The Journal of Transport and Land Use http://jtlu.org
Vol. 12 No. 1 [2019] pp. 303–326

JTLU

Examining the effects of proximity to rail transit on travel to non-work destinations: Evidence from Yelp data for cities in North America and Europe

Zhiqiu Jiang
University of Virginia
zj3av@virginia.edu

Andrew Mondschein
University of Virginia
mondschein@virginia.edu

Abstract: Urban planners often seek to establish land use patterns around transit stations that encourage non-auto travel. However, the willingness of travelers to use different modes in the vicinity of transit remains understudied, in part because of the lack of spatially-precise data on destination and mode choices. Using transportation content extracted from Yelp, a location-based social network (LBSN), we investigate how travel mode to non-work destinations is influenced by proximity to transit. We use textual analysis to analyze travel for non-work activities in seven cities across North America and Europe. Mixed-effect and binomial logistic models show how reported travel by mode varies by distance to rail transit stations. We find that for most non-work activity purposes, reported rail use is highly sensitive to proximity to stations, but some purposes are more amenable to rail use overall. Additionally, compared to non-US cities, US cities are far more parking-dependent near rail stations. The results suggest that not all activities elicit the same levels of non-auto travel, and transit-oriented planning should account for specific activities and regional factors that may modify willingness to travel by different modes. While subject to limitations, LBSNs can illuminate local travel with greater spatial specificity than traditional surveys.

Keywords: Rail transit, social media, accessibility, travel experience, transit-oriented development

Article history:
Received: May 24, 2018
Received in revised form: November 29, 2018
Accepted: January 24, 2019
Available online: May 8, 2019

1 Introduction

Urban planning and design, as applied disciplines, are permeated with best practices that are dictated by experience and long-held, if little-investigated, maxims. One prominent best practice used in transit-oriented planning and design is a tolerable "walking distance to transit." Though it has varied somewhat, the scale of transit-oriented development is often predicated on a limit to the distance the average traveler is willing to walk between a transit station and a neighborhood destination. Values generally hover between 300 meters/one-quarter mile and 750 meters/one-half mile (Dittmar & Ohland, 2012; Gruen, 1964; Guerra, Cervero, & Tischler, 2012). We propose destination proximity to transit not just as a determinant of walkability but also travel by other modes. Particularly for non-work activities, such as meals, shopping, recreation, and socializing, travel patterns are less routinized than commuting, and

http://dx.doi.org/10.5198/jtlu.2019.1409
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

standardized walkable distances may not fully address variability in traveler preferences and behaviors when seeking to access a range of activities near transit (Chatman, 2008; Walle & Steenberghen, 2006). Therefore, we ask how travel mode choices vary as distance from transit increases, and whether factors such as activity purpose and regional differences significantly modify the effect of distance. A better understanding of these relationships can help identify the conditions under which local land use effectively contributes to non-auto mode choices around transit stations.

To address our research questions, we use the spatially-precise activity data included in a location-based social network (LBSN). LBSNs, also called geosocial media, consist of shared human experiences, often with textual content, associated with geographic locations (Crampton et al., 2013; Kelley, 2013; Rybarczyk, Banerjee, Starking-Szymanski, & Shaker, 2018). We use geosocial media data to investigate how traveler mode varies systematically not just based on proximity but also city and activity type for a wide range of non-work activities. An ever-increasing segment of the population makes use of social media in their daily lives. It also plays a significant role in many aspects of daily travel behaviors, especially in information search, decision-making (before the trip), experiences/resources share (during the trip), and post-travel evaluation (after the trip) (Chung & Koo, 2015; Munar & Jacobsen, 2013; Sedera, Lokuge, Atapattu, & Gretzel, 2017; Xiang & Gretzel, 2010). Yelp is a crowdsourced LBSN used for rating and describing non-work activity experiences. The activity reviews posted in Yelp not only provide information on experiences while at a business but can also indicate how reviewers travel to and from an activity (Mjahed, Mittal, Elfar, Mahmassani, & Chen, 2017; Mondschein, 2015).

We use a Yelp dataset with approximately 3 million reviews for seven metropolitan areas in North America and Europe with rail transit systems to examine how proximity to a transit network influences multimodal travel experiences to diverse non-work destinations. The density and spatial precision of observations allow us to categorize non-work activities in urban areas as well as allow us to examine the effect of transit networks on multimodal behavior in a way that traditional travel surveys cannot. There are three major benefits to using Yelp in our study: (1) it is a relatively complete businesses dataset across multiple cities and countries of non-work destinations with precise latitude/longitude attributes; (2) it has a significant number of reviews describing travel experiences; and (3) text-mining methodologies can be applied into the analysis of online reviews in order to identify specific behaviors, including mode choices. By contrast, traditional travel surveys usually focus on a single city with a much smaller number of trips captured by the survey, at lower densities across a given region. To the authors' knowledge, this study is the first to use these more spatially-precise LBSN data to investigate how travel choices vary relative to distance from transit stations.

In this paper, we use textual analysis to extract travel modes associated with reviewed activities across each city. We then present a modeling framework using mixed-effect and binomial logistic models to highlight the factors associated with transportation choices, when accessing non-work destinations. The results show how different modes, whether walking, transit, driving or parking (as an access mode), are differentially affected by proximity to urban rail networks. The results demonstrate the complex relationship between traveler mode and destination proximity to rail transit, across both city and activity purpose. We conclude with a discussion of the implications for public transportation, urban planning, and neighborhood design, highlighting how travel experiences as revealed in activity reviews can indicate the effectiveness of local transportation infrastructure in encouraging reduced driving and increased use of alternative modes.

2 Literature review

2.1 The role of proximity to transit in travel behaviors

Transportation planners have emphasized the critical role that public transport networks play in providing sustainable, congestion-resilient accessibility in metropolitan areas (Lierop, Maat, & El-Geneidy, 2017; Murray, Davis, Stimson, & Ferreira, 1998; Vuchic, 2017). Transit-oriented development (TOD) is predicated on reducing distances between transit network stations and the varied destinations that individuals seek to access, with the expectation that TODs should facilitate increased walking and biking to transit and reduced driving and parking around transit stops (Boarnet & Compin, 1999; Chatman, 2013; Guerra et al., 2012). The proximity of destinations to transport services has generally been found to be an important factor in encouraging non-auto access to different activities, provide more mobility options (Litman, 2011), and increase active travel (Rissel, Curac, Greenaway, & Bauman, 2012). With more transit trips beginning and ending with a walking or biking trip rather than a car trip, TODs are characterized by a higher level of pedestrian or bicyclist activity and lower levels of automobile travel (Ewing, Tian, Lyons, & Terzano, 2017; Hong, Boarnet, & Houston, 2016). Researchers have found that low density and separated land uses create high auto dependency, while urban forms with relatively high density and mixed land use near transit places in order to encourage more walking and transit trips (Clifton, Currans, Cutter, & Schneider, 2012; Reilly, O'Mara, & Seto, 2009). Many newly built TODs have generally been commercially successful and nearby communities have actively planned for mixed use development within tolerable walking distance of rail transit stations (Noland, Weiner, DiPetrillo, & Kay, 2017).

Previous research has provided ways for estimating an average or maximum walking distance to transit (Hoback, Anderson, & Dutta, 2008; Olszewski & Wibowo, 2005; O'Sullivan & Morrall, 1996), suggesting walkable distances for well-planned or designed TODs. However, these studies are focused on walking mode and pedestrian-friendly design specifically. As one expectation of TOD in a metropolitan area is to incorporate a mix of shopping, service, and recreation activities at urban centers combined with high quality of transit. Daily non-work trips accessing TOD areas vary across travel mode choices, activity purposes, and built environments (Greenwald & Boarnet, 2001; Nelson & Niles, 1999). Prior research does not address how multiple modes – transit use, driving, parking as an access mode, biking, and walking – may vary in proximity to transit, and how they may complement or substitute for one another depending on local urban form and traveler activity purpose. In addition, these studies are limited primarily to case studies of single cities.

2.2 Travel behavior analysis using Location-Based Social Network data

In recent years, there has been rapid growth in LBSN services, such as Yelp, Twitter, Foursquare and Facebook, which have attracted an increasing number of users and greatly enriched their daily urban experiences (Choe, Kim, & Fesenmaier, 2017; Evans & Saker, 2017). Exploring the capacity of new data sources such as social media to measure travel activity has become an emerging research area in the planning and design of urban transportation systems. For example, many transportation issues and behaviors can be linked with the volunteered geographic information in Twitter posts. Collins, Hasan, and Ukkusuri (2013) use about 500 twitter texts to evaluate transit riders' satisfaction with a Sentiment Strength Detection Algorithm. Andrienko et al. (2013) extract the geotagged twitter information about everyday life of people – activities, habits, travel behaviors and experience – to understand movement patterns. Kurkcu, Ozbay, and Morgul (2016) examine the spatial and temporal characteristics of human activity and mobility patterns and compare trip characteristics with satisfactory quantities using geolocated Twitter data. Kovacs-Györi et al. (2018) develop a methodology using tweets to extract visitors' spatiotemporal patterns along with the sentiments embedded in the text of tweets.

