Understanding Mobility and Active Transportation in Urban Areas Through Crowdsourced Movement Data

by

Lindsey Conrow

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

Approved April 2018 by the
Graduate Supervisory Committee:

Elizabeth Wentz, Chair
Trisalyn Nelson
Siân Mooney
Christopher Pettit

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

Factors that explain human mobility and active transportation include built environment and infrastructure features, though few studies incorporate specific geographic detail into examinations of mobility. Little is understood, for example, about the specific paths people take in urban areas or the influence of neighborhoods on their activity. Detailed analysis of human activity has been limited by the sampling strategies employed by conventional data sources. New crowdsourced datasets, or data gathered from smartphone applications, present an opportunity to examine factors that influence human activity in ways that have not been possible before; they typically contain more detail and are gathered more frequently than conventional sources. Questions remain, however, about the utility and representativeness of crowdsourced data. The overarching aim of this dissertation research is to identify how crowdsourced data can be used to better understand human mobility. Bicycling activity is used as a case study to examine human mobility because smartphone apps aimed at collecting bicycle routes are readily available and bicycling is under studied in comparison to other modes. The research herein aimed to contribute to the knowledge base on crowdsourced data and human mobility in three ways. First, the research examines how conventional (e.g., counts, travel surveys) and crowdsourced data correspond in representing bicycling activity. Results identified where the data correspond and differ significantly, which has implications for using crowdsourced data for planning and policy decisions. Second, the research examined the factors that influence cycling activity generated by smartphone cycling apps. The best predictors of activity were median weekly rent, percentage of residential

land, and the number of people using two or more modes to commute in an area. Finally, the third part of the dissertation seeks to understand the impact of bicycle lanes and bicycle ridership on residential housing prices. Results confirmed that bicycle lanes in the neighborhood of a home positively influence sale prices, though ridership was marginally related to house price. This research demonstrates that knowledge obtained through crowdsourced data informs us about smaller geographic areas and details on where people bicycle, who uses bicycles, and the impact of the built environment on bicycling activity.

ACKNOWLEDGMENTS

First I would like express my gratitude to my doctoral advisor and committee chair, Dean Professor Elizabeth A. Wentz. Without her guidance and encouragement, this whole thing would have been a great deal more difficult and way worse. Second, I would like to thank my committee members Trisalyn A. Nelson, Sian Mooney, and Christopher Pettit for their invaluable advice and guidance throughout my dissertation research. During my Ph.D. studies, I received support, assistance, and encouragement from friends and colleagues at Arizona State University. I was fortunate to have friends including Ashlee Tziganuk, Angela Sakrison, Kelli Larson, and all other queers and homos in situ. I would also like to mention the members of Wentz lab - Heather Fischer, Joanna Merson and Qunshan Zhao, for providing a friendly and positive working environment. I would also like to sincerely acknowledge the feedback and ideas I've sussed out with Deborah Salon and Michael Kuby in TransportLab. Thank you as well Alan Murray for your patience and guidance in collaborating as well as stanching any fear I had of doing mathematics in front of a room of students who aced the quant section of the GRE. Special thanks as well to First Draft book bar for providing a fun and stimulating environment where I completed a great deal of writing and thinking. I would like to acknowledge the funding I received while I finished my doctoral studies. Thank you College of Liberal Arts and Sciences for granting me a first-generation student fellowship. This research would also have not been possible without the PLuS Alliance, Transport for New South Wales, Maricopa Association of Governments, and AURIN.

TABLE OF CONTENTS

CHAPTER

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

As cities continue to expand and densify, people are tasked with getting to the places they want and need to go in the amount of time they have to get there. This is one aspect of human mobility and accessibility that is studied so that people are able to go more places and reduce overall travel times. While development of many western urban areas has been focused on human mobility via personal motor vehicles, planners and other stakeholders are now investing more in alternative forms of transport. This shift seems to relate to planners' desire to reduce carbon emissions, increase safety for riders, and to decrease motor vehicle trips. The face of human mobility has changed with the advent of city planning specifically aimed at serving people who choose to use bicycles as a mode of transport. The push for bicycle related infrastructure such as bicycle lanes and routes has facilitated bicycling's uptake as an activity and form of transportation. This supportive infrastructure means that bicyclists now have choices about where, when, and how they ride. Cyclists may represent a broad range of ages and skill levels from young and old to cautious riders, people out for a casual Sunday ride, or those coined as "mamil" - middle aged men in lycra - who may be interested in using bicycling as a mode of fitness and competition, recreation, or transportation. A particular cyclist may choose to ride a direct route through several busy intersections or may go out of their way to stay on bicycle-supportive infrastructure.

In light of the push for increased bicycle-friendly infrastructure, city planners and other stakeholders are pressed to understand where, when, and what type of infrastructure to install. Different types of infrastructure might support recreational versus transportation bicyclists and knowledge of where, when, and what type of bicyclists are accessing cycling is needed so that informed decisions can be made. One problem in generating this knowledge lies in the ways that bicycling activity data are usually collected; the data often do not have enough information to understand the rider and street level details needed in decision making. More detail is needed to inform assessments of existing infrastructure as well as planning for installing infrastructure in the future. Informed urbanism in the form of well-planned and assessed infrastructure is

key in understanding the economic impact, overall ridership, and data sources associated with bicycling activity.

1. Problem Statement

Basic research questions in analysis of human mobility pertain to identifying popular places and the routes that people use to travel between them. Outcomes from these analyses characterize human mobility, such as defining transportation modes, identifying reasons for travel (e.g., commuting or recreation), or describing travel paths, and may illuminate drivers that influence those behaviors. The patterns and characteristics associated with mobility including routes taken, purpose for travel, or duration and distance of trips can be used to predict and estimate travel demands on networks or can also be used to examine travel among different modes, such as personal vehicles, public transit, and pedestrian activity. (Kitamura et al., 2000; Hiribarren and Herrera, 2014; Tao, Rohde, and Corcoran, 2014).

Findings in mobility studies usually show that regional travel characteristics can be discovered within the data using a variety of partitioning, flow, and cluster methods, though detailed mobility information is typically left out. For example, specific routes, street level detail, and information about demographic subgroups are often ignored when regional, or origin-destination, data are used. Understanding the drivers of human mobility is key in generating policies and planning that can effectively predict, manage, influence, and facilitate improved human mobility and can illuminate where effective changes or improvements can be made so that congestion, excessive travel time, restricted accessibility and other problems do not occur.

Current sampling methods and data available for planning and analyzing non-motorized travel are limited in detail and scale, and robust systems that account for and measure pedestrian and cyclist travel are needed (Lindsey, Chen, and Hankey 2013). Active transport, or non-motorized travel, is associated with increased health behavior (Sahlqvist et al., 2013), reduced traffic congestion, and potential reduction of emissions associated with vehicular travel (Frank et al., 2006), though it must be a feasible transport option for those benefits to be realized. Increased use of and access to active transport is one aspect of transit planning and design that may improve problems associated with

human mobility. Regional travel surveys may not address non-motorized travel activity at all, and if they do it has been convention to assume that people take the shortest distance path between their origin and destination. This approach overlooks other network features and built environment contexts that may influence movement behavior such as infrastructure, other geographic features like slope or land use, and trip purposes (Broach, Dill, and Gliebe, 2012). These are the networks and contexts that may influence whether a cyclists chooses going out of their way to stay on bicycle specific infrastructure rather than take a direct route or vice versa.

Collection and analysis of activity data is central to the success of human mobility studies. Conventional methods to collect human movement and mobility data include sensor detectors, manual counts, travel surveys and travel diaries. These sources provide information about trip origins, destinations, travel volume, and flows between different areas. They have three major limitations including high costs, coarse scale, and small sample sizes. High costs may be associated with dedicated infrastructure, like loop detectors and sensors, that capture movement behavior (Hiribarren and Herrera, 2014; Leduc, 2008). The sampling strategies of these data are also coarse in that they only collect information about volumetric flows at particular locations, so coverage and detail are limited (Leduc, 2008). While broad movement patterns of travel around regions can be characterized, smaller group or individual travel cannot be distinguished. Since volumes are only collected at particular locations with these technologies, they do not allow actual route information to be observed. Travel surveys and diaries may be similarly limited as only trip origin and destination locations are collected and route information is not often included. Further, these data are typically aggregated to some larger spatial unit to protect privacy, so actual origin and destinations may not be retained. Some analyses are possible at the individual level with travel survey data, though finer scale examinations at sub-regional levels are limited by sample sizes in particular regions because there may be too few respondents in any one area. Since data collection periods often typically span one or two days, it is also possible that certain travel behaviors could be missed depending on the days that are selected. Participants using travel diaries may also fail to record trips, especially short trips that are not taken

from their home location (Wolf, Guensler, and Bachman, 2001). GPS recordings can be used to supplement travel diaries and overcome problems, such as trips that are not recalled or recorded by participants, though they are still associated with small sample sizes and short data collection periods.

Sampling strategies for mobility data using these conventional methods are one factor that has limited detailed analysis of movement and mobility activity. Sampling methods particularly limit analysis of mobility in the area of active transport, or non-motorized travel like walking and cycling for practical purposes rather than leisure (Saunders et al., 2013). Since non-motorized trips tend to be shorter, they necessarily require detailed finer resolution data for analysis (Cervero, 2003). Non-motorized travel usually occurs on local streets so developing and planning its infrastructure are typically within the domain of local governments, though increased federal and state funding has led to increased interest in monitoring pedestrian and cycling activities (Lindsey, Chen, and Hankey 2013).

Greater knowledge about the infrastructure and network contexts that positively influence utilization of active transport modes will inform policy and design so those modes could be adopted and accessed by more users. Physical or environmental features such as transport infrastructure, land use, and the social, cultural, and institutional characteristics of neighborhoods are known to shape human behavior and these contexts influence mobility by offering a setting wherein people are able to engage in and access healthy living patterns (Kwan, 2012). Because people may travel through a variety of neighborhoods as they move about to perform their daily activities, individual, person and path-based definitions of movement and methods that account for people's actual movement paths, the places they visit, and the amount of time they spend there are needed (Kwan, 2012). Mobility data that contain information at finer spatial and temporal scales than conventional data are needed to account for people's experiences during movement behavior. These are the kinds of data that will distinguish between different types of travelers as well as the routes they take.

One of the opportunities to improve examinations of active transport lies in new crowdsourced data sets. These data provide an opportunity to examine detailed

movement contexts that influence behavior and mobility by overcoming some of the limitations associated with conventional data. GPS traces, made available through location aware technologies (LATs) and smartphone applications, provide fine scale space-time data related to people's activities and movements (Laube and Purves, 2011; Miller, 2010). These detailed movement datasets contain information that some conventional methods cannot collect such as full travel routes, speeds, and duration of travel at an individual level, along with the benefit of increased sample sizes and the potential for near-continuous data collection. These data could allow for aspects of non-motorized travel, such as how different types of infrastructure, street level environmental contexts, and movement activities interact and influence human mobility, to be measured. Despite potential insights, crowdsourced data may be biased and limiting in terms of generalizable representation of greater populations and few studies have compared the data to conventional sources to determine what they represent or how they may be used to advance our knowledge about human mobility and active transport. Potential limitations in the representativeness of these data need to be examined so their utility can be realized, and appropriate analytical methods can be applied to them.

The goal of this research is to advance the understanding of human mobility using crowdsourced data. To that end crowdsourced data are used to incorporate geographic context into three studies that examine bicycling activity in urban areas. The following sections explain the background literature that relates to human mobility, the conventional data used to examine it, and explains the use of bicycling as a case study for mobility analysis. The remainder of the dissertation is organized as three separate but related research papers. The first paper compares crowdsourced and conventional data while the second paper examines factors that influence bicycling activity using crowdsourced data. Finally, the third paper uses crowdsourced data to determine whether bicycle ridership is related to house prices.

CHAPTER 2
BACKGROUND LITERATURE AND RESEARCH GOALS

The goal of this literature review is to outline human mobility and the data used to examine it; it provides a synthesis of the background literature and the specific research goals of the dissertation. Conventional methods for collecting mobility data are discussed as they have key shortcomings in representing human mobility. Bicycling is justified as a case study for mobility since it occurs at the fine spatio-temporal scales that are relevant to understanding geographic context and active transportation. The benefits and drawbacks of using crowdsourced data are also outlined. The chapter then presents the specific research goals of the dissertation and discusses how those goals will be addressed. This section describes the research problem context, the goals of the research, as well as the outline of the remainder of the dissertation.

2.1 Literature Review

2.1.1 Human Mobility

Human mobility is studied to determine important places and the routes and modes people use to travel between them, as well as the influence of neighborhoods activity (Hirsch et al., 2014). As people move through space and time, what they experience and how they move is influenced by where, when, and why they move, for how long, and in or between which places (Kwan, 2012; Kwan, 2013). Where and when (i.e., time of day, for how long) people move constitutes the geographic context, and its influence must be considered and accounted for in examinations of active transportation. The geographic context needs to be examined in greater detail so that the factors that influence how and why people move around urban areas can be better understood. Greater understanding of the influence of context on mobility will aid policy and planning decisions like where, when, and what types of infrastructure to install.

Differing delineations of geographic context, however, also influence conclusions that are made about movement behaviors (Kwan, 2012). To accurately assess the impacts of geographic context on mobility, context needs to be examined as a dynamic concept

that considers people's actual movement paths (Kwan, 2013). Larger scale studies that examine regional trends in mobility often ignore the specific street level contexts within which people move because existing data is limited in detail; in examinations of dynamic contexts and human mobility, detailed mobility data are needed. There are limitations associated with some forms of mobility (e.g., walking and bicycling) that reduce the amount of information gleaned from said examinations because they occur at small spatial scales. Outside of small scale and localized travel studies, few conventional sources (e.g., counts and travel surveys) generate the level of detail necessary for studies of where and when people engage in mobility and active transport. Since crowdsourced data do often have detail like routes or counts indicative of popular routes, they have the potential to overcome these limitations and allow for analysis related to the geographic contexts through which people move.

2.1.2 Active Transportation

Active transport has several benefits though its use remains low in many urban areas. This section examines the research related to both bicycling activity as well as studies of crowdsourced data that relate to it. First, the benefits of active transport are discussed. Then, more details on the drivers and implications of bicycling as a mode of active transportation are summarized.

Active transport modes are associated with a range of individual and population-level benefits, including decreased risk of adverse health outcomes, like hypertension and diabetes, as well as reduction in carbon emissions (Saunders et al., 2013; van Heeswijck et al., 2015; Kuzmyak and Dill, 2014). Using active transit modes is often associated with neighborhoods that are walkable, or characterized by dense, connected areas with mixed-use land development and attractive destinations for pedestrians (Frank et al., 2006). Neighborhood walkability is associated with increased active transit and both decreased rates of and risk of obesity (Frank et al., 2006). Increased active transport in walkable areas, with a corresponding decrease in motorized transport, is associated with health benefits related to less air pollution and carbon dioxide emissions reduction (Woodcock et al., 2009; Frank et al., 2006). Despite benefits, levels of bicycling for commute and

recreation remain low in many urban areas, representing low percentages relative to all trips among other modes (Kuzmyak and Dill, 2014; Pucher, Garrard, and Greaves, 2011).

Bicycling research addresses the role of road infrastructure, path preferences, and rider demographics. Current studies that utilize conventional data from regular cyclists have shown that they prefer to ride on bicycling specific infrastructure and well-connected streets (Dill, 2009), prefer short paths and ride an average of about 30 minutes when commuting (Dill, 2008, Plaut, 2005).

Surveys aimed at cyclist preferences might only collect route information from participant recall, if at all, and may have small sample sizes. Travel surveys that target bicycling activity are typically limited in temporal scope and might only offer data on a single day or single week as well (Dill, 2009). Further, questionnaires may limit the information collected by only asking about typical behaviors and leaving out the details of where bicycling actually occurred (Dill, 2009). Travel surveys that utilize tracking technologies, such as GPS devices, gather detailed route and trip information but also generate noise and uncertainty associated with the data (Siła-Nowicka et al., 2016). Participants may not use devices when they are expected to or devices may run out of power and not collect data about relevant trips (Siła-Nowicka et al., 2016). Larger sample surveys comprised of more respondents, like the United States Census Bureau's American Community Survey and National Household Transport Survey, typically do not collect any route information and may miss bicycling activity altogether as it is relatively rare in comparison to other transport modes (Dill, 2009).

Bicycle count surveys are one of the most common methods for gathering bicycling data, though they suffer from a number of limitations. While counts provide bicycling travel volumes in particular locations, additional conclusions about bicycling behavior are difficult to make. Typically no additional route information is collected, so full travel behavior is not captured and cannot be inferred and geographic contexts are not determined (Kuzmyak and Dill, 2014). Count sites are not selected randomly and might only be based in locations where counts have occurred historically, popular or known bicycling routes, or where known problems with bicycling (e.g., crashes) occur (Ryus et al., 2014). Further, manual counts are temporally limited, only conducted annually or

semi-annually, for one day over a period of a few hours (Kuzmyak and Dill, 2014). Automated count technologies that passively collect travel volumes as riders pass by or over them have become more popular, but are associated with some of the same limitations as manual counts (Kuzmyak and Dill, 2014). Once installed, these tools can collect count data over longer periods of time though they have the same problems as manual counts in terms of collecting data from areas that are not randomly selected. Individual technologies are also associated with their own limitations, for example walking, bicycling, and skateboarding behavior may all be recorded as the same type of movement by an infrared counter (Kuzmyak and Dill, 2014). These limitations have caused a gap in research relating to the geographic contexts that cyclists travel through.

Despite knowledge of these factors, little is known about the broad generalizability of these findings because sample sizes are small and localized. Further, while findings address infrastructure, they do not address the larger geographical contexts (i.e., specific neighborhood features) within which people ride bicycles. Overall, these conventional approaches for bicycling data collection might be suitable for discovering general bicycling trends though they remain limited in spatial and temporal scope.

2.1.3 Crowdsourced Data

Crowdsourced bicycling data which often originate from smartphone applications have the ability to overcome data limitations to provide more information on human mobility. The individual trip level movement paths generated from new data sources have fine scale detail and a great number of users in many different areas. With consideration of their representativeness and scope, these data may complement or replace conventional data collection methods for bicycling activity. Further, analytical approaches that have not been possible with conventional data because of its limitations could be applied to these new data to discover new movement behaviors. There remains a gap, however, in examinations of mobility using these data; few studies have examined bicycling activity using crowdsourced data. It remains unknown whether crowdsourced data lead to the same or differing conclusions as conventional data when studying bicycling activity.

In addition to providing more detail and greater data volume, crowdsourced data help overcome difficulties associated with public participation. Traditionally, public

participation in planning processes has been limited despite being considered important to successful planning efforts (Misra et al., 2014). Though information is easier to access, public participation is declining because it still often relies on physical presence in the process wherein certain groups may be excluded because of their inability to attend (Misra et al., 2014). Crowdsourced data overcome this limitation by providing a way for citizens to engage in civic decision making and public advocacy and not requiring citizens to be physically present at a particular time and place (Le Dantec et al., 2015). Using data generated from crowdsourced means, city planners and decision makers have new ways to consider developing projects and transportation planning in particular. Sustainable decision making requires both that the data be produced and that it be shared and used by relevant stakeholders (Le Dantec et al., 2015). Cycle Atlanta is one bicycling app that the city has used to make decisions about transport. The app provides information about routes as well as barriers riders encounter (e.g., pot holes). The city can then use the data to make knowledge based and data driven decisions about infrastructure developments to support bicycling activity. Cycletracks is another app that lends data on bicycling movement to the city of San Francisco and the app provides data for cyclists as well; the service is aimed at public participation and civic engagement and motivated by a lack of data on bicycling activity (Lee et al., 2014; Misra et al., 2014).

Crowdsourced data are useful in a transportation context because they easily engage users within a region, and in the case of bicycling activity, are more cost effective than conventional survey means (Misra et al., 2014). Depending on the platform, geographically spread and diverse types of users can lend data to the system which may not be the case in traditional public participation formats where attendance at specific meetings is required (Misra et al., 2014). Though smartphones are increasingly present, their use is not as popular with certain age, socioeconomic, and ethnic groups and therefore input based on them can be biased (Windmiller et al., 2014).

In general, smartphone samples tend to under-sample females, older age groups, and low income populations while oversampling particular minority ethnic populations (Windmiller et al., 2014; Blanc et al., 2016). Even with smartphone ownership, familiarity with its capabilities or availability of apps could hinder user input from

particular groups (Blanc et al., 2016). With bicycling activity in particular, it must also be considered that many people who ride or rely on bicycles for transport will not also be motivated to log those rides on a smartphone app. Despite these potential sampling biases, they are only a concern in conclusions we make about bicycling activity if the activity represented in crowdsourced data differ from those represented in conventional data. Further, the opportunity crowdsourced data bring to lend new insight also brings new challenges related to their large volumes and potential low quality (Leao et al., 2017).

To date, few studies have examined crowdsourced data related to bicycling and those that have been conducted have typically used aggregated data from the Strava fitness application. One such study found that bicycling infrastructure was only moderately associated with the density of bicycling activity, though fitness-oriented cyclists using Strava may not seek out urban areas where infrastructure is located to support commute activity (Griffin and Jiao, 2015). In another study, categorical volumes (low, medium, high) of crowdsourced cyclists had some correspondence with manual counts, though presence of bicycling infrastructure was not predictive of bicycling volumes (Jestico, Nelson, and Winters, 2016). Recreational cyclists using the Strava app also tended to ride around residential land and along short, connected streets with low traffic volume (Sun et al., 2017). Further, recreational ridership tended to be outside the city center (Sun and Mobasheri, 2017). Individual-level crowdsourced data may overcome limitations related to this research design by providing more detail on both routes and trip purposes.

Crowdsourced data that provide detailed information about how people travel through urban areas might lend new insights into the relationship between urban structures and commute times for non-motorized modes, given examinations that make use of those data. While walking as a transport mode has been studied, bicycling remains under examined in comparison (Winters et al., 2016). The features of neighborhoods and networks that support walking or bicycling share some common characteristics, for example short street blocks or street level destinations (e.g., first floor retail), though bicycling involves additional features that may be relevant in supporting its uptake as an

activity (e.g., bicycle lanes) (Winters et al., 2016). Results from studies on cyclist behaviors and preferences have revealed some characteristics associated with bicycling activity. In general, commuting cyclists show preference for shorter, direct (i.e., few turns) routes, lower traffic volume and speed, and bicycling specific infrastructure (e.g., separated bicycle paths) and areas with more bicycling infrastructure tend to have more cyclists (Broach, Dill, and Gliebe, 2012; Winters et al., 2010; Winters et al., 2013). Many cyclists ride along bicycling infrastructure for a high proportion of their ride, though bicycle infrastructure might comprise a proportionally low amount of distance among the full street network (Dill, 2009). This indicates that cyclists may be willing to travel further distances to ride on bicycling infrastructure, though analysis comparing shortest distances between origins and destinations to actual travel paths is needed to determine whether this is the case (Dill, 2009). Other features of the built environment that influence bicycling include land use, density, and workplace accessibility (Handy and Xing, 2011). In contrast to general findings regarding the positive association between amounts of bicycling activity and bicycling infrastructure, Dill et al. (2014) found no increase in bicycling activity where new bicycling boulevards were installed. Discrepancies such as this require further examination and the high volume, detailed GPS traces of bicycling trips generated from crowdsourced data may provide clearer insights into bicycling activity and the infrastructure that supports it.

2.1.4 Data Usage in Mobility Studies

The goal of this section is to discuss human mobility and the conventional data used to examine it. Key features of new crowdsourced data and their related findings on human mobility are explained. Finally, limitations of both conventional and crowdsourced data are discussed.

Recording mobility, in both conventional and crowdsourced data, consists of a record of a moving object's path through a discretized series of spatial locations at particular time intervals (Long and Nelson, 2013). Specifically, mobility data can be recorded as a simple {ID, XY location, time} sequence for an object or represented as just the origin and destination of that path (often referred to as OD or O-D trajectories). For example, travel diaries and mobile phone data can be used to create movement paths

for individuals that indicate their mobility across an area. In addition to the spatial location of movement, additional features of activities including trip origins and destinations, modes of travel, routes taken, times, and trip purposes can be included.

There are three conventional methods that are used to collect human movement and mobility data: sensor detectors, travel surveys, and travel diaries. Sensor detectors estimate or measure vehicle traffic flows. The technologies include inductive loop detectors or pneumatic road tubes, which count vehicles passing over them at particular points along roadways (Leduc, 2008). There are also less intrusive methods to measure traffic such as manual counts, infrared, and other sensors. Manual counts utilize observers to record information about vehicles and pedestrians that pass by their location (Leduc, 2008). Other types of sensors detect moving objects that pass through their field and may be able to record the type of vehicle and speed in addition to basic counts.

Travel surveys are larger scale (e.g., nation, state, county) data collection efforts that are generated by surveying randomly selected, representative samples. Data are collected regarding all trips a participant takes during the study period, which is usually one day or less than one week (Freeth, 2000; Pucher et al., 2011). Travel diaries are similar to surveys but typically collect more detailed information on travel paths and destinations. Diaries involve recording detailed daily travel from a sample group of study participants or all members of a household. Some travel diaries are collected on their own, while others may be used as a supplement to a larger travel survey. Information about each trip a participant takes during the study period is recorded including origin and destination location, time of travel, mode of travel, and trip purpose (Hirsch et al, 2014; Wolf, Guensler, and Bachman, 2001). More recent efforts at travel diary studies have used GPS recording alongside the traditional written diary (Hirsch et al, 2014).

In addition to the conventional data sources for mobility data, there are newer data collection methods, driven by the ubiquity of GPS and other mobile technologies. These new, high-resolution movement data sets are often generated from Big Data sources, volunteered geographic information (VGI), or crowdsourced data. The term “Big Data” refers to information that is collected frequently, with great volume, from a wide variety of sources (Goodchild, 2013). In terms of mobility data, many of these sources are

associated with automatic data collection from transport services use such as smart card data from public transport and transport hire services (e.g., taxi or bicycle sharing schemes) as well as location-aware devices such as mobile phones.

VGI is user-generated spatial information that includes websites like OpenStreetMap which allow users to create content with spatial or geographic information, as well as smartphone applications where users record information about trips they take while participating in activities like walking or bicycling. These geosocial networking applications (e.g., Strava, RiderLog), a subset of VGI, utilize the GPS functionality of smartphone devices to allow users to record activity locations and often have a social component in terms of connecting with other users (Elwood, et al, 2012). Typically the location information, such as a GPS trace, is gathered and shared automatically by the device and application after a user has opted into the service (Elwood, et al, 2012). The detailed information about activities that people undertake is possibly one of the most valuable aspects of VGI (Goodchild, 2007). In general, it is gathered more frequently and is less expensive than conventional data as well (Goodchild, 2007; Goodchild and Li, 2012). Detailed GPS data sets can be used to examine both individual and group level movement behaviors, which allows for broader application contexts (Meijles et al., 2014). For example, GPS and personal itinerary information were used to supplement one another to determine different groups of park users (Meijles et al., 2014), and VGI has been used to map and describe different areas (e.g., Wikimapia, OpenStreetMap) (Elwood, 2012).

