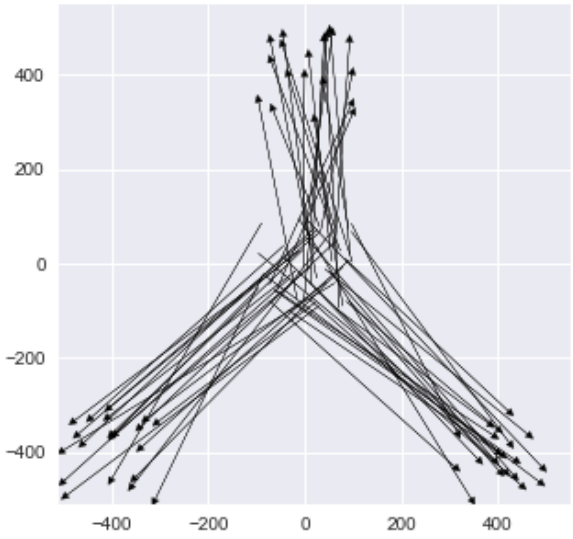


Figure 100: 50 positively autocorrelated vectors from the perspective of destinations.

power and size, two sets of 50 vectors were generated: the first has random origins and destinations (figure 99 top left) and the second has a pattern that is known to result in a positive VMI from the destination perspective (figure 100). Each set of vectors was randomly permuted 99 times for each permutation method and the pseudo p-values were computed. This was then repeated 1000 times in order to obtain a distribution of p-values that could be used to compute the power (using the correlated vectors) and the size (using the random vectors) of the significance tests.

For the positively autocorrelated vectors, the expectation is that the significance tests for both permutation methods will produce primarily small p-values that indicate we should reject the null hypothesis of no autocorrelation and can meaningful interpret the associated VMI statistic. The number of realizations that adhere to this particular outcome, divided by the total number of realizations yields the power of the test. The results of this procedure for permutation method (1) and permutation method (2) are given in figure 101 on the left and right, respectively. Figure 101 shows that for

method (1) all of the p-values are either zero or near-zero, which results in a power of 1, since we would always reject the null hypothesis in favor a pattern of vector autocorrelation at the 95% confidence interval (or even the 99.9% confidence interval). In contrast, the p-values that pertain to method (2) are trending towards a uniform distribution (figure 101 right), which results in a power of 0.054 and means that this technique is rarely able to reject the null hypothesis of random vectors even when there is a clear pattern. Overall, method (1) has an appropriate power while method (2) has little-to-no power.

For the random vectors, the expectation is that the p-values will be uniformly distributed since there is no pattern in the data and any p-value has an equal chance to occur. This also means that we should only reject the null hypothesis by random chance at a rate equivalent to our critical value. Here, we choose $\alpha = 0.05$ for a 95% confidence interval so that we expect no more than 50 out of 1000 realizations to produce a p-value less than or equal to 0.025 ($\frac{0.05}{2}$ because the statistic ranges from -1 to 1). Dividing the number of p-values that less than or equal to 0.025 by the total number of realizations produces the test size, which is expected to be about the same as α. Figure 102 shows that the test size for method (1) is much larger than 0.05 (0.863 is observed) with most of the p-values being clustered at values smaller than 0.05 rather than being uniformly distributed as expected. In contrast, method (2) produces the expected uniform distribution of p-values for the 1000 realizations and results in a test size of 0.054, which is close to the expected value of 0.05. Overall, method (2) has an appropriate test size while method (1) does not have a satisfactory test size.

Combing the results for test power and size, two trends become clear. First, method (1) is too liberal and tends to reject the null hypothesis whether there is a

pattern present or not. Second, method (2) is too indecisive and only rejects the null hypothesis by random chance, whether there is a pattern present or not. Additional experiments were carried out from both the origin and destination perspective, using a square study area and Manhattan as a study area and while varying the distance-decay (-10.0, -2.5, -2.0, -1.75, -1.5, -1.25, -1.0, -0.5) of the spatial weight and the sample size (90, 180, 360). A third permutation method was also explored, which starts with method (1) and then removes randomized vectors out that are outside of the study area. None of these variations produced drastically better results and often when the size was improved for method (1) it was at the cost of the power and vice versa for method (2). The only alteration that was able to produce acceptable test properties was adopting the pattern from figure 99 (top right) where all of the origins are clustered in a single location and destinations are random as the null hypothesis of no vector autocorrelation for method (1). The result was a test power of 1 and a test size of 0.045 (figure 103), which is perhaps not surprising since this pattern also achieved the expected null VMI value. However, this pattern of clustering all of the vector origins at a single location does not align with an intuitive notion of random vectors across a study area. Since no sufficient permutation method can be defined, it was decided to forego applying this statistic on empirical data since it would not be possible to make informed interpretations from the results.

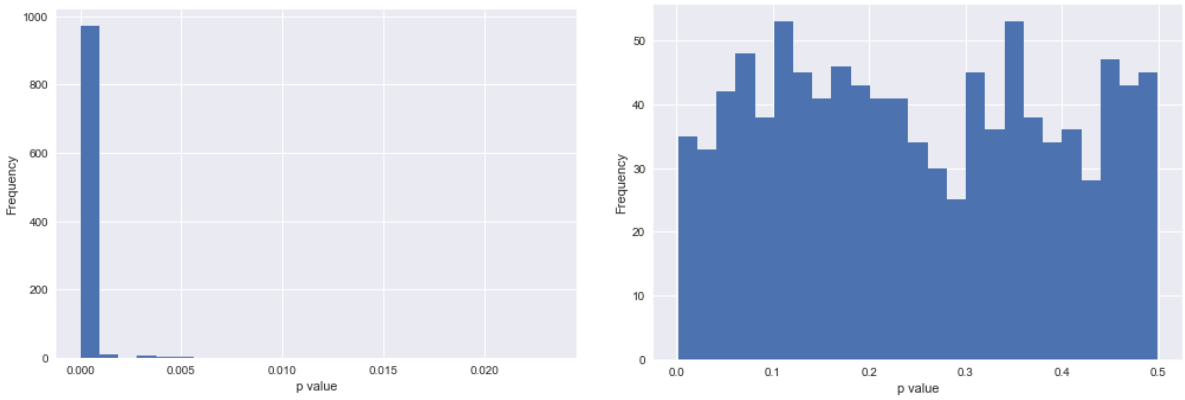


Figure 101: Distribution of p-values for VMI statistic applied to 1000 realizations of sets of vectors known to have positive autocorrelation, which are representative of the test power. Distribution to the left corresponds to method (1) and distribution to the right corresponds to method (2).

