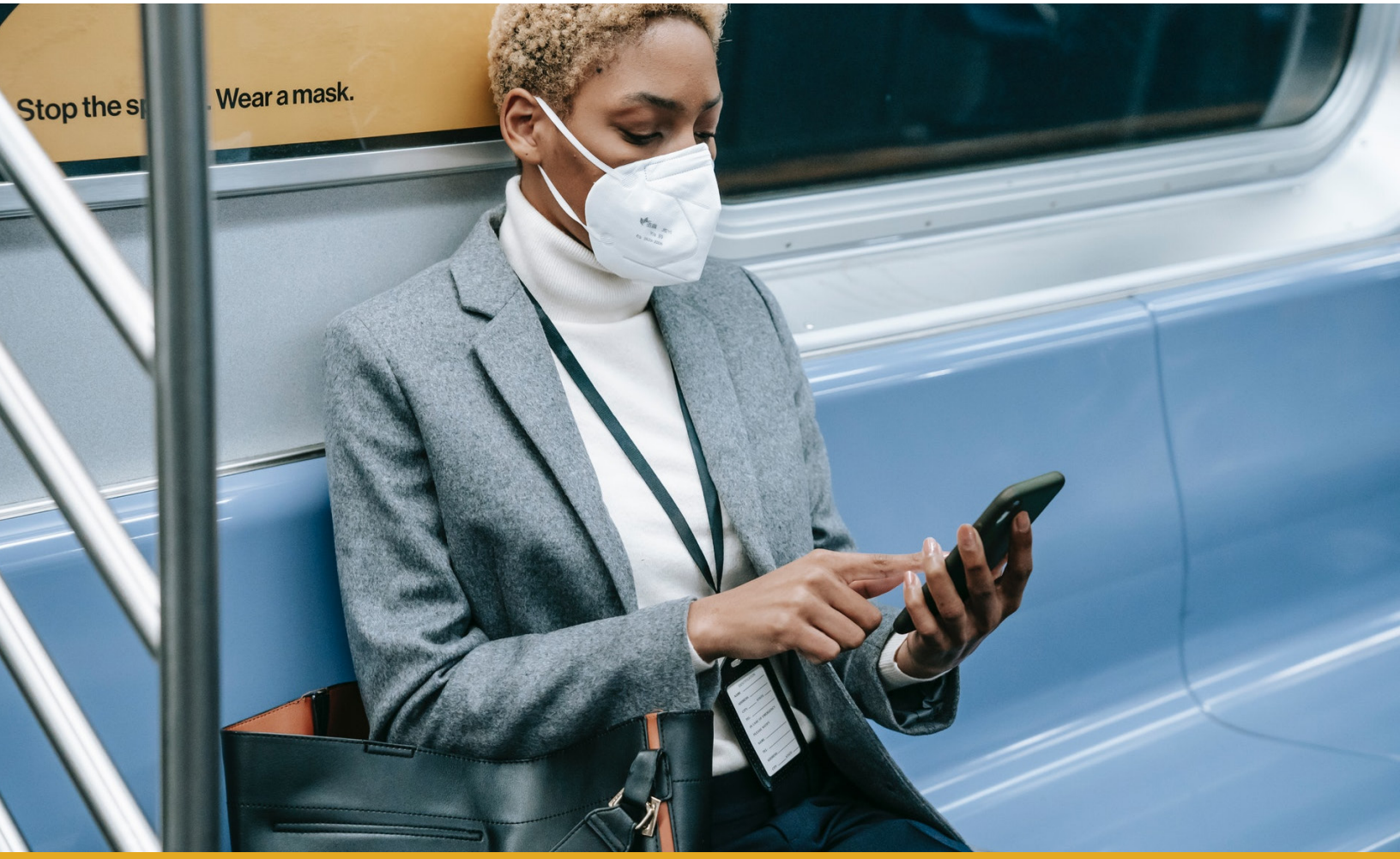
SJSU SAN JOSÉ STATE UNIVERSITY

MTI MINETA TRANSPORTATION INSTITUTE

Comparing Twitter and LODES Data for Detecting Commuter Mobility Patterns

Jochen Albrecht, PhD, GISP
Andreas Petutschnig
Laxmi Ramasubramanian, PhD, AICP
Bernd Resch, PhD
Aleisha Wright



MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Transportation Mobility (MCTM) funded by the U.S. Department of Transportation and the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce Development

To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 21-11

Comparing Twitter and LODES Data for Detecting Commuter Mobility Patterns

Jochen Albrecht, PhD, GISP
Andreas Petutschnig
Laxmi Ramasubramanian, PhD, AICP
Bernd Resch, PhD
Aleisha Wright

May 2021

A publication of
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 21-11	**2. Government Accession No.**	**3. Recipient's Catalog No.**	
4. Title and Subtitle Comparing Twitter and LODES Data for Detecting Commuter Mobility Patterns		**5. Report Date** April 2021 **6. Performing Organization Code**	
7. Authors Jochen Albrecht, PhD, GISP, Andreas Petutschnig, Laxmi Ramasubramanian, PhD, AICP, Bernd Resch, PhD, and Aleisha Wright		**8. Performing Organization Report** CA-MTI-2037	
9. Performing Organization Name and Address Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		**10. Work Unit No.** **11. Contract or Grant No.** 69A3551747127	
12. Sponsoring Agency Name and Address State of California SB1 2017/2018 Trustees of the California State University Sponsored Programs Administration 401 Golden Shore, 5th Long Beach, CA 90802 U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology University Transportation Centers Program 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report **14. Sponsoring Agency Code**	
15. Supplemental Notes DOI: 10.31979/mti.2021.2037			
16. Abstract Local and regional planners struggle to keep up with rapid changes in mobility patterns. This exploratory research is framed with the overarching goal of asking if and how geo-social network data (GSND), in this case, Twitter data, can be used to understand and explain commuting and non-commuting travel patterns. The research project set out to determine whether GSND may be used to augment US Census LODES data beyond commuting trips and whether it may serve as a short-term substitute for commuting trips. It turns out that the reverse is true and the common practice of employing LODES data to extrapolate to overall traffic demand is indeed justified. This means that expensive and rarely comprehensive surveys are now only needed to capture trip purposes. Regardless of trip purpose (e.g., shopping, regular recreational activities, dropping kids at school), the LODES data is an excellent predictor of overall road segment loads.			
17. Key Words Planning Activities leading to information generation Communication Interdisciplinary studies Methodology	**18. Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 41	**22. Price**

DOI: 10.31979/mti.2021.2037

Mineta Transportation Institute College of Business
San José State University San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924- 7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2037

ACKNOWLEDGMENTS

The authors would like to thank our anonymous reviewers. A journal article based on our research has been submitted to the journal *Transportation Research Part C, Emerging Technologies*. We responded to reviewer comments and have resubmitted it on March 1st, 2021. It is currently under review.

CONTENTS

List of Tables

List of Figures

Executive Summary

Local and regional planners struggle to keep up with rapid changes in mobility patterns. This exploratory research is framed with the overarching goal of asking if and how geo-social network data (GSND), in this case, Twitter data, can be used to understand and explain commuting and non-commuting travel patterns. Statistics capturing human mobility are expensive to obtain and deteriorate quickly as existing mobility patterns change and new ones emerge. Planners have been relying on US Census LODES data, which explicitly captures only commuting trips, and seems unsatisfying because only some 16.6% of all vehicle trips are work-related (FHWA 2017). GSND potentially offers a solution, as data derived from repeat origin-destination pairs of the same Twitter ID indicate trips regardless of purpose. We set out to answer the following research questions:

1. Is it possible to extract travel flow patterns in the Bay Area from GSND and if so, how can this be done efficiently?
2. To what degree do commuter flow patterns identified in GSND correlate with official LODES commuting data?
3. Can GSND be used to explain non-commuting trips?

Approximately 33 million geo-referenced Bay Area tweets were harvested for the study period from 2010 until early 2020. They were filtered by repeat occurrences of origin/ destination (O/D) pairs and categorized by time of day and day of week. Each of these pairs, as well as all LODES O/D's were then routed as shortest paths on the Open Street Maps network of roads. This study is limited to road trips only; further research should apply routing procedures for transit trips as well. For the GSND, we attributed trip purpose by the dominant land use in the O/D census blocks.

We then compared the road segment loads of the two input data sets and found not only incredible high rates of correlation but also nearly complete spatial randomness among their differences, which suggests that the findings below are scale-independent and applicable in all parts of the study region. Twitter's 2015 geolocation policy change resulted in a dramatic reduction of available GSND. Since then, smaller temporal samples have shown to be a poor predictor of local traffic loads.

GSND are suitable to capture the over 80% of non-commuting trips that keep our roads busy. GSND *is not* suitable for characterizing real-time or short-term commuting patterns but is complementary to exiting commuting data. Translated into road segment loads, GSND and LODES data are virtually indistinguishable, which means that LODES data are an excellent substitute for overall transportation demand.

The research project set out to determine whether GSND may be used to augment LODES data beyond commuting trips and whether it may serve as a short-term substitute for commuting trips. It turns out that the reverse is true and the common practice of employing LODES data to extrapolate to overall traffic demand is indeed justified. This means that expensive and rarely comprehensive surveys are now only needed to capture trip purposes. Regardless of trip purpose (e.g., shopping, regular recreational activities, dropping kids at school), the LODES data is an excellent predictor of overall road segment loads.

Keywords: Urban Planning, Commuter Mobility, Twitter Mobility, Collective Movement

I. Introduction

Historical land-use and development patterns, coupled with federal, state, and local policies, have resulted in sprawling metropolitan regions and severe imbalances between jobs and housing in many US metropolitan areas (Ihlanfeldt and Sjoquist, 1998). The results are apparent in every major American metropolitan area: traffic nightmares, long commute times, and rising housing costs. In 2019, before the pandemic, drivers in the San Francisco Bay urban area lost an average of 47 hours over the year, just sitting in traffic, earning the region an unenviable 7th place in the US congestion rankings (Inrix, 2020). Good transportation planning strives to increase people's mobility by reducing the friction of distance (Rodrigue, 2020). Policymakers have streamlined and simplified the complexities of travel behavior by focusing on work commutes because commuting to and from work remains one of the primary reasons why people travel, even as scholarly research has consistently acknowledged the influence and importance of non-work trips (Giuliano and Small, 1993; Kockelman, 1997). Transportation planning is a both data-hungry and resource-intensive endeavor. This research is a pilot study to investigate two related ideas – first, if and whether geo-social network data (in our case geo-tagged Tweets) can be used as a reliable and relatively affordable data source to provide information about travel patterns for planning purposes and second, to assess the extent to which this can data can be useful in understanding and explaining non-commuting travel patterns. The reason to focus on non-commuting travel is that most "official" census data focuses on the journey to work.