Researchers have developed multiple approaches to extracting information from LBSNs to track and analyze human movements. The development of data mining and machine learning allows travel experience information such as trip preferences and sentiments to be captured from LBSNs such as Twitter or Yelp (Hu & Liu, 2012; Ignatow & Mihalcea, 2016; Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017; Senaratne, Mobasheri, Ali, Capineri, & Haklay, 2017). Spatial analysis and visualization techniques also enable the user-generated LBSN to be used to identify the most appreciated Points of Interest (POIs) and landmarks in a study area as well as to retrieve trip origins and destinations, durations, inferring activity types or classifying transportation modes (Chaniotakis, Antoniou, Aifadopoulou, & Dimitriou, 2017; Nikšič, Campagna, Massa, Caglioni, & Nielsen, 2017).

LBSNs can include travel attributes as well as land-use variables and socio-demographic attributes (Stieglitz, Mirbabaie, Ross, & Neuberger, 2018). Transportation researchers have increasingly used LBSN data to investigate travel choices to non-work destinations (Manca, Boratto, Morell Roman, Martori i Gallissà, & Kaltenbrunner, 2017; Mjahed et al., 2017; Mondschein, 2015). These data have the potential to address documented limitations with traditional travel surveys: declining sample sizes (Stopher & Greaves, 2007), under-reporting of trips (Forrest & Pearson, 2005), imprecision or absence of locations and times (Stopher, Jiang, & FitzGerald, 2005), and infrequently updated content (Chen, Gong, Lawson, & Bialostozky, 2010). Compared to traditional survey data, LBSNs like Yelp may allow investigation of variability in travel behavior for a single travel mode, or modal mix, that are linked with non-work destinations across different cities and countries with more extensive data across and increased spatial precision. While big data offer opportunities to address questions regarding human activities and mobility, LBSNs also have limitations: for example, many social media datasets do not have associated demographic data, and participants may be biased toward specific demographic groups. General limitations need to be considered beforehand: such as representativeness, objectivity, accuracy, quality, and context-sensitivity, which requires caution when applying LBSN data and analytical methods (Boyd & Crawford, 2012; Hargittai, 2015, 2018; Manovich, 2012; Wu, Zhu, Wu, & Ding, 2014). However, the usage of internet, mobile devices and geotagging messages has greatly increased, suggesting that more detailed and representative analyses are possible (Gilbert & Karahalios, 2009). In the following section, we describe how we address limitations of LBSN data while answering a question not readily addressed by standard travel survey data.

3 Data and methods

3.1 Using the Yelp dataset and addressing its limitations

Our analysis uses Yelp reviews to answer whether travel mode varies significantly by distance from transit stations, controlling for city and activity type. Yelp is an LBSN where reviewers rate "businesses," including non-commercial destinations, and contribute long-form text reviews so that users can make more informed non-work activity and destination choices. As of 2017, Yelp has over 26 million unique reviewers (Yelp, 2017c). Yelp is one of the most comprehensive business databases and review sites in the US and worldwide, though there are other LBSNs that may be useful for similar purposes. There are limitations to the Yelp reviews that may potentially influence the results of our analysis. Foremost, Yelp reviewer demographics are not sampled or weighted to be representative of urban populations. Comparing the demographics of Yelp reviewers from a Quantcast survey in 2017 (Quantscast, 2017) to data from the US Census 2016 (U.S. Census Bureau, 2016) and Canada Statistics 2016 (Statistics Canada, 2016) for the general population, we observe that Yelp users are more female (61% of users), and Yelp users' households are wealthier and more educated on average than households in the US and Canada. In addition, Yelp users are not only local residents but also travelers with distinct travel behaviors.

These demographic biases may result in a spatial mismatch between all business locations and locations with reviews, as Yelp reviewers will likely prefer businesses catering to their demographic, all else being equal, which may include specific neighborhoods and business types. In the case of this analysis, the primary effect of this bias may be that we do not capture the travel experiences of on-average lower-income travelers, including patronizing businesses that may have a different relationship to transit proximity than the businesses most reviewed in Yelp. The inclusion of travelers in the dataset may be a benefit compared to traditional travel surveys, as non-residents are often an important subset of patrons in commercial and mixed-use districts whose travel and activity choices have significant impact on those neighborhoods. Regardless, given the limitations of the dataset, the results should be interpreted with the expectation that the patterns observed may not be shared by lower income travelers that may be more likely to use transit, and the businesses that serve them, all else equal.

Because travel mode to access a destination is not required content in a review, we cannot directly confirm that the rate at which modes are included in reviews is consistent with actual behavior. Overall, we follow Mjahed et al. (2017), who regard a specific mode mentioned in a review as a positive recommendation of that mode for accessing a particular business. Put another way, Yelp review content is "pre-trip" information that travelers can use to make mode choices, and they find that these data are locally correlated with mode choices. Thus, while transportation content in Yelp reviews is not a direct measure of travel behavior as would be measured in a traditional travel survey, it can be readily understood as a measure of each mode's relevance to accomplishing an activity. The modal mix for a given location as derived from the Yelp reviews should reflect how important a reviewer believes each mode is to activities in that area. Our analysis reveals relative differences in reported travel by mode and other factors, even if absolute behavioral differences are not measured by this study. These relative differences address our research questions, which examine the differential effects of activity purpose and city on reported modes near transit stations.

For this analysis, we use the 2017 release of the Yelp Academic Dataset (Yelp, 2017b), which provides full text reviews and the precise latitude and longitude of each reviewed business (Evans & Saker, 2017). In addition to the destination's geographic coordinates, each review is timestamped in terms of when the review as submitted (not when the activity took place). We label the businesses using Yelp's reported 10 "big categories," which are Active Life, Arts, Automotive, Health, Hotels & Travel, Nightlife, Other, Restaurants, Service, and Shopping, transforming each business from multi-label to single-label by training a single-label classifier for each label so that it classifies the business to its matching main category within 10 big categories pool (Trajdos & Kurzynski, 2018). Businesses in the Yelp dataset are automatically categorized using a multi-label classification approach (Tung, 2015) with nearly 1,000 categories (Yelp, 2017a). From the Academic Dataset, we selected seven metropolitan areas: "Charlotte, North Carolina," "Cleveland, Ohio," "Las Vegas, Nevada," and "Phoenix, Arizona" in the US, "Montreal, Quebec" in Canada, "Edinburgh" in Scotland, UK and "Stuttgart" in Germany. There are approximately 3 million reviews total in the seven selected cities. Each has a rail transit system, appropriate for answering our key empirical research question: How is transportation mode influenced by proximity to rail transit, as it varies by cities and business types?

3.2 Overall methodology

Figure 1 illustrates the sequence of analyses included in this paper. First, extract the transportation content of Yelp reviews using textual analysis. We use a Jaccard similarity index to examine the association between transportation content in Yelp and urban transportation behaviors. Second, we estimate the network distance of destinations – and the reviews associated with each destinations – to the nearest rail transit station using ArcGIS Network Analyst (Curtin, 2007; ESRI, 2018). We then use a Generalized

Linear Mixed-Effects (GLME) model is used to examine how destination distance to transit affects reported rail travel in the Yelp dataset, across cities and business types. The final analyses use binomial logistic regression to investigate how reported experiences by multiple modes vary around rail transit stations. Methods used for extraction of transportation content from the reviews, as well the GLME and binomial logistic models are described in greater detail in Sections 3.3, 3.4, and 3.5. Results of each analysis follow in Section 4.

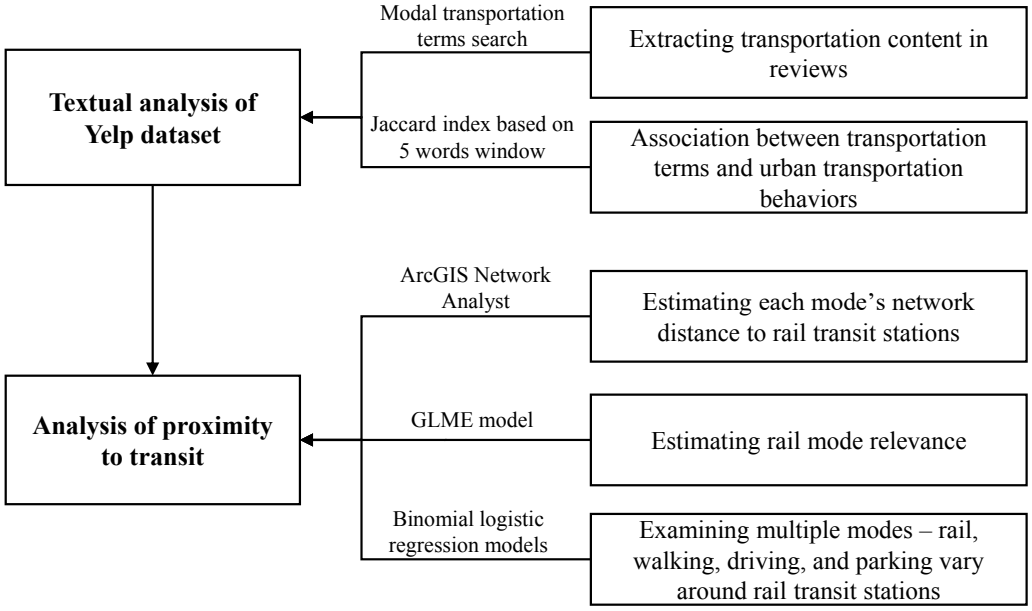

Figure 1. Methodology framework

3.3 Extracting transportation content in reviews

Yelp reviews frequently include transportation content, describing the travel experience to or from a business or other destination. Examples from the dataset:

> *"They have their own free parking lot...very cool."*
> *It's a great place for running, biking, walking, etc. It's a great way to travel on bike between Old Town and Arcadia."*
> *"There's parking validation for the structure adjacent to the theater, so that's cool. It's a bit of a walk, for handicapped, elderly, or lazy people."*
> *"I'm a fan of this place because of the light rail convenience and the low prices."*
> *"Hopefully the cities ramp up interests in their mass transit systems."*

To analyze the large number of reviews with transportation content, we use a text mining approach by identifying and extracting the mentions of a particular travel experience (Hu & Liu, 2012; Krippendorff, 2012). We seek specifically modal experiences within a given review, generating measures of transportation mode experience frequency. Table 1 summarizes the frequencies of modal transportation terms in the seven cities' Yelp reviews. Since the reviews in these seven cities consists of multilingual texts, we additionally use French and German mode terms for text mining in Montreal and Stuttgart. The textual analysis method is from Mondschein (2015). A mode is defined by multiple terms, such as "drive"

and "drove," or "parked" and "parking." 21.4% of reviews in the dataset have identified transportation content. Note that this may be an underestimate, since not all possible terms related to transportation may be included in the selected set of terms.