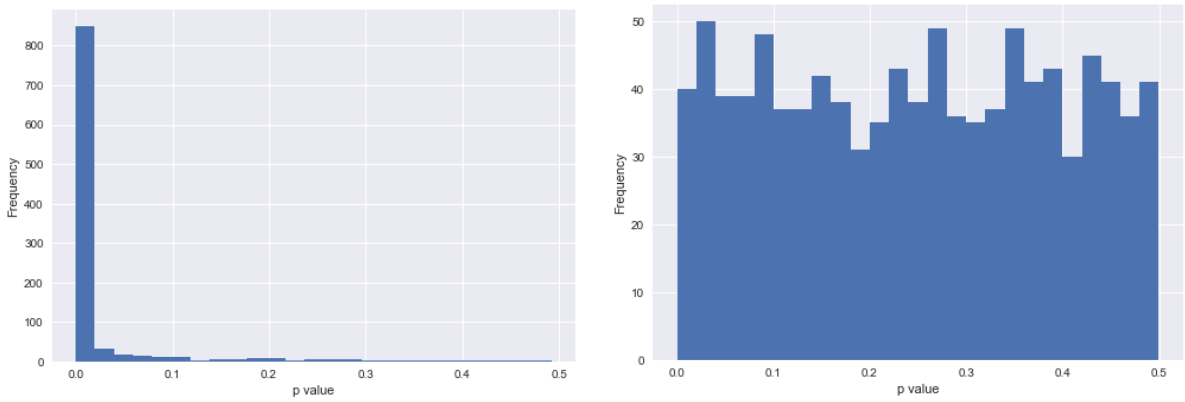


Figure 102: Distribution of p-values for VMI statistic applied to 1000 realizations of sets of vectors with random origins and destinations, which are representative of the test size. Distribution to the left corresponds to method (1) and distribution to the right corresponds to method (2)

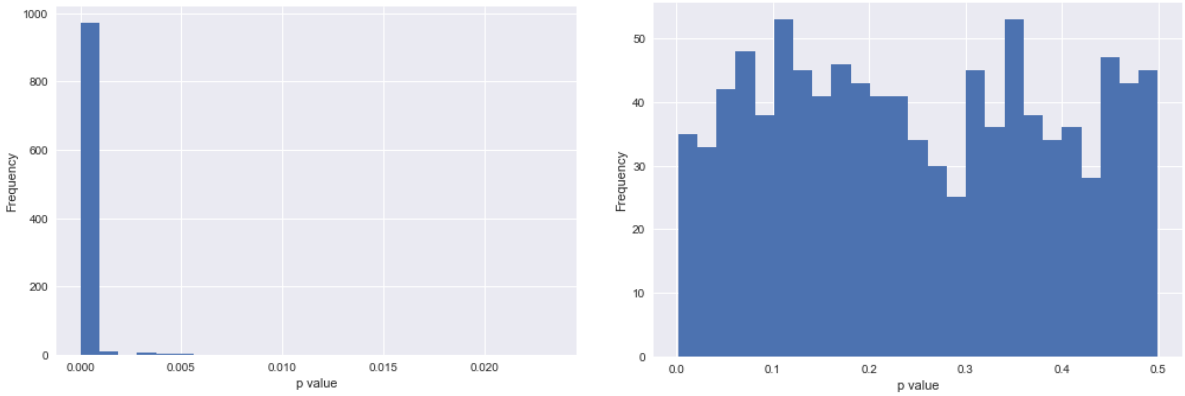


Figure 103: Distribution of p-values for VMI statistic using method (1) applied to 1000 realizations of sets of vectors known to have positive autocorrelation, which are representative of the test power (left) and distribution of p-values for VMI statistic applied to 1000 realizations of sets of vectors with highly clustered origins and random destinations, which are representative of the test size (right).

APPENDIX C

SIMULATION EXPERIMENT FULL RESULTS (CHAPTER 5)