Smart cards are transit passes that are used as automated fare collection rather than traditional paper tickets or magnetic stripe cards (Bagchi and White, 2005). Cards are typically tapped or swiped on a receiver when boarding and alighting public transport which provides an electronic record of where and when passengers use public transport modes (Bagchi and White, 2005). Movement data from taxi trips typically include, at minimum, origin and destination locations and the date and time of each trip, though full GPS traces could be included in some cases (Guo et al., 2013). For cycle hire schemes (i.e., bicycle share), data typically include the origin, destination, date, and time where a customer checked out a bicycle for use. Mobile phone data typically consist of individual

call records or aggregated call volume data (Gao et al., 2013). Call record data includes user information, receiver and base-station location, date/time and duration (Gao et al., 2013). The data are recorded as users communicate via calls or messages over the cellular network and each communication is associated with the latitude/longitude position where the base-station is located (Gao et al., 2013). Broad mobility patterns, flows, and clusters can be detected using the origin and destination points associated with these data and some individual travel behavior can be examined in systems where user information is collected. For example, Gao et al. (2013) used mobile phone records to discover interaction patterns between different groups; results showed that despite the ability to use a mobile phone anywhere, groups still tended to communicate within close spatial proximity. Guo et al. (2013) were similarly able to distinguish regions using graph partitioning of vehicle trajectories. Results from that study identified clusters of travel within 10 regions that allow for spatio-temporal patterns to be better understood (Guo et al., 2013).

While the previously discussed conventional sources of mobility data have been used a great deal to examine broad travel activities, two key limitations appear to include coarse spatio-temporal scale and limited sample sizes. In terms of active transportation, representation in conventional data collection methods suffers especially from coarse scale and limited sampling. Walking and bicycling tend to occur over shorter distances and durations than motorized travel, so conventional data that rely on counts or origin-destination locations do not contain enough detail to examine activity at the localized scales in which it occurs and assumptions about factors that influence movement behavior may not hold. Since walking and bicycling are also associated with their own infrastructures, and different levels of interaction with the built environment and neighborhood contexts, finer scale data are needed to examine those factors that might influence movement behavior. While walking and walkability has been examined, little research has given detailed insight into where, when, and how people use bicycles at fine geographical scales. The finer-scale data and GPS traces collected through crowdsourced data could generate new insights on movement by allowing for detailed examination of full travel paths, rather than being restricted to general movement behaviors or broad

patterns. In terms of limited sampling, conventional methods might only collect volumes at particular locations or origin-destination locations, crowdsourced data provide new information about the actual routes people take to move between locations. This kind of detail is needed to generate knowledge about the network, neighborhood, and geographic contexts that drive the smaller-scale movement behavior associated with non-motorized modes. Detailed GPS data sets can be used to examine both individual and group level movement behaviors, which allows for broader application contexts (Meijles et al., 2014).

Just as conventional data collection methods have limitations for representing human mobility, crowdsourced data too have limitations by potentially being biased thereby limiting their use to generalizable representation of greater populations. Users who generate data using smartphone technology are generally self-selected and those who also have access to the technologies and resources (e.g., money, time) required to take part (Goodchild, 2007; Heipke, 2010). Participants also have varied motivations for contributing to these data sources, for example convenience of sharing information, personal satisfaction or promotion, or in some cases passive contribution when data are collected and disseminated in some way as part of a service (e.g., consumer information collected from store loyalty cards or activities utilizing public transport) (Goodchild, 2007). As such, the data potentially represent only the users specific to their application and not the general public. Depending on the source, commuters, students, children, and occasional or recreational non-motorized activity could be missed completely. For these reasons, comparisons of conventional and crowdsourced, examinations of what the data represent are needed. Critical examination of the features characteristic of these novel data will ensure that they are applied to problem contexts in such a way that biases are minimized (Elwood et al., 2012; Jackson et al., 2013). While crowdsourced data have the potential to lend new insights into human mobility because of their increased detail, and active transportation in particular, there is risk in relying on data that are not authoritative and potentially biased. In a transportation planning and decision making context, it is important to ensure that no one group or area has a disproportionate share of the associated costs or benefits. It is possible that relying on information generated only by

people contributing to crowdsourced and other urban movement data that require access to the resources they depend on could lead to increased inequities in transportation planning and policy.

2.1.5 Summary of the Literature

The environments through which people move, or geographic context, influences their mobility and differing methods of defining those environments may change conclusions made about activity. Active transportation is one form of mobility activity that is influenced by geographic context, though detailed examinations have been limited by the conventional sampling strategies used to gather data. These conventional strategies are primarily limited by coarse spatio-temporal scale and relatively small sample sizes and crowdsourced data may overcome these limitations. Understanding of bicycling activity in particular stands to be improved by using novel crowdsourced datasets. Despite their potential utility, there is a gap in understanding both how crowdsourced data compare to conventional data as well as how they can be used to improve our understanding of mobility in urban areas.

2.2 Research Objectives and Dissertation Plan

Research shows that better and more detailed information about where and when people use bicycles is needed so the processes that influence riding can be better understood. Research that generates detailed information about the specific neighborhood and street level contexts in which people bicycle can aid planning decisions that will increase bicycling rates to improve the health and sustainability of urban areas. A better sense of the contexts that facilitate bicycling as a practical transport option will inform improvements that would encourage or facilitate bicycling. Those contextual features that facilitate bicycling behavior may also highlight factors that prevent people from adopting bicycling as a transport mode. Network and other street level built environment features from areas with low bicycling activity can be compared to those with higher levels to determine which factors drive movement behavior.

The overarching objective of this dissertation research is to identify how crowdsourced data can be used to better understand human mobility, and bicycling activity in particular, and built environment context in urban areas. In light of the gaps in

research relating to bicycling activity and crowdsourced data, the research herein seeks to address human mobility, and bicycling in particular, using crowdsourced data. Three related research questions broadly compare crowdsourced and conventional data, examine the factors that underlie bicycling activity, and address the economic value of bicycling infrastructure and bicycle ridership. These questions relate in using crowdsourced data to incorporate geographic context in understanding bicycling in urban areas. First the research addresses the gap in understanding how crowdsourced and conventional data correspond in representing bicycling activity. It compares crowdsourced and conventional data by asking how do crowdsourced and conventional bicycle data correspond in representing bicycling activity? Related, what are the factors that underlie that bicycling activity? Second the research aims to understand human mobility, bicycling activity in particular, through the use of crowdsourced data. The specific research question is what are the factors that influence where bicycling data are represented in crowdsourced data and what is missed when we use crowdsourced samples? Further, are those factors the same between differing datasets or do different factors drive activity among them? Finally, the research examines the economic value of both bicycling infrastructure as well as neighborhood ridership volume. While density or proximity of some bicycling facilities have been linked to higher home prices, there is a gap in understanding the influence of actual bicycle ridership in those areas. This research seeks to answer what is the impact of bicycle ridership on residential housing prices?

2.2.1 Dissertation Plan

The questions posed are addressed by three independent but related research papers. The papers align by examining urban mobility using crowdsourced data. The first paper is entitled "Comparing Spatial Patterns of Crowdsourced and Conventional Bicycling Datasets". As stated, conventional data on bicycling activity are limited in their spatial and temporal representation; crowdsourced data may overcome those limitations by providing more detailed information on ridership. Little research has addressed whether crowdsourced and conventional data correspond, or whether the new crowdsourced data are representative of broader bicycling activity. This paper explores

the gap in understanding how crowdsourced and conventional data compare in representing bicycling activity using local spatial analysis.

In order to fill the gap in understanding bicycling activity through crowdsourced data, the second paper, entitled "Factors that influence cycling activity density in Sydney using varied crowdsourced datasets", examines the contextual factors that determine where crowdsourced data are contributed. Systematic examination of several sources is needed to determine not only what activities the data represent, but the social and built environment drivers of the areas where those activity data are collected (i.e., positive space). The aim of this paper is to assess how the physical environment, bicycling infrastructure, and sociodemographic features of Statistical Areas Level 2 (SA2) in the Greater Sydney region predict where crowdsourced bicycling data are (i.e., positive space) and are not (i.e., negative space) contributed.

The third paper, entitled "Impact of Bicycle Ridership on Residential Housing Prices" addresses the economic impact of both bicycle facilities and bicycling ridership on local home prices. While the effects of transit on housing prices have been well studied, the influence of bicycle ridership has not been examined. This is partly because conventional data have been limited in the spatial scale necessary to measure bicycle ridership. Using crowdsourced data, this study examines how neighborhood ridership influences housing prices to fill the gap in knowledge about the economic value of neighborhood ridership.

CHAPTER 3
COMPARING SPATIAL PATTERNS OF CROWDSOURCED AND CONVENTIONAL BICYCLING DATASETS*

This chapter is derived from a manuscript published in Applied Geography with coauthors Elizabeth Wentz, Trisalyn Nelson, and Christopher Pettit. The full citation is: Conrow, L., Wentz, E., Nelson, T., & Pettit, C. (2018). Comparing spatial patterns of crowdsourced and conventional bicycling datasets. Applied Geography, 92, 21-30.

3.1 Abstract

Conventional bicycling data have critical limitations related to spatial and temporal scale when analyzing bicycling as a transport mode. Novel crowdsourced data from smartphone apps have the potential to overcome those limitations by providing more detailed data. Questions remain, however, about whether crowdsourced data are representative of general bicycling behavior rather than just those cyclists who use the apps. This paper aims to explore the gap in understanding of how conventional and crowdsourced data correspond in representing bicycle ridership. Specifically, we use local indicators of spatial association to generate locations of similarity and dissimilarity based on the difference in ridership proportions between a conventional manual count and crowdsourced data from the Strava app in the Greater Sydney Australia region. Results identify where the data correspond and where they differ significantly, which has implications for using crowdsourced data in planning and infrastructure decisions. Fourteen count locations had significant low-low spatial association; similarity was found more often in areas with lower population density, greater social disadvantage, and lower ridership overall. Five locations had high-high spatial association, or were locations of dissimilar rank values indicating that they did not have a strong spatial match. Higher coefficients of variation were associated with population density, the number of bicycle journey to work trips, and percentage of residential land use for the significant locations of dissimilarity. The Index of Relative Socio-economic Disadvantage (IRSD) and bicycle infrastructure density were lower than the locations that were not significantly dissimilar. For the significant locations of similarity, all coefficient of variation measures were lower

than the locations that were not significant. Areas where ridership show locations of similarity are those where it may be suitable to substitute conventional data for the more detailed crowdsourced data, given further investigation into potential bias related to rider demographics.

3.2 Introduction

Progress in planning and research for active transportation, or non-motorized transport modes like walking and bicycling for practical purposes, is limited by a lack of data related to where and when people use those modes. Since active transport trips tend to be shorter in duration and occur at finer-scale levels of movement, they necessarily require detailed fine-resolution data for analysis at the neighborhood level where the activity, and therefore policy and planning decisions, occur (Cervero and Duncan, 2003). The lack of reliable data and knowledge about non-motorized travel has limited measures of accessibility and examinations related to human mobility using those modes (Iacono et al., 2010). Further, there is a lack of data that can be used to link non-motorized transport behavior with infrastructure and other network features that may influence it (Broach et al., 2012). This research has two goals: first to determine how crowdsourced and conventional bicycling data correspond in representing bicycling activity and second to determine how that correspondence helps understand factors that underlie bicycling activity.

For bicycling in particular, data limitations are associated with the sampling strategies for conventional data collection. The primary conventional methods for collecting bicycling travel activity data are manual bicycle counts, automated bicycle counts, regional travel surveys, and direct questionnaires. Manual bicycle counts are one of the most common methods for gathering bicycling data; they are conducted by counting travel volumes at specific locations for all riders who pass the location. The advantages of counts are that they do not depend on user participation, though they also have significant disadvantages. No additional information is collected, so route information, cyclist demographics, and reasons for the trip are not included (Kuzmyak and Dill, 2014). Another disadvantage is the representation of the sample is limited both spatially (e.g., count locations may not be spatially distributed in such a way to better

understand problems) and temporally (e.g., typically conducted annually or semi-annually for a few hours on one or two days) (Ryus et al., 2014; Kuzmyak and Dill, 2014). One alternative to manual counts is automated count technologies that continuously collect travel volumes as riders pass the counter (Kuzmyak and Dill, 2014). While these counters solve the problem of temporal sampling, the spatial sampling problem and the lack of associated travel data remain unsolved. Regional travel surveys tend to be more comprehensive and include all travel modalities including automobile and public transportation. Travel surveys typically sample individual travel activities over a short period of time, such as a day or a week's worth of travel and as such, are able to gain more insight on route information and reasons for travel (Dill, 2009). Despite the additional information that is collected, travel surveys are often still limited in overall sample and detail. Since bicycling constitutes a comparatively small percentage of modal share, the activity may be missed completely, and route information may not be collected. If route information is inferred later, many surveys assume a cyclist takes the shortest path between an origin and destination (van Heeswijck et al., 2015) though this may not be the case as cyclists are willing to go out of their way to avoid traffic and stay on bicycling infrastructure (Dill, 2008). Information about bicycling activity may also be gleaned from direct questionnaires to cyclists. Direct questionnaires have been used to examine how bicyclists view urban design (Forsyth and Krizek, 2011), perceptions of risk while bicycling (Lawson et al., 2013; Møller and Hels, 2008), bicycle facility planning (Rybarczyk and Wu, 2010; Dill, 2009), and route-choice modeling (Broach et al., 2012; Dill and Gliebe, 2008). These questionnaires generate valuable demographic and experience or opinion based data, but they also have a tendency toward limited spatial coverage and small sample sizes.

Novel crowdsourced data from smartphone apps have the potential to improve on the resolution of conventional data collection methods. Data collected from personal mobile devices overcome limitations in spatial and temporal scope by both providing finer-scale specificity about the actual route and not depending on the timing of a survey. These smartphone-based geosocial networking apps utilize built-in GPS functionality to allow users to record activity locations and often have a social component in terms of

connecting to, competing with, or sharing information among other users (Elwood et al., 2012). The finer scale detail from near-continuous time frames is needed to generate knowledge about the neighborhood and network contexts that drive behavior associated with non-motorized modes. Detailed GPS data sets can be used to examine both individual and group level movement behaviors, which allows for broader application contexts as well (Meijles et al., 2014). The challenges with using crowdsourced bicycling data are the inherent biases due to self-selected participation. Users who generate data using smartphone technology are limited to those who have access, have the motivation to participate, and who have the resources (e.g., money, time) required to take part (Goodchild, 2007; Heipke, 2010). This means that crowdsourced data sources may be biased and limiting in terms of extensive and generalizable representation of greater populations, despite potential benefits associated with the detail and information they may provide. Groups such as commuters, students, children, and average recreational riders could be missed completely. This is problematic because relying on biased information could lead to increased inequities in transportation planning and policy.

There are pertinent questions related to the effectiveness of these crowdsourced data for understanding the drivers of bicycling behavior because the data may be biased or of poor representative quality. Studies that have compared crowdsourced bicycling data to conventional bicycling data have found similar ridership volumes with correspondence closest when volumes were grouped categorically such as low, medium, and high volumes or according to peak hours, suggesting that spatial patterns between them may be similar (Jestico et al., 2016). For example, all riders in an urban downtown may use similar routes because of limited choices, which helps explain the relationship between crowdsourced and conventional count data. Using bicycle count and survey data, increased ridership is usually associated with increased bicycling infrastructure (e.g., bicycle lanes, separated cycle paths) (Broach et al., 2012). In the case of crowdsourced data, presence of bicycling facilities was not predictive of ridership volumes (Jestico et al., 2016). The poor association between crowdsourced ridership volume and bicycling infrastructure may be related to the manual count locations in known areas of bicycling activity. Another study using Strava data found that bicycling infrastructure was only

moderately associated with the density of bicycling activity; it is possible that fitness-oriented cyclists using mobile apps such as Strava may not seek out urban areas where infrastructure is located to support commute activity (Griffin and Jiao, 2015).

Crowdsourced data also show wide potential for examinations of urban areas and urban transportation. For example, ridership volumes collected by Strava have also been used to show how changes in bicycling infrastructure influence bicycling activity over the short-term (Heesch and Langdon, 2017). Though changes in ridership volumes occurred around bicycling infrastructure improvements, conventional data were still needed to adjust volumes across the region because of variations in app use across the area (Heesch and Langdon, 2017). Since all riders in the area do not use the app, their differing preferences for particular routes or types of infrastructure may influence conclusions if volumes are not adjusted based on the larger bicycling patterns from all riders in a region. Further, Strava ridership have been associated with particular neighborhood characteristics like highly connected streets within residential areas, though the characteristics and riding environments of cyclists using the Strava app may differ from those who are not (Sun et al., 2017a; Sun, 2017). Health and exposure characteristics have also been explored using these data. Researchers found differences between recreational and commuting riders in terms of where they ride and how much pollution to which they are exposed (Sun and Mobasheri, 2017). Since recreational riders tended toward the outskirts of an urban area, they were potentially exposed to less air pollution than commuters in the same region (Sun and Mobasheri, 2017). These studies help indicate where infrastructure changes could be made to positively influence cyclist health, safety, and overall ridership.

As a step toward developing a method for conflating conventional and crowdsourced bicycling data, we seek to explore the, as yet, understudied area of understanding how crowdsourced and conventional data correspond in representing activity. Specifically, we first ask how do manual count data compare with Strava crowdsourced data in terms of bicycling activity volume? We conduct this exploration by analyzing the spatial pattern of crowdsourced data as compared to data collected through conventional methods in the Greater Sydney area. Spatial pattern analysis is used in this

context as a way to measure how the ridership in crowdsourced and conventional data correspond and show differences among the included datasets. The analysis provides greater precision of correspondence in the comparison as the approach avoids binning ridership into just low, medium, and high areas. While previous studies examined the strength of associations between manual counts and crowdsourced data, they did not examine the spatial associations and ridership patterns between locations for the differing data sources. Specifically, we compare manual bicycle count data with crowdsourced data using local Moran's I_i.

Spatial pattern analysis is a commonly used analytical tool to identify where in a study area there are highly correlated areas of activity. Since highly correlated areas of activity may give some indication about the processes that underlie them (Nelson and Boots, 2008), we then also explore socio-economic demographics and infrastructure in the area to determine their explanatory value related to the patterns of data correspondence and differences we discover. Between discovering areas of high and low correspondence and examining the contests that underline them, we will better understand the analytical potential that crowdsourced data may lend in studies of bicycling activity. Areas with greater population density should be more likely to have similar ridership patterns between datasets as cyclists in dense urban areas may tend to take the same routes (Jestico et al., 2016); the same mechanism is thought to drive ridership in areas with a great deal of bicycling infrastructure and residential land and areas where there are more people using bicycling as a journey to work mode. Since disadvantaged areas are associated with lower ridership overall (Cervero and Duncan, 2003), the datasets should show similar rates of low ridership in areas with greater disadvantage**.**

Beyond weak correlation to bicycling infrastructure, the analytical potential for understanding bicycling behavior with crowdsourced data remains unknown; comparison of ridership among conventional and crowdsourced data sets is needed to determine how they correspond. Their correspondence will reveal whether they are interchangeable, complementary, or display differing bicycling activity in the same spatio-temporal contexts. Examining covariates known to correlate with bicycling activity will lend

insight into whether areas of similarity or dissimilarity have common underlying contexts. If there are contexts common to those areas, we can have confidence in determining where Strava data can and cannot be used in place of conventional bicycling data. Considering that bicycling infrastructure was not associated with crowdsourced ridership volumes, and that bicycling infrastructure in urban areas tends to be located in and around denser city areas where people travel to work, it is possible that a fitness-oriented app like Strava is not as well suited for collecting bicycling activity for transport purposes. Bicycling for transport is the analytical focus in an active transport context because both recreational and transport bicycling have health benefits, but the goals of sustainable transport and traffic congestion reduction are associated with people using bicycling for trips where they would have otherwise used motorized means.

3.3 Methods

3.3.1 Study Area

The Greater Sydney area in New South Wales, Australia is used as a case study for the analyses herein (Figure 3.1, insets). The region has a population of 4.92 million people as of 2015 and though bicycling rates remain relatively low, the number of people bicycling to work has increased 50% in the Sydney area since 2006 (Transport for NSW, 2013). Sydney serves as the capital of New South Wales and is a major core of economic activity including finance, business, manufacturing, agriculture, among others (NSW, 2017). The area encompasses 12,368 square kilometers (ABS, 2017) with a variable climate within the area that typically ranges between a summer average of 23°C (73°F) and a winter average of 9°C (48°F) (NSW, 2017).

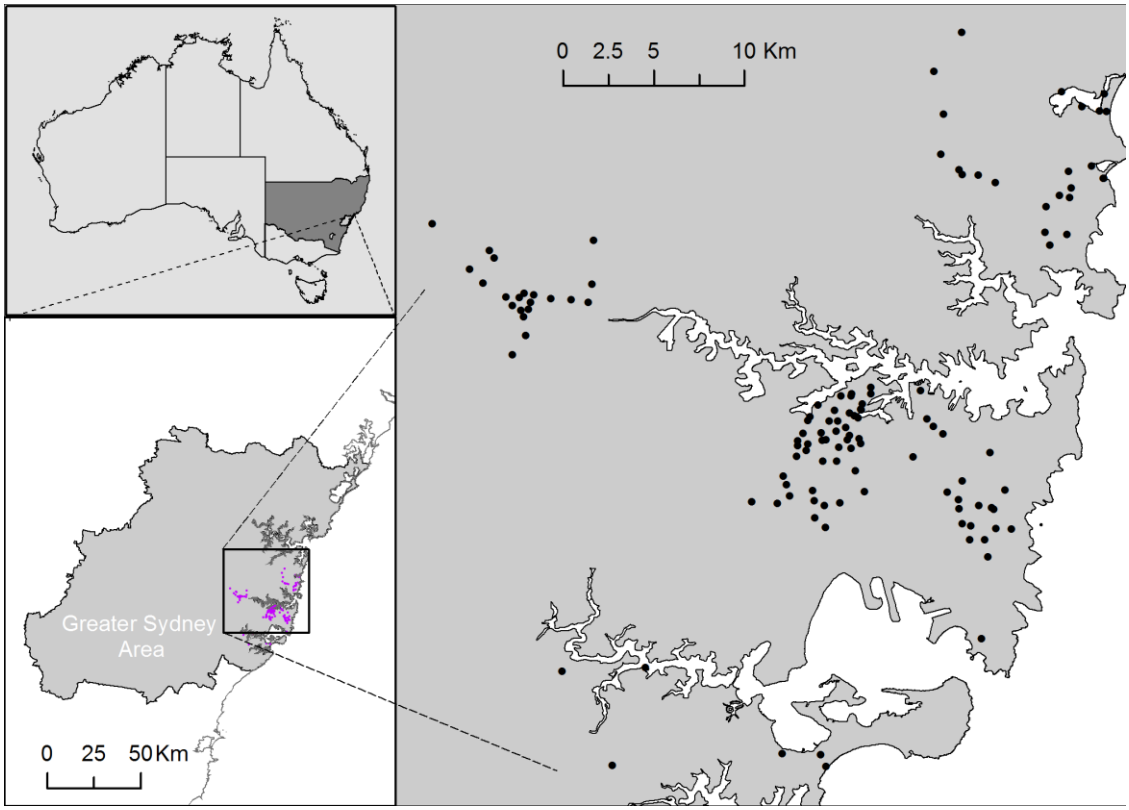

Figure 3.1. Greater Sydney area, New South Wales Australia, and the 122 count locations.

In New South Wales, 70% of residents indicated a willingness to bike for transport provided it was a safer and more convenient option (Transport for NSW, 2013). Transport for New South Wales (2013) has identified the need for providing improved infrastructure and connectivity to encourage increased use of bicycling for transport. Bicycling is a feasible transport option for many people in Sydney, given the large number of daily commute trips within 10 km and a majority of short trips have travel times that are comparable to motorized trips (BITRE, 2016; Ellison and Greaves, 2011).

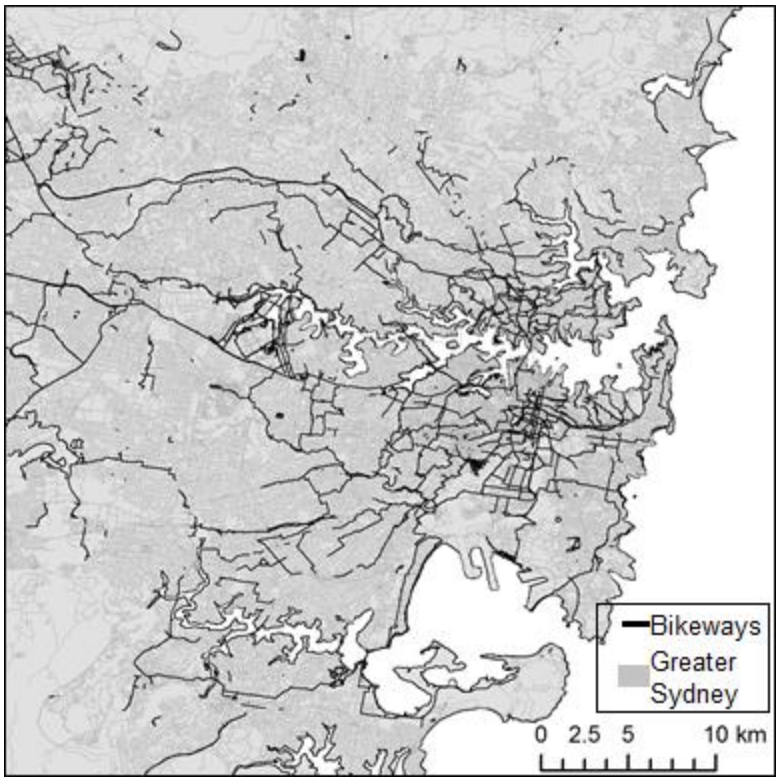

Figure 3.2. Bicycling facilities within the study area.