Table 1. Transportation terms frequency for seven metropolitan areas

Category	Terms	Total		Charlotte		Cleveland		Edinburgh		Las Vegas		Montreal		Phoenix		Stuttgart	
		N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Auto	**car**	93587	3.44	4224	2.29	3678	2.09	432	0.94	38810	2.6	2009	1.82	44056	6.27	378	1.14
	drive, drove	111154	4.08	5424	2.94	5468	3.11	274	0.59	46986	3.2	992	0.9	51886	7.39	124	0.38
	parking, parked	77486	2.85	8097	4.39	5762	3.28	309	0.67	30233	2.1	1721	1.56	31027	4.42	337	1.02
	traffic	9045	0.33	785	0.43	456	0.26	114	0.25	4431	0.3	239	0.22	2963	0.42	57	0.17
Public Transport	**rail, train**	12350	0.45	905	0.49	612	0.35	685	1.48	4743	0.3	452	0.41	4796	0.68	157	0.47
	bus, streetcar	10362	0.38	398	0.22	414	0.24	690	1.5	5821	0.4	549	0.5	2398	0.34	92	0.28
	transit	779	0.03	42	0.02	47	0.03	4	0.01	192	0	285	0.26	185	0.03	24	0.07
Active Travel	**bike, biked, biking, bicycle**	8777	0.32	507	0.27	525	0.3	174	0.38	1755	0.1	494	0.45	5244	0.75	78	0.24
	walk, walked, walking	249265	9.16	13921	7.55	12292	6.99	4470	9.69	121318	8.3	7900	7.17	85569	12.2	3795	11.5
Total Auto		291272	10.7	18530	10.1	15364	8.73	1129	2.45	120460	8.2	4961	4.5	129932	18.5	896	2.71
Total Public Transport		23491	0.86	1345	0.73	1073	0.61	1379	2.99	10756	0.7	1286	1.17	7379	1.05	273	0.83
Total Non-auto		281533	10.3	15773	8.55	13890	7.9	6023	13.1	133829	9.1	9680	8.79	98192	14	4146	12.5
Total Transportation Terms		572805	21	34303	18.6	29254	16.6	7152	15.5	254289	17	14641	13.3	228124	32.5	5042	15.3
Total Reviews		2722484		184430		175916		46148		1470288		110126		702516		33060	

"Walk" and "Drive" are the most frequent modes in this dataset with 9.16 and 7.85% of all reviews, respectively. Note that for the analysis, we divide auto-based terms into "driving" categories including "car," "drive," "drove," and "traffic," and a "parking" category including, "parking" and "parked." Public transit, bike, and bus are mentioned less frequently through the average percentages of 7 cities, though Edinburgh, Montreal, and Phoenix have a higher rate of public transport terms.

In order to understand the usage of transportation terms in the reviews, we use the Keyword-in-Context (KWIC) technique (Ignatow & Mihalcea, 2016; Jockers, 2014; Vinithra, Selvan, Kumar, & Soman, 2015), an approach examining the associations among each transportation keyword and the words that surround it – specifically, five words to the left and right of the transportation term. The analysis is completed with KH Coder software (Higuchi, 2012, 2014), using a Jaccard index of "shared phrases/all phrases" (Markov & Larose, 2007; Mondschein, 2015) to reflect the strength of concordance. Given a review's content, the similarity between every pair of noun or adjective within 5-word window and the transportation term is measured by Jaccard similarity coefficient, a statistic commonly used for comparing the similarity and diversity of sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

The index $J(A,B)$ is the ratio of the number of reviews including both A word and B word over the number of reviews including either A word or B word. The Jaccard similarity coefficient ranges from 0 to 1. The strength of concordance is the association level between the target word and the substantive word. If $(A,B)=0$, this means A word and B word are totally unassociated. If $(A,B)=1$, this means A word and B word are fully co-occurring.

3.4 Estimating rail mode relevance using a Generalized Linear Mixed-Effect model

A Generalized Linear Mixed-Effect (GLME) model is an extension of the corresponding classical linear regression model for cross-sectional data by introducing both fixed and random effects in the model (Faraway, 2016; Mcculloch & Neuhaus, 2001). We use GLME to examine how distance affects the importance of rail to reviewers in accessing non-work activities, across cities and business types.

Figure 2(a) shows that rail-term counts (absolute values) significantly vary within cities. Beyond 2.5 km, rail reviews are very infrequent and don't vary substantially as distance increases. Therefore, we set the 2.5 km as a cutoff for looking at how the frequency of rail reviews correlates with other factors in the model. Figure 2(b) illustrates the relative share of rail reviews to all transport-content reviews within 2.5 km. The variability in rail share among cities, even within a short distance from the rail network, is consistent with the idea that rail usage will vary not just based on proximity but other local factors such as land use, mobility networks, and socioeconomic factors. Therefore, our GLME model adds "city" as a random effect to resolve the non-independence of observations within each city.

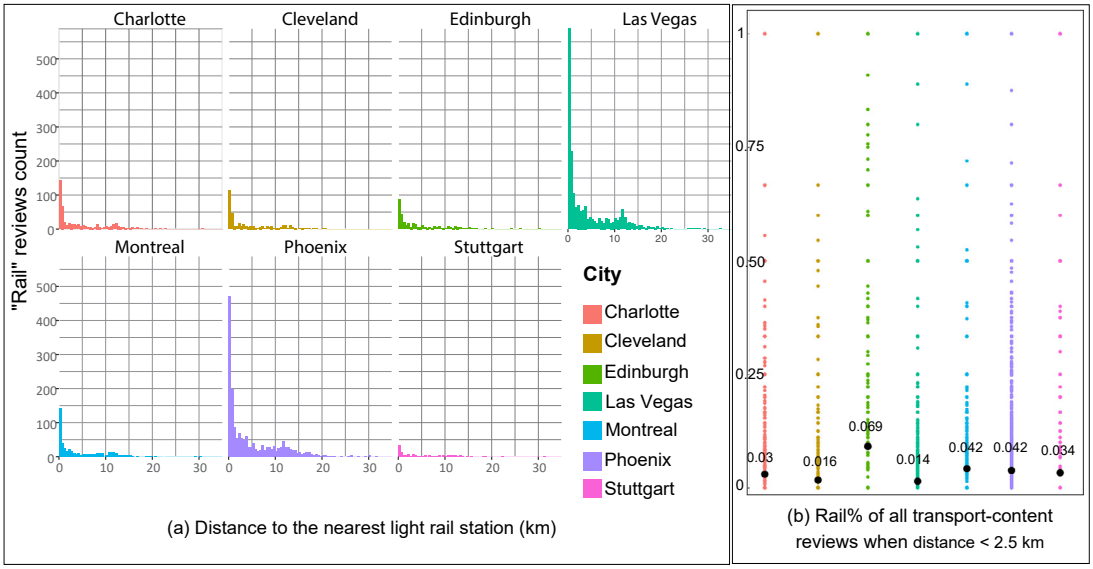

Figure 2. Frequency of "rail" reviews by distance and city and rail %

A linear mixed-effect (LME) is represented as:

$$\begin{aligned} Y_i &= X_i\beta + Z_i b_i + \epsilon_i \\ b_i &\sim N(0, D) \\ \epsilon_i &\sim N(0, R_i) \end{aligned} \tag{2}$$

where i is the subscript. X_i are the predictors. β are the fixed effect coefficients. Z_i is a subset of predictors. b_i are the random effect predictions. The ϵ_i are the random errors. There are two matrices of error structures: One for the random effects D and one for the random errors R_i, where D is referred to as between-group variation while R_i is within-group variation.