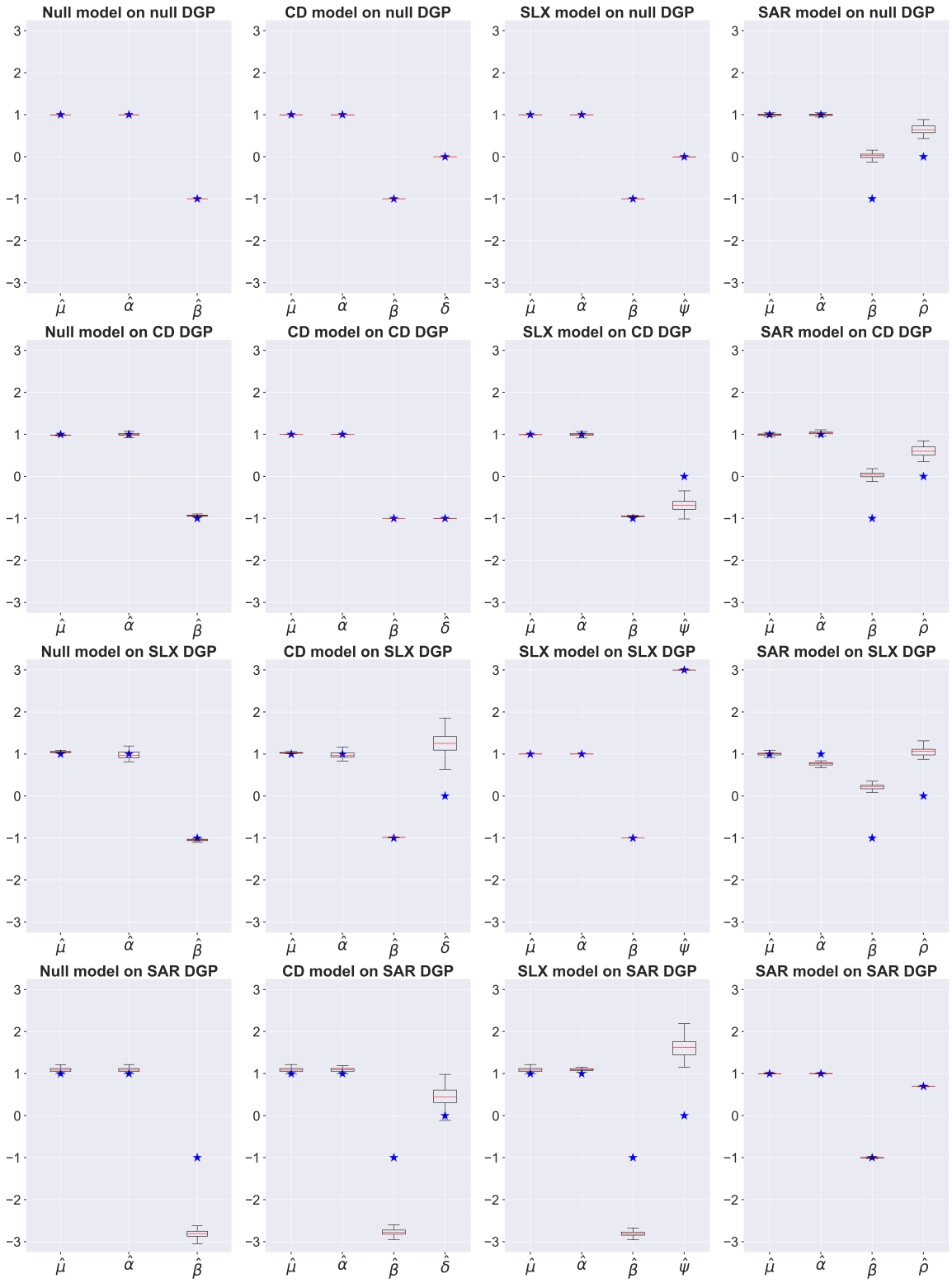


Figure 104: Results for models calibrated on datasets with uniform points.