This study focuses on the San Francisco Bay Area, which includes the nine counties shown in Figure 1: Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma, all of which form the San Francisco Bay region. The area is home to prestigious universities and high-tech industries, and it also offers natural beauty and cultural diversity, making it an attractive destination. The area has experienced rapid population growth over the past two decades and currently houses over 7.75 million people across 101 municipalities in 2020, with a projected additional 1.1 million jobs and 2.1 million people anticipated by 2040 (Mackenzie et al., 2017). The subsequent rising demand in housing, together with topographic and regulatory constraints, has led to significant land-use changes throughout the area and has also exacerbated congestion and related environmental concerns (Cervero, 1996; Cervero and Duncan, 2006). The historic settlement patterns in and around the region and the highly individual growth trajectories of each county have resulted in imbalances between jobs and housing (Chapple and Zuk, 2015), which, in turn, has caused long commutes and an overloaded traffic system (Nguyen and Stivers, 2012). The Bay Area's geography and its bridge crossings create bottlenecks which cause further commute delays.

Figure 1. Study Area Overview

Local and regional planners struggle to keep up with the rapid changes in mobility patterns: statistics capturing human mobility are expensive to obtain, and they deteriorate quickly as existing mobility patterns change and new ones emerge. The currently most detailed example of such statistics is the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) (online, 2019a) published by the US Census Bureau. LODES tables describe commuter flows on the census block level for the entire country. Because of their high quality, spatial granularity, and coverage, they are an important tool for informed decision making in city and regional planning. One significant drawback of the LODES data is that they pertain to commuting only: i.e., they do not cover any other types of trips such as leisure trips or shopping. Yet, based on data from the National Household Travel Survey conducted by the Federal Highway Administration (FHWA, online, 2017), only 16.6% of all vehicle trips are work-

related. Nevertheless, LODES data have been used as a stand-in for all forms of mobility in the US.

In this study, we use geo-social network data ("tweets") from the social media platform Twitter to identify mobility patterns, which we correlate with LODES data. Tweets can be obtained quasi-continuously via an application programming interface (online, 2020d). Each tweet used in this study is linked to a single location and timestamp, which means that the data can be aggregated spatially and temporally at any required level. We derive weighted connectivity information from the tweets by counting the number of connections between regularly visited regions, which we refer to as *flows*. Throughout this paper, we refer to flows derived from Twitter data as Twitter flows. Their data structure is identical to that of the LODES data, which also represent (commuter) flows. The research is framed with the overarching goal of asking if and how geo-social network data can be used to understand and explain commuting and non-commuting travel patterns. To accomplish this goal, we ask the following four research questions:

1. Is it possible to extract travel flow patterns from geo-social network data and if so, how can this be done efficiently?

2. Focusing on commuting trips, to what degree do commuter flow patterns identified in geo-social network data correlate with official LODES commuting data?

3. At which spatial level (scale) can commuter flows extracted from geo-social network data most accurately match official LODES commuting flows?

4. Can geo-social network data be used to explain non-commuting trips?

To answer these research questions, we compare LODES and Twitter flows using two approaches. First, we explore the flows' spatiotemporal characteristics, such as changes in flow magnitude and distribution over time, flow connectivity of aggregated regions, and how cyclical temporal phenomena such as seasons or time-of-day impact the flows. We also integrate parcel-level land-use data to determine which pairs of land-use classes are connected by flows. Second, we use a region-based approach to correlate origin-destination (OD) data to assess the association between Twitter and LODES flows on different spatial scales. Additionally, we map the flows onto a street graph to compare the flows on the individual street segments for the two data sets.

1.1 Related Work

Twitter and other geo-social network data sources have been used in numerous human mobility studies focusing on the detection of events and traffic disruptions (Steiger et al., 2016) or detection and visualization of mobility patterns (Gao, 2015), as well as a variety of other applications in the context of mobility, urban activity, or urban planning (Martí et al., 2019). In a county-level study in the New York City area, researchers used Twitter data to estimate human activity and mobility

patterns based on about 6.5 million tweets collected over six months. They concluded that Twitter data provide a suitable basis for modeling human mobility (Kurkcu et al., 2016). Similarly, a study based in Madrid shows that Twitter data can be used to model commuter mobility by identifying users' home and work locations based on their temporal usage patterns (Osorio-Arjona and García-Palomares, 2019). Furthermore, an Australian study uses Twitter data in combination with call detail records (CDR) to provide evidence that Twitter data are a suitable proxy for human mobility (Jurdak et al., 2015).

The problem of identifying work-related mobility and activity in Twitter data has been addressed using semantic text analysis and spatial autocorrelation methods in temporal bins (Steiger et al., 2015). Using geo-social network data in urban planning not only allows us to conduct traditional studies with alternative data sources, but the high temporal granularity of the data also enables studies on virtually any temporal scale (Batty, 2013).

When interpreting the results of this study, it is critical to take into account the potentially skewed results caused by the temporally and spatially heterogeneous nature of geo-social network data (Li et al., 2013; Zhang and Zhu, 2018). Different people tweet for different purposes and at different occasions and only few tweet continuously throughout the day. In larger geographic areas these difference balance each other out but for smaller area studies like around sports stadiums or in a bar district, the data would be skewed. It is one of the features of this study that it allows for both geographic and temporal differentiation. The scale question alluded to as aim #3 of the study is important to keep in mind when interpreting the results of this study.

One of our research questions pertains to the amount of traffic generated by non-commute trips, which is widely unknown due to the costs involved in capturing this kind of data. We are aware of only a handful of related work, including a Walnut Creek (CA)-based study (online, 2020g) and a small national study (Convery and Williams, 2019). In a Chicago study, social media data were used to develop a gravity model using a classification of destination points of interest (Yang et al., 2015). Similarly, a study based in New York City describes training a neural network model with Twitter data to augment a traditional gravity model (Pourebrahim et al., 2018). Social network geo-data have also been used to model travel demand (Lee et al., 2016). Probably closest to the work presented here is work that estimates local commuting patterns from geolocated Twitter data based on frequently visited locations: the paper documents similar success rates for using Twitter data to estimate census-based commuting data but it does not deal with non-work-related trips (McNeill et al., 2017).

Historically, transportation planners have relied on commuting data to extrapolate to non-commuting trips and everything that flows from that. Officially commuting data arrives with significant delays. This report investigates if geo-social network data can fill in the gaps to add robustness to conclusions about both commuting and non-commuting trips. It sheds light into the under-researched area of non-commuting trips by presenting a quantitative and replicable approach.

II. Data Description and Processing

The LODES data were downloaded from the US Census website as plain text files that represent a sparse matrix in long-form format. For the 710,485 census blocks in California, the file lists 15,327,971 work block to home block flows with the number of jobs at the origin block of each of those flows (16,566,140 jobs for all of California). The LODES data do not specify the modal split, i.e., the means of transport between the blocks. The closest equivalent for that is provided as a regular census table at the census tract level (online, 2019b). For our nine-county study area, we worked with 2,972,821 flows between 109,228 census blocks representing 3,252,286 individual commutes. The original Twitter data consists of 44,812,476 tweets posted between October 8, 2010, and April 19, 2020. All observations are within the study area. Each includes a timestamp, a message (text), and a geographic point defined through a pair of coordinates. In this study, only tweets written manually by humans are of interest, which is why we removed the remaining tweets by identifying user accounts that were used to post tweets with unusually high frequencies and accounts featuring little editing distance[1] among their tweets like those seen from advertisements or public weather stations (Petutschnig et al., 2020). Filtering based on those criteria reduced the dataset to 33,755,914 tweets. For temporal data clustering, we considered the time of day rather than the absolute timestamps of the tweets. This way, we were able to cluster tweets that were sent at roughly the same time of day, even if there were long periods of inactivity between them.

The tweets include geographic point coordinates. We projected these coordinates onto a Cartesian coordinate system so we could assume consistent, metric distances between all points for the spatial clustering process. We used the outlines of the census blocks, census tracts, and counties obtained from the United States Census Bureau (online, 2019a). Table 1 shows some descriptive statistics of the study area and Twitter data. We see a high variation in the number of tweets and areas of the tracts and blocks on the census level, as opposed to the county level.

Table 1: Study Area Description.

	County	*Census Tract*	*Census Block*
Number of regions	9	1,584	109,228
Mean area [km^2]	2,357.7	13,4	0.2
Median area [km^2]	2,126.5	1.6	0.02
Mean number of Tweets	3,750,657	21,310	352
Median number of Tweets	3,013,390	15,503	54

An additional dimension of the flows is the trip purpose. To understand possible motives for a particular trip, we identified the land-use in the destination block using parcel-level land-use

[1] 'Editing distance' is a similarity measure. For example, a weather station would send out a tweet every hour that is identical to the previous one (all the same words in the same sequence) with only some numerical values changing gradually. A human being would not do that. This way, non-human generated tweets can be filtered.

designation in a commercially available dataset from Boundary Solutions (online, 2020b). Most census blocks are made up of multiple parcels. To integrate the land-use data into the analysis, we aggregated the parcel-level data to the census blocks that contain the respective parcels. We did so by assigning one land-use class to each census block. If a census block intersected with a single parcel, we assigned that parcel's land-use class to it. If a census block consisted of multiple parcels, we assigned it the land-use class of the parcels covering the largest part of its area.