In our GLME model, we let the linear predictor, η, be the combination of the fixed and random effects excluding the residuals.

$$\eta = X_i\beta + Z_i b_i \tag{3}$$

The generic link function $g(.)$ relates the outcome Y to the linear predictor η . $\eta=g(\mu)$ where $\eta_i=g(\mu_i)$, i=1,…, n is unconstrained. $h(.)=g^{-1}(.)$= *inverse link function*. For - ∞ < η < ∞, thus, Y:

$$Y = h(\eta) + \epsilon_i \tag{4}$$

We tested a series of combinations of the predictors and random effects, such as (I). one predictor (distance) + one level of grouping (city); (II). one predictor (distance) and two levels of grouping (business category, city); (III). multiple predictors (distance, business category) and random effects (business category, city). We considered both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) together as a method for assessing the quality of fitted models (Dziak, Coffman, Lanza, & Li, 2017). The AIC or BIC for a GLME model is usually written in the form *-2logL+kp* , where L is the likelihood function, p is the number of parameters, k is 2 for AIC and *log(n)* for BIC. AIC estimates the relative information lost by a given model. In other words, it is based on the *deviance* which is a measure of goodness of fit of a GLME model. BIC is an estimate of a function of the posterior probability of a model being true, under a Bayesian setup, the smallest BIC is considered to be more likely to be the true model. First, we used an Anova check using Wald test to compare the performances of all the candidate models in order to select the predictors and random effects. Then, we fitted the GLME model by creating different combination of fixed effects and random effects and compare to see which fits the best (Faraway, 2016). Finally, we selected the model that has the smallest AIC and BIC as our best model. The final fitted generalized linear mixed-effect model is:

$$Y = \beta_0 + \beta_1 x_i + \beta_2 x_i + (\beta_1 * \beta_2) x_i + b_i + \epsilon_i, i = 1, 2, \dots, n \tag{5}$$

where Y is the percentage of rail reviews of business i, x_i is business i. b_i is a random effect for business i, representing the city of business i, assuming that an intercept that is different for each business i. ϵ_i is the random error of business i. β_0 is the intercept. β_1 and β_2 are fixed effects, β_1 is the business i's walking distance (km) to its nearest rail transit station, and β_2 is business i's business category. ($\beta_1 \times \beta_2$) is the interaction term, can be interpreted as the additional effect of business category and walking distance of business i. Same, the random effect b_i and the error e_i are independent and both follow normal distributions with zero means, i.e., $b_i \sim N(0, D)$, $\epsilon_i \sim N(0, R_i)$. Since the number of reviews varies for each business, so to optimize the estimates, we set the number of total transportation reviews as the weights in our GLME model. Data processing is performed under the R statistical program package (R Core Team, 2017), version 3.4.1 with lme4 package for GLME modeling (Bates et al., 2017).

3.5 Rail, walking, driving, and parking by distance from transit: Binomial logistic regression models

In addition to estimating the relevance of rail over distance to non-work activities, we also examine how multiple modes or travel phenomena—rail, as well as walking, driving, and parking—vary around rail transit stations. If urban planners seek to not just increase walking and transit use but reduce driving, we should examine how modes either complement or substitute for one another as distance from rail increases. In other words, do driving and parking experiences increase as transit and walking decrease? Using the transportation terms from the Yelp dataset, we estimate the likelihood that a business will host reviews with each mode using binomial logistic regression. The y_i (dependent variable) is defined as:

$$y_i = \begin{cases} 1, & \textit{if the target transportation term occurs in the business's reviews} \\ 0, & \textit{otherwise} \end{cases} \quad (6)$$

Explanatory variables are a set of $X = (X_1, X_2, \ldots, X_k)$. Taking a single variable X, the model is represented as:

$$\pi_i = \Pr(y_i = 1 | X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad (7)$$

Where, x_i is the observed value of the explanatory variables for observation i.

Logistic regression estimates the coefficients for observed outcomes using the maximum-likelihood (MLE) technique rather than ordinary least squares (OLS), and thus relies on large-sample approximations(Cole, 1991; Czepiel, 2002). The maximum-likelihood for (β_0, β_1) is obtained by finding $(\widehat{\beta_0}, \widehat{\beta_1})$ that maximizes:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{N} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} = \prod_{i=1}^{N} \frac{\exp\{y_i(\beta_0 + \beta_1 x_i)\}}{1 + \exp\{y_i(\beta_0 + \beta_1 x_i)\}} \quad (8)$$

Because the likelihood of observing 1 or 0 outcomes in our dataset imbalanced, we use an "undersampling" strategy for rebalancing our dataset to improve estimation with a *10–fold cross*-validation (Refaeilzadeh, Tang, & Liu, 2009) for our training and testing sets to diagnose model fit (King & Zeng, 2001). For example, the "rail" term ratio of "1 or 0" outcomes is an imbalanced 0.097:0.903. Other independent variables including the distance to the nearest light rail station and a range of business characteristics (business type, average stars, is open/closed, city) are considered for the model fit. To validate the fitted binominal logistic regression models, prediction performances are compared across two types of tests: (I) one to five independent variables (distance, business type, average stars, is open/closed, city), no interactions among these variables; (II) different combinations of independent variables, including interactions among variables. In the modeling fitting process, we used Wald tests (W) (Hosmer, Lemeshow, & Sturdivant, 2013) of deviance residuals as a criterion for inclusion or removal of independent variables. Regarding overall goodness-of-fit, we calculated the accuracy in classification, as a measurement to reflect the percentage of matched predicted and observed transportation term occurrences. In addition, we examined the Sensitivity and Specificity of the classification model according to the confusion matrix. Sensitivity is the percentage of the percentage of 1's correctly predicted by the model, while Specificity is the percentage of 0's correctly predicted. By examining diagnostics of each model, our final fitted binomial logistic regression model is selected based on the highest prediction accuracy after cross-validation:

$$logit(\pi_{ijk})_{mode\gamma} = \alpha + business\ category_i + distance_j + city_k + (distance \times city)_{jk} \quad (9)$$

$logit(\pi_{ijk})_{mode\gamma}$ is the existence of γ transportation mode of business *i*, where $\gamma \in$ (*Driving, Parking, Rail, Walking*). The parameter α is the intercept. $business\ category_i$, $distance_j$, and $city_k$ are the business category, walking distance (km) to the nearest transit station based on the real road network, and the city of business *i*. $(distance \times city)_{jk}$ is the interaction term, can be interpreted as the additional effect of distance and city of business *i*.

4 Analysis and results

4.1 Association between transportation terms and urban transportation behaviors

Table 2. Top word associations with key transportation terms, nouns and adjectives by Jaccard similarity index

Parking				**Rail**				**Walk**				**Drive**			
Assoc. nouns[a]		Assoc. adjectives[b]		Assoc. nouns		Assoc. adjectives		Assoc. nouns		Assoc. adjectives		Assoc. nouns		Assoc. adjectives	
Word	JI[c]	Word	JI	Word	JI	Word	JI	Word	JI	Word	JI	Word	JI	Word	JI
lot	0.243	free	0.062	stop	0.035	light	0.230	casino	0.051	short	0.070	minute	0.050	worth	0.051
garage	0.087	easy	0.057	station	0.034	convenient	0.015	strip	0.050	long	0.041	order	0.044	long	0.039
car	0.079	nice	0.052	train	0.033	public	0.015	hotel	0.047	easy	0.029	window	0.043	quick	0.026
spot	0.077	small	0.049	downtown	0.030	central	0.014	parking	0.038	nice	0.026	location	0.042	friendly	0.026
street	0.072	little	0.048	block	0.020	accessible	0.014	room	0.037	quick	0.024	car	0.041	short	0.025
area	0.064	good	0.047	ride	0.016	east	0.011	minute	0.035	clean	0.023	home	0.039	great	0.024
valet	0.063	great	0.046	bus	0.016	close	0.011	street	0.033	little	0.021	test	0.035	new	0.024
place	0.059	clean	0.042	light	0.015	uptown	0.010	elevator	0.032	convenient	0.020	hour	0.034	fast	0.024
location	0.058	busy	0.039	city	0.015	easy	0.010	park	0.030	free	0.019	time	0.033	good	0.024
hotel	0.058	friendly	0.038	airport	0.014	right	0.009	garage	0.027	great	0.019	way	0.032	little	0.023
space	0.056	big	0.034	transit	0.013	19th	0.009	area	0.026	small	0.019	strip	0.032	best	0.023
room	0.054	sure	0.032	walk	0.013	easier	0.007	floor	0.026	main	0.019	service	0.031	sure	0.022
time	0.053	bad	0.032	access	0.013	urban	0.007	pool	0.026	beautiful	0.019	line	0.031	nice	0.021
people	0.052	long	0.030	distance	0.013	outdoor	0.006	location	0.025	right	0.018	food	0.030	better	0.021
night	0.050	best	0.030	metro	0.013	7th	0.006	lot	0.025	big	0.018	customer	0.029	wrong	0.021

[a]"Assoc. nouns" is short for "associated nouns."
[b]"Assoc. adjectives" is short for "associated adjectives."
[c]"JI" is short for "Jaccard Index."

Demonstrated by the Jaccard analysis (see Table 2), the usage and intent of transportation terms are revealed in the words proximate to the transportation terms. Consistent with previous findings (Mondschein, 2015), for "parking," reviewers associate nouns like "lot," "spot," and "car," and adjectives like "convenient," "accessible," "free," and "ample." The top fifteen "Rail" associated adjectives are: "accessible," "convenient," "close," "right," "central," "east," "easy," "nearest," "short," "uptown," "ample," "cheap," "clean," "good." "Walk" association words are somewhat more diverse, but the majority of nouns and adjectives associated with "walk" are related to outdoor walking experiences. We exclude bicycle terms specifically because they often refer to bicycle shops.

Note that this conceptualization is supported by an examination of transportation content over time. We observe that the review-derived mode split is very stable within each city when examined year-to-year, except in the case of Phoenix, where the opening of a light rail line during the period revealed a significant increase in rail reviews. This responsiveness to a major change in the network supports the linkage between modal recommendations in the reviews and travel behavior. Importantly, the modal categories presented here are not necessarily mutually exclusive – those who mention "parking" almost certainly drove, and those who mention "rail" almost certainly walked or biked. However, distinctions between these modal terms allow us to identify what is most important when accessing destinations in rail-adjacent neighborhoods.