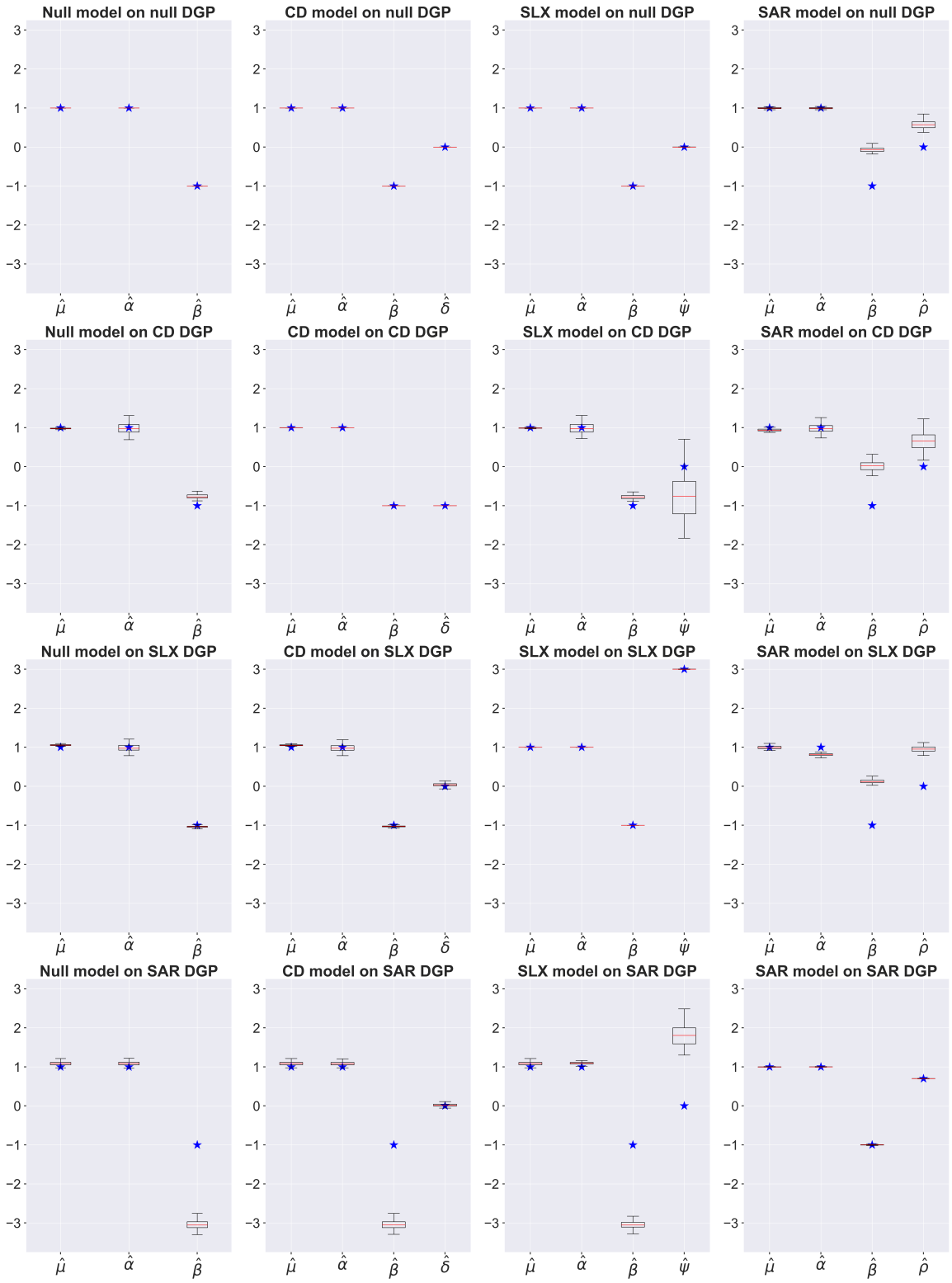


Figure 105: Results for models calibrated on datasets with random points.

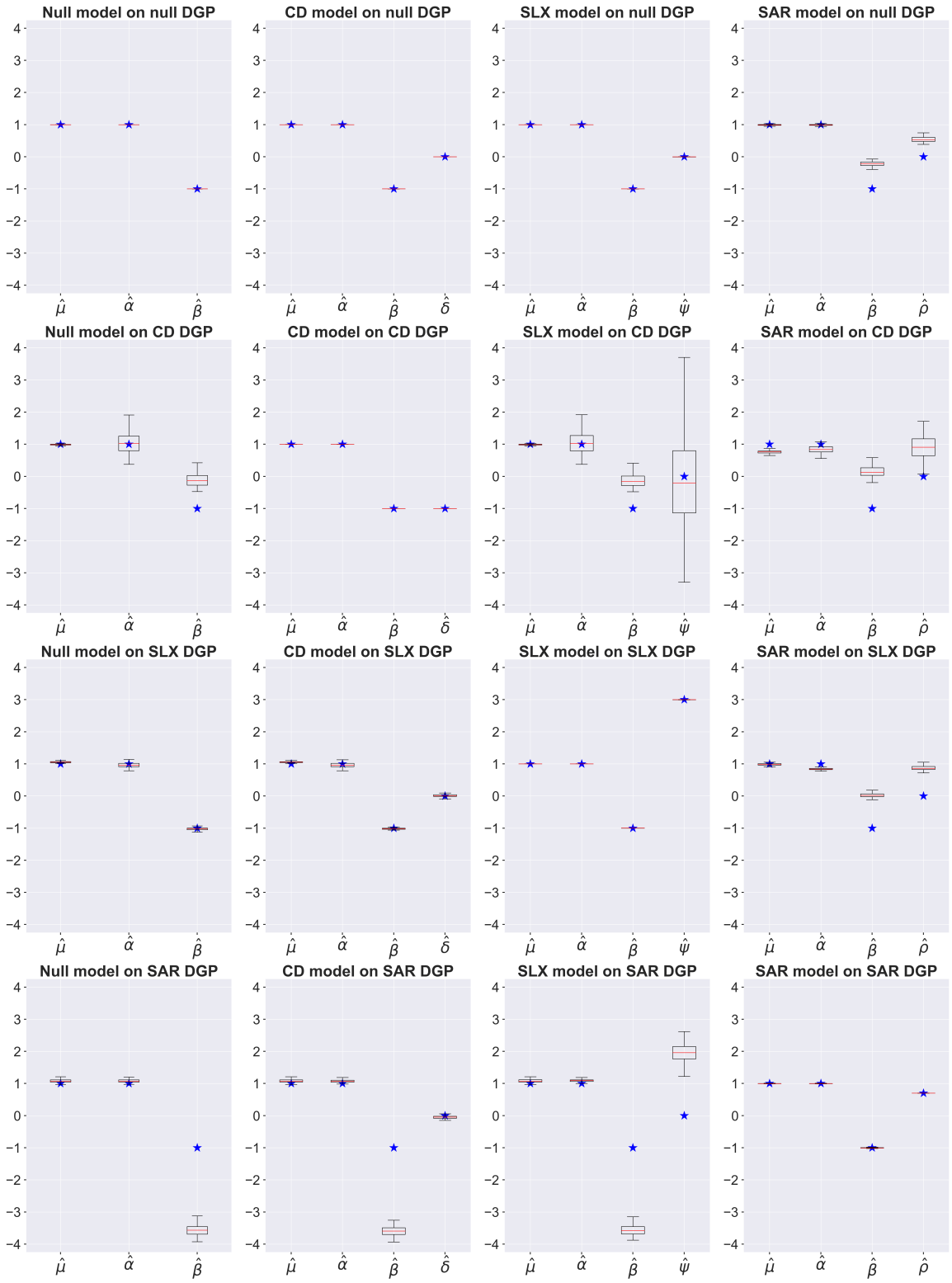


Figure 106: Results for models calibrated on datasets with clustered points.