Some experiments in this study required a street network graph, for which we used graph data provided by OpenStreetMap (online, 2020a). For the network dataset, we extracted all drivable public streets in a 250-km radius around the center of our study area, thus extending it substantially. We dimensioned the graph so liberally to capture routing results even if they include street segments not located within the study area, preferring this to omitting these partial routes from the analysis. We used Dijkstra's least-cost path algorithm for routing, which is implemented in the module NetworkX (Hagberg et al., 2008). The street graph weighting scheme is aimed at car travel, which is a simplifying assumption, given that not all workers commute by car.

Open Street Map (OSM) data has been shown to be highly reliable (Haklay 2010, Zheng & Zheng 2014, Anahid et al. 2016), and this is especially true for the study area, which features an unusually large number of well-trained contributors. whose reliability has been described by an armful of literature. Based on random checks, the commercial land use data seems to be reasonably reliable and its reliability only increases by its aggregation to census blocks. We also compared it with land use information derived from OSM points of interest and found them matching well. The only cause for concern is mixed land uses, where buildings have commercial land use on the ground floor and residential land use on upper floors. In those cases, we erred on the side of commercial uses.

The technologies used in this study are PostgreSQL (online, 2020d) databases with the spatial extension PostGIS (online, 2020c) for data storage and analysis, the programming environments R (online, 2020i) and Python (online, 2020e) for analysis and visualization, and the geographic information system QGIS (online, 2020f) for visualization.

III. Study Methods

The overall methodological workflow of the research presented in this paper is illustrated in Figure 2. Starting from the upper left, the flowchart depicts the preprocessing and analysis steps applied to the raw Twitter and LODES data to produce two comparable OD matrices. The right box shows how the street network and the land-use data were integrated to produce intermediate outputs (flow graphs and trip purpose datasets). The box on the bottom shows how these intermediate outputs and origin-destination (OD) data feed into the final analytical steps, which final outputs they produce (correlations and visualizations), and which research questions the results refer to.

Because we use *flows* to model movement in this study, we focus on regularly occurring trips rather than one-off visits to a chance location. To single out candidate locations that are part of such regular trips, we use the DBSCAN clustering algorithm (Ester et al., 1996) to identify the spatiotemporal tweet clusters of each user. DBSCAN requires two parameters: *minpts* for the minimum number of points per cluster and ε to limit the search radius in the clustering process. We chose *minpts* = 5, ε_s = 100m for the spatial clustering process and ε_t = 30min for temporal clustering. We chose the values for *minpts* based on manual inspection of the data and $_s$ and $_t$ by identifying visible "elbows" in k-distance-graphs (Schubert et al., 2017).

The DBSCAN algorithm detects clusters based on density. In our case, a set of tweets is considered a dense cluster if at least five of them meet the requirements specified through the parametrization. A cluster has no defined maximum extent, instead it is restricted by the data distribution. There is also no requirement to specify the number of expected clusters for the algorithm. This is an advantage, as we do not need to make assumptions about the number of regularly visited locations of Twitter users.

We only consider movements between such candidate locations, therefore a user to be considered, they must have produced enough tweets meeting our cluster requirements in at least two locations. The trip detection assumes that, occasionally, users send consequitive Tweets from two separate candidate locations. Assuming the user has consistent usage behavior for their pairs of candidate locations, their movement patterns emerge over time.

To identify individual trajectories describing the trips of a single user, we need to first establish what we consider a trip. We defined a trip as the movement between two census blocks if they contain a user's cluster centroids and the user travelled between the blocks within three hours. We consider this time frame the maximum duration it takes to move from one corner of the study area to the furthest other point under reasonable traffic conditions. The reason for this time cap is that we focus on trips made directly (without longer interim stops) between two regions. If there is a longer time span between origin and destination, there may be intermediate stops in between,

which does not comply with our definition of a direct trip in this research. The connections are directional, based on the chronological order in which the blocks were visited.

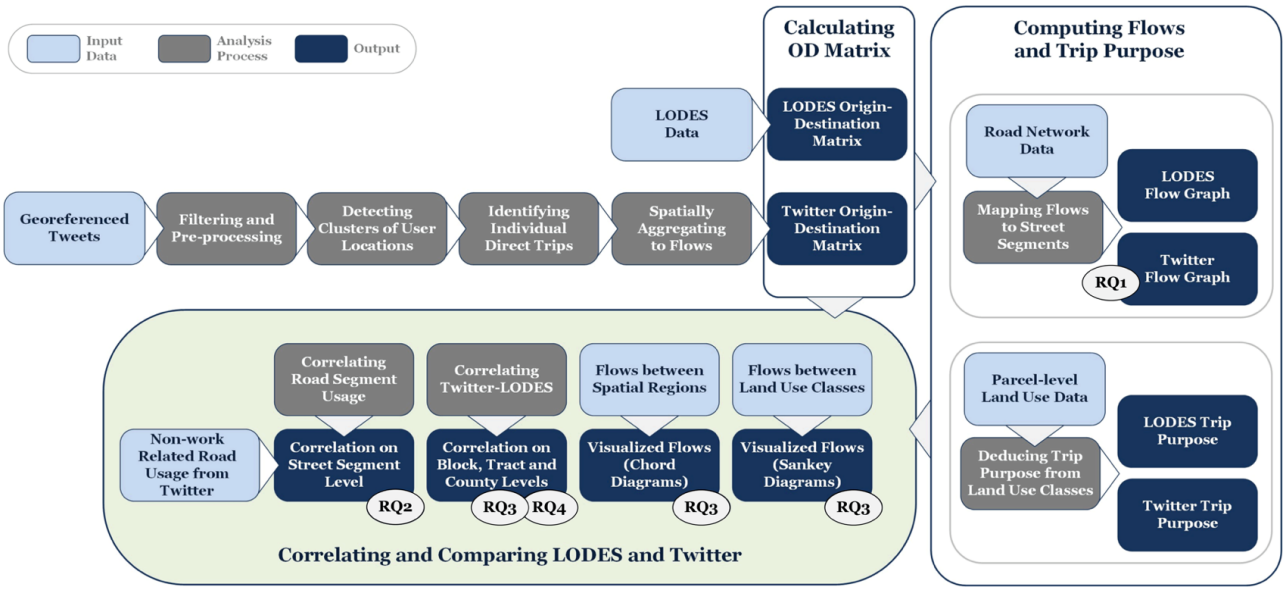


Figure 2. Schematic Workflow Illustrating Input Data, Analytical Steps, and Outputs

By summing up users' connections grouped by blocks, we generate an adjacency matrix which represents the number of connections between all pairs of blocks. We refer to the aggregated connections of regularly frequented locations as flows. The structures of the Twitter and LODES adjacency matrices are identical, allowing us to compare them. By aggregating the data to this level, we also obscure individual users' data. This effect is desirable because it constitutes a vital privacy protection measure complying with best-practice recommendations (Kounadi and Resch, 2018; Kounadi et al., 2018).

To explore whether the Twitter flow data exhibit different patterns during and outside of traffic-heavy times, and to account for longer-term trends and spatial scale differences, we partitioned the Twitter data spatially and temporally. To observe longer-term trends, we split the data into chunks of two years, because two-year timeframes are long enough to be robust against outlier years but still short enough to observe trends developing over longer time periods. The cyclical temporal phenomena of time of day and day of week govern day-to-day activity like working hours and therefore human mobility: to account for these daily and weekly variations, we partitioned the data by whether connections were made during the weekend (Saturday or Sunday) and whether they were made during typical Bay Area rush hour times (6:00–8:00 AM or 3:00–5:00 PM).

In terms of spatial aggregation, we partitioned the flow data on the census block level, but we also require the data on census tract and county levels for some of our experiments and visual outputs. The three scale levels are hierarchically organized administrative divisions where *census block* ⊂ *census tract* ⊂ *county*, which makes aggregation simple. We aggregated the data by tallying the flows based on travelers' administrative division membership.

Each datum of the LODES and Twitter OD pairs represents aggregated movement between a pair of census blocks. We compare the LODES pairs with the Twitter flow data and quantify differences between the two datasets. We use correlation coefficients to quantify the relations between LODES and Twitter connections. Because the data deviate substantially from a normal distribution, we chose the nonparametric Spearman's rank correlation coefficient ρ as a measure of association. For every OD pair of two regions, we have two flow counts (for tweets and LODES). This allows us to draw a direct comparison. To show the impact of the spatial scale on the correlation, we performed the calculations on the census block, census tract, and county levels.

We used block-level land-use data to identify the land-use class pairs associated with each LODES and Twitter flow. By aggregating these data by land-use class, flow size, and direction of travel, we can quantify the share of land-use class pairs for the two datasets. As with regional flow aggregations, we can aggregate flows by summing up all flows belonging to the same pair of land-use classes.

LODES data solely contain work trips, whereas Twitter flows contain other trip purposes as well, such as leisure trips or other non-commuting-related mobility. We hypothesized that the trip purpose is reflected in the connected land-use classes of the trips and that we should, therefore, be able to identify differences between the datasets.

With the exception of the city of San Francisco, the typical mode of transportation for commuters in the nine-county region is automobile transportation (McKenzie, 2015). It is therefore appropriate to identify travel routes between origins and destinations by mapping them to the least-cost path on the street network weighted for car usage to obtain a model that resembles the actual road usage patterns. For routing between pairs of census blocks, we had to specify exactly where on the street graph the routes begin and end. We chose the graph's node closest to the centroid of each block. For each pair of census blocks, the routing routine returns a set of edges of the street graph to represent street segments. Integrating the LODES and Twitter flow data, we calculated how often each individual street segment is used to accommodate the flows. The resulting graph permits direct comparisons between the LODES and Twitter data on each edge. To make the data sources comparable, we converted the results to standard scores to account for the scale differences between both datasets. Thereby, the comparison between LODES and Twitter flows becomes quantifiable for each street segment or area unit comprising a set of segments.

We calculated the correlation of the segment load of LODES and Twitter flows, which are shown in Figure 2. When looking at the whole study region and incorporating all Twitter-derived trajectories with the LODES-based ones, there is no statistically discernible difference between the two datasets (the correlation rates are perfect beyond the highest Z scores).