As shown in Figure 3.2, the Greater Sydney area has a variety of facilities that comprise the bicycling network including painted cycleways, shared vehicle and shared busway lanes, separated on-road, and separated off-road cycle lanes, though shared and painted lanes are the majority (Pucher et al., 2011). Despite infrastructure, the region also has several factors that hinder bicycling as a feasible transport option without continued improvement. The terrain is hilly in areas, with natural barriers and few crossings (Pucher et al., 2011). Planned cycle routes are aimed toward recreational cyclists, such as those paths delineated along coastlines, or serve outer regions with low bicycling demand for commutes (Figure 3.2). While these paths may be used daily commuting by some populations, they are less likely to support direct routes that facilitate bicycling as an inner city transport mode (Pucher et al., 2011).

3.3.2 Data

The data sources used for this study originate from the Super Tuesday manual count and the Strava smartphone app. The manual count data represent a conventional

manual count targeted specifically at capturing commuter bicycling activity. Strava data are crowdsourced data collected voluntarily through a smartphone app. The remainder of this section describes the features of each dataset and how they were integrated for use in this study.

The manual count data were collected between 7 and 9 am on March 1, 2016 at 122 locations across 12 council areas in New South Wales. Of the 12 council areas, only those within the greater Sydney area were included (Figure 3.1); those eight included council areas were Canterbury, Leichardt, Marrickville, Parramatta, Randwick, Sutherland Shire, Sydney, and Warringah. The data collection locations were not distributed evenly across the study area since they depended on participating councils. The data collection points were situated at key intersections or along bicycling specific infrastructure such as cycleways or bike lanes and volunteers recorded counts of cyclist movement along the streets and paths at each location. As shown in Figure 3.1, the count locations were focused within several areas of the Greater Sydney region: the central business district and inner west, Paramatta (west of the Harbor), and northern beaches (northeast of the harbor area). Each count location had a corresponding total number of cyclists that passed by it during the data collection period.

The crowdsourced data originate from the Strava app's commercial data product, Strava Metro, which is marketed for use by city planners, transportation departments, and advocacy groups (Strava, 2016). Rather than providing full paths, users' bicycling activity in a particular area is aggregated to protect privacy. The data product consists of counts within origin and destination polygons, flow volumes through network nodes (intersections), or flow volumes along street segments that are indicative of popular routes. Since the node based intersection flows did not always correspond to the Super Tuesday count locations, the data used for this analysis consist of flow volumes along street segments for weekday bicycling in the month of March 2016. The aggregate month was the finest time scale available for this analysis that did not result in data scarcity at count locations. The volume measure used in this analysis was the total number of riders on a segment, regardless of direction travelled. The overall Strava sample included trips from 84863 unique riders who generated 2,889,139 trips; 77% of the riders were male.

The majority of the sample was between ages 25 and 54 for both males and females, with age 35-44 having the largest number of riders (n = 17032 and n = 2845 for male and female respectively). The median distance travelled was 23.73 km and approximately 39% of the trips were marked as commutes.

To facilitate comparison between the two datasets, the Strava segment flow volumes were converted to counts that match the 122 Super Tuesday count locations. Each manual count location contained a description the streets included in that location's count, along with its direction (e.g., Darling St. [E], Beattie St. [W], Darling St. [NW]). The corresponding street segments were selected from the Strava data to generate the matching count locations; in some cases street segments in the digitized data format required additional cleaning. For example, some bi-directional streets were represented as two separated line feature segments (Figure 3.3) so one of the segments had to be excluded from the set of segments at that location. Specific rules and assumptions as to the inclusion or exclusion of street segments based on issues with matching are listed in Table 3.1 and further illustrated in Figure 3.3.

Table 3.1. Rules and assumptions for matching Strava street segments to Super Tuesday manual count locations.

Matching issue	Figure 3.3 example	Decision	Justification
Bi-directional street represented as separate segments	Segments a and b	Include segment in direction of travel into intersection (e.g. segment b)	Riders in the manual count are recorded according to their direction of travel
Segment separating bi-directional street, within count intersection	Segment d	Exclude segment	False/excess segment
Bi-directional street split into multiple segments, no intersection	Segment e	Include segment closest to count location	Closest segment is where a cyclist enters the intersection to be counted

Roundabout	n/a	Include street segments that enter/exit roundabout, exclude segments that comprise the roundabout itself	Roundabout curves often represented by multiple segments which over-counts cyclists.
Non-conventional intersection/street (e.g., median separated)	n/a	Include street segments that match direction in manual count description	Riders in the manual count are recorded according to their direction of travel

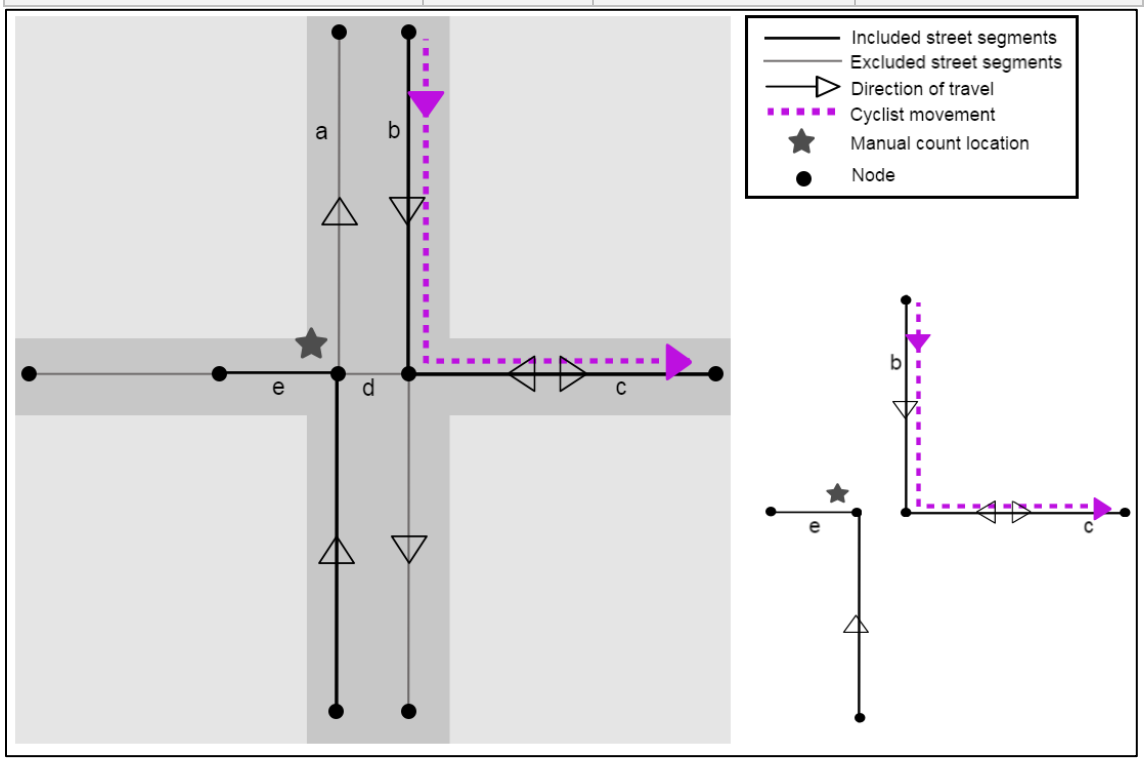

Figure 3.3. Example digitized representation of an intersection at a manual count location (left) and the finalized intersection representation after cleaning (right).

This process resulted in a set of segments that matched the description of streets included in the count at each of the manual count locations. This volume of riders on each segment was aggregated by location to derive a total sum of Strava ridership volume for each of the 122 count locations. This process was conducted using ArcMap 10.2.

While the volumes in the manual counts account for turning activity and therefore only count riders once as they move through an intersection, the Strava street segment volumes represent all riders on each segment so the aggregated segment activity is double counted. For example, Figure 3.3 depicts cyclist movement behavior as the rider turns from segment b to segment c. In this scenario, the manual count at this location would record one rider regardless of turning behavior or the segments crossed. The same turning behavior in the Strava segment volume would count the rider's movement as one at segment b and one at segment c for a total of two. For this reason, the total Strava ridership volume at all count locations comprised of two or more segments was divided in half once they were aggregated since each rider would have approached and departed the location via at least two segments. After data preparation, the manual count data contained 16,554 rides taken by bicyclists and the adjusted Strava data contained 65,062 rides across the 122 count locations.

Information about factors that are potential influencers of bicycling activity were obtained from the Australian Bureau of Statistics' National Regional Profile (NRP), New South Wales Office of Environment and Heritage (OEH), and The Socio-Economic Indexes for Areas (SEIFA) from 2011 at the Statistical Area 2 (SA2) level. The NRP data included estimated resident population density and method of travel to work including car, bus, train, and bicycle. The OEH data included the land use classifications for parcels in the Greater Sydney region. The SEIFA data included the Index of Relative Disadvantage (IRSD) score for each SA2 unit. The IRSD is a summary measure of the socio-economic conditions; low index scores indicate the most disadvantaged areas where higher scores indicate less general disadvantage (ABSb, 2017). The bicycle infrastructure data consisted of a polyline file obtained from OpenStreetMap representing segments of road and separated paths that were designated as bicycle infrastructure. The road network data originated from Strava as part of the included core data. To give some indication of accessibility, bicycle and road network densities were computed by dividing the length of bicycle infrastructure or road segments by the area in square kilometers of each SA2 unit.

3.3.3 Methodological Approach

To establish whether the data represent similar spatial patterns in volume of ridership at each count location, descriptive summaries were used to compare both ridership volumes and proportions among each data set. Local Moran's I_i was then used in this analysis to determine whether spatial patterns in ridership volume differed between the conventional and crowdsourced data; the statistic indicates the degree of spatial clustering of similar values around each observation. Positive values indicate clusters of similar entity values (high or low) and negative values indicate clusters of dissimilar values (e.g., a location with a high value and neighbors with low values) (Anselin, 1995). The entities in this case are the count locations along the network and the attributes are their respective rank difference of ridership volume. Since the aim in this analysis was to quantify and compare ridership between the datasets, a single value representing the difference in ridership volume between them was needed. To derive this value, the ridership volumes for each dataset were first ranked and a rank difference was computed according to equation 1, where R_i = the rank at location *i* for data sets *x* and *y*.

$$RD_i = (R_{xi} - R_{yi})^2 \quad (1)$$

Using rank difference, the statistic will detect clusters among the count locations where differences in rank of ridership volume differ from expectation (Boots, 2003). Detecting clusters is a first step in discovering the process that underlie some phenomenon (Nelson and Boots, 2008). Clusters in this context will indicate areas where soco-economic and infrastructure characteristics may be influencing bicycling activity volumes. The statistic (Equation 2) used for the analysis herein was defined as:

$$I_i = \frac{(n-1)z_i \sum_j w_{i,j} z_j}{\sum_j z_j^2} \quad (2)$$

Local Moran's I_i is computed for a row standardized binary spatial weights matrix based on the eight nearest neighbors for all *i* locations (i.e., a weight of 1 if location *j* is within the set of location *i* nearest neighbors and 0 otherwise). Since the local Moran's I_i values

are relative, standardized z scores and p-values derived from conditional randomization using 10000 permutations are used for interpreting the statistic (Anselin, 1995).The statistic was calculating using Python 2.7.13 and the PySAL library.

The factors related to socio-economic demographics and infrastructure were then explored to determine whether they were associated with the spatial patterns of similarity and dissimilarity between manual and Strava data counts. If the patterns are associated with those factors, they show the potential influence between population density, land use, IRSD, and journey to work modes as well as density of bicycle infrastructure in driving bicycling activity. The coefficient of variation was computed for each factor to indicate the relative variability of the factors between significant and non-significant count locations. QGIS 2.18.10 was used to associate each location with the socio-economic and infrastructure factors as well as to visualize results.

3.4 Results

The overall mean number of cyclists per location in the manual count was 136 for March 1st, 2016, the day the data were collected. The Strava sample mean was 533 cyclists per location for the month of March 2016. The differences in ridership volume between datasets are eliminated when considering the proportions of rides; the mean, standard deviation, and range are similar between the data sets (Table 3.2). Both data sets have a mean proportion of 0.82% as well as showing the highest proportions of ridership were just over 8.0%.

Table 3.2. Summary of Super Tuesday (March 1st, 2016) and Strava (March, 2016) descriptive statistics

	Super Tuesday		Strava	
	Count	**Proportion (as %)**	**Count**	**Proportion (as %)**
Total	16554		65062	
Mean	136	0.82	533	0.82
St. Dev.	227	1.37	810	1.25
Min	0	0.00	0	0.00
Max	1338	8.08	5296	8.14

The spatial distribution of the proportional ridership volumes at each location for each data set is shown in Figure 3.4. The correlation between ridership volumes for both data sets was 0.79, indicating a relatively strong positive correspondence in bicycling volumes across the study area (Figure 3.5). When divided into low, medium, and high bicycling activity the correlations dropped to 0.55, 0.17, and 0.57 respectively. While the ridership proportions between the data sets are similar in many areas, there are also areas where qualitative differences are present. Locations with the highest proportions among the manual count data tended to be near the Sydney central business district and the surrounding area south of the Harbor (Figure 3.4a). The lowest ridership mainly occurred at the southernmost locations, the northern suburbs, and some interspersed with the moderately-low locations immediately west and southwest of Sydney. Like the Super Tuesday sample, the highest ridership proportions in the Strava data were located in the central business district, though locations in the eastern suburbs displayed differing locations of higher ridership (Figure 3.4b). The areas immediately to the west and southwest of Sydney had a greater number of lower ridership sites with fewer moderately-low locations than the Super Tuesday data (Figure 3.4b). The Strava data also showed a greater number of northern suburbs and southernmost locations that had moderately-low ridership compared to low ridership in the same locations among the Super Tuesday data. Since correlation alone shows only the relationship between quantities, rank difference provides a more precise examination of the correspondence between the data sets which is needed to determine whether they contrast or complement one another and can be used interchangeably.

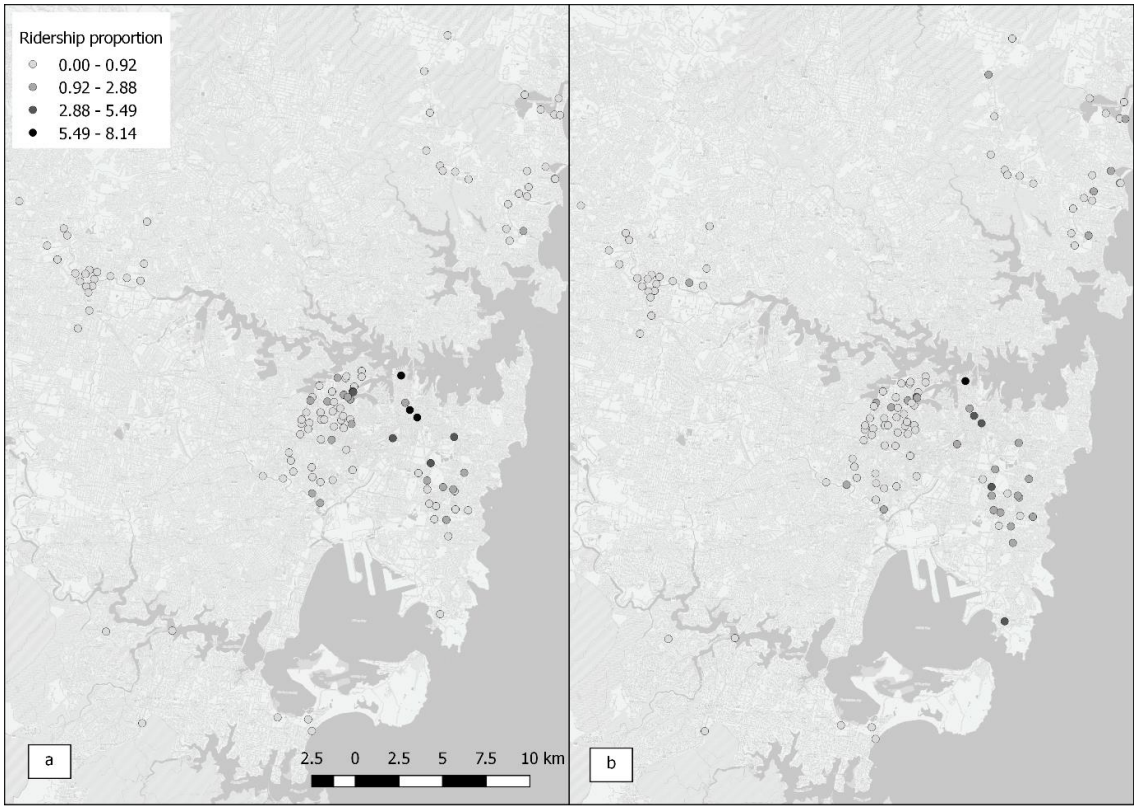

Figure 3.4. Ridership proportions (as %) for (a) Super Tuesday and (b) Strava

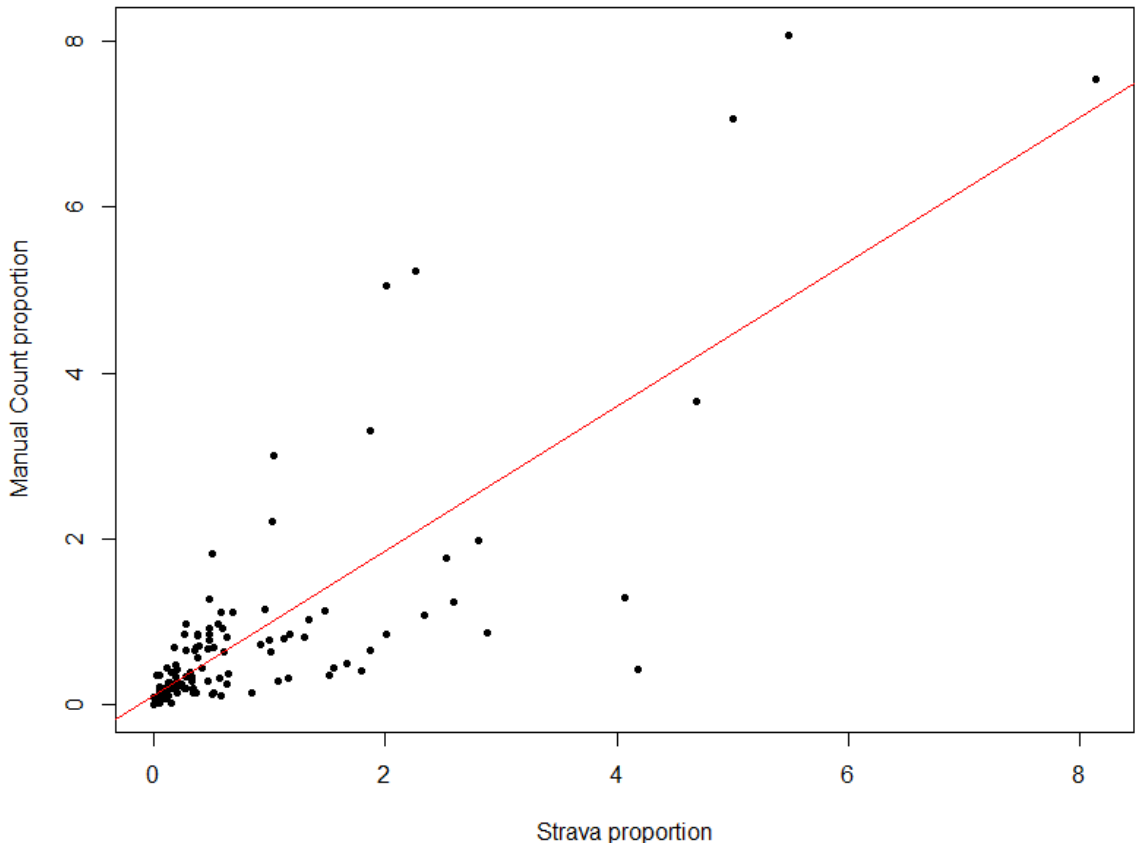

Figure 3.5. Correlation between Strava and Manual Count proportions, 0.79. Correlation for low, medium, and high ridership volumes were 0.55, 0.17, and 0.57 respectively.

In terms of rank difference, lower values indicate that the rank for a location was similar between the two data sources whereas higher values indicate greater difference in rank or said another way, greater mismatch between relative ridership volumes between the data sources. The spatial distribution of rank difference among count locations is shown in Figure 3.6. While count locations in the westernmost region of the study area had proportionally low ridership in both the Super Tuesday and Strava data, the rank difference values show differing levels of mismatch among them. Similarly, few areas in the northern suburbs or eastern suburbs show similar rank values. The Sydney central business district, where both data sets had relatively high proportions of ridership, also has low rank difference values. The area immediately west of Sydney shows a mix of similarly ranked locations where proportional ridership was low or moderately low for

both data sets, or where ridership was moderately low in one data set and moderately high in the other.

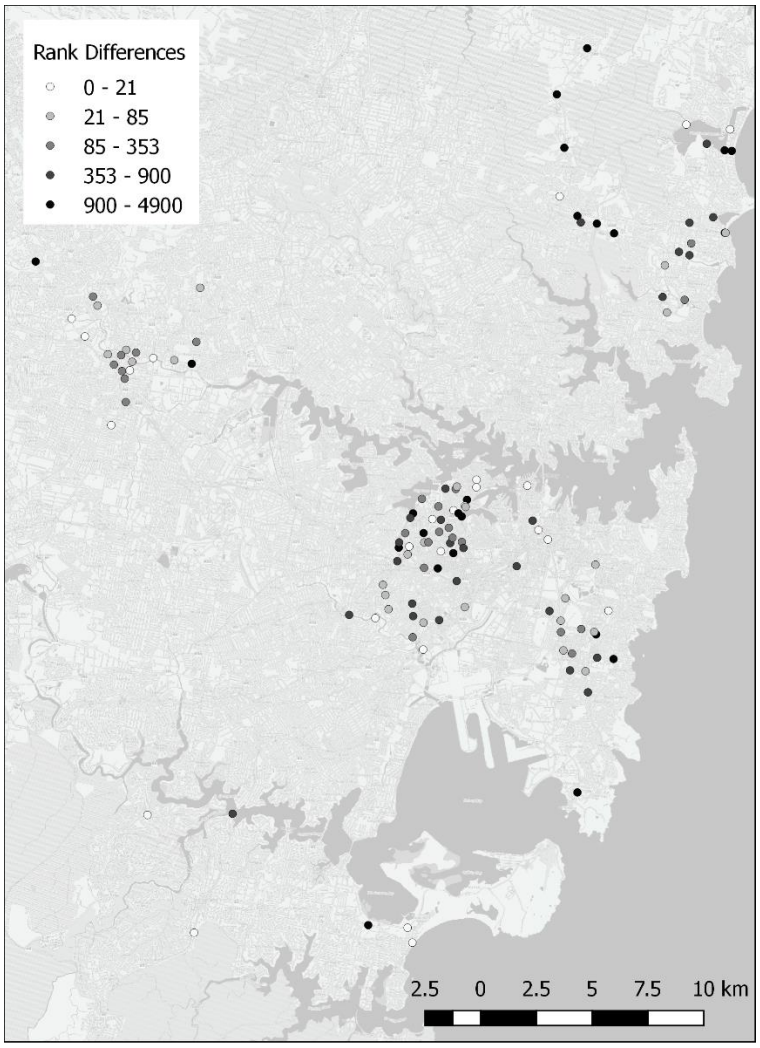

Figure 3.6. Spatial distribution in rank difference among count locations.

Figure 3.7 shows the results of the Local Moran's I_i analysis as the significant locations and the quadrant within which they fell. There were 14 locations that had significant low-low spatial association, or locations of similarity, (quadrant 3) while five locations show a high-high spatial association (quadrant 1), or locations of dissimilarity. There were also three locations with mixed spatial associations; two locations had high ridership ranks with low neighbors (quadrant 4) and one had a low rank with high neighbors (quadrant 2). Table 3.3 shows the relative variability for the rank difference locations of

dissimilarity and similarity. Higher coefficients of variation were associated with population density, the number of bicycle journey to work trips, and percentage of residential land use in the area for the significant locations of dissimilarity. IRSD and bicycle infrastructure density were lower than the locations that were not significantly dissimilar. Comparing significant locations of dissimilarity and similarity measures, the pattern in coefficient of variation is the same. For the significant locations of similarity, all coefficient of variation measures were lower than the locations that were not significant.

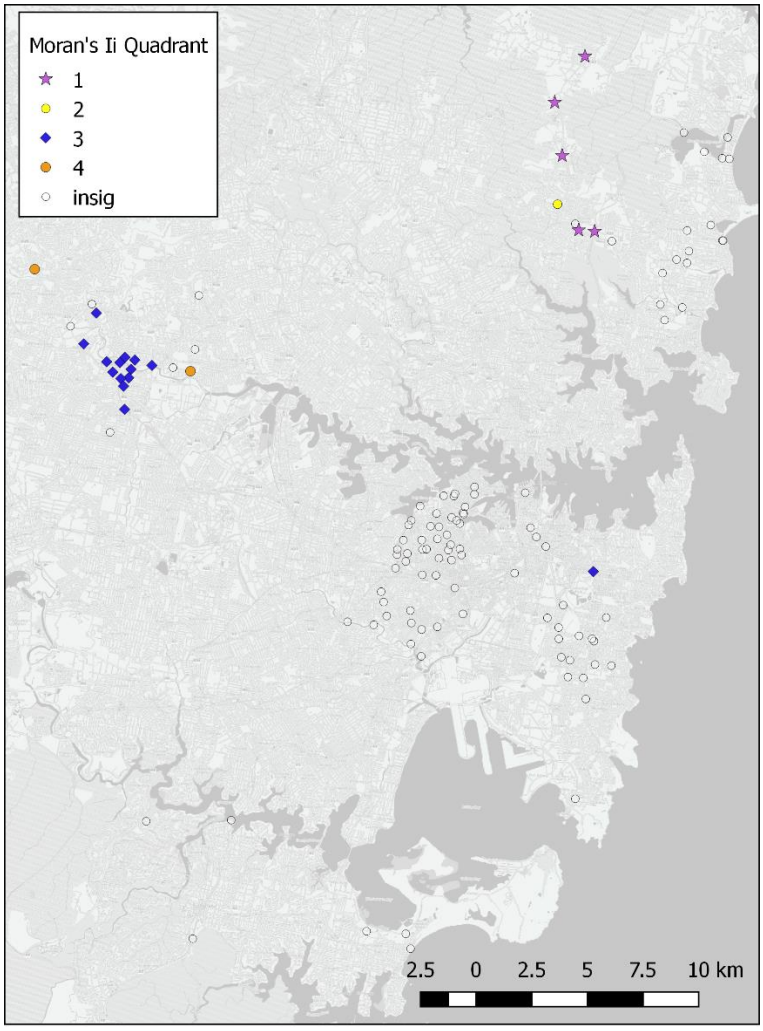

Figure 3.7. Location and quadrant of significant rank difference.