4.2 Estimating each mode's distance to rail transit stations

Geographic Information System (GIS) analysis enables estimation of mode share variation across the study cities, in terms of actual walking distance based on road network from stations along urban rail transit infrastructure. Because of their relatively low frequency, we exclude bus and bike terms from subsequent analysis. Figure 3 displays the share of driving, parking, rail, and walk modes by distance to rail stops. Overall, the closer to stations, the higher percentages of reviews mentioning rail and walking, while the lower for drive and parking.

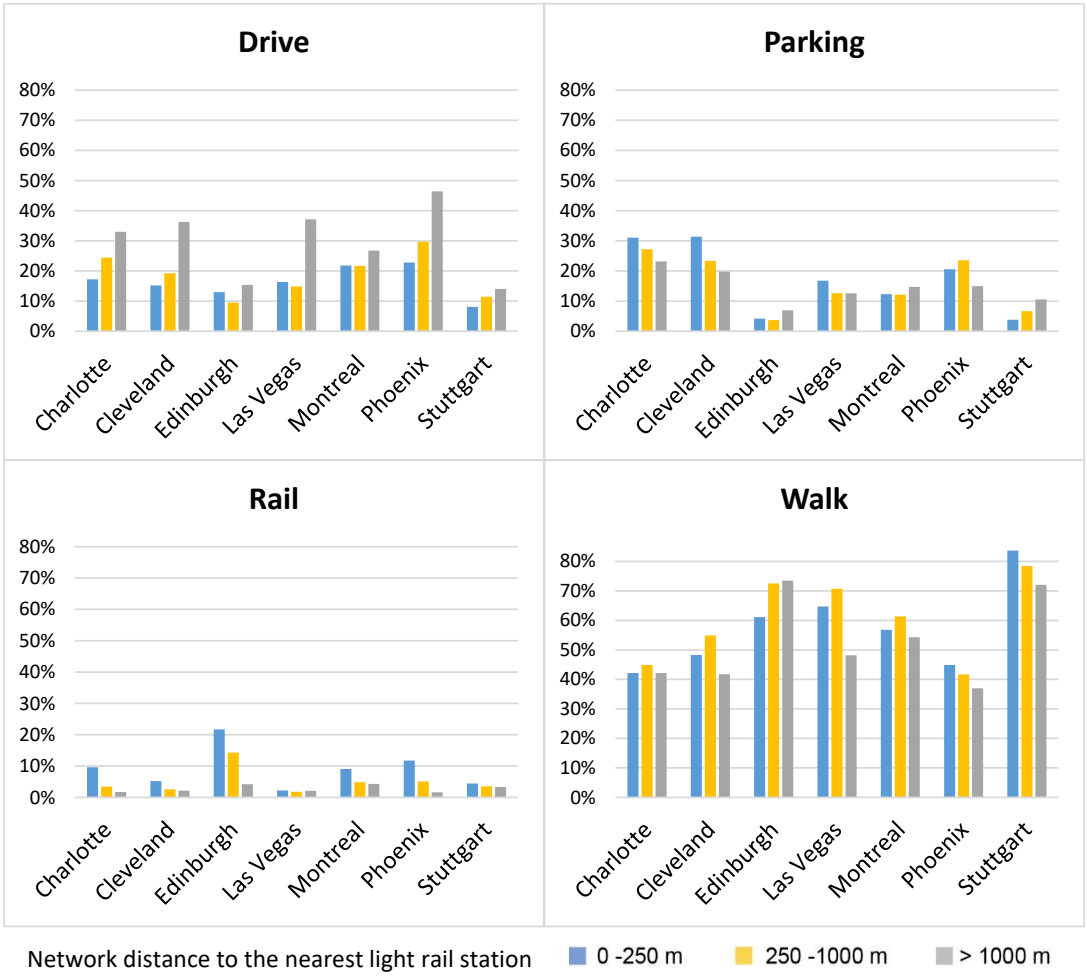

Figure 3. Mode share of all transport-content reviews by city in distance ranges

4.3 GLME models for rail mode relevance

Results of the best fitted model are shown in Table 3. The coefficients are log-odds scaled, shown with standard errors, test statistics (z values) and p-values. In Table 3, we observe that the interactions between distance and some business categories are statistically significant with the rail%. The GLME model results illustrate that the distance, by itself, is not statistically significant to the percentage of rail terms mentioned in the reviews; however, the interactions between activity type and distance are all significant, except for hotel/travel destinations. "Restaurants," "Nightlife," and "Service" destinations have the strongest interaction effects, indicating that rail experiences associated with these activities attenuate

particularly quickly. Activity purposes, by themselves have generally smaller and non-significant effects on rail %, compared to the interaction terms. However, "Automotive" destinations (such as car repair), unsurprisingly, have a significant negative relationship with rail reviews.

The combination of main effects and interactions can be difficult to interpret individually, but the results of the model can be more readily understood through a visualization of their combined effects (see Figure 4). We use the R packages "effects" and "sjPlot" to plot the results and show how rail% predicted probabilities change with variation across independent variables (Fox, 2003; Lüdecke, 2018). Shown in Figure 4(a), rail% predicted values decrease as distance increases conditional on "city." Edinburgh and Phoenix have the largest 2 random intercepts, revealing an increased sensitivity to distance in those cities. Figure 4(b) gives us a direct assessment of how fixed effects "business type" and "distance" affect the rail% predicted values. If the lines were relatively horizontal, that would mean that business type has little effect on predicted rail%, as distance changes. However, we find an intriguing pattern where each business type shapes the relationship between distance and rail%. "Active Life" and "Hotels & Travel" have relatively high rail% shares, but distance is unimportant in the model. On the contrary, and perhaps unsurprisingly, when reviewing "Automotive" businesses, access to rail transit is relatively unimportant. In between, for "Shopping" and "Restaurants," the predicted rail% is primarily influenced by distance. These results show that non-work destinations are quite diverse in how they interact with a rail system, and distance to a station may or not be a significant factor in explanation a person's decision to use rail. Regardless, the "city" effect shows that urban context matters as well, and built environment factors likely play a significant role in how far people may be willing to go to use a rail system.

Table 3. Fitted GLME model summary

Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)					
Family: binomial (logit)					
Formula (in R format): rail% ~ network distance * business category + (1| city)					
Weights: the total number of transportation-content reviews					
AIC	BIC	Log likelihood	deviance	df. residual	
23244.5	23408.3	-11601.2	23202.5	18013	
Scaled residuals					
Min	1Q	Median	3Q	Max	
-6.331	-0.413	-0.252	-0.157	27.739	
Random effects					
Groups	Name	Variance	std.Dev		
City	(Intercept)	0.251	0.501		
Number of obs: 18034, groups: city, 7					
Fixed effects:					
	Estimate	std.error	statistic z value	p.value	sig
(Intercept)	-2.825	0.245	-11.544	0	***
Distance	0.097	0.104	0.931	0.352	
Arts	0.185	0.184	1.008	0.313	
Automotive	-1.645	0.283	-5.804	0	***
Health	-0.471	0.309	-1.527	0.127	
Hotels & Travel	0.382	0.159	2.395	0.017	*
Nightlife	0.172	0.190	0.905	0.365	
Other	0.029	0.187	0.155	0.877	
Restaurants	0.129	0.157	0.819	0.413	
Service	0.931	0.219	4.256	0	***
Shopping	-0.173	0.180	-0.963	0.336	
distance:Arts	-0.588	0.152	-3.859	0	***
distance:Automotive	-0.601	0.207	-2.899	0.004	**
distance:Health	-0.418	0.267	-1.570	0.116	
distance:Hotels & Travel	-0.020	0.109	-0.181	0.856	
distance:Nightlife	-1.041	0.163	-6.369	0	***
distance:Other	-0.782	0.152	-5.163	0	***
distance:Restaurants	-1.027	0.111	-9.250	0	***
distance:Service	-1.096	0.231	-4.739	0	***
distance:Shopping	-0.829	0.140	-5.908	0	***
ANOVA analysis of variance					
	df	Sum sq	Mean sq	F value	
Distance	1	161.78	161.78	161.78	
business category	9	1732.32	192.48	192.48	
distance:business category	9	477.67	53.07	53.07	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;					

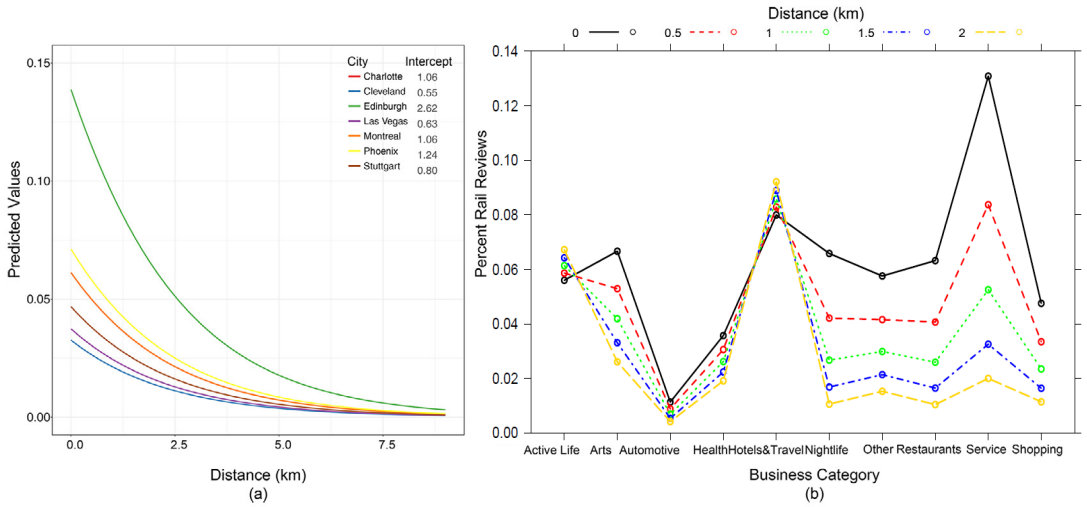

Figure 4. GLME model effect plots for predicted probabilities of random effect and fixed effects