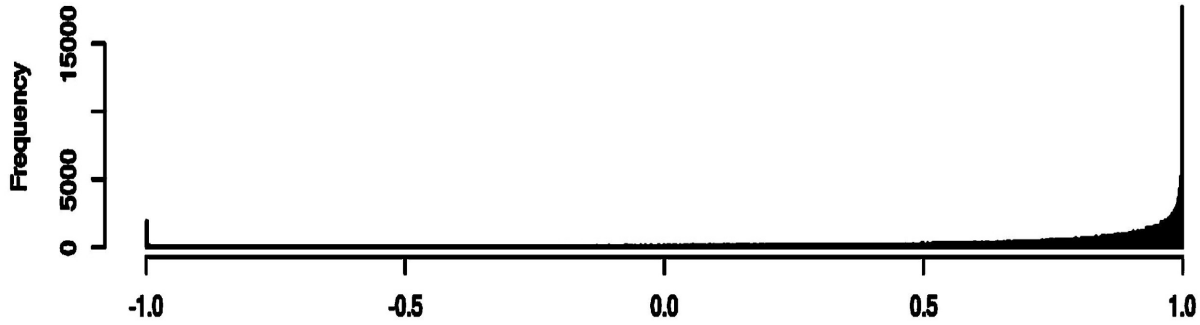


Figure 3. Frequency Distribution of r^2 Values for the Comparison of LODES to All Twitter Data

Twitter and LODES flows represent different facets of mobility, which is why we questioned the informative value of such high correlations and chose a different measure of comparison, creating travel demand surfaces for the LODES and Twitter data. The surfaces are the result of converting the midpoints of over 3.1 million street segments (median length: 92m) with their segment use value into irregularly distributed points. These points were then transformed into grids of cell size 100m (1ha).

The high correlation depicted in Figure 2 can be ascribed to the law of large numbers. The more interesting questions arise from comparing the LODES data with different subsets of the Twitter data such as only those trips that fall within rush hours, or smaller temporal windows to determine whether the Twitter data may be used to update LODES data and to capture behavior changes. Correlating LODES with Twitter-derived street segment usage, we ran spatial regression models on all possible combinations of the LODES data and the whole Twitter dataset as well as temporal subsets (see Tables 1 and 2).

Based on our results, a few conclusions may be drawn. Two-year temporal subsets are relatively poor substitutes for LODES-based street segment use. As one would expect, there is a higher correlation for those same two-year subsets predicting overall Twitter-based street segment usage. This, in turn, suggests that the Twitter data represent different populations compared to the commuters represented by the LODES data, which confirms our initial hypothesis.

Table 2: Explanatory Power of Different Twitter Data Subsets Predicting LODES Street Segment Loads

LODES Prediction	Spatial Error
All Twitter	0.766143
Outside of rush hour	0.770077
Rush hour	0.736184
Weekends	0.772134
Outside of rush hour 2018/19 only	0.425247
Rush hour 2018/19 only	0.320049
Weekends 2018/19 only	0.41174

Table 3: Explanatory Power of Subsets from the 2018/19 Twitter Data Predicting Twitter Street Segment Loads for the Entire Study Period

All Twitter Predictions 2018/19	Spatial Error
Outside of rush hour	0.602277
Rush hour	0.492426
Weekends	0.609862

Trip purposes may be discerned from a semantic analysis of a tweet's content, from the land-use type of the trip destination (sports venue, shopping center, etc.), and from the time of day and day of the week. Using the technique of mapping flows to street segments described above, we created maps of street segment loads during and outside of rush hours, on weekends, and in two-year windows that not only show the expected differences in travel patterns but also match the trip purposes derived from the analysis of destination land-uses.

IV. Findings

We performed some exploratory data analysis to get a sense of the regional differences between Twitter and LODES data. There are notable differences between the data sources. While Twitter flows occur more regionally—the vast majority of them happen within their census tract and county—over 40% of LODES connections cross county borders. Given that LODES data only include commuter flows, whereas Twitter flows include other trip purposes as well, such discrepancies are expected. There is a strong discrepancy between Twitter and LODES on the census tract level: 42.7% of Twitter flows happen within a tract, as opposed to only 3.4% of LODES flows. This suggests that Twitter flows represent shorter trips, which is supported by the descriptive statistics of trip lengths and estimated car travel times outlined in Table 3 (previous chapter).

Table 4: Trip Lengths on the OSM Network

	LODES	*Twitter*
Minimum	0.001 km / 0 min	0.008 km / 0 min
Median	12.660 km / 14 min	2.800 km / 3 min
Mean	22.538 km / 19 min	7.561 km / 6 min
Maximum	259.761 km / 178 min	366.315 km / 264 min

Figure 3 shows the magnitudes of (a) Twitter flows during rush hours, (b) Twitter flows outside of rush hours, and (c) LODES data at the county level. Each pair of counties is represented by an arrow that denotes direction and magnitude of flow. The color and shape of the arrows indicate origin and destination counties, whereas the base width indicates the flow magnitudes.

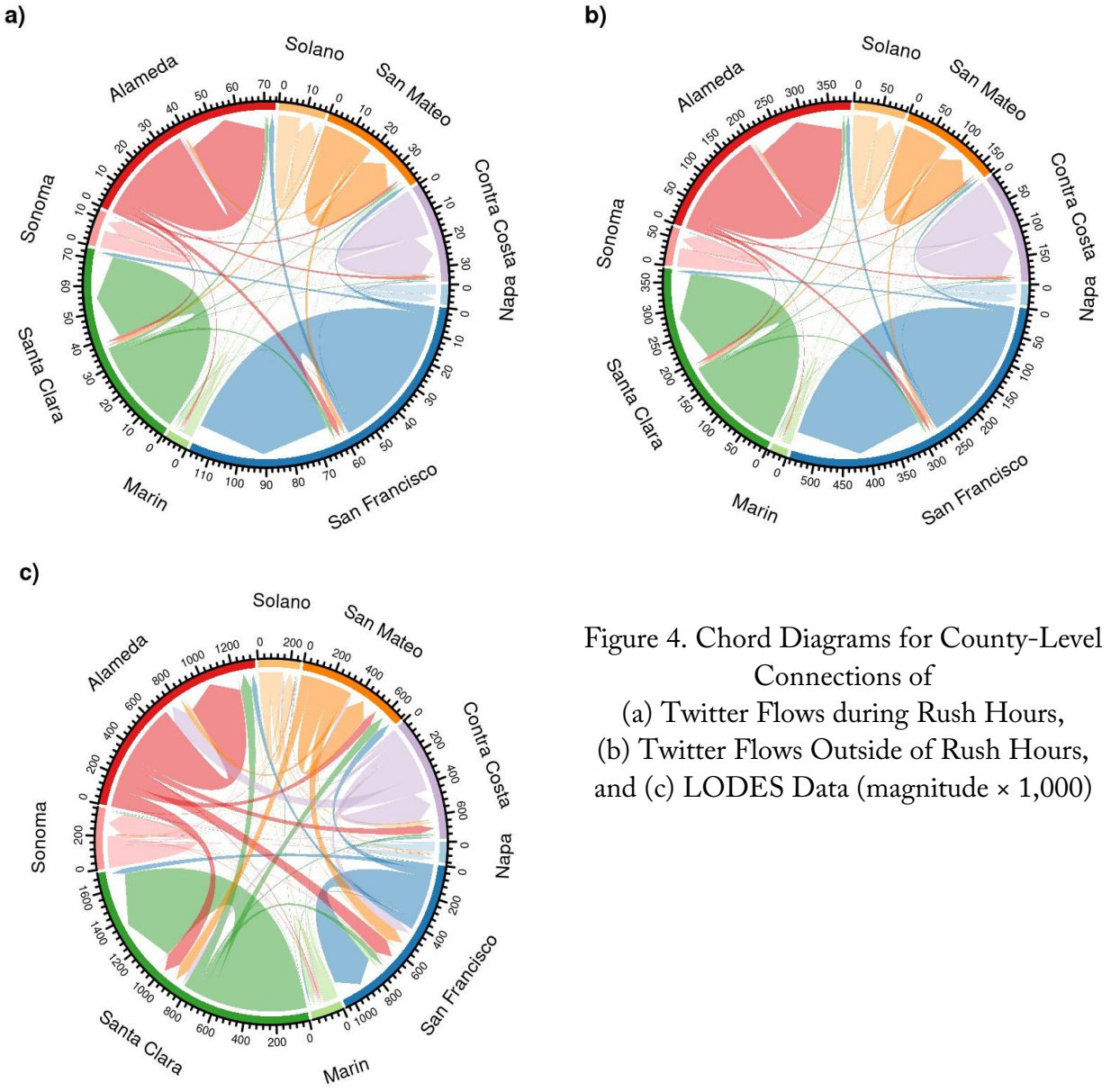


Figure 4. Chord Diagrams for County-Level Connections of
(a) Twitter Flows during Rush Hours,
(b) Twitter Flows Outside of Rush Hours,
and (c) LODES Data (magnitude × 1,000)

The numbers of connections show that the absolute flow magnitudes are higher for the LODES data. Besides that, the LODES data contains a lot more intercounty connections than the Twitter flows. For example, around half of the outgoing connections of Alameda County connect to other counties in the LODES data, whereas only about 15% of Twitter connections are outbound. In accordance with this effect, the LODES data reflect more inbound connections to the counties compared with Twitter. Another visible difference between Twitter and LODES flows is the difference in relative connections between individual counties. In particular, the City and County of San Francisco is represented more strongly in the Twitter flows compared to other counties.

Figure 4 shows which land-use classes are connected by the flow data based on (a) Twitter data during rush hours, (b) Twitter data outside of rush hours, and (c) LODES data. Even though the proportions of origin and destination land use and the size of flows in parts (a) and (b) are very similar especially compared to (c), there are noteworthy differences between them. During rush hours, there are fewer connections between residential areas and more connections between residential areas and work-related land-use classes. Also salient is the similar distribution of land-use classes for origin and destination areas in (a) and (b), in contrast to (c). This is because the LODES data underlying (c) have a clear separation of home and work location, whereas the Twitter data represent trips regardless of the functional context of their origin and destination regions.