Table 3.3. Coefficient of variation for locations of dissimilarity and locations of similarity in rank difference.

Variable	Dissimilar locations, significant	Dissimilar locations, not significant	Similar locations, significant	Similar locations, not significant
Population density	82.22	50.03	29.41	53.22
IRSD	0.49	4.32	3.32	3.91
Number of bicycle JTW	79.99	76.72	55.69	71.15
Bike lane density	0.00	82.52	47.44	87.96
Road density	43.50	36.27	11.69	40.20
Residential land use	88.47	40.14	13.56	44.13

3.5 Discussion

Results from this analysis quantified the patterns in ridership volume among conventional and crowdsourced data to determine how they correspond in representing bicycling activity. Qualitative comparison of the data shows that the spatial distribution of ridership shows some difference between the conventional manual count and the crowdsourced Strava data sources. The finding that both data sources had elevated ridership proportions near the central business district aligns with prior findings that conventional and crowdsourced ridership correspond in urban downtowns where riders, whether recreational or for transport, use similar routes because of limited choices (Jestico et al., 2016). Further, it is possible that the barriers to riding associated with few

crossings along the harbor would necessarily funnel riders to specific areas of access, which would lead to elevated volumes regardless of data sources (Pucher et al., 2011).

Given that the Strava data showed medium and high ridership in the eastern Sydney suburbs (e.g., Randwick), while the Super Tuesday data did not, there are implications for data use cases and planning decisions dependent upon them. In thinking about bicycling for transport, we anticipate high ridership in urban downtown areas where many employment centers are located. The high ridership proportions in the eastern Sydney suburbs are not associated with the expected employment centers, demonstrating that the data reflect differing bicycling patterns in terms of where the populations that are captured prefer to ride. While it would require investigation beyond the scope of this paper, it is possible that Strava users who may be more focused on fitness and competition prefer riding in locations that are more removed from the dense, high traffic areas associated with central business districts (Griffin and Jiao, 2015). Well-developed infrastructure to support bicycling is key in encouraging people to switch from motorized modes to bicycling for transport. Bicycling could be a feasible option for many commuters in the Greater Sydney area, given installed infrastructure that supports it. Switching transport modes to bicycling for short trips will reduce problematic traffic congestion in the area, allowing improved traffic flow and travel times for commuters who must travel further distances (Transport for NSW, 2013).

Despite some relative spatial similarity in the distribution ridership proportions, the significant locations in the local Moran's I_i analysis demonstrate areas of interest in terms of rank difference. Ranks in this context are measuring the relative importance of a location within each dataset; the rank difference then is a measure of how similar the datasets are in terms of each location's standing. By creating locations of similarity and dissimilarity of bicycling activity, we can identify regions of varied bicycling activity beyond what can be generated by highlighting high or low volume point locations alone. This allows further understanding of the underlying factors of the spatial landscape that supports bicycling activities as captured by differing data collection methods.

The 14 locations in the Parramatta area that were locations of similarity (figure 3.7, quadrant 3) had low ridership overall and have significantly low rank differences of

ridership volume, relative to their neighbors. In addition to having low ridership, this low rank difference cluster indicates that the manual count and Strava data represent similar ridership volumes. These locations, therefore, are those where it may be appropriate to utilize the more-detailed Strava data for planning and design decisions if more details about the rider demographics in the area are known (i.e., whether the average Strava rider represents the average non-Strava rider). In terms of socio-economic and infrastructure characteristics that may play a role in generating low bicycling volumes that are similar between conventional and crowdsourced data, the western portion of the study area has few designated bicycle paths which may explain the overall low ridership since higher ridership is associated with installed infrastructure (Dill, 2009; Broach et al, 2012). This area also shows higher rates of using train or car as a journey to work mode as compared to bicycle. It is also of note that this area also had some of the lowest scores on IRSD indicating greater relative disadvantage in terms of social and economic conditions such as low income, education, and/or skill. Low income neighborhoods have been associated with decreased probability of taking bicycling trips (Cervero and Duncan, 2003) so this aligns with general findings on bicycling activity. In a similar vein, the area also has lower percentages of residential land use mix which helps explain the low ridership; Strava riders generally prefer residential land use areas for bicycling activity (Sun et al., 2017).

The five locations of dissimilarity were in the northern suburbs; this association indicates locations that have a high rank difference and whose nearest neighbors also have a high rank difference. All five locations had higher ridership proportions in the Strava data set as compared to the manual count data. Despite their popularity in the Strava dataset, these locations are not located on or near any bicycling infrastructure and have low bicycle infrastructure density relative to the locations of similarity areas (Table 3.3). This indicates again the importance of installed bicycling infrastructure in supporting commute activity rather than just recreational riding (Broach et al., 2012). Since many Strava users are focused on fitness, it is possible that despite not having infrastructure, the roads in this area in some way support riding for purposes other than commuting (Griffin and Jiao, 2015). Interestingly, the one low-high location in this

northern area of high-high rank differences was off the main streets as compared to the high-high points. Since it was more of an out of the way location, this may explain why the Super Tuesday and Strava ranks for it were similar. Like the western portion of the study area, there were low rates of bicycling as a journey to work mode in this northern area. In contrast though to the western area, the locations in the northern area are associated with higher scores on the IRSD which indicates that they are some of the least disadvantaged areas. Strava users are typically associated with higher socioeconomic status areas, so this relationship is in line with previous findings (Griffin and Jiao, 2016). As reflected here, the Strava data are likely biased toward recreational riders. Because the high-high spatial association indicates high rank difference surrounded by other high difference locations, this area is one where caution should be taken in substituting Strava data for conventional data.

The two high-low locations in the data were also located in the western most part of the study area. The high-low association indicates they have high rank difference while their neighbors were low - overall this means those two locations were dissimilar in terms of their position or importance in the data. For the furthest west high-low point, Super Tuesday ridership was higher than Strava, and the opposite is true for the other high-low point. It is important to note here as well that there is some disparity in the IRSD where the westernmost location was more advantaged than the other high-low point. Both points are surrounded by similar areas of medium-level bicycling and street infrastructure.

The vast majority of locations showed no spatial association, meaning that the spatial pattern of rank difference was not unexpected based on a null hypothesis of random processes. Considering that many of the locations with no significant difference were located close to the Sydney central business district, there are several factors that may be at play. First, this area is dense with bicycling infrastructure and since it is also a dense urban area, cyclists may have few choices in terms of bicycling paths so the Super Tuesday and Strava counts show relatively stable rates of activity (Jestico et al., 2016). Second, this area is close to the harbor where cyclists have few options for crossing, so it is possible that bicycling activity is forced to occur along particular corridors where that access is possible. Also, it is not surprising that these locations had some of the highest

scores on the IRSD indicating that they have good relative socioeconomic advantage. These areas also had some of the highest population density in general so the findings may be associated with having a larger total sample in the area. Finally, since the variable at study is rank difference, it is possible that bicycling rates in this area are actually different between the datasets, but that their *relative importance* in the data is stable as indicated by having similar rank difference to their neighbors.

One limitation of the current analysis is that the conventional manual count data collection depend on the councils that choose to participate in the count and are therefore not randomly distributed throughout the study area. The uneven distribution of count sites across the study area could have some implications for spatial pattern analysis results, though they are thought to be minimized by focusing on comparability rather than the pattern result itself. The current analysis is also limited in the demographic information included. Strava metro publishes little data on their sample and no demographics are captured in the conventional count. Future endeavors to examine crowdsourced and conventional data may include intercept surveys to capture more information about the riders who generate said data. Another limitation relates to the time periods for each data set. While the data sets were comparable proportionally, the Strava data presumably reflect more diurnal variation in bicycling activity since they reflect weekday riding for a full month rather than one day. Again, the influence of this variation minimized by using the ordinal values between the datasets rather than focusing on counts or proportions. Future examinations may consider using Strava data to extrapolate diurnal variations to scale one-day manual counts.

3.6 Conclusions

In terms of comparing ridership patterns, these results begin to indicate how crowdsourced data can be used by planners and other stakeholders when designing and planning bicycling facilities. Despite Strava data appearing to offer much more than conventional data in terms of spatial and temporal resolution, it is clear that the patterns of ridership are different even at basic levels. There are differences in ridership patterns related to socioeconomic status and presence or absence of bicycling infrastructure. In terms of planning and design to encourage and facilitate active transport, the high-high

location of dissimilarity cluster area demonstrates where the data may not quite match. In this case, the absence of bicycling infrastructure indicates an area where improvements could be made while accounting for multiple data sources rather than Strava or counts alone. On the other hand, low-low locations of similarity and those that are not significant in terms of spatial association may show areas where crowdsourced and conventional data can be used together or as a substitute without introducing undue bias, given further investigation into the demographic characteristics of the riders themselves. The low areas too would indicate good areas to improve or install new bicycling facilities. While outside the scope of this study, future examinations of the patterns between crowdsourced and conventional data should more deeply examine the infrastructure, neighborhood, and socioeconomic factors that underlie the patterns. The anomalous high-low and low-high locations may be of particular interest as the disparate values in their rank my reveal characteristics that are not present in the other areas. Overall, in consideration of the difference in ridership proportions and the relative importance of locations between crowdsourced and conventional, it is imperative that planners and other stakeholders do not consider crowdsourced data as a singular source of ridership activity data. This could lead to decisions that benefit only the types of riders who use apps to record rides, rather than all or potential riders in a region. Strava is potentially a good substitute for conventional manual count data in areas where ridership is low overall, if rider demographics match those of the typical rider in the area as well. Further investigation into the characteristics of the riders in Strava versus typical riders who do not use the app is needed, however. Additional data collection is needed in areas where social advantage and ridership are high, especially if these areas do not have the type of infrastructure that supports high rates of bicycling activity.

CHAPTER 4

FACTORS THAT INFLUENCE BICYCLING ACTIVITY DENSITY IN SYDNEY USING VARIED CROWDSOURCED DATASETS

4.1 Introduction

City planners, policymakers, and bicycling interest groups aim to understand where and when people use non-motorized transportation so that bicycle-friendly infrastructure can be better planned. Though walking as a transport mode has been studied, bicycling remains under examined in comparison (Winters et al., 2016). The features of neighborhoods and networks that support walking or bicycling share some common characteristics, though bicycling involves additional features that may be relevant in supporting its uptake as an activity (Winters et al., 2016). Physical or environmental features of neighborhoods include sidewalks, parks and other elements or services that facilitate behavior by offering a setting wherein people are able to engage in and access bicycling activity. The importance of considering these factors, which look beyond both characteristics of individuals and their immediate home residence, have been established (Kwan, 2013; Kwan, 2012). Local attitudes related to health, social connections and socioeconomic status are among the additional neighborhood features that might influence behavior (Kwan, 2012).

Results from studies on bicyclist behaviors and preferences have revealed some characteristics associated with bicycling activity. Short term local surveys have shown that commuting cyclists show preference for shorter, direct (i.e., few turns) routes, lower traffic volume and speed, and bicycling specific infrastructure (e.g., separated bicycle paths) and areas with more bicycling infrastructure tend to have more bicyclists (Broach et al., 2012; Winters et al., 2010; Winters et al., 2013). Many cyclists ride along bicycling infrastructure for a high proportion of the distance of their ride, though bicycle infrastructure might comprise a proportionally low amount of distance among the full street network (Dill, 2009). This indicates that cyclists may be willing to travel further distances to ride on bicycling infrastructure (Dill, 2009). Other features of the built environment that influence bicycling include land use, density, and workplace

accessibility (Handy and Xing, 2011). In contrast to general findings regarding the positive association between bicycling activity and bicycling infrastructure, Dill et al (2014) found no increase in bicycling activity where new bicycling boulevards were installed. Discrepancies such as this require further examination and the high volume of bicycling trips generated from crowdsourced data may provide clearer insights into bicycling activity and the infrastructure that supports it. Surveys have shown some factors that are associated with ridership, but conventional bicycle counts and short term surveys are problematic in spatial and temporal scope or cannot provide information about user level bicycling activity. Additionally in terms of crowdsourced data, it remains unknown how these factors vary by app or user type.

Conventional data sources provide some information, but are limited in spatial and temporal scope as well as information about individual users. New crowdsourced data from smartphone applications (apps) may provide more information about where and when people use bicycles. While previous studies have examined the factors that influence bicycle ridership, there is a gap in understanding the contextual factors that determine where crowdsourced data are contributed. Systematic examination of several sources is needed to determine not only what activities the data represent, but the social and built environment drivers of the areas where those activity data are collected. The aim of this paper is to assess how the physical environment, bicycling infrastructure, and sociodemographic features of SA2 areas in the Greater Sydney region predict where crowdsourced bicycling data are (i.e., positive space) and are not (i.e., negative space) contributed. That is to say, what are the factors that influence where people bicycle from crowdsourced data samples and what new information can we glean from them? Further, are those factors the same between datasets or do different factors drive activity among them? The paper begins with a review of crowdsourced and conventional data, using bicycling app data as a case study, then discusses the associated potential for sampling bias. Finally, models comparing two exemplar crowdsourced datasets, Strava and RiderLog, are discussed in the context of the factors that determine where the data are contributed.

There is some evidence that crowdsourced bicycling data do not have the same predictors of ridership as conventional data, and there are mixed results overall as to what urban structures are preferred by different types of cyclists. In one examination of crowdsourced bicycling activity for health purposes, presence of bicycle lanes was not associated with ridership volumes; people bicycling for fitness are likely to access areas with fewer barriers to riding (e.g., stop lights) that are away from the urban center where bicycle infrastructure is located (Griffin and Jiao, 2015). Further, cyclists using the Strava app are more likely bicycling for recreation on short streets with high connectivity and low traffic volume (Sun et al., 2017). Conventional data, however, have shown repeatedly that bicyclists who are using bicycling as a transport mode prefer bicycle lanes and will use them where they are available (Dill and Carr, 2003). For cyclists using bicycling for utilitarian purposes, almost all of it occurs in areas with bicycle facilities (Dill, 2009). From app-based crowdsourced data as well, cyclists show more comfort on separated paths and bicycle boulevards (Blanc et al., 2016). As with prior research, cyclist's comfort levels fell when vehicular traffic was present during their ride. Cyclists in that study (Griffin and Jiao, 2015) also preferred hilly areas which is contrary to findings for cyclists who ride for transport purposes. These findings illustrate that crowdsourced data, when contributed by bicyclists aiming for health and fitness, may provide different information when the data are intended to be used with utilitarian bicycling in mind. Further, bicycling data gathered from smartphones is generally biased toward relatively young male riders, and low-income populations are under sampled (Blanc et al, 2016).

Active transport, or non-motorized travel like walking and bicycling for utilitarian trips rather than leisure, is associated with increased health behavior, reduced traffic congestion, and potential reduction of emissions associated with vehicular travel. Walking and bicycling, however, must be feasible transport options for those benefits to be realized. Since non-motorized travel usually occurs on local streets, developing and planning its infrastructure are typically within the domain of local governments, though increased federal and state funding has led to increased interest in monitoring pedestrian and bicycling activities (Lindsey et al., 2013). Robust accounts of pedestrian and cyclist

travel are needed to ensure we plan for pedestrian safety and comfort so that more users could adopt active transport modalities (Transport for NSW, 2013). Current tools and data available for planning and analyzing non-motorized travel are limited in detail and spatial and temporal extent (Saunders et al., 2013)., and robust systems that account for and measure pedestrian and cyclist travel are needed (Lindsey et al., 2013).

Manual bicycling counts are limited in spatial and temporal coverage, and regional travel surveys may not address non-motorized travel activity at all. Since non-motorized trips tend to be shorter, they necessarily require detailed finer resolution data for analysis (Cervero and Duncan, 2003). This lack of reliable data and knowledge about non-motorized travel has limited measures of accessibility and examinations related to human mobility using those modes (Iacono, Krizek, and El-Geneidy 2010). Further, there is a lack of data that can be used to link active transport modal behavior with infrastructure and other network features (Broach et al, 2012). Since data availability is a limitation in understanding active transport, there has been increased demand for using crowdsourced data in examinations of activity (Blanc, Figliozzi, and Clifton, 2016; Heesch and Langdon, 2016). Greater knowledge about the infrastructure and network contexts that positively influence utilization of active transport modes will inform policy and design so those modes could be adopted and accessed by more users. Physical or environmental features such as transport infrastructure, land use, and the social, cultural, and institutional characteristics of neighborhoods are known to shape human behavior and these contexts influence mobility by offering a setting wherein people are able to engage in and access healthy living patterns (Kwan, 2012).

Crowdsourced data from smartphone apps often contain route or origin-destination information, some user demographics, and may be collected more frequently than conventional methods. Many such bicycling apps have been developed for commercial purposes and widespread activity monitoring (e.g., Strava, MapMyRide) or for capturing activity to use in a particular area's or municipality's planning efforts (e.g., RiderLog, CycleTracks) (Leao et al., 2017, Blanc et al., 2016). While these apps differ in focus and function (i.e., fitness or implicit data collection efforts), they have in common the large amount of data they produce at detail levels that are not typically possible

through conventional means. These app-based data may help answer questions about where people ride bicycles, and the factors that underlie that ridership because they contain that extra detail and information that are not captured with conventional sources.

In addition to providing more detail and greater data volume, crowdsourced data help overcome difficulties associated with public participation. Traditionally, public participation in planning processes has been limited despite being considered important to successful planning efforts (Misra et al., 2014). Though information is easier to access, public participation is declining because it still often relies on physical presence in the process wherein certain groups may be excluded because of their inability to attend (Misra et al., 2014). Crowdsourced data overcome this limitation by providing a way for citizens to engage in civic decision making and public advocacy and not requiring citizens to be physically present at a particular time and place (Le Dantec et al., 2015). Using data generated from crowdsourced means, city planners and decision makers have new ways to consider developing projects and transportation planning in particular. Sustainable decision making requires both that the data be produced and that it be shared and used by relevant stakeholders (Le Dantec et al., 2015). Crowdsourced data are useful in a transportation context because they easily engage users within a region, and in the case of bicycling activity, are more cost effective than conventional survey means (Misra et al., 2014). Cycle Atlanta is one bicycling app that Atlanta, Georgia has used to make decisions about transport. The app provides information about routes as well as barriers riders encounter (e.g., pot holes). The city can then use the data to make knowledge based and data driven decisions about needed infrastructure developments to support bicycling activity. Cycletracks is another app that lends data on bicycling movement to the city of San Francisco and the app provides data for cyclists as well; the service is aimed at public participation and civic engagement and motivated by a lack of data on bicycling activity (Lee et al., 2014; Misra et al., 2014).

Differing smartphone apps make for a compelling comparison because of their explicit and implicit nature. Crowdsourced data are generated in two ways, explicitly or implicitly. Explicit crowdsource systems involve generating data and information to solve or provide input on a stated problem (Misra et al., 2014). Some bicycling apps have

been made with explicit data generation in mind, though others provide implicit data. Implicit systems are those where user input generates an indirect data product (Misra et al., 2014). The data can still be used to solve a problem, but said problem is not the reason the data are generated in the first place. Implicit systems, therefore, do not have an explicit agreement with the user as to for what purposes their data will be used. (Misra et al., 2014).

In this paper we focus our efforts on two bicycling apps in particular, Strava and RiderLog. The Strava app has international reach and is focused mainly on fitness and competition. Users choose to log GPS traces of the bicycle rides they take to track progress and compete virtually with other users. In terms of a spatial data product, Strava Inc. releases an aggregated version of users' ridership, Strava Metro, which was made available for analysis by Transport for New South Wales (Strava, LLC., 2016). The rides in a particular area are aggregated to polygons, road intersections (nodes), or roads (edges) to give volumes of rides indicative of popular routes and areas. Strava is used worldwide and over 2.5 million activities are logged each week through the app (Strava, LLC., 2016). Since the spatial data is a secondary product generated from Strava users but the focus of the app itself is fitness and competition, it is an implicit source of crowdsourced data. Riderlog, on the other hand, is marketed specifically toward helping local stakeholders generate data to appeal to policy and planning so it is an explicit case of crowdsourced data. RiderLog is a smartphone app that was developed by the Australian nonprofit organization Bicycle Network (Leao et al., 2017). The app was developed and marketed specifically toward generating data for city planners to use to advocate for bicycle oriented planning, and users are able to monitor both their movements and performance over time (Leao et al., 2017). On the development side, the app collects full GPS traces of rides that users choose to log, along with demographics and ride information such as gender, age, and trip purpose (Leao et al., 2017). The app was launched in May 2010 and has nearly 10 thousand users across Australia.

4.2 Methodological Approach

4.2.1 Study Area

Analysis was conducted at the statistical area level 2 (SA2) level in the Greater Sydney area in New South Wales (NSW), Australia. The city of Sydney is the capital of New South Wales and is the major economic core in the area (NSW, 2017). The population as of 2015 was 4.92 million people and the Greater Sydney area encompasses 12,474 sq. kilometers of land at the eastern coast of NSW to the Blue Mountains in the interior with both urban and rural populations (ABS, 2017; NSW, 2017). The SA2 areas represent medium sized areas of social and economic interaction; they generally have a mean of 10,000 persons and the average sq. kilometers of the SA2 areas was 44.33. Generally the SA2 areas are smaller areas with larger populations in more urban areas and are larger with less population as areas become more suburban and rural (ABS, 2016).

Bicycling rates have increased since 2006 in the Greater Sydney area though overall rates remain relatively low (Transport for NSW, 2013). Despite low rates, a large number of daily commute trips fall within a feasible bicycling trip range (<10 km) (BITRE, 2016) and travel times between bicycles are motorized trips are often comparable (Ellison and Greaves, 2011). The area has a variety of bicycling facilities including painted lanes, mixed traffic and shared busway lanes, as well as separated on-road and off-road lanes (Pucher et al., 2011). Despite this infrastructure, the area has several features that hinder bicycling including hilly terrain and natural barriers with few crossings along the harbor area. Many of the planned cycle routes serve recreational riders as they are located along coastlines or along outer regions with lower demand for commute activity (Pucher et al., 2011). Despite barriers, 70% of residents surveyed in NSW expressed willingness to use bicycling for transport if it were a safer and more convenient mode (Transport for NSW, 2013).

4.2.2 Data

This study used data from two bicycling smartphone apps, characteristics of the road network, as well as sociodemographic data from the Australian Bureau of Statistics. The dependent variables are derived from the bicycling apps while the independent variables consist of infrastructure and socioeconomic features of the Greater Sydney area. The Strava aggregate data consisted of SA2 area volumes for 2016 which show the

number of rides originating, ending, or passing through each SA2 area in the study region; the Strava sample included rides from 84,863 unique riders who generated 2,889,139 trips overall in 2015. RiderLog trip records consist of a user and ride identifier, trip purpose, trip start and end date and time, trip duration, distance, average speed, user gender and age, with a full record of the route via GPS locations taken throughout each recorded trip. The dataset used for the Greater Sydney area included 14,844 trips generated by 1172 riders that were collected between May 2010 and May 2014. The individual routes were aggregated to the SA2 level to derive a count of rides that originated, ended, or passed through for each SA2 area.

The dependent variables for the models herein consists of two measures for Strava and RiderLog. First, a count of rides (ridership volume) that passed through each SA2 area was derived for each dataset. The second dependent variable was a binary of whether an SA2 area was positive for bicycling activity. Whether an area was positive was determined by a threshold based on quartiles in the data. Any SA2 area with less than the first quartile's value was considered negative for crowdsourced data contributions. For the Strava data, the first quartile was 1176 which left 209 areas labelled as positive and the remaining 70 negative. The threshold for RiderLog was 17 leaving 206 areas positive and 73 negative. The positive and negative spaces represented by both Strava and RiderLog are shown in Figure 4.2. Since we aim to determine whether there are asymmetries in our crowdsourced data, we hypothesize that the factors that predict ridership volume and presence for Strava will be different than those which predict the same for RiderLog. Namely, we predict that higher bicycle lane density, lower amounts of residential land, and higher population density areas will be important in predicting RiderLog ridership as those variables are associated with utilitarian bicycling trips; we suggest that RiderLog use will be associated more with commuting than will Strava.

The independent variables used in the modeling herein are comprised of built environment and sociodemographic measures; they were chosen based on previous findings in the literature that relate to road structure and connectivity as well as where cyclists prefer to ride (i.e., residential land areas). The built environment data consist of

the street network, bicycle infrastructure line data and land use zoning information. The bicycle infrastructure include streets and cycleways that were designated as bicycling facilities within the Greater Sydney area. The land use data contain zoning classifications that give details about the type of land use that is permitted within each parcel.

Table 4.1. Independent variables considered for modeling.

Variable	Source	Formulation/ operationalization	Reason
Street Network	OpenStreetMap	$N = {}^{l}/_{s}$ $\beta = \frac{e}{v}$ $\gamma = \frac{e}{3(v-2)}$	The measures of network structure quantify features of the street network that typically influence bicycling behavior.
Bicycle Network	OpenStreetMap	"	"
Land Use (residential, business, industry, recreation, infrastructure (e.g., streets and highways), mixed use, environment, and city/town centres.)	New South Wales Department of Planning and Environment's 2015 Local Environment Plan	Percentage of each within SA2	Strava users prefer residential areas (Sun et al., 2017) and business areas are a popular destination feature.
Population density	Australian Bureau of Statistics' census	SA2 areas	More cyclists in less dense areas.
Median age	"	"	Ridership has been associated with specific age groups. Might influence the potential pool of cyclists in the area.
Index of Relative Socio-economic	"	Low scores indicate relatively greater social disadvantage while high scores indicate the least	These variables were meant to represent the idea that cyclists would avoid disadvantaged areas

Disadvantage (IRSD) score		disadvantaged areas.	and that more activity would be present in higher income and higher affluence regions.
Median income	"	"	"
Median weekly rent	"	"	"
Number of people using two modes for journey to work (JTW using two modes)	"	"	Possible greater pool of bicycling commuters

The street network and bicycle infrastructure data were used to derive measures of network structure. Network density is the ratio between the total length l of street segments within an SA2 area to the total area s of that area and measures street network intensity. An area based measurement of density was used rather than graph-based density measure to examine the relative neighborhood street density in an SA2 area to consider the overall context within which people ride. The graph-based metrics of beta index, gamma index, were also computed. Beta index indicates the level of connectivity and complexity of a graph and is defined as the ratio of the number of edges, e (i.e., street segments) to nodes, v (i.e., intersections). Gamma index measures the relationship between actual connections on a network and all possible connections. Since the gamma index is a ratio of the number of edges to the number of all possible edges, it gives an indication of connectivity within the actual network context on a 0-1 scale where 1 is completely connected.