4.4 Binomial logistic regression models for probabilities of rail and alternative modes

Results of four logit models are presented in Table 4, predicting the likelihood of driving, parking, rail, and walking terms included in a business's reviews. Many of the predictors are statistically significant, making interpretation from the tabular results difficult. As with the GLME model, we utilize an effects plot to visualize the results. Figure 5 illustrates how predicted probabilities of driving, parking, rail and walking recommendations change with increasing distance from rail stations. In general, "rail" and "walking" reviews are negatively related to distance and "drive," with the exception of Cleveland, is generally positively related or unrelated (flat) to distance from rail stations, as we might expect.

The relationship of parking to distance, however, varies markedly by city, with an explicit difference between US and non-US cities. Parking recommendations in US cities decrease as distance increases. This effect may due to the car-oriented nature of these typical US cities, where rail transit is still understood as secondary to auto-based access for non-work destinations, and finding parking remains a common preoccupation even in transit-oriented districts. However, in Montreal, Edinburgh, and Stuttgart, parking appears to be less of a concern when travelling to non-work activities, at least in the neighborhoods around rail transit relative to those beyond. For driving recommendations, only Cleveland has a slightly decreasing predicted probability as distance from rail transit increases. In Stuttgart, driving appears unimportant across all distances, and in fact as Table 1 showed, the driving-related reviews are only 2.7% of all reviews, the lowest share of the seven cities. In addition, the predicted probability curve for walking suggests a high degree of walkability in Stuttgart. In general, the results show that though rail and walking generally and consistently decrease at distances from rail stations, regardless of city, driving and parking show very mixed relationships suggesting that travelers do consistently substitute driving and parking for transit and walking, particularly in American cities.

Table 4. Binomial logistic regression model results of existence of "drive," "parking," "rail," and "walking" terms in reviews

Term: Drive				**Term: Parking**			**Term: Rail**			**Term: Walking**		
Deviance Residuals:	Min	1Q	Median 3Q Max	Min	1Q	Median 3Q Max	Min	1Q	Median 3Q Max	Min	1Q	Median 3Q Max
	-2.62	-1.08	-0.33 1.1 2.43	-1.97	-1.03	0.56 0.98 2.64	-2.03	-1.03	0.53 1.05 2.4	-1.78	-1.18	-0.41 1.10 2.23
Coefficients:	estimate	SE	statistic	estimate	SE	statistic	estimate	SE	statistic	estimate	SE	statistic
(Intercept)	-0.7	0.08	-8.42***	1.04	0.09	13.1***	1.63	0.16	10.47***	0.1	0.10	0.92
Arts	0.24	0.1	2.33*	0.64	0.11	4.2***	-0.64	0.18	-3.55***	0.59	0.13	4.52***
Automotive	3.05	0.11	28.45***	-0.74	0.08	-10.57***	-1.76	0.15	-11.34***	-1.49	0.09	-16.14***
Health	-0.2	0.08	-2.52*	-1.38	0.09	-17.78***	-2.09	0.15	-13.08***	0.14	0.09	1.43
Hotels & Travel	1.62	0.09	18.39***	0.74	0.09	7.55***	0.00	0.15	-0.01	-0.01	0.10	-0.13
Nightlife	-0.09	0.09	-0.96	-0.1	0.09	-1.45	-1.00	0.16	-6.2***	0.69	0.12	5.66***
Other	0.15	0.07	2.12*	-0.82	0.07	-13.28***	-2.06	0.13	-16.3***	-0.07	0.08	-0.89
Restaurants	0.89	0.07	13.26***	0.25	0.07	2.04*	-0.88	0.12	-7.42***	0.38	0.08	4.71***
Service	0.26	0.11	2.37*	-0.08	0.11	-1.73.	-1.31	0.20	-6.37***	-0.2	0.13	-1.48
Shopping	-0.02	0.07	-0.32	-0.65	0.07	-10.97***	-1.64	0.13	-12.82***	0.29	0.08	3.47***
distance	0.02	0.01	2.57*	-0.07	0.01	-8.29***	-0.08	0.02	-5.19***	-0.04	0.01	-4.94***
Cleveland	-0.05	0.07	-0.71	-0.49	0.07	-6.19***	-0.75	0.15	-4.96***	-0.2	0.09	-2.21**
Edinburgh	-2.14	0.12	-18.46***	-2.98	0.13	-23.16***	0.41	0.19	2.13*	0.49	0.13	3.81***
Las Vegas	0.07	0.06	1.11	-0.79	0.07	-11.04***	-0.23	0.12	-1.85.	0.28	0.08	3.33***
Montreal	-0.67	0.07	-9.36***	-1.71	0.08	-21.81***	-0.42	0.14	-3.06**	-0.4	0.09	-4.31***
Phoenix	0.48	0.06	7.73***	-0.43	0.06	-6.68***	0.11	0.12	0.93	-0.27	0.08	-3.46***
Stuttgart	-1.82	0.11	-17.19***	-2.52	0.11	-21.69***	-1.19	0.18	-6.67***	0.59	0.13	4.46***
distance:Cleveland	-0.03	0.01	-3.11**	0.04	0.01	3.01**	0.05	0.02	2.81**	0.02	0.01	1.59*
distance:Edinburgh	0.21	0.05	3.83***	0.33	0.07	5.42***	-0.75	0.14	-5.44***	-0.14	0.06	-2.37*
distance:Las Vegas	0.02	0.01	2.79**	0.08	0.01	7.95***	0.06	0.02	3.44***	0.02	0.01	2.26*
distance:Montreal	-0.01	0.02	-0.3	0.14	0.02	6.65***	-0.01	0.03	-0.18	-0.03	0.02	-1.66.
distance:Phoenix	-0.01	0.01	-1.63	0.02	0.01	2.96**	0.02	0.02	1.28	0.04	0.01	4.47***
distance:Stuttgart	-0.03	0.03	-0.89	0.18	0.03	3.83***	0.01	0.06	0.11	0.01	0.03	0.11
AIC: 45562				AIC: 42556			AIC: 12098			AIC: 30131		
Confusion Matrix and Statistics												
Accuracy	95% CI	Sensitivity	Specificity	Accuracy	95% CI	Sensitivity Specificity	Accuracy	95% CI	Sensitivity Specificity	Accuracy	95% CI	Sensitivity Specificity
0.66	(0.65, 0.68)	0.71	0.61	0.66	(0.64, 0.67)	0.67 0.65	0.67	(0.63, 0.68)	0.63 0.68	0.62	(0.60, 0.63)	0.72 0.51