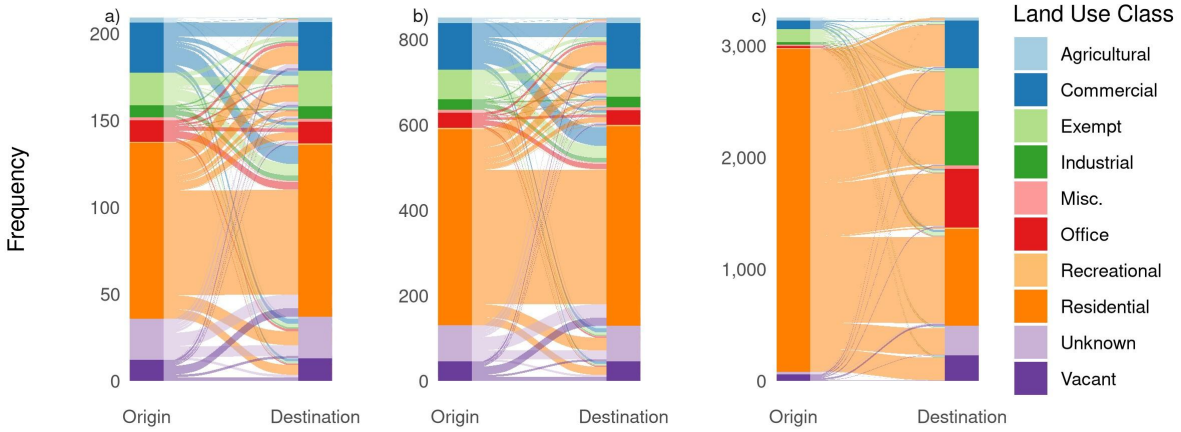


Figure 5. Sankey Diagrams of Land-Use Pairs from (a) Twitter Flows during Rush Hours, (b) Twitter Flows Outside of Rush Hours, and (c) LODES Data (magnitude × 1,000)

The results of the spatial error model in Table 2 show that the rush hour trips are a poorer predictor of LODES trips than the ones taking place outside of rush hour and during the weekends. Actual work-related trips seem to be less likely to be accompanied by tweets than non-commuting trips. This matches our observations that a significant number of Twitter trajectories have residential origins and destinations, which also fits the results of national surveys by the FHWA. Furthermore, we observe a relatively poor predictive capacity of the two-year subsets, although it is important to keep in mind that the one-hectare resolution is a fairly stringent constraint. We endeavor to continue our research towards determining the scale thresholds for such predictions.

Another aim of our research was to assess whether the finer temporal resolution of the Twitter data could be used to improve upon existing datasets such as the LODES datasets To answer this research question, we ran spatial regression models with a two-year subset (2018/19) to explore how well it predicts the much more voluminous Twitter data from previous years. As mentioned above, the constraint of basing predictions on one-hectare cells makes this goal fairly ambitious. However, as the results in Table 5 illustrate, even the much smaller 2018/19 dataset exhibits relatively high correlations, except for rush hour trips. Correlation rates of 0.6 are considered quite

good in the social sciences. The only outlier in this table is the value for rush hour tweets from our two-year window.

Table 5: Robustness of a Two-Year Subset of Twitter Data Predicting LODES Street Segment Usage

Temporal Subset	r^2
Weekends	0.6098
Outside of rush hour	0.6023
Rush hour	0.4924

Our original assumption was that the street segment usage during rush hours provided by our Twitter data should be most similar to the LODES data for the rush hour periods. Our correlation results show that this hypothesis could not be confirmed. Instead, it stands to reason that tweets are relatively rare immediately before or after a trip to work and that those trips completed during rush hour may actually represent other trips such as social calls or for the consumption of goods and services. While this observation is not specifically verified through this research, our claim is supported by the land-use class connectivity results, which contain a high percentage of residential-to-residential trips and residential to commercial trips, as well as the exceedingly short trip lengths, which differ significantly from the mean trip length of the LODES data.

The map in Figure 6 represents quantifiable evidence of trips that are not captured by the traditionally used LODES data. The map displays the areas of negative correlation (up to -0.13) between the LODES data and the Twitter weekend data, which emphasizes the differences between the two sources of movement data, as described in the previous section. The areas depicted show no spatial autocorrelation, i.e., they are randomly distributed. Virtually all of these areas are in residential areas (a small remainder are in remote areas) and do not match known points of interest such as shopping centers, sports venues, state parks, etc. For well-known traffic chokepoints, the street segment loads confirm our expectations set by the LODES data. The ubiquity of the flows beyond these notable points is a new source of information that has hitherto been unavailable. However, it is important to note that the way we selected the Twitter data excludes trips that do not show up repeatedly for the same individual and hence underestimates the amount of non-routine trips.

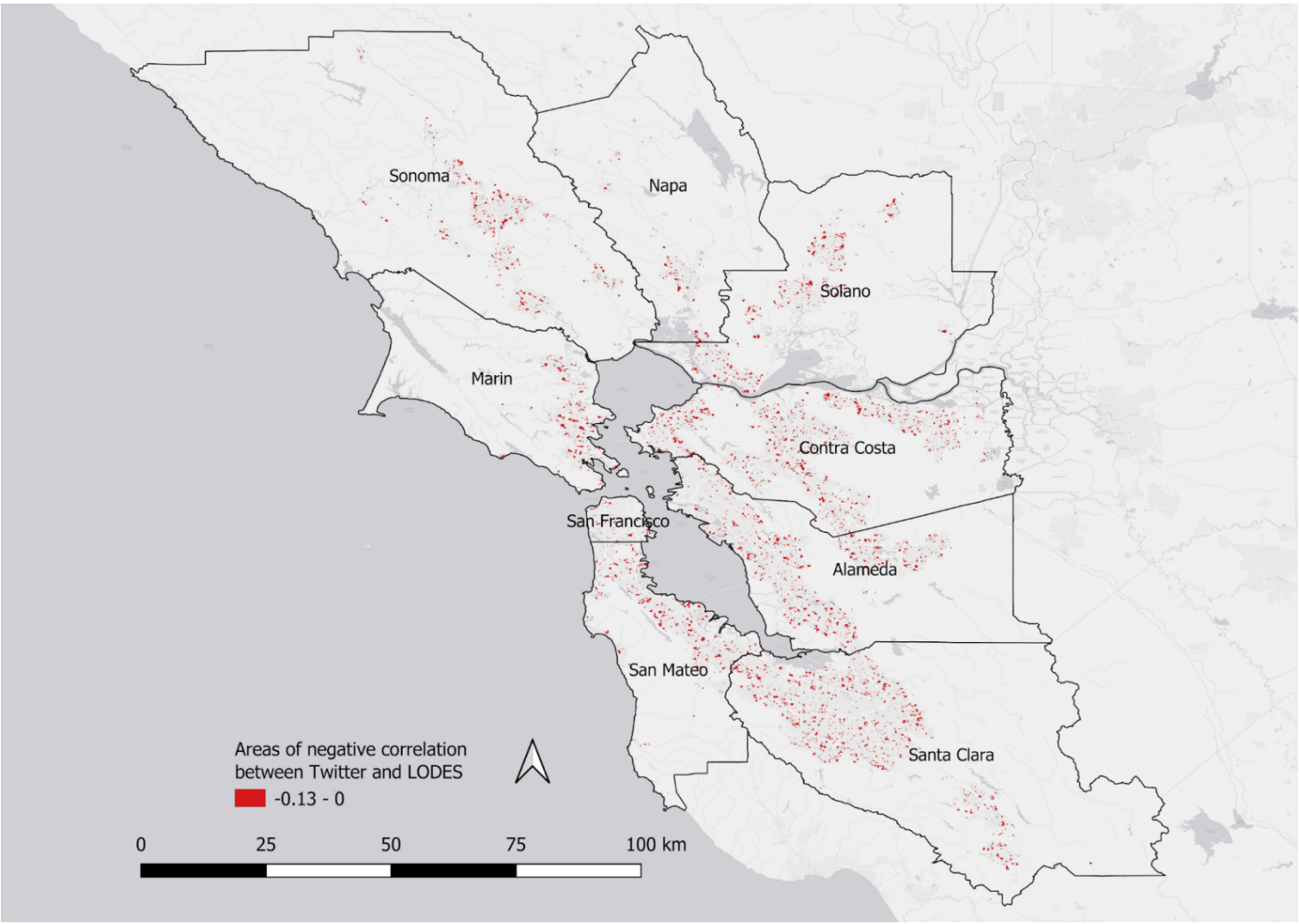


Figure 6. Areas with Negative r^2 for Twitter Data Predicting LODES Street Segment Use

Note to Figure 6. The areas in red are the only areas, where there is a significant difference for all the tweets that were filtered for the purpose of simulating LODES data. The total sum of these areas is small and not spatially autocorrelated.

V. Discussion and Limitations

To address research question 1 (to what degree do commuter flow patterns identified in geo-social network data correlate with official LODES commuting data?), we correlated the OD pairs of our data sources on different spatial scales. We found that they correlate strongly on a relatively small spatial scale. When mapping the flows to the street graph and considering the land-use classes associated with trips, we found indications that a large portion of Twitter flows are not direct work trips. Instead, they tend to be significantly shorter, even if they occur during rush hour times.

To address research question 2 (Can geo-social network data be used to augment flow data to include information about non-commuting trips?), we worked under the assumption that commuting is strongly tied to certain land-use classes, e.g., from residential areas to office spaces, and deviations from this can be attributed to extraprofessional travel. We identified two pieces of evidence in the land-use comparisons supporting our hypothesis. The comparisons of land-use classes associated with the LODES and Twitter flows suggest that there are substantial differences between the two datasets pointing to a strong difference in trip purpose. Secondly, the comparison of connected land-use classes in the Twitter flows during and outside of rush hour times suggests a relationship as well. Per our assumption, we would expect commuter travel to show a stronger linkage between residential and work-related areas than non-commuter travel, which the comparison of Figure 4 panels (a) and (b) confirms.