4.3 Statistical Analysis

Determining which factors predict bicycling activity will reveal differences among the data sets. Understanding where we do and do not have detailed data can help focus efforts to collect data from geographic areas or populations that are underrepresented as well generate strategies for how those data might best be captured.

Using varied data sources also allows for comparison between the spaces generated by them which will help understand how those data can best be conflated with conventional data to get a better picture of bicycling activity. Examining the spaces where bicycling activity is present as well as where it is absent will help determine which infrastructure and sociodemographic contexts are potential drivers of activity or lack thereof.

Descriptive summaries of each of the independent and dependent measures were computed. The correlations between all measures were also computed to ensure that collinear measures were not included in the multivariate models. Bike beta and bike gamma had a moderate negative correlation, while road beta and road gamma had a strong positive correlation. Further, the household specific measures of total households, number employed, and number of people using two modes for their JTW were also moderately to strongly correlated. These measures were tested separately within exploratory models to avoid adding multicollinearity. Extreme outliers in the ridership measures were winsorized using median absolute deviation so that they did not skew statistical estimates.

Four separate models were used to measure the factors that influence bicycling activity; the first two models predicted ridership volume for Strava and Riderlog to determine whether the same factors predict the level of bicycling activity. A Poisson model and subsequent testing indicated that the data were overdispersed, necessitating the use of the Negative Binomial parameterization of the model. Moran's I_i analysis confirmed that significant spatial autocorrelation was present in residuals of the preliminary Negative Binomial models; since spatial data present concerns in terms of independence of both observations and error terms, spatial autoregressive (SAR) models were also used in models three and four to supplement the Negative Binomial findings. Since the SAR model requires a spatial weight, neighbors were defined as any polygons sharing boundary points. Because the estimates of the SAR cannot be interpreted as partial derivatives, the direct and indirect impacts are shown to assess the sign and magnitude of impacts from changing the independent variables.

Models 5 and 6 used linear discriminant analysis to predict the binary whether an SA2 area was labelled positive or negative. The aim of linear discriminant analysis is to

find the combination of variables that gives the best separation between groups; here the groups are the binary positive and negative spaces in the crowdsourced datasets. The two most useful discriminant functions can be discovered for predicting positive space and we can determine whether they are the same or different between data sets. In terms of coefficients, the larger the coefficient, the more important it is in the discriminant function.

4.4 Results

Table 4.2 shows the winsorized minimum, maximum, and median of ridership counts per SA2 area. Despite differences in overall ridership, the data sources show similar distributions.

Table 4.2. Descriptive statistics for ridership counts for Strava (year 2016) and RiderLog (2010-2014)

	Min	Max	Median	St. Deviation
Strava	30	16090	2895	4913
RiderLog	0	432	70	156

Next the factors that underlie ridership were addressed through several models, answering the question which built-environment and sociodemographic factors predict where the data are contributed.

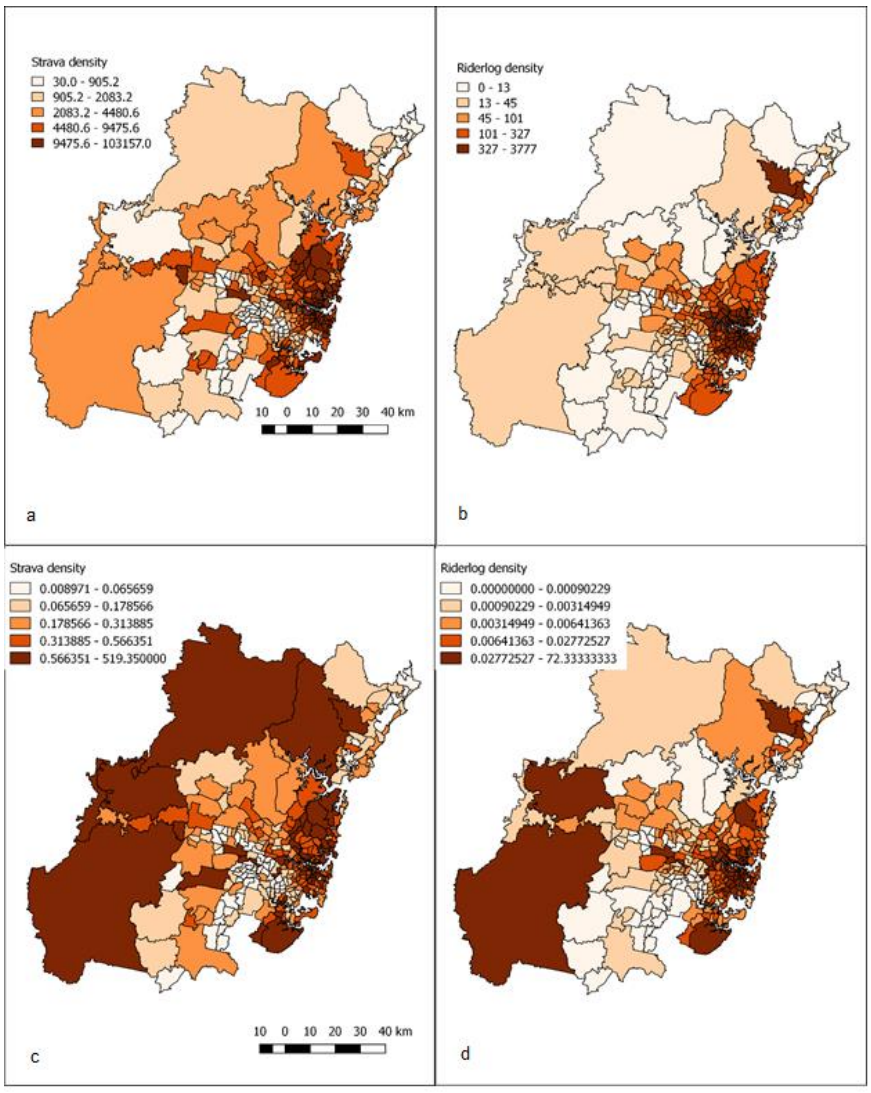

Figure 4.1. Quintile maps of Strava vs. Riderlog ridership frequency (Strava a, Riderlog b), and normalized by population (bottom, Strava c, Riderlog d).

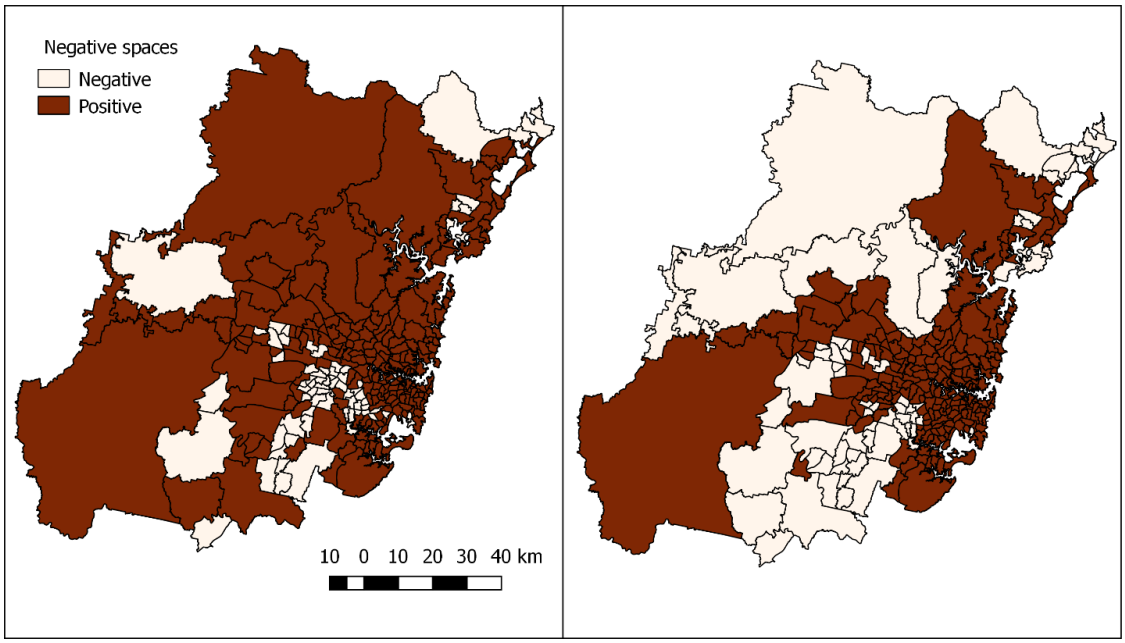

Figure 4.2. Negative and positive space for Strava (left) and Riderlog (right) identified by the lowest quintile in each dataset.

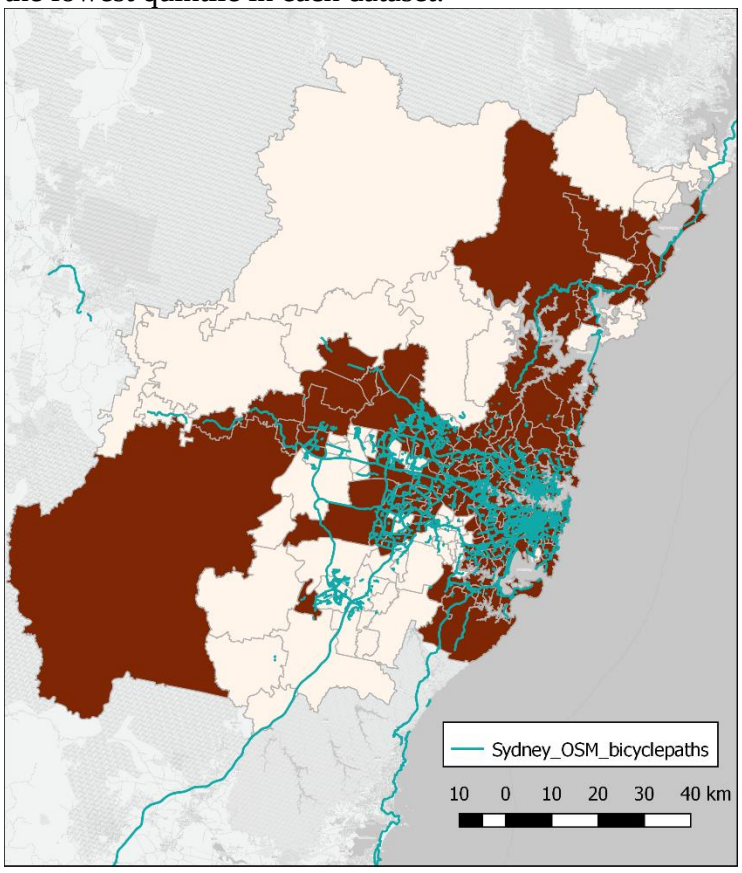

Figure 4.3. OpenStreetMap bicycle lanes for Greater Sydney Area

Table 4.3. Models 1-4 predicting ridership volume (1 and 3) and ridership positive space (3 and 4).

	Dependent variable:			
	Strava	RiderLog	Strava	RiderLog
	Negative Binomial		*SAR*	
	(1)	(2)	(3)	(4)
Median age	-0.007	0.010	-8.241	-0.708
	(0.010)	(0.013)	(32.688)	(0.810)
Median weekly rent ($AUS)	0.005***	0.004***	6.998***	0.008
	(0.001)	(0.001)	(2.380)	(0.057)
IRSD score	-0.001***	-0.002***	-0.744	0.031
	(0.0003)	(0.0004)	(1.126)	(0.028)
Population Density	-0.0001	0.0001*	-0.449***	-0.0005
	(0.00005)	(0.0001)	(0.142)	(0.004)
% Residential land	-0.015***	-0.006**	-36.722***	-0.418**
	(0.002)	(0.003)	(7.031)	(0.170)
% City Centres	-0.026	-0.072***	x	x
	(0.019)	(0.025)		
JTW using 2 modes	0.001***	0.0002	4.629***	0.050***
	(0.0002)	(0.0003)	(0.756)	(0.019)
Road density	60.366***	78.740***	144,721.300**	171.375
	(17.985)	(23.175)	(57,256.890)	(1,419.154)

Bike lane density	2.540	31.328	272,007.000**	11,183.710**
	(38.339)	(49.358)	(125,260.900)	(3,136.209)
Constant	7.488***	3.365***	-1,210.281	-8.715
	(0.374)	(0.484)	(1,236.632)	(30.548)
Observations	279	279	279	279
Log Likelihood	-2,582.398	-1,580.370	-2,614.907	-1,597.445
theta	1.479*** (0.114)	0.897*** (0.070)		
Akaike Inf. Crit.	5,184.796	3,180.740	5,251.814	3,216.890
		$sigma^2$	7,288,107.0	4,484.184
		Wald Test (df = 1)	185.611***	642.882***
		LR Test (df = 1)	114.147***	243.614***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.3 presents the results obtained from the initial Negative Binomial analysis of factors related to bicycle ridership volumes. For the built environment factors in models 1 and 2, percentage of area zoned as residential had a significant negative effect on ridership in both models; as percentage of land zoned as residential decreased, sampled ridership increased. While all land classifications were tested, the percentage of land zoned as city centres was the only other land use variable that had a significant effect in the models. As with residential land, it was negative for both datasets though significant for RiderLog only. Though not significant, it is noted that land zoned as business had a positive relationship with ridership volumes. In terms of predicting ridership volume per SA2 area in models 1 and 2, the other built-environment factors had mixed results. Road density stands out in the models as it is significant and positive for both Strava and RiderLog samples. The effect of roads was still significant when substituting the gamma and beta measures for road density, though the coefficients changed in magnitude they

did not change in direction. Bike lane density was not significant though it was positive as expected.

The measures of sociodemographic context for models 1 and 2 also had mixed results. Median age of the residents in SA2 areas was not significant. Median rent was a significant positive factor for both models. IRSD score was negatively associated with ridership volume for both Strava and RiderLog. Population density was significant for RiderLog only, and interestingly it was positive for RiderLog and negative for Strava. Though it was not significant, median age also had flipped signs with positive and negative for RiderLog and Strava respectively. Using more than two modes of travel for journey to work was positive for Strava meaning that as more people had to use two modes for their JTW, they were more likely to log a ride on Strava.

Results of the SAR models are shown in Table 4.3, models 3 and 4. Spatial lag models were used because of the autocorrelation associated with the linear models. In terms of predicting ridership volumes bike lane density, residential percent, and JTW using two modes were significant for both Strava and Riderlog. Road density, median rent and population density were additionally significant for Strava only. In order to understand the true effects of the variables in the SAR models, direct and indirect effects were computed for both models as shown in Tables 4.4 and 4.5.

The lagged variables in the spatial autoregressive model necessitate interpretation of direct and indirect effects. Direct effects on the dependent variables are those that result from a change in the independent variable for a single SA2 area (Golgher and Voss, 2016). Changes in the dependent variable based on the independent variable for another SA2 area, outside itself, are the indirect effects in the model (Golgher and Voss, 2016). From Table 4.4, changes in median age and IRSD score are not significant which suggests they do not impact Strava ridership counts. The indirect effects of median rent, number of people who JTW using two modes, road density, and bike lane density are all significant and positive, suggesting that increasing any of these in neighboring regions will have an impact in neighboring areas. Because direct impacts are also positive, this suggests that increasing those factors in any area will also increase Strava ridership. Percentage of residential land and population have negative effects for both direct and

indirect impacts indicating that decreasing either factor would increase ridership in any area or neighboring area. For RiderLog ridership, only percentage of residential land, number of people who JTW using two modes, and bike lane density were significant (Table 4.5). Bicycle lane density and JTW using two modes were positive while percentage of residential land was negative. In considering total effect as the sum of direct and indirect impacts, the effects from these factors are mostly comprised of indirect effects.

Models 5 and 6 in Table 4.6 show the discriminant analyses results that predict which factors determine whether an SA2 area is labelled as positive or negative for crowdsourced data contributions. The positive and negative spaces represented by both Strava and Riderlog and shown in Figure 4.2. A large number of the negative Strava spaces occur where there is bicycling infrastructure relatively close or within the western suburbs of the city of Sydney. The result of the model is shown in Figure 4.4; while the models do reasonably well to discriminate among positive and negative spaces, there is still some overlap between groups; the accuracy was 83.15%. The model predicts that 26% of the areas in the data correspond to negative spaces. For both Strava and Riderlog, median weekly rent is the strongest predictor of whether a space is labelled positive. For Strava, the next most useful functions to discriminate are percentage of residential land and JTW using two modes, while for RiderLog it is IRSD score and population density. With the exception of percentage of residential land in the Strava model and median age in the RiderLog model, there was a tendency for the group variables to be negative when areas were labelled positive, and positive when areas were labelled negative.

Table 4.4. Direct and indirect effects for Strava SAR models

Variable	Direct	Indirect
Median age	-9.299	-13.501
Median rent	7.896***	11.4641***
IRSD score	-0.8399	-1.2194
Pop. Density	-0.5066***	-0.7354***
% Residential	-41.4385***	-60.1601***
JTW two modes	5.2232***	7.58299***
Road density	163307.9***	237089.0**

Bike lane density	306940.7**	445614.0**

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.5. Direct and indirect effects for RiderLog SAR models.

Variable	Direct	Indirect
Median age	-.9233	-3.0394
Median rent	0 .01099	0.00362
IRSD score	0 .0409	0.1345
Pop. Density	-0.0006	-0.0019
% Residential	-0.5457**	-1.7964**
JTW two modes	0.0655***	0.2156**
Road density	223.551	735.9081
Bike lane density	14588.64***	48024.44***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.6. Discriminant function results

	Strava	Riderlog
	Discriminant function	
	(5)	(6)
Median age	0.1015	0.0688
Median weekly rent ($AUS)	0.7108	0.9352
IRSD score	0.2494	-0.6287
Population Density	-0.0931	0.3772
% Residential land	-0.6182	-0.0595
% City Centre	-0.0573	-0.3609
JTW using 2 modes	0.4414	0.0954
Road density	0.0598	0.2789

Bike lane density	0.1661	0.2645

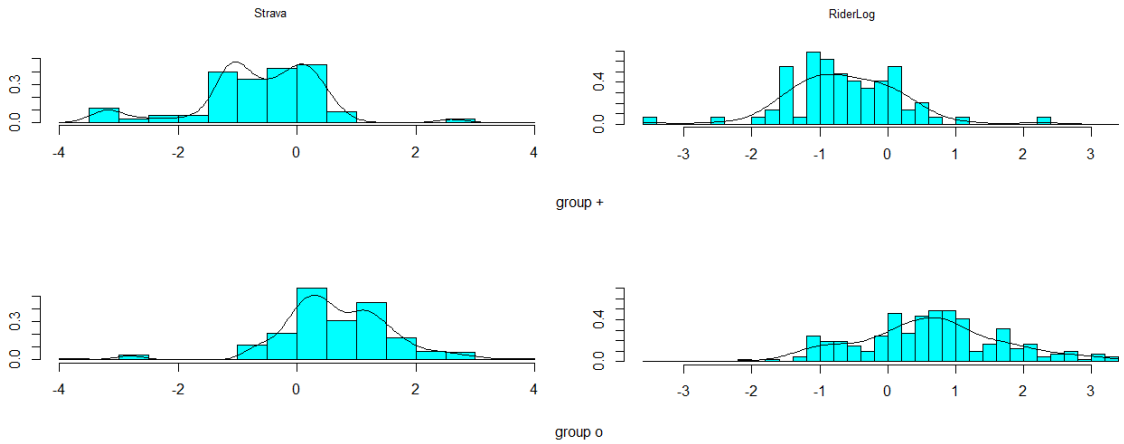

Figure 4.4. Results from Discriminant Analysis for Strava (Left) and Riderlog (Right). Histograms Show the Distribution of Discriminant Scores for Positive and Negative Spaces.

4.5 Discussion

As shown in Figure 4.1, there are qualitative asymmetries in ridership frequency between Strava and RiderLog. Both sources show a great deal of ridership in the most central Sydney area (near the harbor), though there are differences outside of the downtown core. Within the Strava data, ridership appears higher in the areas on the outskirts of the city center as compared to RiderLog; Strava ridership is particularly frequent in the northern Sydney suburbs. RiderLog ridership is comparatively low on the outskirts of the study area and is mainly clustered tightly around the city center. When considering the densities of ridership, Strava and RiderLog have a similar inner-city pattern with some differences moving toward the outskirts. Most notably, Strava and RiderLog both have elevated ridership in the northern suburbs, though in different areas. Overall, the higher RiderLog density is more compact and close to the urban core of the region whereas Strava is more spread through the region. When considering volume of ridership only (Figure 4.1 upper), it is clear that Riderlog is more concentrated around the downtown core with volumes generally fading as they move outward. Strava on the other hand has high volume ridership near the city center but much higher ridership remains

while moving outward and the lower ridership areas occur in the west/southern suburbs. When accounting for the area of the SA2 areas, Strava and RiderLog ridership look similar in the city center, with most ridership concentrating around that area. Strava ridership, however, remains elevated on the outer regions of the study area. While RiderLog has some elevated ridership in the outskirts, the majority occurs within the urban core with a ring of lower ridership between that urban core area and the outskirts. In Figure 4.3 there appears to be some relationship between ridership volume/density and the location of bicycle infrastructure. The RiderLog positive space seems to align more closely with the bicycle infrastructure than does the Strava positive space. Considering that bicycle lane density was a significant positive factor for predicting RiderLog ridership volume whereas it was not significant for Strava, this general conclusion makes sense. Further, bicycle infrastructure has been unrelated to ridership volumes in previous studies using Strava data (Griffin and Jiao, 2015; Jestico et al., 2016). As with prior findings, it appears the Strava users seek areas with few barriers to riding that are away from urban centers.

The overarching question in this research was what are the physical, infrastructure, and socioeconomic factors that predict ridership among two different sources of crowdsourced bicycling data: Strava and RiderLog. The models show that some similarities and differences are present. First median rent, IRSD score, percentage of residential land, and road density were significant in the models predicting ridership volume for both data sources. Since median rent was positive, this lends support to the idea that ridership in crowdsourced data is recorded in higher valued areas which could be an effect of the people who log their rides as well as where people choose to log those rides. Ridership also increased as IRSD score decreased; lower IRSD scores are indicative of areas with higher social disadvantage. This seems contradictory to the finding that higher rental areas were associated with more Strava ridership. Again, users are likely seeking areas away from the city center that would be associated with less overall socioeconomic capital and rents decrease as one moves away from the city center with highest rents south of the harbor and in the northern suburbs.

That percentage of residential land was negative is supported by the previous findings that bicyclists using these apps are seeking areas away from urban cores (Griffin and Jiao, 2015); they were more likely to ride where residential land percentages were lower. In the case of RiderLog, ridership volumes were higher when percentage of city centre land is lower. Considering that the percentage of business-zoned land was positive though not significant, it is possible that riders are either seeking the destinations of the urban core or somehow necessarily required to travel through business heavy areas. This is possible in Sydney as any travel across the harbor is funneled through the few bridge crossings available, and the bridges are near business centered areas (Pucher et al., 2011). Road density was a positive predictor of ridership in both models which aligns with the idea that riders prefer connected streets (Sun et al., 2017). It is also possible that the types of riders that choose to use apps are those that are more bold in their ridership choices and willing to ride on roads that are not specifically marked for bicycle use, since bicycle lane density was not significant.

The remaining factors had mixed results, indicating that there are differences between the ridership in each dataset. Population density was positive and significant for RiderLog indicating again that users are seeking denser urban areas to ride. Population density however, was negatively associated with ridership for the Strava sample (though not significant) which lends support to the idea that Strava users are interested in exercise away from busy corridors (Griffin and Jiao, 2015). Further, RiderLog is not as popular as Strava so one must consider that its uptake may not reach far beyond the areas where it has been marketed, in this case around the urban cores of the Greater Sydney area.

Using two modes during the journey to work in an area was positively associated with Strava ridership. This makes sense as a positive predictor of ridership for Strava volumes as people are likely to use bicycles to connect to public transportation options, and the variable implies a greater potential pool of cyclists. It is possible that Strava users undertake longer commutes or recreational rides that begin in high rental cost areas but also journey to low population density/low SES areas. If these same areas have lower percentages of residential land and high weekly rents, it is possible that they are set further from business cores where people need to travel for work.

After controlling for spatial effects (models 3 and 4), the models support the idea more rides are logged in areas with greater bike lane density and less residential percentage of land, with a greater potential pool of cyclists (i.e., people using two modes on their journey to work may choose bicycling) (Dill and Carr, 2013; Sun et al., 2017). Median weekly rent and population density however, are no longer significant for RiderLog. While JTW using two or more modes may indicate a greater pool of cyclists in an area, it is additionally possible that local attitudes toward bicycling are more popular in those areas out of necessity of connecting to public transportation, which can spur more bicycling behavior (Kwan, 2012). It is also notable that bicycle lane density was not significant in the aspatial model but that it becomes significant in the spatial model. When the effects of variation in bicycle lane density over space are accounted for, it becomes an important predictor of bicycle ridership in both datasets - bicycle lane density is only important when neighboring bicycle lane density is also considered.

That said, it is perhaps more important to examine the direct and indirect effects of the spatial model to get a clearer sense of the true effects of the factors that influence ridership. The only significant direct and indirect effects for RiderLog were percentage of residential land, JTW using two modes, and bike lane density. Road density, population density, and median rent were additionally significant for Strava. Again it seems increases in any of these factors in an SA2 area will increase Strava ridership for that and any neighboring area. It seems however that RiderLog users may only increase activity with increases in bicycle lanes, as they are more likely to stick to the bicycle infrastructure. On the other hand, increasing road density and median rent or decreasing population density would lead to more Strava Ridership. It is possible that the difference between implicit and explicit ridership apps is coming into play. A possible explanation for this is that users of RiderLog, an explicit ridership app, are different from those who use Strava to log rides. Namely, it seems the RiderLog users have more in line with conventional ridership data focused on riders taking utilitarian trips; their trips are logged more frequently where there is low residential land and more bicycle lanes to support bicycling. This could be because of RiderLog's explicit aim of generating data for cities

to use to improve the bicycling experience for riders and their specific marketing in city regions.

When considering just positive and negative space as in models 5 and 6, that median weekly rent was the most important factor between datasets supports the findings that it is a good predictive factor for where bicycling activity occurs. The differences between the remaining coefficients reveal that the factors are of differing importance between the datasets. For example, JTW using two modes was less important for RiderLog than Strava which may indicate that areas where a great deal of people ride as part of their commute are the same where people log rides on Strava. That bike lane density was less important for Strava again lends credence to the notion that Strava riders are bolder and willing to ride in areas with less bicycling-specific infrastructure and again it seems RiderLog users have more in line with traditional commuting cyclists. Overall for the factors that influence bicycling activity levels, it seems median rent is the greatest overall predictor with more people using JTW using two modes and less percentage of residential land as strong secondary predictors. In terms of the differences between Strava and RiderLog, it seems the idea that Strava users seek areas away from urban cores is supported while RiderLog users are more concentrated in city centers on bicycling infrastructure.