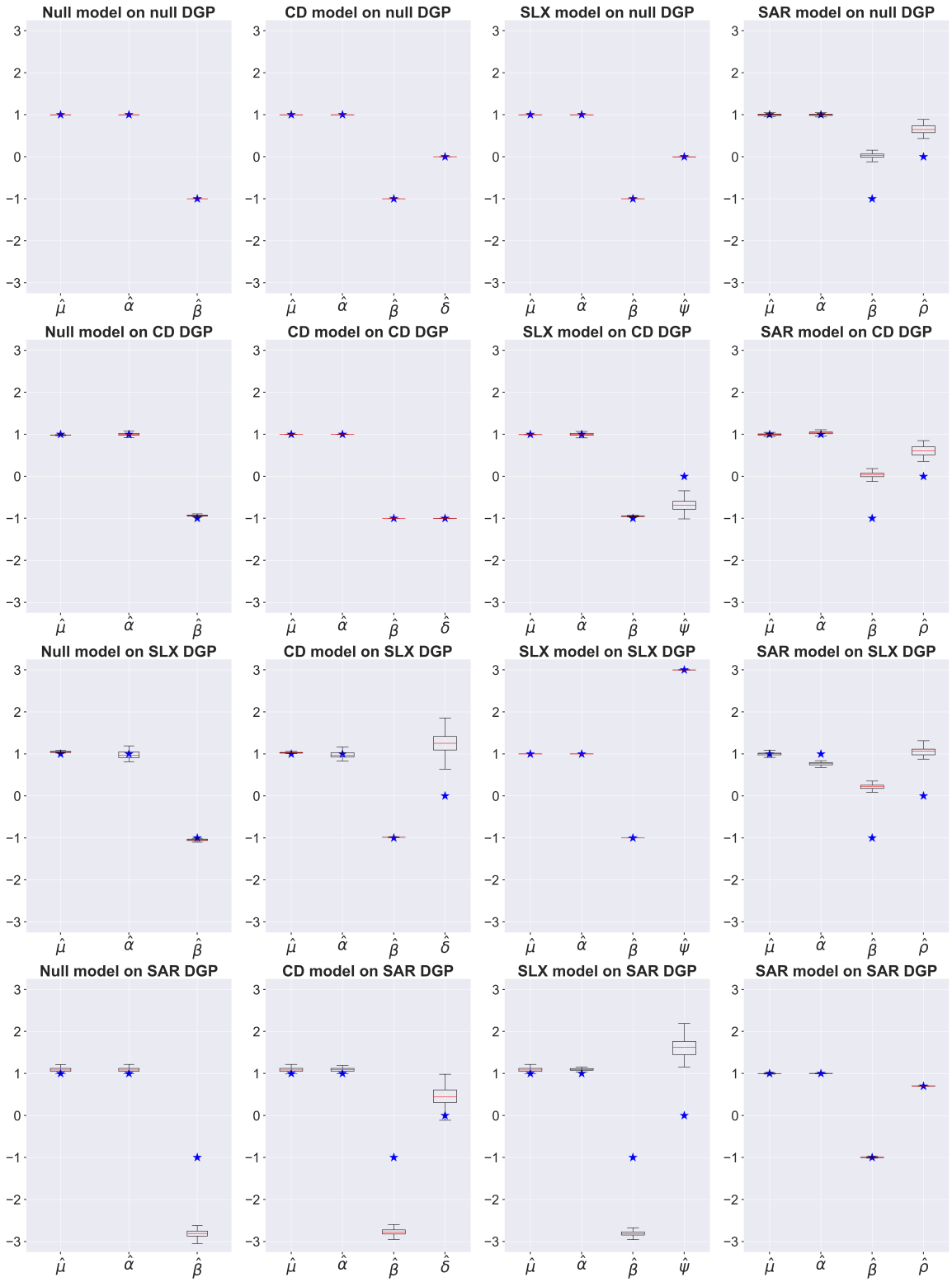


Figure 107: Results for models calibrated on datasets with uniform points and then aggregated to a 24 by 24 grid.