To address research question 3 (At which spatial level (scale) can commuter flows extracted from geo-social network data most accurately match official LODES commuting flows?), we again refer to the correlation coefficients of Twitter and LODES flows at different times and on different spatial scales. The correlation coefficients show that for county-level mobility, Twitter and LODES exhibit robust, high correlations. This is consistent with several nation-wide studies conducted at county-level resolution. At the scale of county-level analyses, Twitter data works very well as a substitute for LODES data as the two are effectively undistinguishable.

5.1 Discussion of Methods

Our definition of flows works from the premise that they should only include regions that a user visits repeatedly. We implemented this premise by only including regions in which we detected spatiotemporal clusters of tweets by a given user. The intention behind using this approach was to adequately represent routine travel behavior, which means that rarely visited locations are likely to be excluded from the analysis. This approach, however, relies on the assumption that frequent visits to a location result in frequent tweets. If a user tweets so rarely at a frequently visited location that we cannot detect a tweet cluster there, the location is falsely excluded from our analysis. For the region-based approach, we consider flow magnitudes, as well as the start and end regions encoded in the OD matrices. The direct comparisons of OD flows via correlation coefficients result in simple summary statistics. On one hand, these are compact and easily comparable, but on the other, they do not provide insights into the spatial characteristics of the results. Another

problem is the inherently binary perspective when comparing OD pairs: two OD pairs are either identical or not. In reality, each OD pair is a simplified representation of what is actually a route along a street network. Two OD pairs in close spatial proximity are likely to share some street segments in their routes even though they are not identical, and they are therefore a mismatch from a region-based perspective. This skews the correlation statistics towards low values, especially for large-scale target regions. The graph-based reasoning methods are better suited to capturing spatially similar but non-identical connections. This effect can be observed when comparing the region-based results in Figure 7 with the graph-based ones from Table 3. Even though the median street segment length of 92m is a finer scale than the census block level, the predictive power is significantly higher.

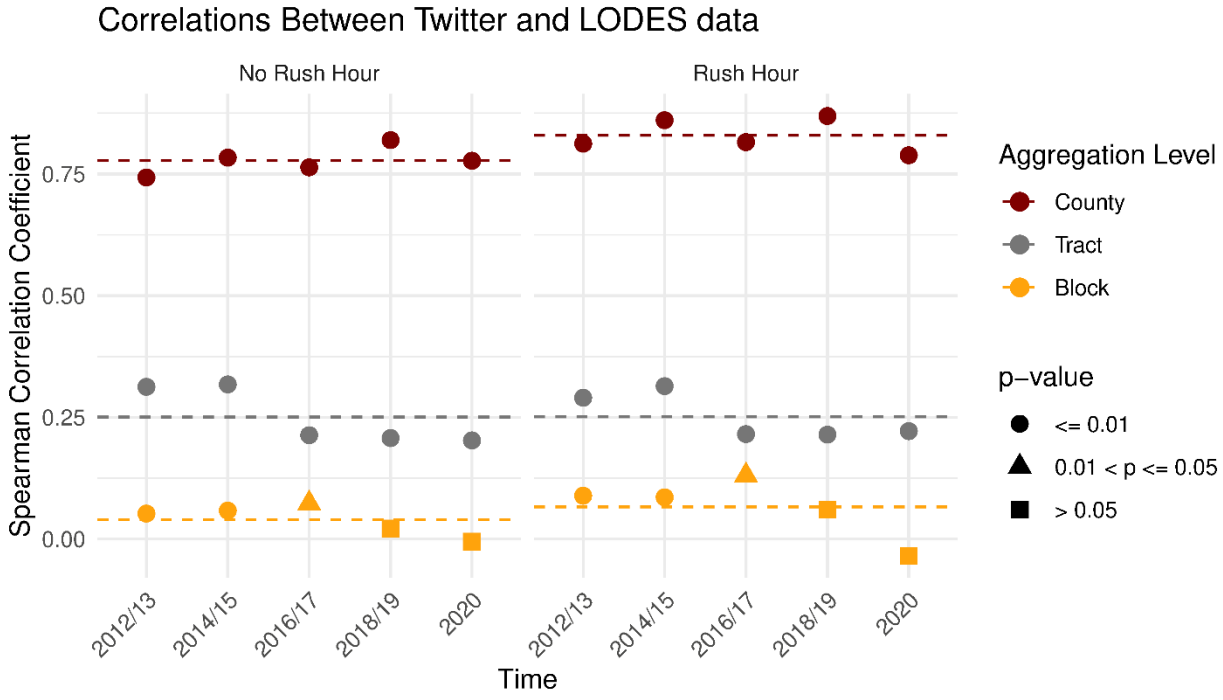


Figure 7. Flow data correlations as a function of scale.

Twitter usage is skewed by demographic and geographic context. Therefore, when deriving flow data from tweets, the population of some regions will be represented more than others. For example, there are residential areas with few active Twitter users but a large working population; there are also places with few permanent residents that attract large numbers of visitors like sports venues or shopping centers. Knowledge about such places should be taken into account when interpreting the results of this study. By clustering tweets for region selection in the flow data, we intended to capture regular travel patterns. However, this approach can lead to an underrepresentation of regions that generate large traffic volume from large amounts of visitors who, individually, visit only on rare occasions, such as sports stadiums or national parks. A study focusing on such regions would have to adjust the flow detection methodology or integrate points of interest as another data source.

While LODES exclusively covers commuter mobility, Twitter flows represent other travel purposes as well. In principle, the difference between the two datasets should pertain only to non-commuting mobility. However, because the commuter flows contained in the Twitter data are likely not represented equally across the pairs of regions, the result of comparison will contain a region-dependent error term that has to be addressed. The resulting error adds to the effect of skewed Twitter usage explained above. The same principle applies to the flows between land-use data classes. It is possible to scale the Twitter origin land-use class distributions to resemble the distributions from LODES and adjust the Twitter destination land-use classes accordingly: this would skew the distribution of Twitter flows towards the LODES data at the cost of introducing an additional error term. Commuter flows are a geographic phenomenon and as such spatially dependent, i.e., the effect varies regionally. We have hence two different spatial error models that combine. The development of such a combined model is beyond the scope of this research.

The clustering process identifying the candidate locations based on which we quantify the Twitter flows is biased towards regions that are visited multiple times by the same user. By doing so, the flow detection becomes robust against outliers caused by one-time trips to otherwise never visited regions. However, by filtering the data like this, we also eliminate locations that are visited only rarely by individuals but attract large numbers of people. Places like holiday destinations, national parks or sports venues potentially fall in this category.

Mobility data can provide intimate insights into people's lives. To protect Twitter users' privacy, we apply the principle of data economy throughout the entire workflow and only disclose results in which the spatial and temporal aggregation prevents the identification of individuals, following the "geo-privacy by design" guidelines by (Kounadi and Resch, 2018; Kounadi et al., 2018).

5.2 Discussion of Results and Relevance for Transportation Planning

This research acknowledges that transportation planning and policy requires long-range planning including the use of demographic forecasting and travel demand modeling to direct infrastructure investments. Researchers use existing data sources like the US Census Bureau's LEHD and its derivative data products such as LODES to support these analyses, but they do not capture non-commuting trips and are unsuitable for short-term fluctuations. For example, the 2020 commuting data will not be available to the researchers until at least 2022, limiting the ability of researchers to draw immediate and meaningful conclusions about the impact of the pandemic. While it is true that large firms and city agencies may rely on data gathered from a variety of sources such as remotely sensed imagery data or data from passive sensors, such data is inaccessible to the majority of transportation planning researchers and professionals.

In recent years, the prevalence of volunteered geographic information and similar data sources made available by commercial providers like SeeClickFix (online, 2020c), Waze (online, 2020e), (Plunz et al. 2019) have assisted planners and city managers to undertake just-in-time planning, usually by making modest adjustments in response to public requests for intervention. These interventions tend to work well in well-defined jurisdictional areas (within city limits, for example).

They also tend to be "reactive", more suited for transportation management, rather than future-oriented planning. In this research, we used geo-referenced tweets as a freely available data source to support and substantiate the data gathered through official sources such as the US Census Bureau. One of the limitations that we wish to note is that while Twitter data are indeed free to access, some technical expertise is necessary to scrape the data using published APIs. In addition, the company's policy changes enacted in 2015, made it easier for twitter users to tweet without sharing their actual geo-location, thereby reducing the volume of geo-referenced tweets available for analysis. In other words, unless a Twitter user turned on their "geo-location", we will not know where they tweeted from, even if they tweet about a particular location (unless we engage in considerably more sophisticated semantic analyses). From an analytical perspective, this absence of location information adds to additional challenges related to reliability of data gathered.

In our research, we compared flow data using different methods, each of which highlighted different features/aspects of the data we examined. We clustered tweets spatio-temporally to filter repeat trips of the same twitter ID, and then created routes for each origin-destination pair. Once we had O-D pairs, we were able to compare LODES data alongside the twitter data to determine road segment loads. We also determined trip purpose by mapping the origin and destination locations of the tweets and linked them with land uses at census block level. Please note that we only used geo-located tweets, which is a much smaller set than all tweets – this can be considered a limitation, but it also adds to data authenticity and reliability for this type of research without having to rely on the semantic interpretation of tweet contents.

When focusing on correlations between flow magnitudes, we found the most influential factor to be spatial scale. When we examined our results at a county by county level of analysis, our model of flows using LODES and Twitter data are quite similar. In other words, the informal (more frequently available) geosocial data and the official (more infrequently available) census-based data confirm that the same road segments are more frequently traversed. When we examine the very same data at a finer scalar resolution – i.e., at looking at flows at the census tract and block, Twitter and LODES flows deviate significantly. Given that we identified large differences between three spatial scales, it may be worth investigating even more scale levels such as using zip code as a unit of analysis to learn more about travel flow dynamics at different spatial scales. What does this variability mean to a planner or policy maker? We propose that geosocial network data can allow us to draw conclusions about travel flows in the absence of census data as long as we are looking at broad flow patterns across counties. It will become increasingly unreliable if we attempt to rely exclusively on geosocial data to draw conclusions about finer grain movement flows such as between census tracts without additional supporting analyses not yet explored through our research. Our research team is interested in exploring another alternative to using administrative units by using regularly spaced grid cells to explore arbitrary spatial scales.