4.6 Limitations

While crowdsourced data provide more detail than many conventional sources, they also have potential bias which may limit their broad generalizability. In general, smartphone samples tend to under-sample females, older age groups, and low income populations while oversampling particular minority ethnic populations (Windmiller et al., 2014; Blanc et al., 2016). Even with smartphone ownership, familiarity with its capabilities or availability of apps could hinder user input from particular groups (Blanc et al., 2016). With bicycling activity in particular, it must also be considered that many people who ride or rely on bicycles for transport will not also be motivated to log those rides on a smartphone app. Despite these potential sampling biases, they are only a concern in conclusions we make about bicycling activity if the activity represented in crowdsourced data differ from those represented in conventional data. Further, the

opportunity crowdsourced data bring to lend new insight also brings new challenges related to their large volumes and potential low quality (Leao et al., 2017). Additionally, while there are many variables we could have included here based on the literature it is notable that slope has been left out as a factor that may influence bicycling activity; the DEMs available did not provide adequate coverage of the study area, particularly around north and western Sydney, so it was best to leave them out. That said, local slopes likely serve as a barrier to uptake of riding in the first place rather than influencing specifically where people ride.

4.7 Conclusions

Crowdsourced data are at the forefront of mobility analysis because these new data may provide more information about activity than conventional data. Questions remain however about what new insights we can gain based on those data that we would not be able to glean from conventional sources. This study has shown that there are similarities and differences in the factors that influence bicycling activity density, depending on the data source. While the Strava data did not contain full paths, it contained the sequence of SA2 areas that riders traversed so that a volume of ridership per SA2 area could be derived. This is finer resolution than would be possible with conventional manual counts or travel surveys with just origin and destination. Similarly, since the RiderLog data contained full paths, a measure that accounts for the total journey in each area could be generated. In terms of the drivers of crowdsourced activity, several factors aligned with previous findings. The findings remain true that cyclists record more activity in areas with higher bicycle lane density. Further, Strava cyclists tended to show more activity in areas with high median weekly rent and low population density which is supported by Figure 4.1 showing moderate to high ridership in areas outside the center of Sydney. RiderLog cyclists on the other hand had more activity where population density was high and bike lane density was high which aligns more with riding close to the city center. Overall it seems the best predictors for crowdsourced ridership in an area are the median weekly rent, an indicator of the income of an area, as well as percentage of residential land and the number of people in that area who are using two or more modes

to commute, though the importance of these variables changes with the data used and the model specifications.

The analysis undertaken here could not be achieved with conventional sources of data. Since count data are static in location, it would not be possible to consider the total volume of bicycling activity through each SA2 area. The spatial scope in general for counts is quite low whereas here we were able to consider the entirety of the region as bicycling occurred across the study area rather than at discrete locations. As a comparison to household travel survey data as well, we were able to capture the actual dynamic context that people were bicycling through rather than just the origin and destination. This improves on previous research by exploring the factors related to bicycling along an entire trajectory rather than discrete locations.

CHAPTER 5

IMPACT OF BICYCLE INFRASTRUCTURE, CONVENIENCE, AND RIDERSHIP PATTERNS ON RESIDENTIAL HOUSING PRICES

5.1 Introduction and Problem Statement

Planners in cities and municipalities are increasingly interested in addressing urban problems such as traffic congestion, pollution, sprawl, and housing availability (Cervero et al., 2002). Solutions to such problems include compact cities, transit oriented design (TOD), accessibility to public transportation, and downtown revitalization. TOD in particular focuses on providing mixed use (i.e., business and residential) designs near transit, often accompanied by more compact living environments that include pedestrian and bicycle friendly facilities as a mechanism to reduce vehicle trips and increase accessibility (Cervero et al., 2002; Bartholomew & Ewing, 2011). As cities make such changes to address complex urban problems, the intended and unintended consequences need to be evaluated. One dimension that requires evaluation is how changes to transportation infrastructure and design impact the local economy and the livelihood of individuals residing there. Additional information on the economic costs and benefits of changes can be reported to taxpayers, voters, and other stakeholders.

One challenge in estimating the benefit of bicycle facilities relates to the lack of market data identifying the economic impact or value of bicycle lanes as well as dearth of information on their use and how use is related to market values and economic activity (Krizec, 2007). While several prior studies have demonstrated that the presence of bicycle facilities influence property values in some contexts (Liu and Shi, 2017), the volume of cyclists on said facilities or the volume of cyclists in a given area is an additional factor not typically measured in modeling contexts and may give additional insight into the question. The volume of cyclists is presumably indicative of neighborhood and bicycle facility features that attract cyclists and where the volume (bicycle use) may relate to property values independently from the existence of bicycle facilities themselves. Crowdsourced data can be used to overcome limitations of conventional models as they provide estimates of the ridership on local streets.

Home values are affected by home characteristics but are also a reflection of buyers' perceived value of amenities that are both structural and neighborhood based (Bartholomew & Ewing, 2011). Changes in transit infrastructure, for example additional bike paths and safe walking routes, might affect local economies and tax revenues through their influence on property prices. Bicycle facilities can be valued by current and potential future users; home buyers may pay a premium to be near facilities because they may want to use them at some point (Krizec, 2007). Despite plans to improve bicycling infrastructure in many cities, additional research is needed to clarify the linkages between a city's bicycling infrastructure and the economy of that city. If bicycling infrastructure has a positive effect on property values this can provide additional support for transit oriented projects (Nicholls and Crompton, 2005). One limitation of studies that measure bicycle infrastructure in relation to property values and economy is that while they account for presence of said facilities, they do not measure the number of riders that are actually using the infrastructure, or cyclist volume.

The goal of this paper is to evaluate the role bicycling infrastructure and usage, measures to combat traffic congestion and air pollution, have on home values in Tempe Arizona. This study fills a research gap in understanding the impacts of bicycling infrastructure, access, and bicycling activity on property values using crowdsourced data. Similar to prior studies, the paper starts by examining how bicycling infrastructure and proximity to that infrastructure affects property prices. The analysis is then expanded to include other composite measures such as walkability and bikeability and then address actual bicycling activity through ridership volumes. Inclusion of novel measures of bicycling and crowdsourced data into the analysis of property values is a significant contribution provided by this paper.

5.2 Literature Review

Many studies have used hedonic pricing analysis to estimate the economic effects of infrastructure associated with TOD through their effects on property prices. Findings show that factors like proximity to bicycle infrastructure, trails, light rail, and other TOD features influence housing prices. In this section, we review their analysis techniques,

variables commonly examined as well as their results and focus on their findings related to transit specific variables.

Bicycle infrastructure has been studied in terms of its impact on housing prices, though there are mixed findings and different ways of operationalizing bicycle facility variables. Bicycle infrastructure includes bicycle lanes, buffered bike lanes, and bicycle boulevards and all may have differing impacts on housing prices (Krizek, 2006). Bicycle lanes are typically painted lanes which share a regular vehicle traffic lane. Buffered bicycle lanes are also proximal to traffic, but are separated from vehicular traffic by a buffer such as bollards, a parking lane, or a berm/landscaping. Bicycle boulevards are separated from traffic and are sometimes multi-use trails; they may be paved or unpaved.

A study that examined the impact of advanced bicycle facilities, or separated bike lanes, buffered bike lanes, and bike boulevards, found that closer proximity and higher density of facilities were associated with higher property values (Liu & Shi, 2017). The facilities were operationalized as the distance from each property to the nearest bicycle lane and density of bicycling infrastructure was measured within a ½ mile radius of each property (Liu & Shi, 2017). Housing prices similarly increased as distance from the nearest regional multi-use path decreased in another study (Welch et al., 2016). Off-street facilities were similarly associated with housing prices, though the variable was not significant. Overall, proximity to on-street bicycle facilities was associated with lower house prices; city and suburban prices were most affected by roadside trails. (Welch et al., 2016). While contrary to other findings, the authors attribute the negative effect of on-street paths to high volume traffic arterial streets which are less attractive to buyers (Welch et al., 2016). Using distance to each type of facility, roadside bicycle facilities, or those separated from the vehicular roadway by a berm, affected city and suburban home prices negatively while off-street trails were associated with higher home prices (Krizek, 2006). Proximity to bicycle facilities overall was associated with lower house prices; city and suburban prices were most affected by roadside trails. Proximity to on-street and off-street trails did not affect house prices (Krizek, 2006). Independent bicycling infrastructure variables have typically been operationalized as either dummy indicating whether there was a bicycle facility in the area, or as network distance to a nearest

bicycle facility. Within these definitions, presence of trails in the neighborhood, proximity to trails, and proximity to bicycle lanes were all positively associated with house prices.

Prior housing price research has also examined the impacts of similar features e.g., greenways, trails, associated with pedestrian infrastructure and estimated impacts on property prices. In general, development density, land use mix, and pedestrian infrastructure are associated with higher residential property values (Sohn et al., 2012). Greenbelts are open corridors that typically follow a natural landscape feature such as riverfronts or stream valleys, or manmade features like canals or railways that are no longer in use (Nicholls & Crompton, 2005). Trails are typically defined as multi-use paths that are separated from road traffic and may be paved or unpaved.

Other studies have addressed the impacts of other TOD features, such as light rail, on property prices. Light rail may influence house prices through several mechanisms. On the positive side, light rail provides greater access and may be associated with reduced commute times for those who live in proximity to it. Retail establishments may be attracted to station areas and built up around them (Bowes and Ihlanfeldt, 2001). Negative impacts include potential noise or neighborhood aesthetic issues as well as possible increased crime at station locations (Bowes and Ihlanfeldt, 2001). That said, proximity to light rail has been associated with higher residential home prices. For properties within a ½ mile of a station, property values increased by $2.31 for each foot closer the house was to a light rail station (Hess and Almeida, 2007). Additionally however, high and low income areas had opposite property price effects; high-income areas had positive effects while low-income areas had negative effects (Hess and Almeida, 2007; Bowes and Ihlanfeldt, 2001). The value of properties in close proximity to the light rail were affected negatively while those between one and three miles away from a station had positive price effects (Bowes and Ihlanfeldt, 2001).

Walkability and bikeability are used to operationalize the transit-friendliness of an area; they measure how accessible an area is for either foot or bicycle traffic. While these scores summarize the infrastructure present in an area, they cannot reflect actual connectivity, especially unmarked roads that cyclists still use, or total activity by

pedestrians or cyclists (Pivo & Fisher, 2011). Walkability is usually measured as an area's Walk Score which is produced by www.walkscore.com. Walk Score measures how car-dependent people are in residential areas (Boyle et al., 2013). The score is determined by the distance to amenities such as schools, retail, food, and recreation as well as street connectivity; it assigns an area a score from 0-100 based on distance to said amenities and the level of connectivity (Pivo & Fisher, 2011). Greater walkability has been associated with higher office, retail, and apartment values (Pivo & Fisher, 2011). In contrast, once heterogeneity of Walk Score was accounted for, there was no relationship between walk score and house prices in another study (Boyle et al., 2013). Neighborhood walkability, rather than walkability associated with the specific property, may be a better measure of how valuable it is as a price premium for an area (Boyle, et al., 2013). Improving walkability may lead to higher residential prices, though highest payoff is in neighborhoods that are already walkable (Li et al., 2015).

Related to walk score, Bike Score measures the bikeability of an area. Bike Score values range from 0-100 and are comprised of three measures relating to bikeability: bike lane score, hill score, and destinations and connectivity score (Winters et al., 2016). The bike lane score measures the amount of bicycling infrastructure in an area, though does not include sharrows, bicycle parking, or bicycle sharing. The hill score is derived from the steepest slope grade within a 200m area and the destinations score is the same as the area's walk score (Winters et al., 2016). Lane score is weighted as 50% while the hill and destinations scores are weighted 25% each, which overcomes one of the drawbacks of using derived measures like Walk Score.

Built environment features near residences can influence whether people engage in bicycling activity; an important consideration when valuing home prices in a neighborhood. The origins of bicycle trips are strongly influenced by density, diversity, and design, or the core elements of the built environment (Cervero and Duncan, 2003). Off-road bike paths and greenway trails may be associated with more greenspace overall; on-street bicycle lanes are associated with higher connected street areas with low speeds and lower overall traffic volume, which are characteristics that homebuyers may value rather than the bike lanes themselves. Well-connected streets with small block distance,

mixed land use, and access to retail activity spurs walking and bicycling activity (Cervero and Duncan, 2003; Sun et al., 2017). An increase in residential density and access to parks and attractive destinations was associated with uptake of transport-related bicycling whereas street connectivity was associated with starting recreational bicycling activity after people moved residences (Beenackers et al., 2012).

While studies have examined the associations between walkability, trails, bike lanes and property costs they do not directly measure activity; there is no guarantee that proximity to walk or bike friendly amenities will actually increase pedestrian activity (Li et al., 2015). Some bicycle activity occurs on bicycle specific infrastructure, but neighborhood ridership volumes may relate more to street connectivity or other built environment features than bicycle infrastructure alone; riders may utilize streets that are not specifically marked as bicycle lanes. Related, many low speed and low traffic volume streets are bicycle activity friendly, but because they are naturally conducive to bicycling activity they are not separately labelled as such.

Ridership in particular is associated with certain neighborhood features that appeal to cyclists and residents alike. For example in Strava, a smartphone app for recording bicycling trips that is focused on fitness and competition, recreation trips occur on short length streets, with high connectivity, and in areas nearer to residential land (Sun et al., 2017a). Further, cyclists have some preference for areas with traffic calming features (Broach et al., 2012) and low traffic speeds which could be seen as a beneficial neighborhood features. Greater safety for motorists and cyclists alike has also been associated with presence of bicycle lanes (Brady et al., 2010), which could contribute to overall neighborhood safety. Increased active transport in walkable areas, with a corresponding decrease in motorized transport, is associated with health benefits related to less air pollution and carbon dioxide emissions reduction (Woodcock et al., 2009; Frank et al, 2006). Overall, walkable neighborhoods are associated with greater rates of bicycling activity (Dill and Carr, 2003; Krizek et al., 2009; Nelson and Allen, 1997; Reynolds et al., 2009; Van Dyck et al., 2010). These characteristics that are associated with increased bicycle ridership are the same that may be desirable neighborhood features and therefore relate to higher property values, as market values of properties should

reflect attraction to particular developments by displaying a price premium (Bartholomew & Ewing, 2011). We specifically examine ridership volume herein as it also captures where infrastructure is not located but ridership still happens.

Greater community health benefits may result from self-selection of places suited to pedestrian activity rather than overtly designing neighborhoods for said activity (Cervero and Duncan, 2003). Walking and bicycling have obvious health benefits related to increased physical activity, but also have also been shown to improve hypertension, type 2 diabetes, and overall mortality (Saunders et al., 2013). Further, adopting walking or bicycling on larger scales can improve traffic congestion and carbon emissions if the overall number of vehicular trips is reduced. These are factors that may contribute to positive neighborhood features that appeal to home buyers.

5.3 Hedonic Price Analysis

Hedonic pricing is commonly used to infer values for non-traded environmental or public goods from observed market transactions for related private goods, such as housing. Housing is a differentiated product that includes attributes such as environmental quality or access to public goods, e.g. bicycle infrastructure. Observed market transactions for housing implicitly carry price signals about these non-market attributes that are part of the bundle of attributes that consumers consider when purchasing housing (Palmquist, 2002; Freeman, 2003). Hedonic modeling is a revealed preference technique that uses information about housing costs to identify premiums buyers are willing to pay for perceived amenities or to avoid disamenities. While hedonic models are difficult to generalize across regions, characteristics that are repeatedly significant in the models point to features that are more consistently valued by home buyers (Sirmans et al., 2005). Hedonic studies that investigate residential properties typically examine characteristics related to the structure of the house itself as well as the location and surrounding neighborhood, which includes proximity to a number of potential amenities and disamenities. The structural features may include square footage, number of rooms, presence of a garage and pool, and age of the house among other attributes (Bartholomew & Ewing, 2011; Sirmans et al., 2005). Neighborhood and proximity features may include factors related to the population like median age or racial

makeup or other location characteristics like proximity to parks, and greenspaces, business districts, school quality, and access to transit (Bartholomew & Ewing, 2011; Sirmans et al., 2005). Hedonic models are used to extract the value of these non-market characteristics through the market for housing (Palmquist, 2002). In this way, hedonic pricing is one way to evaluate livability or the aesthetic value of bicycling activity and bicycle facilities (Krizec, 2007).

5.4 Empirical application: Study area Tempe, Arizona

Residents and city officials in Tempe, AZ are engaged in a debate as to the economic value of bicycling infrastructure, with some residents arguing that the presence of bicycle lanes will reduce home values. Because of the interest in this question of value, we select Tempe as our study area. Our analysis of the influence of bicycle lanes on housing prices will inform this debate, and provide robust analysis to help city planners and citizens alike make informed decisions on installing bicycle infrastructure

The city of Tempe is part of the Phoenix metro area in Maricopa county and has a population of over 174,000 people as of 2016, growing from 161,719 in 2010 (US census factfinder). This population growth has been absorbed into the 40.2 sq. mi. existing land area of the city. Land annexation for Tempe to accommodate growth halted when the city boundaries joined with the cities of Mesa, Scottsdale, Chandler, and Phoenix (Figure 5.1). This has led to urban densification and an emphasis on mixed land use development. This type of development requires mixed modal transportation that includes active transportation such as bicycling. Bicycling is made easier due to the generally flat topography, low rainfall amounts (average annual precip is less than 10 inches a year), and year round warm temperatures (annual average high temperature in Tempe is 87.3 F and the average low is 55.3). The moderate fall, winter, and spring temperatures as well as little rainfall make Tempe bikeable year-round.

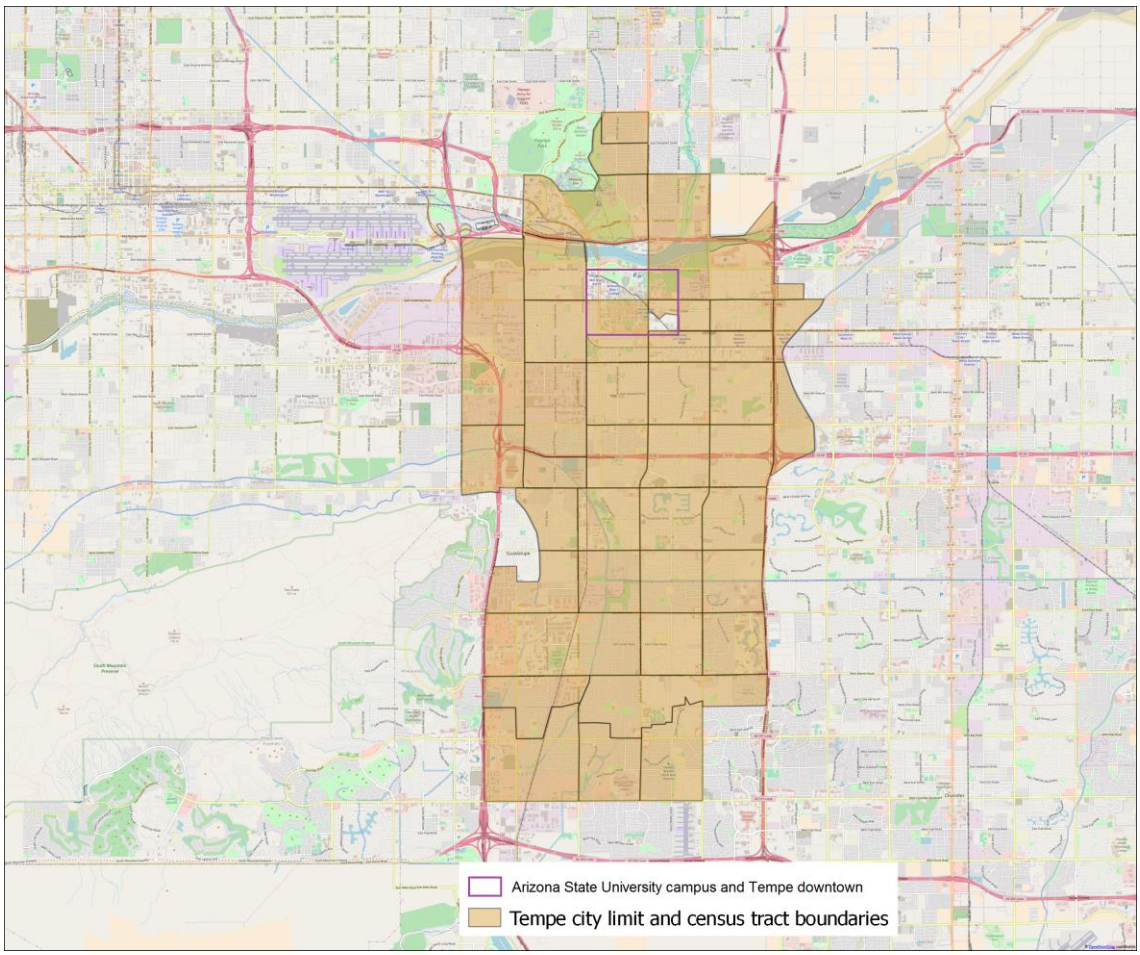

Figure 5.1. Tempe city limits and census tract boundaries, situated between Phoenix in the west and Mesa/Chandler in the East.

One of the significant drivers of urban growth and economic development in the City of Tempe is the largest campus of Arizona State University (ASU). Situated in the north central part of the city near the Tempe Town Lake, ASU brings thousands of people to campus daily. ASU students and employees either reside in Tempe or commute to the campus by car, public transportation (light rail and bus), or active transportation (e.g., walking, skateboarding, bicycling).

To support both ASU and the residents of Tempe, the city's master transportation plan includes improvements to bike lanes, addition of buffered and protected bike lanes, and development of bicycle boulevards. Tempe's bicycle infrastructure includes over 175 miles of bicycle lanes, routes, paved and unpaved multi-use paths, and paved shoulders

(Figure 5.2). Approximately 52% of Tempe's facilities are bicycle lanes, or designated portions of roadways that have preferential or exclusive use for bicycles. An additional 27% of the facilities are multi-use paths, which are completely separated from motorized traffic. Bicycle routes comprise 15% of the infrastructure and are designated by signs only, typically on residential streets (TMP, 2015). Currently, a greater percentage of Tempe residents bike and walk to work as compared to other cities in the Phoenix Metropolitan area (TMP, 2015). The American Community Survey (ACS) 2012 revealed that 4.2% of residents bike to work as compared to the Maricopa County average of 0.8%. Figure 5.2 shows the distribution of bicycle lanes throughout the study area. In addition, the Tempe area supports three bicycle share systems which are known to increase rates of bicycling activity.

Despite the relatively high use of bicycling as well as the planning and advocacy for more bicycling infrastructure, residents at town hall meetings debate the economic value of bicycling infrastructure, with some residents arguing that presence of bicycle lanes will reduce home values. Our analysis of bicycle lane influences on housing prices will help city planners and citizens alike make informed decisions about the value of bicycle infrastructure. Our study therefore focuses on 37 of the 39 2010 census tracts in the Tempe area with a mean size of 1.14 sq miles (2.95 sq km). Two census tracts within the study area were excluded as they contained no single family houses; the excluded tracts corresponded to areas largely containing ASU and the Tempe downtown area. We based our analysis on 5437 single family properties within these 37 census tracts.

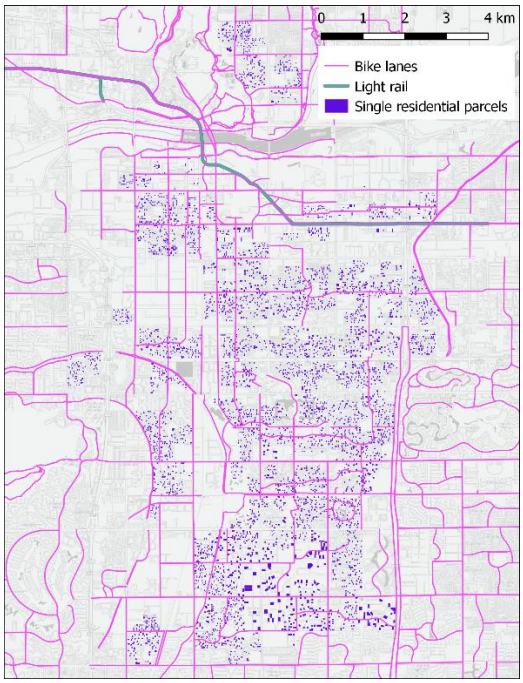

Figure 5.2. Location of bicycle lanes, light rail line, and single family homes in the study area.

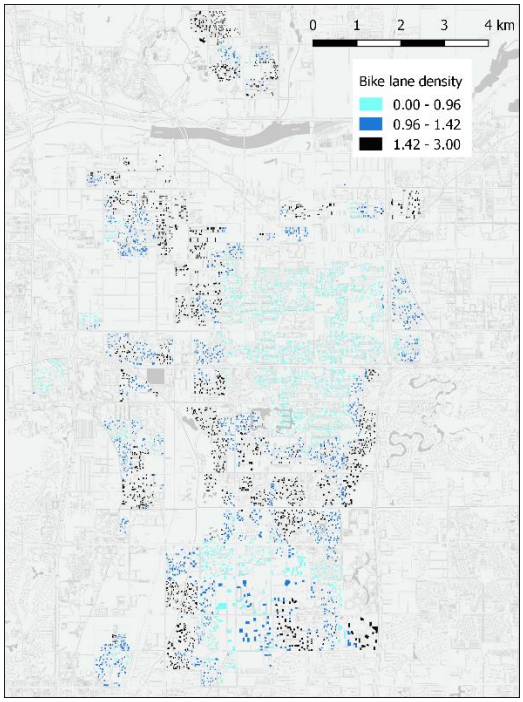

Figure 5.3. Low, medium, and high quantiles of bicycle lane density, where density is the total distance of bicycle lanes within a ½ mile of each parcel.