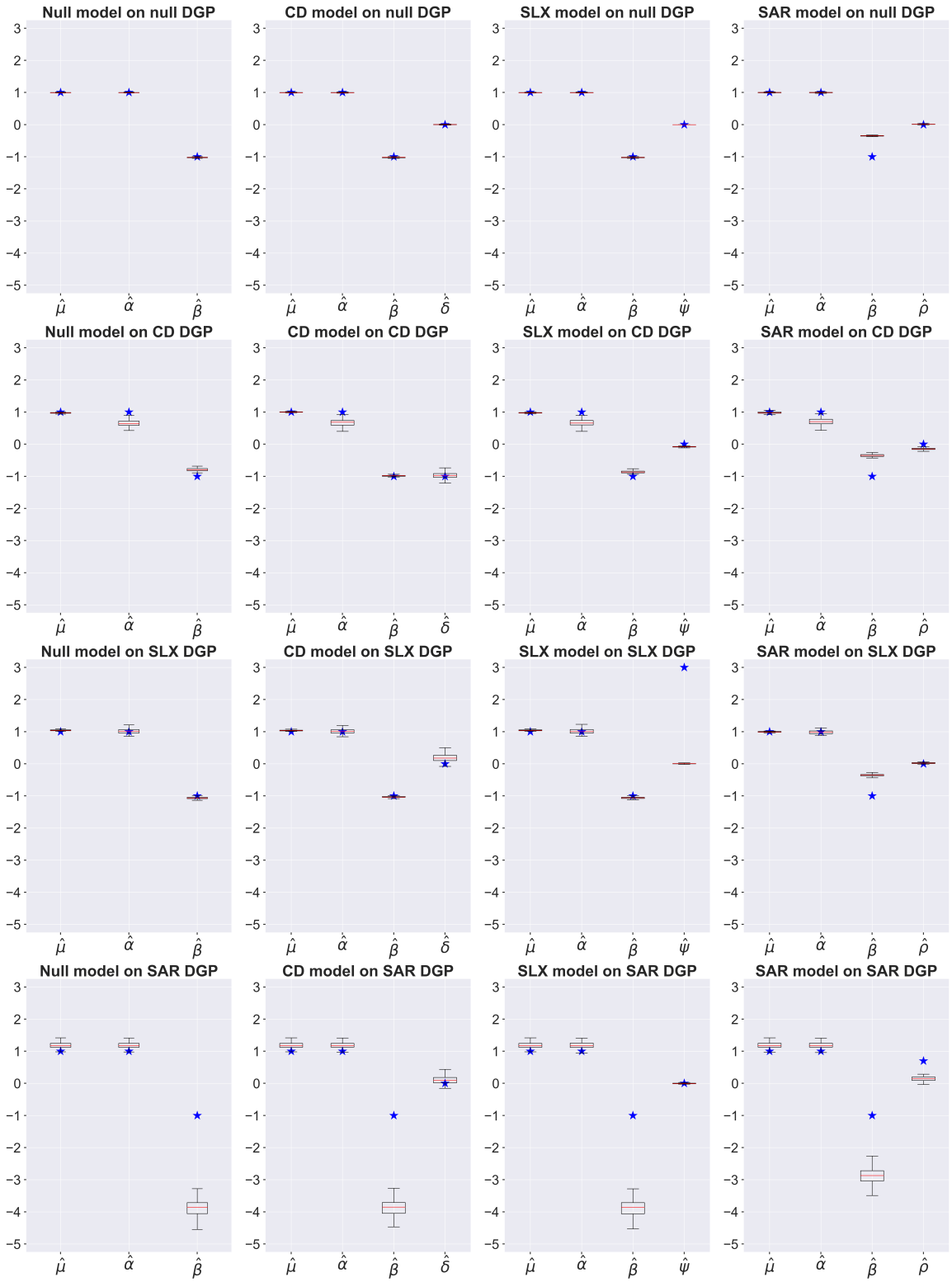


Figure 108: Results for models calibrated on datasets with random points and then aggregated to a 24 by 24 grid.

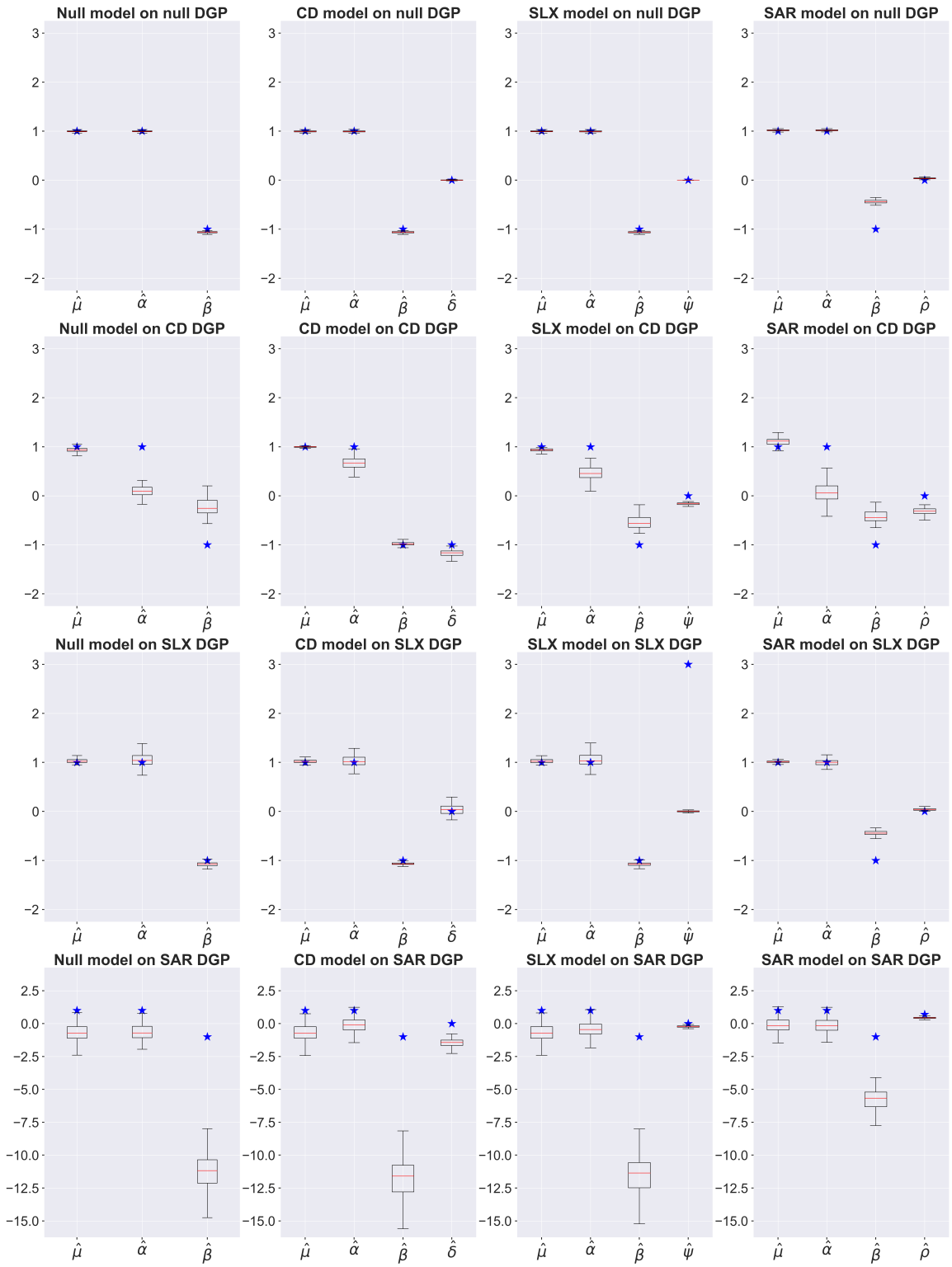


Figure 109: Results for models calibrated on datasets with clustered points and then aggregated to a 24 by 24 grid.