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

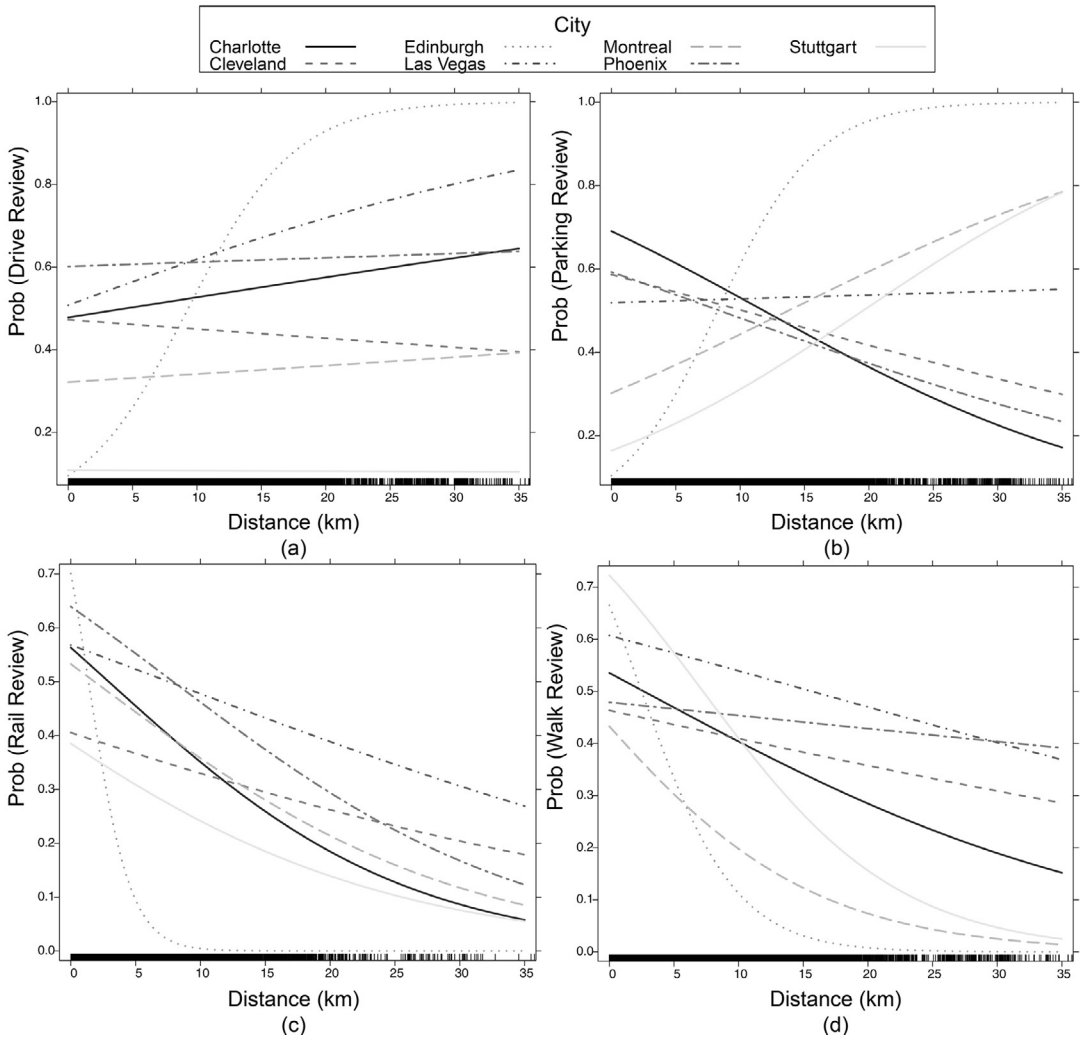

Figure 5. Predicted transportation term existence probabilities with distance change by city split

How does business type affect transportation mode? We present these effects in Figure 6, showing the correlation between the transportation terms' predicted probabilities and different business categories. "Automotive" (unsurprisingly) and "Hotels & Travel" have the highest predicted probabilities for including driving-related reviews, while "Nightlife" and "Shopping" have lower predicted probabilities of driving mentions. The probability of people recommending driving for accessing "Restaurants" is around 0.58, roughly average among business types.

Importantly, some business types reveal high levels of travel content across all modes. For example, in addition to driving, "Hotels & Travel" is also highly predictive of parking- and rail-review content. For this type of destination, access is unfamiliar and challenging, and reviewers may seek to describe means of reaching lodging at high rates across multiple modes. Rail terms are less likely to be mentioned in "automotive" and "shopping" activities, while walk terms are most frequently associated with "Nightlife", "Arts," "Restaurants" and "Shopping" businesses. In summary, "Automotive" has a tight linkage with driving and parking modes. "Hotels & Travel" is an activity that requires people to seek access by multiple transportation modes, from driving to transit. Finally, "Restaurants," the most popular non-work activity in Yelp reviews, remains more associated with driving or parking recommendations than rail and walking recommendations.

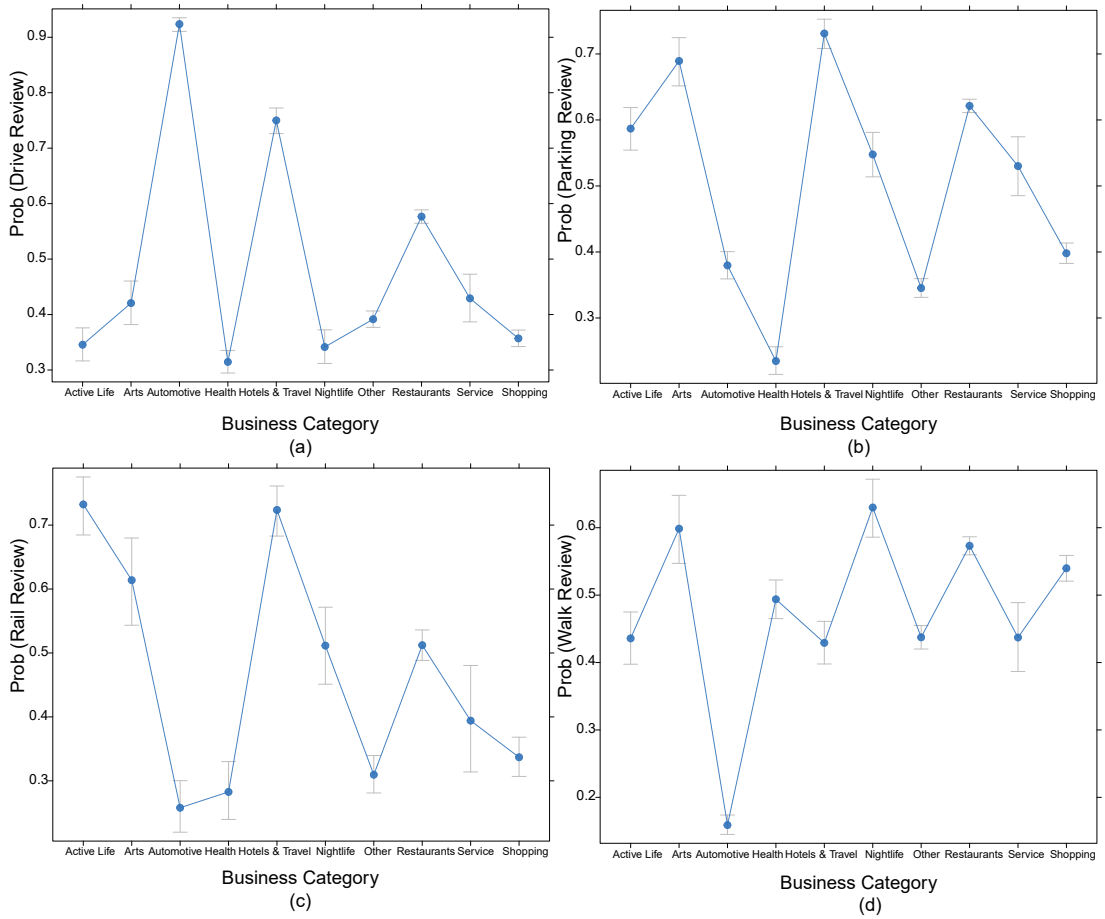

Figure 6. Predicted transportation term probabilities by business category

5 Conclusion

This investigation examines the relationship between proximity to rail transit and the relative importance of multiple travel modes for different types non-work activities. Broadly, the results show that while proximity matters, urban context and activity purpose matter too. For "everyday" activities such as shopping, services, eating out, and going to bars, distance to rail transit has a significant effect on whether people recommend it as an access mode. For other activities, such as hotels, rail transit is relatively important regardless of distance. Similarly, when examining multiple modes in terms of their relevance to a business, we observe that walking is commonly recommended near transit, and driving usually follows the opposite pattern. Parking, however, reveals a complex interaction with city and distance, where it plays its largest role in reviewer recommendations near rail in US cities, but is most relevant away from rail in the non-US cities in the sample.

The findings demonstrate that travelers are significantly sensitive to proximity when making mode choices/recommendations accessing varied types of non-work destinations. Rather than focus on a "standard" walking distance around transit, we find empirical evidence of more complex patterns of reported mode use by distance, depending on the activity type. With further validation from additional studies, the differences in mode use by activity type could be used by urban planners and designers to anticipate how particular mixes of activities in a transit-oriented commercial district are likely to be accessed, ceteris paribus. Some activities are very sensitive to distance from rail, such as restaurants and

personal services, while users of hotels and travel services, most likely due to their use by visitors, are far more inelastic with high rates of reported rail use. Still, the findings for parking, in particular, highlight that the presence of a rail network does not guarantee reductions in demand for auto access, and American cities in particular remain reliant on car-based trips even around transit stops. Overall, our findings here can be used as indicators of what types of activities are more responsive to rail investment, and at what distances from stations.

The mode split of reported travel in the reviews varies significantly between US and non-US cities, which is likely explained not just by factors around the destinations (businesses in the Yelp dataset), but also factors at traveler origins (people's homes or workplaces). Even with transit proximity at the destination, a lack of transit access at the trip origin may lead to more cars coming to these ostensibly transit-oriented areas. While the Yelp data do not tell us about origins directly, the inclusion of city as a random effect in the model allows us to assess not only the usage differences in transit and micro-level land use at the destination, but also suggest that regional land-use patterns result in different modal experiences in American and non-American cities. Put another way, it may be hard to start a rail trip in many US cities, even when headed to a TOD, so demand for parking remains critical for users, even when a rail station is nearby.

Yelp and similar LBSN data have limitations, including a demographically-biased set of respondents and self-reporting of travel experiences. However, LBSN data can be used within those limitations to understand how travel may vary at fine geographic scales, controlling for factors of interest such as activity purpose and regional differences. Our findings are also an example of how experiential data, in this case Yelp reviews, can be used as inputs for transportation planning research. Spatially-precise information from these data can provide insight both toward highly local, as well as regional transportation and land-use relationships. Future research could integrate additional factors, including population density, land use, employment, parking facilities, or local socioeconomic factors, to further examine what causes individuals to assign value to the modes they prioritize when going out. In addition, analysis of other datasets or Yelp data for additional cities with extensive transit networks such as New York or San Francisco, can extend the findings in this paper.