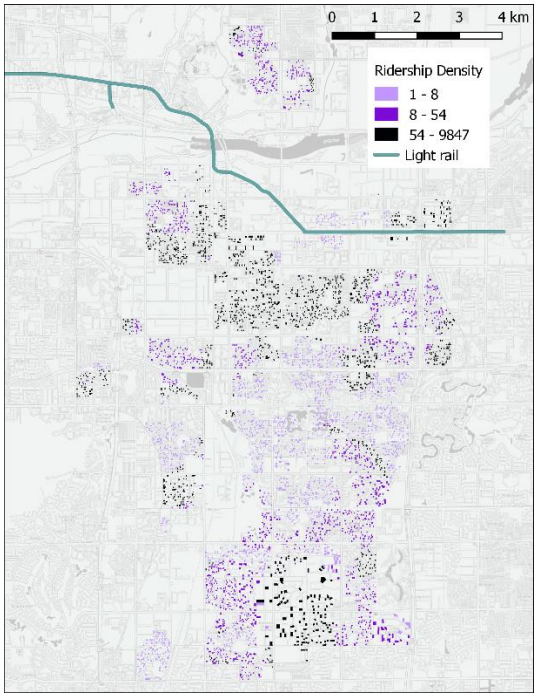

Figure 5.4. Low, Medium, and high quantiles (equal count) of ridership count, where count is the total number of riders within a ½ mile of each parcel

5.5 Data

We construct a dataset to conduct a hedonic pricing analysis using information about property characteristics, locational and demographic characteristics, bicycling amenities and other relevant variables. Figure 5.2 shows the distribution of bicycle lanes and residential properties throughout the study area. The bicycle lanes and single family houses appear to be situated ubiquitously throughout the study area. The light rail runs from the northwestern corner of the study area to the northeastern side; housing is sparser along the light rail corridor as compared to the rest of the study area. As shown in Figure 5.2, the light rail runs through the northernmost region of the study area so a large distance for most properties is expected. The lowest house prices occurred near the light rail and in the westernmost part of Tempe while the highest house prices are situated in the southern extent. Figure 5.3 shows that the density of bike lanes varies, with a notable area of low bicycling infrastructure near the central portion of the study area. Related to bicycle density, ridership density is shown in Figure 5.4. The general qualitative patterns

between bicycle infrastructure and ridership density differ. Ridership is high in the northern and south-central parts of the study area; the part of the study are that corresponds to the least dense bicycle infrastructure also has low ridership, though there are some pockets of high and medium ridership within it.

We constructed a dataset to for a hedonic pricing analysis using information about property characteristics, neighborhood features, amenities, and features related to bicycling. This section describes the dependent and independent variables for our models and the data sources for each. We obtain single-family residential property sales and associated housing characteristics from the Maricopa County Assessor's office sales affidavits st42025 files from 2017. Actual market sales data from housing market transactions are collected as the dependent variable (*P*); sales price was used because it represents the actual market transactions. We limited the period of analysis to the years 2013 and 2016 inclusive, representing a period of relative market stability at the study location. This period does not span events that could create significant structural changes, such as recession, in the housing market. We also restricted transactions to only single-family residential property and omitted sales that were not arms-length or included non-standard financial arrangements resulting in 5,437 observations. Although inflation was very low during this period, we also adjusted nominal prices to reflect 2016 real dollars. This lead to increases of 5.91, 4.09, and 2.21 percent for the years 2013 to 2015 respectively.

Independent variables reflecting structural attributes, neighborhood characteristics including demographics, and the bicycling infrastructure and usage were acquired to analyze the factors that influence housing values in Tempe. The Maricopa Assessors data contained information about livable square feet and lot size for each property as well as other characteristics such as pool, garage, and the year the house was built. Pools were considered as a binary whether a property had a pool. House age was the number of years since the house was originally built. The average house in the study was 656.3 sq. ft. on a lot of 9059 sq. ft. and was 41 years of age (Table 2). Of the 5437 properties, 2576 had a pool and the vast majority (94%) had a garage or carport.

Using bicycle lane locations acquired from Maricopa Association of Governments (2014) we also calculated the distance from each property to the nearest bicycle lanes. Similarly, using parks and other greenspaces (e.g., golf courses), we computed distance from each property to the nearest park. The mean distance to a bike lane was 567 ft. whereas the mean distance to the light rail was nearly 3 miles. In terms of greenspace, the mean distance to parks was 1432 ft.

Neighborhood variables reflect the neighborhood characteristics representative of all within neighborhood and neighborhood amenities, which are opportunities available to those within the neighborhood. Neighborhood level variables were derived from the US Census's American Community Survey 2015 and include population density, median age, median household income, percent rented households, and percent minority at the census tract level. Each single family residence in our study was assigned the value of the tract within which it was situated. We recognize that ecological fallacy tells us that these values that are assigned to individual households are deduced from group observations and are not necessarily the actual observed value at that household. The variables do however reflect the neighborhood characteristics, which is the intent of the study.

For neighborhood amenities, we collected data on school quality and eight variables on transportation amenities. In order capture school quality, each property was assigned to the elementary, middle, and high school enrollment catchments that contained them. The GreatSchools rating from greatschools.org was then used to rate each school. The overall rating accounts for test scores as well as socioeconomic, racial, and attendance factors; ratings are on a 1-10 scale where 10 is the highest and a 4-7 is about an average performing school ("About GreatSchools' Ratings", 2018).

Since bicycle ridership is associated with bike lanes but riders do still ride on both roads with no markings as well as roads with "sharrows" (not included in our set of marked bicycle lanes) we have also controlled for the road network in two ways. In terms of road access, we first calculated the road density within the census tract, then we also calculated the beta measure for the roads in each census tract for each property. The beta index is defined as $\beta = e/v$ where e = edges (streets) and v = nodes (intersection) in a graph and it is a measure of connectivity. Complex connected roads would have a higher

value of beta where lower values indicate a less connected network. Access to light rail was computed as the distance from each property to the light rail line.

The set of transit-related variables were developed in several ways and include accessibility indices and transit infrastructure variables. First Bike Score was used to indicate bikeability of an area and was derived from the WalkScore.com website. While the official measures of walkability and bikeability are proprietary to the site, points are given to areas according to their distance to amenities where those within a ½ mile (approximately a five minute walk) are given the maximum points and no points are given for walks that would be in excess of 30 minutes ("Walk Score Methodology", 2018). Census tracts defined neighborhoods in this study, so the Bike Score at the centroid of each tract was collected and assigned to each property using spatial join. The Walk Score for each census tract was also recorded as a measure of the walkable access for each neighborhood. In addition, Bike Score was not available for five of the census tracts which led to too many properties with unknown values. Bike Score and Walk Score were moderately correlated (0.74) so Walk Score was used in place of Bike Score in this analysis

Second, bike lane density was computed two ways; a shapefile was obtained from Maricopa Association of Governments that displayed a polyline file of all bicycle lanes. The first density measure was the distance of bicycle lanes within the area of each census tract (density = distance/tract area). The second measure of density computed the length of bicycle lanes within a ½ mile buffer of each property (density = length/buffer area). Therefore the densities are measured as a property's immediate neighborhood as well as the individual surrounds of the property itself. We chose ½ mile as the distance a person is willing to travel for everyday activities. Next bicycle infrastructure variables specific to the property were developed; proximity to bicycle lanes was derived by determining the distance in feet from each property to the nearest bicycle facility. Since greenspaces are associated with higher home values (Lindsey et al, 2004; Asabere and Huffman, 2009), we also computed distance from each property to the nearest park to control for the existence and location of those spaces. Ridership was operationalized as the number of riders on each street segment within a ½ mile buffer of each property. The ridership data

originated from the Strava smartphone application and consisted of volumes of riders on each street segment in the study area.

5.6 Empirical modelling

The sale price (*P*) of a house is a function of differentiated property attributes such as the characteristics of the lot and the physical housing structure, neighborhood characteristics, amenities/accessibility in the area, as well as bicycling related infrastructure, equation (1).

$$P=P(\boldsymbol{H},\boldsymbol{N},\boldsymbol{A},C) \quad (1)$$

Where:
H = vector of housing characteristics (e.g., lot size, house size and age)
N = vector of neighborhood characteristics e.g. school quality
A = vector of amentities and accessibility in the area e.g., road density
C = vector of bicycling characteristics e.g., bike lanes, ridership volume

Taking the first derivative of (1) with respect to any of the bundle of attributes yields their marginal prices at the given level of consumption. For example, $\partial P/\partial c$, represents the marginal value of a given bicycling infrastructure variable *c* i.e. the increase in expenditure required to obtain one more unit of attribute *c*, all else equal (Freeman, 2003). These relationships can be determined statistically by estimating the hedonic pricing function and parameter values, which represent the marginal values. Economic theory lends latitude in model estimation and there are few restrictions on functional form, data characteristics are generally used to guide our model parametrization (Owusu-Ansah, 2011).

We utilize hedonic price modeling technique to understand the impact of various characteristics on single-family residential homes in Tempe, AZ. We start initially with an OLS specification

$$P_i=\beta_0 + \beta_1 H_i + \beta_2 N_i + \beta_3 A_i + \beta_4 C_i + \varepsilon_i \quad (2)$$

Where

P_i = adjusted property sale price

H_i = vector that includes house/property characteristics (e.g., square footage, lot size, age, pool binary)

N_i = vector that includes neighborhood sociodemographic characteristics (e.g., % renters, school score)

A_i = vector that includes access to amenities such as light rail, roads, and parks (e.g., distance to light rail, road density)

C_i = vector of bike facility and ridership characteristics (e.g., bicycle lane density, ridership volume)

In the OLS model (Model 1), we include characteristics specific to the transaction, the property, the neighborhood, and our set of biking-related variables. Physical property attributes such as lot size (in sq. ft.), livable square footage of the house, and age of the house in years were also obtained from the Maricopa County assessor's office along with a shapefile showing the location and footprint of each property throughout the study area. Housing prices are positively skewed and were logged in Model 2 to correct the skewness yielding the following specification:

$$lnP_i=\beta_0 + \beta_1 H_i + \beta_2 N_i + \beta_3 A_i + \beta_4 C_i + \varepsilon_i \quad (3)$$

As prior models have suggested spatial dependence may occur between units (Shir, 2007) or in model residuals (Liu and Shi, 2017). Moran's I_i analysis was conducted to determine whether spatial autocorrelation in housing prices was present in the study area, as we can expect that housing prices are influenced by the price of neighboring homes. Results indicated that adjusted housing prices were similar among the nearest neighbors (0.68,

p=0.01) within the study area. For these reasons we include spatial specifications of the model alongside the general ordinary least squares model. The general spatial autoregressive model then is specified as:

$$y=\rho Wy+X\beta+\epsilon \quad (4)$$

Where ρWy specifies a spatially lagged dependent variable parameter using a spatial weighting matrix; ρ is the spatial lag parameter, W is the weight matrix, and X is the vector of independent variables in the model in equation 1. For this study we used k-nearest neighbor weighting as it can be thought of as a weighted averaged of the response variable in a neighborhood, and it allows for a neighborhood that is not of a fixed width (Owusu-Ansah 2011). As an important consideration for the model, we have chosen to specify *W* as the 8 nearest neighbors to account for the spatial arrangements and reduce bias while maintaining variance.

We anticipate that lot size, livable square footage, and whether a house has a pool will all contribute positively to house price whereas price will decrease with house age. In terms of neighborhood level variables, we expect that higher median household income and school scores will be associated with higher house prices as these are desirable neighborhood features; higher percentage of renters will be associated with lower prices. Road density should be negatively associated with house prices as we believe homeowners value secluded home locations with longer blocks and fewer connecting streets, including development areas with cul de sac style streets. Bicycle lane density should be positively associated with house price and we further anticipate that ridership volume will also be associated with higher prices; riders prefer low traffic volume areas with greenspace which are also considered housing amenities.

5.7 Results

Three different models were specified. Model 1 shows the regular OLS approach while Model 2 shows the log-linear OLS model. We performed several tests to check that the models herein did not violate the assumptions of OLS. First the model is linear in parameters and the mean of the residuals is zero (1.728143e-12). Figure 4 shows the residuals vs. fitted values plot indicating no severe decreasing or increasing trend and

therefore homoscedasticity. The Durbin-Watson test results were 1.98 for both models 1 and 2 with a p-value of 0.2889 and 0.1848 respectively indicating that there was no autocorrelation among the residuals. Similarly, correlation tests showed that the independent variables were not correlated with the residuals and all independent variables had variance greater than 0. All variance inflation factor values were less than four indicating that there was no multicollinearity present in the models. Finally, QQ plots (Figure 5) showed that the residuals in models 1 and 2 were represent a heavy tailed normal distribution. Table 5.1 shows the results of the ordinary least squares models; the coefficients are given with standard errors in parentheses beneath.

Table 5.1. Descriptive statistics for independent variables

Statistic	N	Mean	St. Dev.	Min	Max
PriceAdj (USD)	5,437	$286k	$143k	$26.5k	$4.3 mil
Property level:					
Livable SqFt		1889	686.2	480	9650
Lot Size (sqft)		9059	5550.4	1133	128k
House Age (years)		41.9	12.83	1	94
Neighborhood:					
Pop. Density (ppl per sq mi)		4,867.9	2,354.7	515.6	13,033.1
% Minority pop.		34.6	10.6	12.9	65.8
Med. HH income (USD)		66,136	27,578.8	19,221	127,879
% Renters		43.8	22.4	3.8	96.6
School Score (max)		5.5	1.9	1.5	8.5
Infrastructure:					
Road density (ft/buffer area)		21.8	5.3	11.8	35.5

Road beta	1.7	0.15	1.45	2.14
LTR distance (ft)	15,194.5	9,251.5	285.3	36,680.1
Walk score	38.99	14.17	8	74
Bike score	69.734	14.524	0	100
Bike lane density (ft/buffer area)	6.0	2.6	1.6	16.1
Distance to bike lane (ft)	567.3	521.3	0.0	3,121.7
Ridership density	171.2	507.8	1	9,847

As expected lot size, livable square feet, and whether or not the property had a pool were all highly significant positive predictors of house price ($p<0.01$). House age was also significant where house price decreased as age increased.

Ultimately, only percentage of renters at the census tract level was used to characterize neighborhood level features. At the neighborhood (tract) level median age, median household income, and percentage of renters in the tract were considered, though subsequent variance inflation factors were moderately high (>4) indicating that they were highly correlated in the models. Percentage renters was a better predictor of house price so it was retained in subsequent models. According to the model results in Table 3, house price increases as the percentage of renters in the neighborhood decreases. As anticipated, distance to light rail from each property was also highly significant and negative in the models. The maximum elementary school score was also significant and contributed positively to house price.

When examined independently Walk Score and Bike Score were significant and negative meaning that house price increased as the scores decreased, though they were not significant when other bicycling and accessibility measures were included in the specification. Modeling distance to nearest bike lane and neighborhood (tract based) bike lane density showed they were negative and positive respectively. The signs align as house price would decrease the further a house was from a bike lane and higher

neighborhood bike lane density also contribute to higher house prices. As predicted, neighborhood road density was significant and negative. Similarly, bike lane density contributes positively to house price; for each unit increase in bike lane density, we would expect a 1.4% increase in house price. Similarly, bicycle ridership was significant and positive $p<0.01$. (Table 3). We would expect very marginal (<1%) increases in house price for every unit increase in neighborhood ridership density.

Table 5.3, model 3, shows the results of the spatial autoregressive models, with coefficients listed with standard error in parentheses. Again, spatial models were used as there was spatial autocorrelation present in the house prices across the study area. Lot size, livable sq. feet, age of house, and whether the house had a pool and garage remained significant. As in the aspatial model, distance from light rail was negative and significant indicating that house price rose as distance from light rail decreased. Percentage of renters and median household income remained positive neighborhood level predictors of house price. Even when controlling for road density in the area, bike lane density remained positive and significant indicating that house prices were higher in areas with higher bike lane density. If bike lane density was controlled for in the model, the amount of ridership in the buffered neighborhood was not significant.

Table 5.4 shows the direct and indirect effects associated with the SAR model, since the coefficients cannot be directly interpreted based on the lagged variables. The direct effects result from a change in the respective independent variable for each property whereas the indirect (spillover) effects are based on changes within a property's neighbors. Direct and indirect impacts that are positive suggest that increasing a given variable will have an impact in the immediate or neighboring area. Lot size, livable square feet, pool binary, and school score are all positive impacts whereas age, distance to LTR, percentage renter, and road density are negative; decreasing any of these factors would increase house price in neighboring and immediate properties.

Table 5.2. OLS (model 1 & 2) and SAR (model 3) results

	Dependent variable:		
	OLS		SAR
	PriceAdj	log(PriceAdj)	log(PriceAdj)

	(1)	(2)	(3)
Lot Size (sq ft)	5.741***	0.00001***	7.1406e-06***
	(0.201)	(0.00000)	(0.00000)
Livable Sq.Ft.	141.463***	0.0004***	2.7468e-04***
	(1.906)	(0.00001)	(0.00000)
Age (years)	-906.520***	-0.003***	-3.3048-03***
	(114.976)	(0.0003)	(0.0003)
Pool binary	5,372.378***	0.070***	6.2827e-02***
	(2,053.161)	(0.006)	(0.00525)
Distance to lightrail (ft)	-0.628***	-0.00000***	-5.0012e-06***
	(0.197)	(0.00000)	(0.00000)
Percent Renter	-414.338***	-0.003***	-5.2553-03***
	(66.816)	(0.0002)	(0.00019)
Max. elem. school score	5,992.084***	0.004**	3.9397e-02***
	(742.686)	(0.002)	(0.00200)
Road Density (ft/buffer area)	-2,875.302***	-0.010***	-4.4234e-03***
	(200.313)	(0.001)	(0.00054)

Ridership volume	7.958***	0.00002***	5.9534e-06
	(1.914)	(0.00001)	(0.00000)
Bike Lane Density (ft/buffer area)	5,533.627***	0.014***	1.1606e-02***
	(405.695)	(0.001)	(0.00105)
Constant	12,226.320	12.094***	6.0560***
	(11,887.300)	(0.033)	(0.199)
Observations	5,437	5,437	5,437
R^2	0.770	0.750	0.787 (pseudo)
Adjusted R^2	0.769	0.749	
Residual Std. Error (df = 5426)	68,655.900	0.191	
F Statistic (df = 10; 5426)	1,813.820***	1,625.357***	
Log Likelihood			1739.967
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 5.3. Direct and indirect effects associated with the SAR.

	Direct	Indirect	Total
Lot Size	0.000005641326	0.000004077118	0.000009718445
Livable Sq.Ft.	0.000259443047	0.000187505547	0.000446948595
Age (years)	-0.002239891125	-0.001618821607	-0.003858712732
Pool Binary	0.061458648461	0.044417600021	0.105876248482
Distance to LTR	-0.000003277067	-0.000002368413	-0.000005645479
% Renter	-0.001031601086	-0.000745562188	-0.001777163274
School Score	0.002852229089	0.002061372549	0.004913601638
Road Density	-0.005107507601	-0.003691314980	-0.008798822581
Ridership vol.	0.000008396792	0.000006068557	0.000014465349
Bike ln. Density	0.010085373054	0.007288934554	0.017374307608

5.8 Discussion

Considering debates in Tempe as to the value of bicycle lanes in residential areas, the current study demonstrates that there is some value associated with the presence of

infrastructure that supports bicycling activity. Our findings align with previous research linking property values to bicycle lanes (e.g., Liu, 2017). House price increased as percentage of renters in a neighborhood increased; this relationship is possibly indicative of areas where people are willing to pay a premium to live or that owning a house in a high rentership area costs a premium. Decreasing the percentage of renters around any given property would increase house prices, based on the direct and indirect impacts. The highest rates of renting were in the northern and northwestern areas of the study areas where housing prices were medium or low which correspond to the university and business areas of Tempe; the lowest numbers of renters occurred where the highest house prices occurred. Since we did not distinguish between owner-occupied single family residences, it is also possible that the same high cost homes are being rented rather than occupied by the owners. It makes intuitive sense that neighborhoods with higher incomes have higher cost homes.

House prices generally increased as distance to light rail decreased, though the highest price homes in the study area are also furthest from the light rail. While some of the least expensive houses immediately surround the light rail (Figure 5.2), several high home price areas are located within the average distance (3 miles) to the light rail. While there have been mixed results as to whether transit benefits property prices (Bartholomew & Ewing 2011), there is a clear and strong relationship in this study. Overall the finding aligns with previous research showing that properties close to light rail see negative effects while those further from it see positive effects (Bowes and Ihlanfeldt, 2001).

The results for bike lane density align with previous findings in that buffered bike lane density is positive and significant in home prices. We expect that the neighborhood features associated with bike lanes are the same ones that people are willing to pay a premium for e.g., small blocks, low traffic speeds. While Liu and Shi (2017) examined the impacts of bike lanes on home prices, they did not consider all types of bike lanes as we have done here. Further, we have shown that ridership is also significant and positive, thought its effect size is small. It likely that rather than bicycle ridership spurring higher home prices, the features of neighborhoods that are associated with higher ridership are those for which home buyers are willing to pay a premium. For example, wide and

connected streets with low traffic volume and speed are those that bicyclists tend to choose. Increased traffic leads to more traffic noise, which is associated with lower home prices (Theebe, 2004). Surveys of bicyclists have shown that striped bicycle lanes lead to greater perception of safety that those without and similarly pavement conditions play a role in cyclist perceptions of the quality of the built environment (Landis et al., 1997). These may be the same factors, as part of the bundle of property attributes, that home buyers examine while they decide to purchase and decide the premium they are willing to pay for better conditions.

This study has used crowdsourced data to examine whether ridership is related to property values. Crowdsourced data from smartphone apps has allowed this analysis to overcome prior limitations related to the economic valuation of bicycle infrastructure. Rather than examine the presence of bicycle lanes, which are not necessarily widely used in all areas of Tempe, despite their presence, we have demonstrated that rider volume near properties is related to house prices. In a similar vein, there is dense ridership in places where there is no bicycle infrastructure (Figures 5.3 and 5.4) which demonstrates that infrastructure and ridership are separate factors that both need to be examined. Future work should further assess the influence of said ridership by examining additional, street level factors that are known to influence bicycle ridership volumes. For example, street shade and vegetation and attractiveness of areas could be examined to determine the relationship between those features, ridership, and property values.

5.9 Conclusion

The transit factors that influence house prices in Tempe AZ include distance to light rail and bicycle lane density. The results support the idea that transit oriented design, and bicycle lanes in particular, increase house prices nearby. It is possible and certainly plausible that the same characteristics that come with bicycle lanes are those that are desirable to potential homeowners. The lower traffic speeds and lower traffic volumes/noise associated with bike-friendly areas are the same features that contribute to neighborhood safety. Further, community health benefits as presence of bicycle lanes can spur active transport behavior. In this case, it seems that bicycle lanes and bicycle

ridership can be considered neighborhood amenities and an attractive selling point for potential buyers.

Transit oriented design, including built environment features that support walking and bicycling, is of interest to planners and stakeholders in many cities. TOD aids problems associated with mobility and accessibility including congestion, pollution, and sprawl. Studies show the economic valuation of such TOD features, including potential positive influences on property prices. Results from this study can be used to inform debates on the positive economic effects associated with bicycle friendly infrastructure and ridership.

CHAPTER 6
CONCLUSION

Analysis of human mobility is primarily focused on popular places and how people travel to and between them. Active transport, or walking and bicycling, is associated with both personal and community level benefits including reduced risk of adverse health outcomes and reduced carbon emissions. Despite benefits, levels of walking and bicycling remain low in many urban areas. Detailed information about the contexts that support and encourage walking and bicycling as practical transport options will aid planning and policy decisions. Understanding the factors that influence mobility is key in creating effective planning and policy to predict and manage human movement behavior.

This detailed information however, depends on the data that are available for research and analyses. While conventional methods used to collect mobility data capture some information about movement such as traffic flows and volumes, they have a number of limitations. These limitations include high operational costs, coarse spatio-temporal scale, and small sample sizes. Conventional data are particularly limiting in analysis of active transport like walking and bicycling because these movement behaviors occur on finer spatial scales and necessarily require finer scale data. Travel surveys and studies are limited in spatial and temporal scope and may have comparatively small sample sizes. Bicycle counts are commonly conducted, but they suffer some of the same spatio-temporal limitations as surveys but are additionally limited in that no route information can be deduced from their results. Lack of data has limited analysis of walking and bicycling as mobility modes. Further, there has traditionally been a dearth of data that can link active transport modes to the built environment and networks in which they occur.

One of the opportunities to improve analysis of active transport lies in crowdsourced data that originate from smartphones and other location aware technologies. Crowdsourced bicycling data provide more information on human mobility. These detailed movement datasets often contain full routes, speeds and durations of

travel, and have the benefit of increased sample sizes and near-continuous data collection. Few studies have critically examined crowdsourced data related to bicycling activity. Those that have focused on the Strava fitness app and found that Strava users tended to ride in the outskirts of urban areas, on short, connected streets. The neighborhood characteristics that support bicycling are those associated with short, direct routes, lower traffic volume and speeds, and bicycling specific infrastructure that supports riding. In smaller scale studies, cyclists prefer bike lanes and boulevards and are willing to go out of their way to stay on bicycling specific infrastructure though it remains to be seen whether conclusions like this hold when considering crowdsourced rather than conventional data.

The gap in the research lies in the under examination bicycling activity using crowdsourced data as a mobility and active transportation data source. Chapter 3 addressed the gap by first examining crowdsourced and conventional data sources to determine how they correspond in representing bicycling activity. Local indicators of spatial association were used to generate locations of similarity and dissimilarity; these locations were based on the difference in ridership proportions between a conventional manual count and crowdsourced data in the greater Sydney region. Similarity was found more often in areas with lower population density, greater social disadvantage, and low ridership overall. Dissimilarity was found among five locations; these locations were among low bicycle infrastructure density areas though they were popular locations in the Strava dataset.

Next, Chapter 4 assessed how the built environment, infrastructure, and sociodemographic factors in Sydney predict where bicycling activity occurs. The analysis further examined whether the factors that predict bicycling activity were the same between two datasets, one from Strava and one from RiderLog. Median rent, percentage of residential land, road density and number of people using two or more modes to travel to work were all significant predictors of ridership volume in both datasets. Overall, more rides were logged in areas with greater bike lane density and a lower percentage of residential land. In terms of the difference in riders between datasets, Strava users tended to ride near the city center but also in the outskirts of the study area whereas RiderLog

users were much more concentrated in the city center only. Further, bicycling infrastructure density was more important for RiderLog users than Strava users.

Chapter 5 examined the influence of ridership and bicycle lane density on home prices. Findings showed that bicycle lane density with a ½ mile neighborhood of a home positively influenced adjusted sale price. Similarly, ridership volume within the same neighborhood radius also positively contributed to house price. This is an important distinction as neighborhoods may have bike lanes but lack in ridership, or may not have bicycle lanes but have other built environment features that spur bicycle use in the area. It is probable that rather than ridership directly influencing house price, it is an indicator of neighborhood features that facilitate bicycle riding. Some of these factors may be low traffic speeds and volumes, wider streets, and green spaces. Bicycle lanes and neighborhood activity may also spur uptake of bicycling as an activity, which will help realize the associated community health benefits.