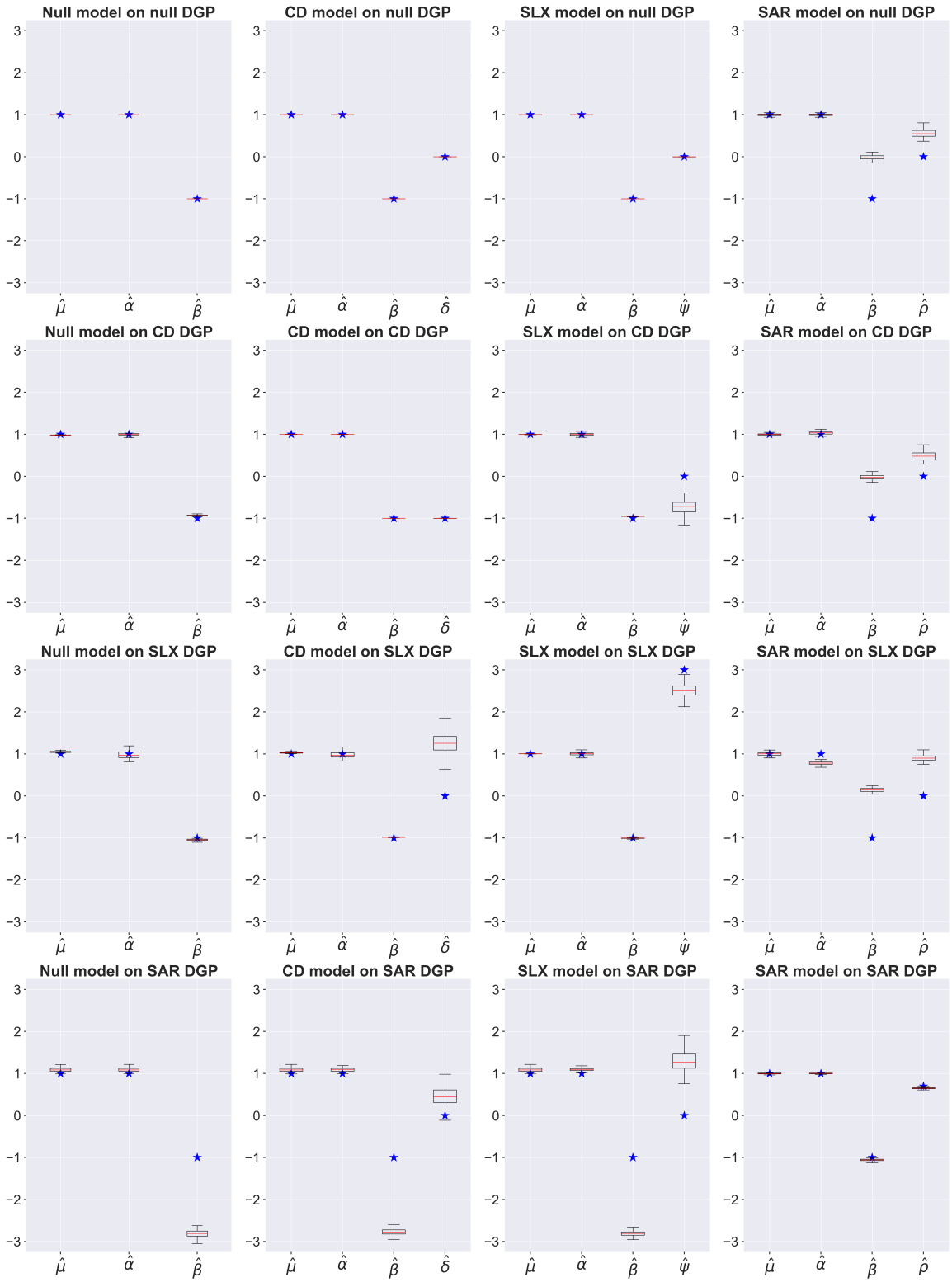


Figure 110: Results for models calibrated on datasets with uniform points and then aggregated to a 12 by 12 grid.