One of the strengths of our study is that we acknowledged that it is not useful for transportation planners to focus only on the start and end points of the origin-destination matrix. Travel movement occurs along road networks and our analysis is anchored and linked to existing road networks. Vehicular travel on road networks accounts for a majority of trips in the Bay Area and

our study incorporates these trips. In the future, this research can be expanded to include transit routes. By integrating the street graph, we were able to circumvent the constraint of simply matching OD pairs to one another and comparing the individual street segments used to facilitate the flows allowed us to compare different flows with much higher granularity. We found that the street segment data deviate significantly, especially during rush hours. This suggests that Twitter flows capture regular trips with purposes other than commuting. We also used the street-level data to show that a large portion of Twitter flows cover very short distances compared to LODES. We attribute these to trip purposes other than direct travel to work, even though many of them fall into rush hour times. The differences between LODES and Twitter data in the land-use class connections support this interpretation. They show that LODES data contain fewer connections between residential areas compared to Twitter flows. Aside from residential-to-residential connections, the proportions of the remaining land use classes are also very dissimilar for the two datasets, which is yet another indicator of different trip purposes.

Our original goal was to use the finer temporal grain of twitter data (minute by minute rather than year to year) available in a densely populated and tech-savvy region such as the Bay Area could support a different kind of planning: planning in close-to-real time, to allow for decision-making related to modest capital improvements and other planning and policy interventions that are likely to benefit the public. For the reasons stated earlier, we were not able to make this case to our complete satisfaction. We were able to fill in information gaps in the LODES data by using twitter data in two-year time partitions but not for shorter temporal windows because of the data availability challenges resulting from relying on geolocated tweets only. Other methods of passive data sensing may address this problem, but those approaches were not the focus of our investigation.

A significant portion of the land-use class connections are residential to residential trips. This is not a surprising result for the Twitter flows, since we expected movement between different private residences as part of day-to-day social interactions. In the LODES data, however, this was unexpected, since we did not expect many residential areas to function as workplaces. We suggest possible reasons for this unexpected observation: areas classified as residential areas in our land-use data could in fact be compound areas of different land-use classes. Also, by integrating the land-use data on the census block level, compound areas would have been aggregated to the most dominant land-use class, thereby obscuring some commercial land-use class parcels.

LODES data are available for the entire US, as are Twitter data. Therefore, it is possible to transfer our study design as-is to other geographical areas, although local variations in Twitter usage patterns will determine the explanatory value of the results. Depending on the area of interest, different mobility data like mobile phone usage statistics or data from other geo-tagged social network platforms may be better suited for the task. Another factor to be taken into account in this discussion is the representation of the underlying population in the data. Population groups who are not willing or able to participate in social media are likely to be underrepresented and need to be included by means of different data sources like surveys.

We now have the answers to our four research questions.

1. It is relatively straight forward to extract travel flow patterns from geo-social network data. Even without any further semantic analysis, there is enough voluntarily georeferenced data available to arrive at tens of thousands of tweets for a study area the size of the none-county Bay Area. With temporal constraints added (repeat tweets by the same user, multiple tweets having to fall within a 3-hour window from distinct locations within the study area), we can extract movement information.

2. The geo-social network data evidently does not represent commuter flows as provides by LODES data. The traffic patterns derived from tweets consist of significantly shorter (repeat) trips.

3. This is especially true for large-scale (neighborhood level) flows. Only at the scale of overall county-to-county flows, which are captured by the longer trips, can we find a relatively high match rate between the LODES and the geo-social network data.

4. Virtually all geo-social network-derived trips are complementary to the LODES-based commuter trips. As such, this new source of data forms an excellent addition to the sparse survey data that has so far been used to infer about non-commuting trips.

   As determined by the FHWA, only 16.6% of vehicle trips on US streets are work-related, which means these trips are accounted for by LODES data. The remaining 83.4% of trips are not, however, covered by this dataset. We showed that non-work-related traffic flows have different spatiotemporal characteristics from work-related ones, which makes traffic models purely based on LODES insufficient for a wide range of applications. We therefore see the need for data complementary to LODES to cover the remaining flows at a comparable spatial granularity. We see the methods presented in this paper as a step toward the development of such a dataset.

Our research was originally aimed at determining to what degree geo-social network data could be used to substitute for or augment LODES data at finer temporal grains. The answer to this endeavor is a clear no; geo-social network data is a poor substitute for the LODES data, especially for smaller temporal windows. However, quite to our surprise, it turns out that extremely high correlation rates of street segment use between the two data sources suggests that although LODES data is intended to capture only commuter flows, it is actually an excellent predictor of overall traffic loads for non-rush hour and weekend trips. This should be a relief for the transportation planning community because it validates the existing practice of using LODES data as a stand-in for all kinds of traffic demands.

The question of whether a geo-social network data-derive trip is a commute can be addressed by integrating various types of information. In this study, we chose land-use data as an additional source of information. However, it is also possible to extract semantic information from the tweets

by analyzing their text content (Steiger et al., 2015). It would be worthwhile to compare the merits of these two approaches.

We began this research while the nation was navigating the demanding challenges imposed by the pandemic. Travel flows were altered dramatically. We anticipate that we will be able to assemble and review geosocial network data from 2020 to 2022 and compare with previous years in order to draw conclusions about changes in travel flows and their potential impacts. Our research and our approaches have prepared us to conduct this new research which would allow for a more robust discussion about where people were going during the pandemic year, which routes were used more frequently. By conducting sentiment analyses of the tweets, we could also draw conclusions about why they traveled at all. Given that the much of the Bay Area region observed a shelter-in-place order for 2020, this may yield interesting and new findings about non-work trips which continues to be a neglected component of conventional transportation flow analyses.

Abbreviations and Acronyms

DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FHWA	Federal Highway Administration
LEHD	Longitudinal Employer-Household Dynamics
LODES	LEHD Origin-Destination Employment Statistics
OD	Origin-Destination
OSM	Open Street Map

Bibliography

Anahid, B., M. Jackson, P. Amirian, A. Pourabdollah, M. Sester, A. Winstanley, T. Moore and L. Zhang. "Quality assessment of OpenStreetMap data using trajectory mining", *Geo-spatial Information Science* 19 (2016):1, 56-68, doi: 10.1080/10095020.2016.1151213.

Batty, M. "Big Data, Smart Cities and City Planning." *Dialogues in Human Geography* 3 (2013): 274–279. doi:10.1177/2043820613513390

Boeing, G. "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks." *Computers, Environment and Urban Systems* 65 (2017): 126–139. doi:10.1016/j.compenvurbsys.2017.05.004

Chapple, K., and M. Zuk. Case Studies on Gentrification and Displacement in the San Francisco Bay Area. Technical Report. University of California Berkeley, 2015. https://escholarship.org/content/qt1pn8t7rz/qt1pn8t7rz. pdf (accessed November 13, 2020).

Cervero, R. "Jobs-Housing Balance Revisited: Trends and Impacts in the San Francisco Bay Area." *Journal of the American Planning Association* 62 (1996): 492–511. doi:10.1080/01944369608975714

Cervero, R., and M. Duncan. "Which Reduces Vehicle Travel More: Jobs Housing Balance or Retail-Housing Mixing?" *Journal of the American Planning Association* 72 (2006): 475–490. doi:10.1080/01944360608976767

Chapple, K., and M. Zuk. Case Studies on Gentrification and Displacement in the San Francisco Bay Area. Technical Report. University of California Berkeley, 2015. https://escholarship.org/content/qt1pn8t7rz/qt1pn8t7rz. pdf (accessed November 13, 2020).

Convery, S., and B. Williams. "Determinants of Transport Mode Choice for Non-Commuting Trips: The Roles of Transport, Land Use and Socio Demographic Characteristics." *Urban Science* 3 (2019): 82. doi:10.3390/urbansci3030082

Cuba, N. "Research Note: Sankey Diagrams for Visualizing Land Cover Dynamics." *Landscape and Urban Planning* 139 (2015): 163–167. doi:10.1016/j.landurbplan.2015.03.010

Ester, M., H.P. Kriegel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

Gao, S. "Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age." *Spatial Cognition and Computation* (2015). doi:10.1080/13875868.2014.984300

Giuliano, G., and K.A. Small. "Is the Journey to Work Explained by Urban Structure?" *Urban Studies* 30 (1993): 1485–1500. doi:10.1080/00420989320081461

Hagberg, A.A., D.A. Schult, and P.J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX." 7th Python in Science Conference (SciPy 2008), pp. 11–15.

Haklay M. "How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. " *Environment and Planning B: Planning and Design* 37 (2010);682-703. doi:10.1068/b35097.

Ihlanfeldt, K.R., and D.L. Sjoquist. "The Spatial Mismatch Hypothesis: A Review of Recent Studies and their Implications for Welfare Reform." *Housing Policy Debate* 9 (1998): 849–892. doi:10.1080/10511482.1998.9521321

Jurdak, R., K. Zhao, J. Liu, M. AbouJaoude, M. Cameron, and D. Newth. "Understanding Human Mobility from Twitter." *PLOS ONE* 10 (2015): e0131469. doi:10.1371/journal.pone.0131469

Kockelman, K.M. "Travel Behavior as Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from San Francisco Bay Area." *Transportation Research Record: Journal of the Transportation Research Board* 1607 (1997): 116–125. doi:10.3141/1607-16

Kogan, N.E., L. Clemente, P. Liautaud, J. Kaashoek, N.B. Link, A.T. Nguyen, F.S. Lu, P. Huybers, B. Resch, C. Havas, A. Petutschnig, J. Davis, M. Chinazzi, B. Mustafa, W.P. Hanage, A. Vespignani, and M. Santillana. "An Early Warning Approach to Monitor COVID-19 Activity with Multiple Digital Traces in Near Real-Time." 2020. ArXiv:2007.00756

Kounadi, O., and B. Resch. "A Geoprivacy by Design Guideline for Research Campaigns That Use Participatory Sensing Data." *Journal of Empirical Research on Human Research Ethics* 13 (2018): 203–222. doi:10.1177/1556264618759877

Kounadi, O., B. Resch, and A. Petutschnig. "Privacy Threats and Protection Recommendations for the Use of Geosocial Network Data in Research." *Social Sciences* 7 (2018): 191. doi:10.3390/socsci7100191

Kurkcu, A., K. Ozbay, and E.F. Morgul. "Evaluating the Usability of Geolocated Twitter as a Tool for Human Activity and Mobility Patterns: A Case Study for NYC." In *Transportation Research Board's 95th Annual Meeting*, 2016, pp. 1–20.