The studies from Chapters 3 and 4 overcome a gap in the knowledge base related to understanding bicycle ridership and crowdsourced data. Because conventional count data are static in location, it would not be possible to determine the factors that influence bicycling activity across the study area while considering general movement paths. Similarly, household travel surveys and studies often only include origin and destination information so analysis of the factors that influence movement across the study area would not be possible with those data. The major link between Chapters 3, 4, and 5 is the neighborhood and built environment. In Chapters 3 and 4, the neighborhood and built environment play a role in facilitating bicycle activity. In Chapter 5, bicycle lanes influence, or are at least associated with, neighborhood valuation of amenities. Overall, this research aims to explore gaps in knowledge related to the relationships between ridership and the factors that underlie that activity, and using crowdsourced data to explore those gaps. The dissertation research confirms that there are differences in the activity represented by conventional and crowdsourced datasets, as well as differences in users from differing data sources. These differences must be considered and accounted for when using crowdsourced data to examine the factors that underlie bicycling activity. Further, this research aids in understanding how we can use both conventional and

crowdsourced data sets to fill gaps in knowledge and answer questions that have not been possible before.

As planners and other stakeholders invest more in bicycling specific infrastructure, research that includes detailed information about where and when people ride is of importance in making informed decisions related to said infrastructure's location and type. Informed decisions about infrastructure will facilitate bicycling activity so that the benefits of it (e.g., reduced traffic congestion and emissions) can be realized. Similarly, as bicyclists have more choices about where and how they ride, research that includes rider and street-level detail remains important in understanding bicycling activity; if bicycling activity is better understood, planning efforts can be improved to develop safe, efficient, and desirable routes.

The research herein used detailed bicycling data to make conclusions about activity that can be used by planners and bicycling interest groups to both make decisions and appeal for bicycling specific infrastructure. For example, the findings related to locations of similarity and dissimilarity in Chapter 3 can inform policy and planning. Locations of dissimilarity can be used to direct further data collection efforts in order to understand what population of bicyclists (e.g., recreation or transport) is being served by the infrastructure so that future planning efforts better accommodate different types of bicycling activity. The finding that Strava and RiderLog users were different in Chapter 4 highlights the need to consider that crowdsourced data does not universally represent the same types of riders; riders may be more transport or recreation oriented. It is important to understand the types of riders that are represented as policy and planning efforts need to consider the populations that produce different types of activity as infrastructure design choices may vary based on who is riding where and for what purpose. Recreational riders may be better served by long bicycle boulevards on the outskirts of the city whereas transport oriented cyclists may need direct, connected routes through the dense city center. Finally, the link between infrastructure and economic value can be used by planners and interest groups to appeal for increased bicycling infrastructure in residential areas. Overall this research informs on the economic impact, overall ridership, and different data sources associated with bicycling activity.

6.1 Limitations and Future Work

While the crowdsourced data utilized in these studies allowed for analysis that would not be possible with conventional data, the data are likely biased in user input which was not corrected. As mentioned, crowdsourced samples tend to under sample females, older age groups, and low income groups; we did not make any corrections for this under sampling as information about the underlying sample was not available. In a similar vein, comparisons of the crowdsourced samples were limited to those councils or areas that participate in data collection.

This work will allow planners and other stakeholders to make informed, proactive decisions about bicycling infrastructure design. Creating enjoyable, safe, and carefully planned places for people to engage in bicycling activity will increase bicycling rates, which benefits both personal and environmental health. Further, it allows for us to understand how we can do better with conventional data collection strategies. Understanding where data are and are not captured will help direct new data collection efforts.

REFERENCES

"About GreatSchools' Ratings". (2018, March 14). Retrieved from https://www.greatschools.org/gk/summary-rating/

Australian Bureau of Statistics (ABS). (2017). Greater Sydney: Region Data Summary. Viewed 05/16/2017. < http://stat.abs.gov.au/itt/r.jsp?RegionSummary&region=1GSYD&dataset=ABS_REGIONAL_ASGS&geoconcept=REGION&datasetASGS=ABS_REGIONAL_ASGS&datasetLGA=ABS_NRP9_LGA&regionLGA=REGION&regionASGS=REGION>

ABS (2013). Accessed 10/24/2017. Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011 <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2033.0.55.001main+features100052011>

ABS (2016). Australian Statistical Geography Standard (ASGS): Volume 1 - Main Structure and Greater Capital City Statistical Areas, Accessed 9/16/17 <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.001~July%202016~Main%20Features~Statistical%20Area%20Level%202%20(SA2)~10014>

Asabere, P. K., & Huffman, F. E. (2009). The relative impacts of trails and greenbelts on home price. *The Journal of Real Estate Finance and Economics*, *38*(4), 408-419.

Bartholomew, K., & Ewing, R. (2011). Hedonic price effects of pedestrian-and transit-oriented development. *CPL bibliography*, *26*(1), 18-34.

Beenackers, M. A., Foster, S., Kamphuis, C. B., Titze, S., Divitini, M., Knuiman, M., ... & Giles-Corti, B. (2012). Taking up cycling after residential relocation: built environment factors. American journal of preventive medicine, 42(6), 610-615.

Bureau of Infrastructure, Transport and Regional Economics (BITRE) (2016). Lengthy commutes in Australia, Report 144, Canberra ACT. < https://bitre.gov.au/publications/2016/files/rr_144.pdf>

Blanc, B., Figliozzi, M., & Clifton, K. (2016). How representative of bicycling populations are smartphone application surveys of travel behavior?. *Transportation Research Record: Journal of the Transportation Research Board*, (2587), 78-89.

Blanc, B., & Figliozzi, M. (2016). Modeling the Impacts of Facility Type, Trip Characteristics, and Trip Stressors on Cyclists' Comfort Levels Utilizing Crowdsourced Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2587), 100-108.

Boots, B. (2003). Developing local measures of spatial association for categorical data. *Journal of Geographical Systems*, 5(2), 139-160.

Bowes, D. R., & Ihlanfeldt, K. R. (2001). Identifying the impacts of rail transit stations on residential property values. Journal of Urban Economics, 50(1), 1-25.

Boyle, A., Barrilleaux, C., & Scheller, D. (2014). Does walkability influence housing prices?. *Social science quarterly*, *95*(3), 852-867.

Brady, J., Loskorn, J., Mills, A., Duthie, J., Machemehl, R., Beaudet, A., Barrea, N., Wilkes, N., Fialkoff, J., (2010). Effects of shared lane markings on bicyclist and motorist behavior along multi-lane facilities. *tech. rep., Center for Transportation Research*, University of Texas, Austin, TX.

Broach, J., Dill, J., & Gliebe, J. (2012). Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, *46*(10), 1730-1740.

Cervero, R., & Duncan, M. (2003). Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *American journal of public health*, *93*(9), 1478-1483.

Cervero, R., Ferrell, C., & Murphy, S. (2002). Transit-oriented development and joint development in the United States: A literature review. *TCRP research results digest*, (52).

Dill, J. (2009). Bicycling for transportation and health: the role of infrastructure. *Journal of public health policy*, *30*(1), S95-S110.

Dill, J., & Carr, T. (2003). Bicycle commuting and facilities in major US cities: if you build them, commuters will use them. *Transportation Research Record: Journal of the Transportation Research Board*, (1828), 116-123.

Dill, J., & Gliebe, J. (2008). Understanding and measuring bicycling behavior: A focus on travel time and route choice. Accessed 10/10/2017 < http://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1027&context=usp_fac>

Dill, J., McNeil, N., Broach, J., & Ma, L. (2014). Bicycle boulevards and changes in physical activity and active transportation: Findings from a natural experiment. *Preventive medicine*, *69*, S74-S78.

Ellison, R., & Greaves, S. (2011). Travel time competitiveness of cycling in Sydney, Australia. *Transportation Research Record: Journal of the Transportation Research Board*, (2247), 99-108.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the association of American geographers*, *102*(3), 571-590.

Forsyth, A., & Krizek, K. (2011). Urban design: is there a distinctive view from the bicycle?. Journal of Urban Design, 16(4), 531-549.

Frank, L. D., Sallis, J. F., Conway, T. L., Chapman, J. E., Saelens, B. E., & Bachman, W. (2006). Many pathways from land use to health: associations between neighborhood walkability and active transportation, body mass index, and air quality. *Journal of the American Planning Association*, *72*(1), 75-87.

Freeman, A.M. (2003). The measurement of environmental and resource values: Theory and methods (Second Edition). Resources for the Future, Washington DC.

Golgher, A.B. & Voss, P.R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography*, 4(3), 175-205.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Griffin, G. P., & Jiao, J. (2015). Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *Journal of Transport & Health*, *2*(2), 238-247.

Handy, S. L., & Xing, Y. (2011). Factors correlated with bicycle commuting: A study in six small US cities. International Journal of Sustainable Transportation, 5(2), 91-110.

Heesch, K. C., & Langdon, M. (2017). The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health promotion journal of Australia*, *27*(3), 222-229.

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(6), 550-557.

Hess, D. B., & Almeida, T. M. (2007). Impact of proximity to light rail rapid transit on station-area property values in Buffalo, New York. Urban studies, 44(5-6), 1041-1068.

Hiribarren, G., & Herrera, J. C. (2014). Real time traffic states estimation on arterials based on trajectory data. *Transportation Research Part B: Methodological*, *69*, 19-30.

Hirsch, J. A., Winters, M., Clarke, P., & McKay, H. (2014). Generating GPS activity spaces that shed light upon the mobility habits of older adults: a descriptive analysis. *International journal of health geographics*, *13*(1), 51.

Iacono, M., Krizek, K. J., & El-Geneidy, A. (2010). Measuring non-motorized accessibility: issues, alternatives, and execution. *Journal of Transport Geography*, *18*(1), 133-140.

Jestico, B., Nelson, T., & Winters, M. (2016). Mapping ridership using crowdsourced cycling data. *Journal of transport geography*, *52*, 90-97.

Kitamura, R., Chen, C., Pendyala, R. M., & Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation*, *27*(1), 25-51.

Krizek, K. J. (2006). Two approaches to valuing some of bicycle facilities' presumed benefits: Propose a session for the 2007 national planning conference in the city of brotherly love. *Journal of the American Planning Association*, *72*(3), 309-320.

Krizec, K. J. (2007) Estimating the economic benefits of bicycling and bicycle facilities: An interpretive review and proposed methods. *Essays on transport economics*. Physica-Verlag HD, 219-248.

Krizek, K. J., Handy, S. L., & Forsyth, A. (2009). Explaining changes in walking and bicycling behavior: challenges for transportation research. Environment and Planning B: Planning and Design, 36(4), 725-740.

Kuzmyak, R.J., & Dill, J. (2014). Walking and Bicycling in the United States. *Accident Analysis and Prevention*, 65, 63-71.
http://onlinepubs.trb.org/onlinepubs/trnews/trnews280www.pdf

Kwan, M. P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, *102*(5), 958-968.

Kwan, M. P. (2013). Beyond space (as we knew it): toward temporally integrated geographies of segregation, health, and accessibility: Space–time integration in geography and GIScience. *Annals of the Association of American Geographers*, *103*(5), 1078-1086.

Landis, B., Vattikuti, V., & Brannick, M. (1997). Real-time human perceptions: toward a bicycle level of service. *Transportation Research Record: Journal of the Transportation Research Board*, (1578), 119-126.

Lawson, A. R., Pakrashi, V., Ghosh, B., & Szeto, W. Y. (2013). Perception of safety of cyclists in Dublin City. *Accident Analysis & Prevention*, *50*, 499-511.

Le Dantec, C. A., Watkins, K. E., Clark, R., & Mynatt, E. (2015, August). Cycle Atlanta and OneBusAway: Driving innovation through the data ecosystems of civic computing. In *International Conference on Human-Computer Interaction*, 327-338, Springer, Cham.

Leao, S. Z., Lieske, S. N., Conrow, L., Doig, J., Mann, V., & Pettit, C. J. (2017). Building a National-Longitudinal Geospatial Bicycling Data Collection from Crowdsourcing. *Urban Science*, *1*(3), 23.

Lee, J. H., Hancock, M. G., & Hu, M. C. (2014). Towards an effective framework for building smart cities: Lessons from Seoul and San Francisco. *Technological Forecasting and Social Change*, *89*, 80-99.

Li, W., Joh, K., Lee, C., Kim, J. H., Park, H., & Woo, A. (2015). Assessing benefits of neighborhood walkability to single-family property values: A spatial hedonic study in Austin, Texas. *Journal of Planning Education and Research*, *35*(4), 471-488.

Lindsey, G., Man, J., Payton, S., & Dickson, K. (2004). Property Values, Recreation Values, and Urban Greenways. *Journal of Park & Recreation Administration*, *22*(3).

Lindsey, G., Chen, J., & Hankey, S. (2013). Adjustment factors for estimating miles traveled by nonmotorized traffic. In *Transportation Research Board 92nd Annual Meeting* (No. 13-4082).

Liu, J. H., & Shi, W. (2017). Impact of Bike Facilities on Residential Property Prices. *Transportation Research Record: Journal of the Transportation Research Board*, (2662), 50-58.

Mammen Jr, M. P., Pimgate, C., Koenraadt, C. J., Rothman, A. L., Aldstadt, J., Nisalak, A., ... & Getis, A. (2008). Spatial and temporal clustering of dengue virus transmission in Thai villages. *PLoS medicine*, *5*(11), e205.

Meijles, E. W., de Bakker, M., Groote, P. D., & Barske, R. (2014). Analysing hiker movement patterns using GPS data: Implications for park management. *Computers, Environment and Urban Systems*, *47*, 44-57.

Misra, A., Gooze, A., Watkins, K., Asad, M., & Le Dantec, C. (2014). Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record: Journal of the Transportation Research Board*, (2414), 1-8.

Møller, M., & Hels, T. (2008). Cyclists' perception of risk in roundabouts. *Accident Analysis & Prevention*, *40*(3), 1055-1062.

Nelson, T. A., & Boots, B. (2008). Detecting spatial hot spots in landscape ecology. *Ecography*, *31*(5), 556-566.

Nelson, A. C., & Allen, D. (1997, January). If You Build Them, Commuters Will Use Them: Cross-Sectional Analysis of Commuters and Bicycle Facilities. In Transportation Research Board, 76th Annual Meeting, Washington, DC.

New South Wales Government (NSW) (2017). Accessed 10/10/2017. Regional profile - Greater Sydney Local Land Services. < http://greatersydney.lls.nsw.gov.au/our-region/region-profile>

Nicholls, S., & Crompton, J. L. (2005). The impact of greenways on property values: Evidence from Austin, Texas. *Journal of Leisure Research*, *37*(3), 321.

Owusu-Ansah, A. (2011). A review of hedonic pricing models in housing research. *Journal of International Real Estate and Construction Studies*, *1*(1), 19.

Parent, O., & Vom Hofe, R. (2013). Understanding the impact of trails on residential property values in the presence of spatial dependence. *The Annals of Regional Science*, *51*(2), 355-375.

Palmquist, R. B. (2002). Hedonic Models. Chapter 53 in Ed. J.C.J.M van den Bergh. *Handbook of Environmental and Resource Economics.* Edward Elgar Publishing LTD, UK.

Pivo, G., & Fisher, J. D. (2011). The walkability premium in commercial real estate investments. *Real Estate Economics*, *39*(2), 185-219.

Plaut, Pnina O. (2005). Non-motorized commuting in the US. *Transportation Research Part D* 10:347-356.

Pucher, J., Garrard, J., & Greaves, S. (2011). Cycling down under: a comparative analysis of bicycling trends and policies in Sydney and Melbourne. *Journal of Transport Geography*, *19*(2), 332-345.

Reynolds, C. C., Harris, M. A., Teschke, K., Cripton, P. A., & Winters, M. (2009). The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. Environmental health, 8(1), 47.

Romanillos, G., Zaltz Austwick, M., Ettema, D., & De Kruijf, J. (2016). Big data and cycling. *Transport Reviews*, *36*(1), 114-133.

Rybarczyk, G., & Wu, C. (2010). Bicycle facility planning using GIS and multi-criteria decision analysis. Applied Geography, 30(2), 282-293.

Ryus, P., Ferguson, E., Laustsen, K. M., Schneider, R. J., Proulx, F. R., Hull, T., & Miranda-Moreno, L. (2014). *Guidebook on pedestrian and bicycle volume data collection* (No. qt11q5p33w). Institute of Transportation Studies, UC Berkeley.

Sahlqvist, S., Goodman, A., Cooper, A. R., & Ogilvie, D. (2013). Change in active travel and changes in recreational and total physical activity in adults: longitudinal findings from the iConnect study. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 28.

Saunders, L. E., Green, J. M., Petticrew, M. P., Steinbach, R., & Roberts, H. (2013). What are the health benefits of active travel? A systematic review of trials and cohort studies. *PLoS One*, *8*(8), e69912.

Sirmans, S., Macpherson, D., & Zietz, E. (2005). The composition of hedonic pricing models. *Journal of real estate literature*, *13*(1), 1-44.

Strava LLC. (2016) Strava Metro Comprehensive User Guide 3.0.

Sun, Y. (2017). Exploring potential of crowdsourced geographic information in studies of active travel and health: Strava data and cycling behaviour. In: ISPRS Geospatial Week 2017, Wuhan, China, 18-22 Sep 1357-1361

Sun, Y., Du, Y., Wang, Y., & Zhuang, L. (2017). Examining Associations of Environmental Characteristics with Recreational Cycling Behaviour by Street-Level Strava Data. *International Journal of Environmental Research and Public Health*, *14*(6), 644.

Sun, Y., & Mobasheri, A. (2017). Utilizing Crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava Data. *International journal of environmental research and public health*, *14*(3), 274.

Tao, S., Rohde, D., & Corcoran, J. (2014). Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap. *Journal of Transport Geography*, *41*, 21-36.

Tempe Transportation Master Plan (TMP) (2015). *www.tempe.gov/transportationplan.*

Theebe, M. A. (2004). Planes, trains, and automobiles: the impact of traffic noise on house prices. *The Journal of Real Estate Finance and Economics*, *28*(2-3), 209-234.

Transport for NSW. Sydney Cycling Future. December, 2013. Accessed 10/10/2017 < https://www.transport.nsw.gov.au/sites/default/files/media/documents/2017/sydneys-cycling-future-web.pdf>

Van Dyck, D., Cardon, G., Deforche, B., Sallis, J. F., Owen, N., & De Bourdeaudhuij, I. (2010). Neighborhood SES and walkability are related to physical activity behavior in Belgian adults. Preventive medicine, 50, S74-S79.

van Heeswijck, T., Paquet, C., Kestens, Y., Thierry, B., Morency, C., & Daniel, M. (2015). Differences in associations between active transportation and built environmental exposures when expressed using different components of individual activity spaces. *Health & place*, *33*, 195-202.

"Walk Score Methodology". (2018) https://www.walkscore.com/methodology.shtml

Welch, T. F., Gehrke, S. R., & Wang, F. (2016). Long-term impact of network access to bike facilities and public transit stations on housing sales prices in Portland, Oregon. *Journal of Transport Geography*, *54*, 264-272.

Windmiller, S., Hennessy, T., & Watkins, K. (2014). Accessibility of Communication Technology and the Rider Experience: Case Study of Saint Louis, Missouri, Metro. *Transportation Research Record: Journal of the Transportation Research Board*, (2415), 118-126.

Winters, M., Teschke, K., Brauer, M., & Fuller, D. (2016). Bike Score®: Associations between urban bikeability and cycling behavior in 24 cities. *International journal of behavioral nutrition and physical activity*, *13*(1), 1.

Winters, M., Teschke, K., Grant, M., Setton, E., & Brauer, M. (2010). How far out of the way will we travel? Built environment influences on route selection for bicycle and car travel. *Transportation Research Record: Journal of the Transportation Research Board*, (2190), 1-10.

Winters, M., Brauer, M., Setton, E. M., & Teschke, K. (2013). Mapping bikeability: a spatial tool to support sustainable travel. *Environment and Planning B: Planning and Design*, *40*(5), 865-883.

Woodcock, J., Edwards, P., Tonne, C., Armstrong, B. G., Ashiru, O., Banister, D., & Franco, O. H. (2009). Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *The Lancet*, *374*(9705), 1930-1943.

Yamada, I., & Thill, J. C. (2010). 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.

APPENDIX A

DIAGNOSTIC PLOTS FOR HEDONIC PRICING ANALYSIS

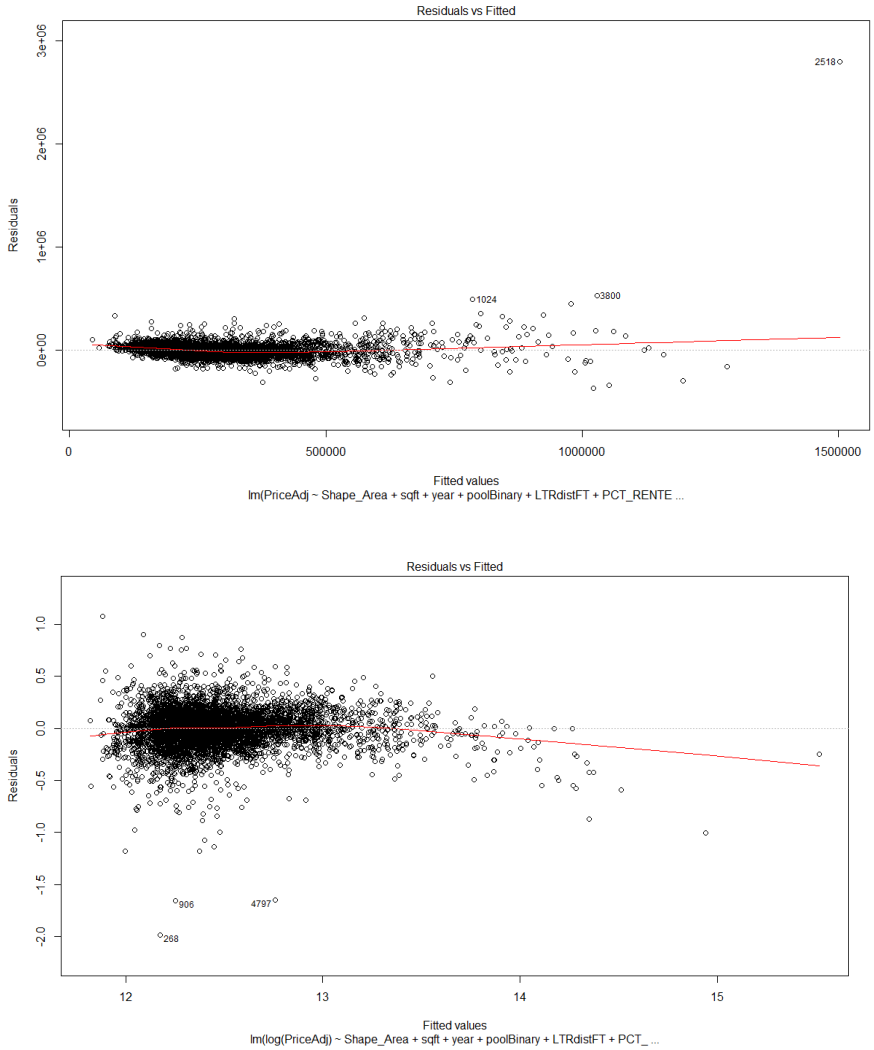

Figure 4. Residuals vs. Fitted plot for model 1(top) and model 2 (bottom) showing homoscedasticity

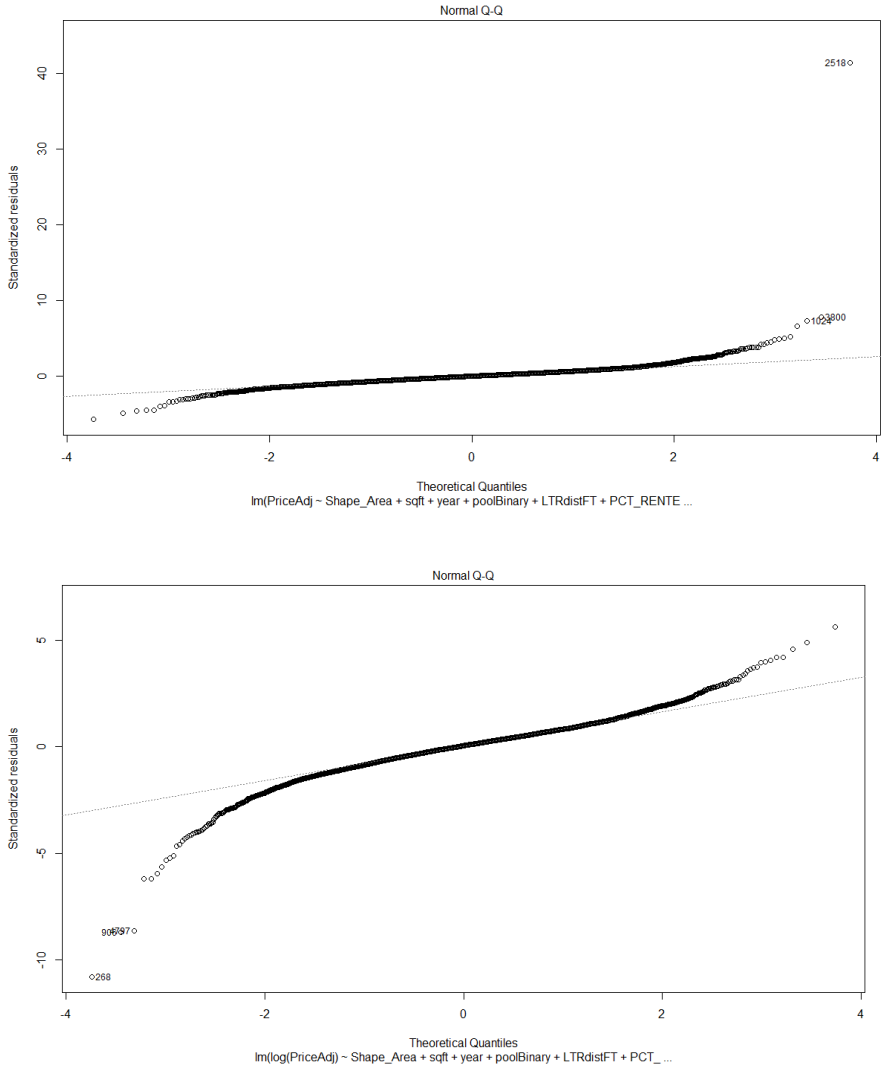

Figure 5. QQ plots showing normal residuals with extreme values for models 1 (top) and 2 (bottom)

Chapter 2 reproduced from a published paper with permission from all authors.