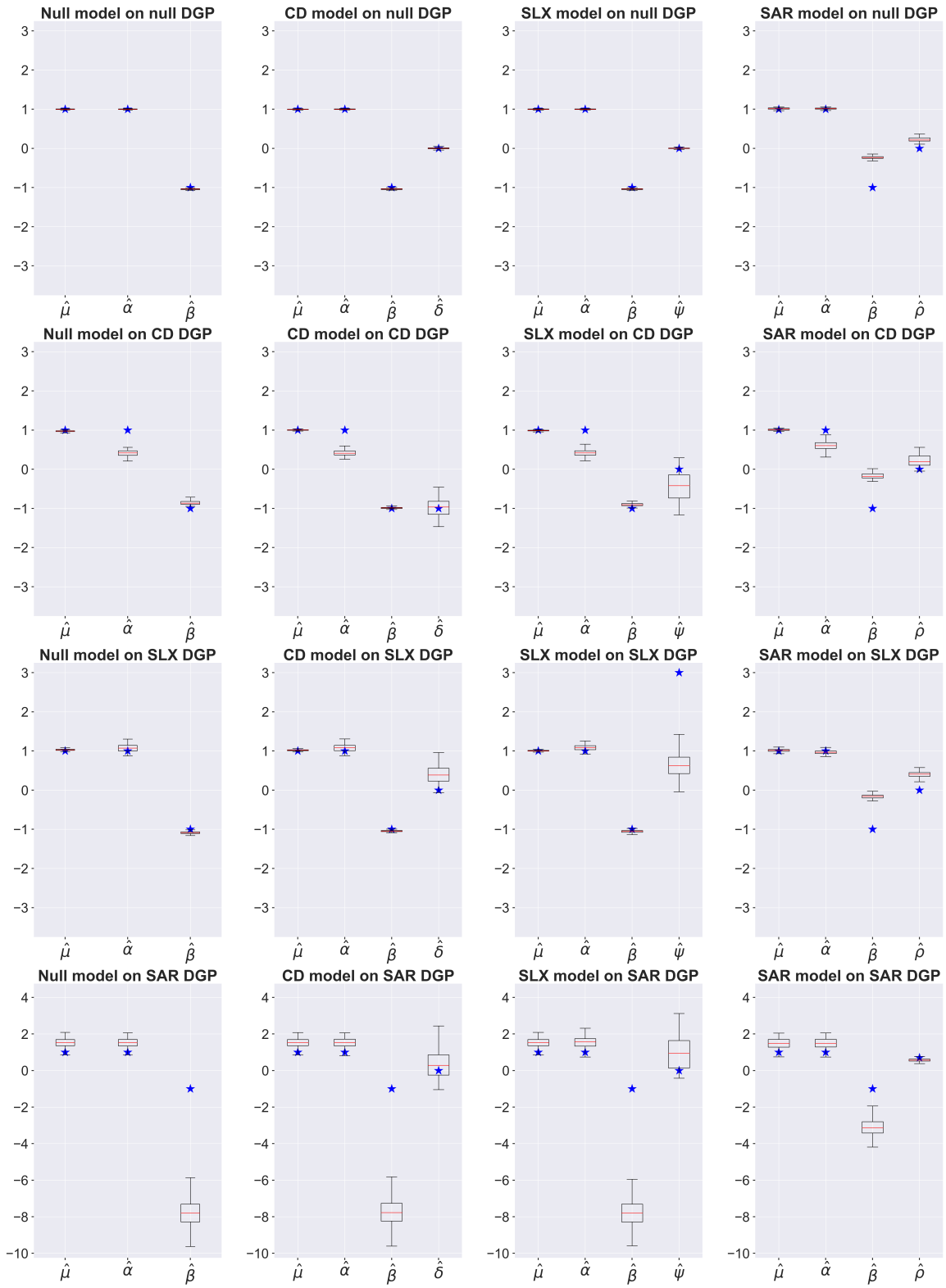


Figure 111: Results for models calibrated on datasets with random points and then aggregated to a 12 by 12 grid.

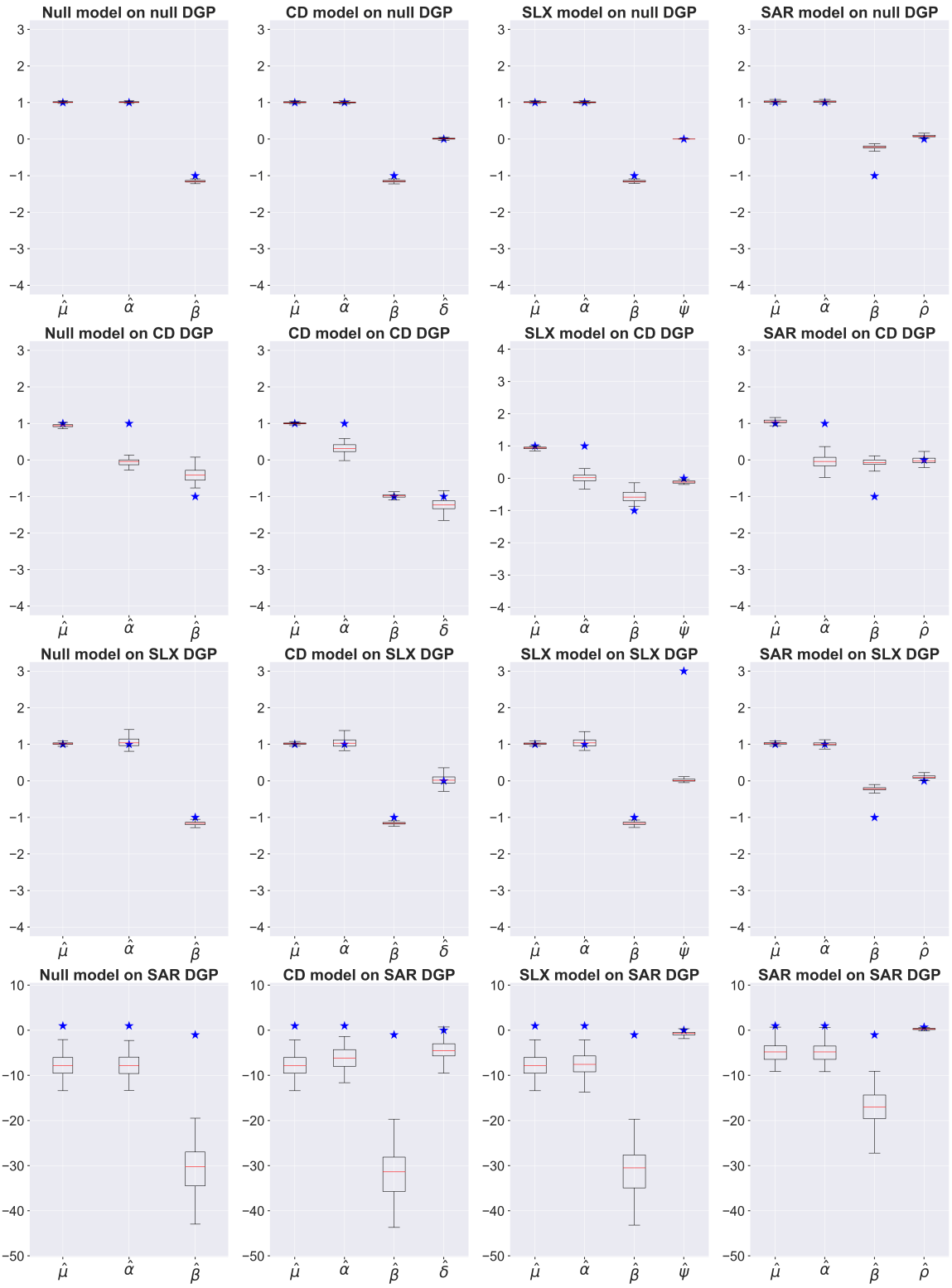


Figure 112: Results for models calibrated on datasets with clustered points and then aggregated to a 12 by 12 grid.