Lee, J.H., A.W. Davis, S.Y. Yoon, and K.G. Goulias. "Activity Space Estimation with Longitudinal Observations of Social Media Data." *Transportation* 43 (2016): 955–977. doi:10.1007/s11116-016-9719-1

Li, L., M.F. Goodchild, and B. Xu. "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr." *Cartography and Geographic Information Science* 40 (2013): 61–77. doi:10.1080/15230406.2013.777139

Mackenzie, J., T. Azumbrado, D. Connolly, C. Dutravernaci, A.W. Halsted, L. Schaaf, W. Slocum, A.R. Worth, C.J. Pierce, M.L. Gibbons, M.G. Scharff, A.C. Washington, L.G. Mcelhaney, M.B. Halliday, H. Alameda, and M.L. Garcia, M.L. 2017. *Plan Bay Area 2040*. http://2040.planbayarea.org/forecasting-the-future (accessed November 13, 2020).

Martí, P., L. Serrano-Estrada, and A. Nolasco-Cirugeda. "Social Media Data: Challenges, Opportunities and Limitations in Urban Studies." *Computers, Environment and Urban Systems* 74 (2019): 161–174. doi:10.1016/j.compenvurbsys.2018.11.001

McKenzie, B. Who Drives to Work? Commuting by Automobile in the United States: 2013. American Community Survey Reports. 2015. https://www.census.gov/library/publications/2015/acs/acs-32.html (accessed November 13, 2020).

McNeill, G., J. Bright, and S.A. Hale. "Estimating Local Commuting Patterns from Geolocated Twitter Data." *EPJ Data Science* 6 (2017): 24. doi:10.1140/epjds/s13688-017-0120-x

Nguyen, V.B., and E. Stivers. MOVING SILICON VALLEY FORWARD. Technical Report. Urban Habitat, 2012. http://www.reimaginerpe.org/files/MovingSiliconValleyForward1.pdf (accessed November 13, 2020).

online, 2017. Vehicle Trips Data. https://nhts.ornl.gov/vehicle-trips (accessed November 13, 2020).

online, 2019a. Geographic Region Outline Data. https://www.census.gov/cgi-bin/geo/shapefiles/index. php?year=2019 (accessed November 13, 2020).

online, 2019b. LODES data directory. https://lehd.ces.census.gov/data/lodes/ (accessed November 13, 2020).

online, 2019c. Means of Transportation to Work by Selected Characteristics. https://data.census.gov/cedsci/table?q=S0802&tid=ACSST1Y2019.S0802 (accessed November 13, 2020).

online, 2020a. OpenStreetMap Contributors. https://www.openstreetmap.org (accessed November 13, 2020).

online, 2020b. ParcelAtlasUSER MANUAL. https://www.boundarysolutions.com/ParcelAtlas/ParcelAtlasUserManual.pdf (accessed November 13, 2020).

online, 2020c. PostGIS. https://www.postgis.net (accessed November 13, 2020).

online, 2020d. PostgreSQL. https://www. postgresql.org (accessed November 13, 2020).

online, 2020e. Python. https://www.python.org (accessed November 13, 2020).

online, 2020f. QGIS. https://www.qgis.org (accessed November 13, 2020).

online, 2020g. Rethinking Mobility. http://www.rethinkingmobilitywc.com/ (accessed November 13, 2020).

online, 2020h. SeeClickFix. https://seeclickfix.com/ (accessed November 13, 2020).

online, 2020i. The R Project for Statistical Computing. https://www.r-project.org (accessed November 13, 2020).

online, 2020j. Twitter Developer API v1.1. https://developer.twitter.com/en/docs/twitter-api/v1 (accessed November 13, 2020).

online, 2020k. Waze. https://www.waze.com/ (accessed November 13, 2020).

Osorio-Arjona, J., and J.C. García-Palomares. "Social Media and Urban Mobility: Using Twitter to Calculate Home-Work Travel Matrices." *Cities* 89 (2019): 268–280. doi:10.1016/j.cities.2019.03.006

Petutschnig, A., C.R. Havas, B. Resch, V. Krieger, and C. Ferner. "Exploratory Spatiotemporal Language Analysis of Geo-Social Network Data for Identifying Movements of Refugees." *GI Forum* 1 (2020): 137–152. doi:10.1553/giscience2020_01_s137

Plunz, R., Y. Zhou, M.I. Carrasco Vintimilla, K. Mckeown, T. Yu, L. Uguccioni, M.P. Sutto. "Twitter sentiment in New York City parks as measure of well-being". Landscape and Urban Planning 189 (2019):235-246. doi.org:10.1016/j.landurbplan.2019.04.024.

Pourebrahim, N., S. Sultana, J.C. Thill, and S. Mohanty. "Enhancing Trip Distribution Prediction with Twitter Data: Comparison of Neural Network and Gravity Models." In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI 2018, pp. 33–42. doi:10.1145/3281548.3281555

Rodrigue, J.P. *The Geography of Transport Systems*, 5th edition. Abingdon, Oxon; New York, NY: Routledge, 2020. doi:10.4324/9780429346323

Schubert, E., J. Sander, M. Ester, H.P. Kriegel, and X. Xu. "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN." *ACM Transactions on Database Systems* (2017). doi:10.1145/3068335

Steiger, E., R. Westerholt, B. Resch, and A. Zipf. "Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data." *Computers, Environment and Urban Systems* 54 (2015): 255–265. doi:10.1016/j.compenvurbsys.2015.09.007

Ward, M.D., and K.S. Gleditsch. Spatial Regression Models. Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage Publications. 2008. doi:10.4135/9781412985888

Yang, F., P.J. Jin, Y. Cheng, J. Zhang, and B. Ran. "Origin-Destination Estimation for Non-Commuting Trips Using Location-Based Social Networking Data." *International Journal of Sustainable Transportation* 9 (2015): 551–564. doi:10.1080/15568318.2013.826312

Zhang, G., and A.X. Zhu. "The Representativeness and Spatial Bias of Volunteered Geographic Information: A Review." *Annals of GIS* 24 (2018): 151–162. doi:10.1080/19475683.2018.1501607

Zheng, S. And J. Zheng. "Assessing the Completeness and Positional Accuracy of OpenStreetMap in China." In: *Lecture Notes in Geoinformation and Cartography* (2014). Springer, Cham. https://doi.org/10.1007/978-3-319-08180-9_14

About the Authors

Jochen Albrecht, PhD

Jochen Albrecht is a professor of computational geography at Hunter College, City University of New York. His general research centers on modelling and analysis of spatio-temporal phenomena. On the transportation side, Jochen specializes on multi-model modeling in realistic (complex) settings that allow decisionmakers to explore policy options. He is renowned for his work on standardizing geospatial workflows, which forms the basis of his third monograph in addition to his 56 peer-reviewed publications. Jochen serves on the board of directors for the Urban and Regional Information Systems Association, the GIS Certification Institute, and the California Geographic Information Association.

Andreas Petutschnig

Andreas Petutschnig is a PhD candidate at the Department of Geoinformatics (Z_GIS) at the University of Salzburg, Austria. He received his BEng in Cartography and Geomedia-technology at the University of Applied Sciences in Munich, Germany, and his MSc in Applied Geoinformatics at the University of Salzburg, Austria. His research revolves around the analysis of geo-social network and mobile sensor data, mostly focusing on the spatiotemporal analysis of human mobility and other mobile phenomena, including visualization.

Laxmi Ramasubramanian, PhD

Laxmi Ramasubramanian is professor and chair of the Department for Urban and Regional Planning at San José State University. Her research examines the complexities, considering both the opportunities and the constraints, associated with planning with advanced digital technologies. She is a pioneer in the development of concepts and methods to advance participatory GIScience research. Her recent work emphasizes investigations of community resilience in the wake of large-scale social-ecological disruptions and the development of a critical planning education to advance a more just and equitable society. She has published two books, several peer-reviewed publications, and has presented her work at national and international conferences. In 2019, she received the Dale Scholar Prize for advancing reflective practice from CalPoly Pomona. She is a past president of the University Consortium of Geographic Information Science and the incoming president of the Association of the Collegiate Schools of Planning.

Bernd Resch, PhD

Bernd Resch is an Associate Professor at the University of Salzburg's Department of Geoinformatics (Z_GIS) and a Visiting Scholar at Harvard University. Bernd did his PhD in the area of "Live Geography" (real-time monitoring of environmental geo-processes) jointly with the University of Salzburg and MIT. His research interest is understanding cities as complex systems through analyzing a variety of digital data sources, focusing on developing machine learning algorithms to analyze human-generated data like social media posts and physiological

measurements from wearable sensors. The findings are relevant to a number of fields including urban research, disaster management, epidemiology, and others. Bernd received the Theodor Körner Award for his work on urban emotions. Amongst a variety of other functions, he is an Editorial Board Member of *IJHG*, *IJGI*, *PLOS ONE*, and *Urban Planning*, a scientific committee member of various international conferences (having chaired several conferences), an Associated Faculty Member of the doctoral college GIScience, and an Executive Board member of Spatial Services GmbH.

Aleisha Wright

Aleisha Wright is a Master's student and Research Assistant in the Urban Planning Program at San José State University. She received her BA in Environmental Studies with a minor in Psychology at Winthrop University in South Carolina. Her research focuses on improving sustainable public transportation in the San Francisco Bay Area with an emphasis on expanding affordable Transit-Oriented Development in San José.

MTI FOUNDER

Hon. Norman Y. Mineta

MTI BOARD OF TRUSTEES

Directors

MTI MINETA TRANSPORTATION INSTITUTE