For urban planners and public transit providers, these data, or similar social media data, allow a deeper look at how travelers are using rail transit networks after they step off the train, and whether transit and walking are effective substitutes for auto-based mobility. We use this new dataset to illustrate differences travel experience across small geographic scales, and identify future research directions that could help more fully understand and anticipate trip-making associated with a new transportation and land-use patterns. Our study demonstrates methods that can be used strategically in future studies to evaluate the success or failure of sustainable planning initiatives, and our findings provide important insights into how rail transit investment has the potential to help achieve policy goals related to increased public transit preferences.

References

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science Engineering, 15*(3), 72–82. doi: 10.1109/MCSE.2013.70

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Green, P. (2017). lme4: Linear mixed-effects models using "Eigen" and S4 (Version 1.1-13). Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Boarnet, M. G., & Compin, N. S. (1999). Transit-oriented development in San Diego County: The incremental implementation of a planning idea. *Journal of the American Planning Association, 65*(1), 80–95. doi: 10.1080/01944369908976035

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. doi: 10.1080/1369118X.2012.678878

Chaniotakis, E., Antoniou, C., Aifadopoulou, G., & Dimitriou, L. (2017). Inferring activities from social media data. *Transportation Research Record: Journal of the Transportation Research Board, 2666*(1), 29–37. doi: 10.3141/2666-04

Chatman, D. G. (2008). Deconstructing development density: Quality, quantity and price effects on household non-work travel. *Transportation Research Part A: Policy and Practice, 42*(7), 1008–1030. doi: 10.1016/j.tra.2008.02.003

Chatman, D. G. (2013). Does TOD need the T? *Journal of the American Planning Association, 79*(1), 17–31. doi: 10.1080/01944363.2013.791008

Chen, C., Gong, H., Lawson, C., & Bialostozky, E. (2010). Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice, 44*(10), 830–840. doi: 10.1016/j.tra.2010.08.004

Choe, Y., Kim, J., & Fesenmaier, D. R. (2017). Use of social media across the trip experience: An application of latent transition analysis. *Journal of Travel & Tourism Marketing, 34*(4), 431–443. doi: 10.1080/10548408.2016.1182459

Chung, N., & Koo, C. (2015). The use of social media in travel information search. *Telematics and Informatics, 32*(2), 215–229. doi: 10.1016/j.tele.2014.08.005

Clifton, K. J., Currans, K. M., Cutter, A. C., & Schneider, R. (2012). Household travel surveys in context-based approach for adjusting ITE trip generation rates in urban contexts. *Transportation Research Record: Journal of the Transportation Research Board, 2307*(1), 108–119. doi: 10.3141/2307-12

Cole, T. J. (1991). *Applied logistic regression*. In D. W. Hosmer & S. Lemeshow (Eds.), New York: John Wiley & Sons. https://doi.org/10.1002/sim.4780100718

Collins, C., Hasan, S., & Ukkusuri, S. (2013). A novel transit rider satisfaction metric: Rider sentiments measured from online social media data — National Center for Transit Research. *Journal of Public Transportation, 16(*2), 21–45.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science, 40*(2), 130–139. doi: 10.1080/15230406.2013.777137

Curtin, K. M. (2007). Network analysis in geographic information science: Review, assessment, and projections. *Cartography and Geographic Information Science, 34*(2), 103–111. doi: 10.1559/152304007781002163

Czepiel, S. A. (2002). *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*. Retrieved from czep. net/stat/mlelr.pdf

Dittmar, H., & Ohland, G. (2012). *The new transit town: Best practices in transit-oriented development.*

Washington, DC: Island Press.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2017). *Sensitivity and specificity of information criteria.* Manuscript submitted for publication. doi.org/10.7287/peerj.preprints.1103v3

ESRI. (2018). *ArcGIS network analyst.* Retrieved from https://desktop.arcgis.com/en/arcmap/latest/extensions/network-analyst/what-is-network-analyst-.htm

Evans, L., & Saker, M. (2017). *Location-based social media: Space, time and identity.* New York: Springer.

Ewing, R., Tian, G., Lyons, T., & Terzano, K. (2017). Trip and parking generation at transit-oriented developments: Five US case studies. *Landscape and Urban Planning, 160*, 69–78. doi: 10.1016/j.landurbplan.2016.12.002

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models, second edition*. Boca Raton, FL: CRC Press.

Forrest, T., & Pearson, D. (2005). Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. *Transportation Research Record: Journal of the Transportation Research Board, 1917*, 63–71. doi: 10.3141/1917-08

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27.

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 211–220). New York: ACM. doi: 10.1145/1518701.1518736

Greenwald, M. J., & Boarnet, M. G. (2001). *The built environment as a determinant of walking behavior: Analyzing non-work pedestrian travel in Portland, Oregon*. Retrieved from https://escholarship.org/uc/item/9gn7265f#metrics

Gruen, V. (1964). *The heart of our cities: The urban crisis, diagnosis and cure*. New York: Simon and Schuster.

Guerra, E., Cervero, R., & Tischler, D. (2012). Half-mile circle. *Transportation Research Record: Journal of the Transportation Research Board, 2276*, 101–109. doi: 10.3141/2276-12

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 63–76. doi: 10.1177/0002716215570866

Hargittai, E. (2018). Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 089443931878832. doi: 10.1177/0894439318788322

Higuchi, K. (2012). Quantitative content analysis or text mining by KH Coder. Retrieved from https://sourceforge.net/p/khc/wiki/KWIC%20Concordance/

Higuchi, K. (2014). *KH Coder* (Version 2.00 beta. 32). Retrieved from http://khcoder.net/en/

Hoback, A., Anderson, S., & Dutta, U. (2008). True walking distance to transit. *Transportation Planning and Technology, 31*(6), 681–692. doi: 10.1080/03081060802492785

Hong, A., Boarnet, M. G., & Houston, D. (2016). New light rail transit and active travel: A longitudinal study. *Transportation Research Part A: Policy and Practice, 92*, 131–144. doi: 10.1016/j.tra.2016.07.005

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. Hoboken, NJ: Wiley.

Hu, X., & Liu, H. (2012). *Text analytics in social media. In Mining text data* (pp. 385–414). New York: Springer. doi: 10.1007/978-1-4614-3223-4_12

Ignatow, G., & Mihalcea, R. (2016). *Text mining: A guidebook for the social sciences*. Thousand Oaks, CA: SAGE Publications.

Jockers, M. (2014). *Text analysis with R for students of literature.* New York: Springer.

Kelley, M. J. (2013). *The emergent urban imaginaries of geosocial media. GeoJournal, 7*8(1), 181–203. doi: 10.1007/s10708-011-9439-1

King, G., & Zeng, L. (2001). Explaining rare events in international relations. *International Organization, 55*(3), 693–715. doi: 10.1162/00208180152507597

Kovacs-Györi, A., Ristea, A., Kolcsar, R., Resch, B., Crivellari, A., & Blaschke, T. (2018). Beyond spatial proximity—classifying parks and their visitors in London based on spatiotemporal and sentiment analysis of Twitter data. ISPRS International Journal of Geo-Information, 7(9), 378. doi: 10.3390/ijgi7090378

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: SAGE.

Kurkcu, A., Ozbay, K., & Morgul, E. F. (2016). Evaluating the usability of geo-located Twitter as a tool for human activity and mobility patterns: A case study for New York City. In *TRB 95th Annual Meeting Compendium of Papers.* Retrieved from https://trid.trb.org/view/1393445

Lierop, D., Maat, K., & El-Geneidy, A. (2017). Talking TOD: Learning about transit-oriented development in the United States, Canada, and the Netherlands. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 10*(1), 49–62. doi: 10.1080/17549175.2016.1192558

Litman, T. (2011). *Evaluating accessibility for transportation planning: Measuring people's ability to reach desired goods and activities*. Victoria, BC: Victoria Transport Policy Institute.

Lüdecke, D. (2018). sjPlot-package: Data visualization for statistics in social science in sjPlot. Retrieved from https://CRAN.R-project.org/package=sjPlot

Manca, M., Boratto, L., Morell Roman, V., Martori i Gallissà, O., & Kaltenbrunner, A. (2017). Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. *Online Social Networks and Media, 1*, 56–69. doi: 10.1016/j.osnem.2017.04.002

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). Minneapolis: University of Minnesota Press. doi: 10.5749/minnesota/9780816677948.003.0047

Markov, Z., & Larose, D. T. (2007). *Data mining the web: Uncovering patterns in Web content, structure, and usage.* Hoboken, NJ: John Wiley & Sons.

Mcculloch, C., & Neuhaus, J. (2001). *Generalized linear mixed models.* Hoboken, NJ: John Wiley & Sons.

Mjahed, L. B., Mittal, A., Elfar, A., Mahmassani, H. S., & Chen, Y. (2017). Exploring the role of social media platforms in informing trip planning. *Transportation Research Record: Journal of the Transportation Research Board, 2666*, 1–9. doi: 10.3141/2666-01

Mondschein, A. (2015). Five-star transportation: Using online activity reviews to examine mode choice to non-work destinations. *Transportation, 42*(4), 707–722. doi: 10.1007/s11116-015-9600-7

Munar, A. M., & Jacobsen, J. K. S. (2013). Trust and involvement in tourism social media and web-based travel information sources. *Scandinavian Journal of Hospitality and Tourism, 13*(1), 1–19. doi: 10.1080/15022250.2013.764511

Murray, A. T., Davis, R., Stimson, R. J., & Ferreira, L. (1998). Public transportation access. T*ransportation Research Part D: Transport and Environment, 3*(5), 319–328. doi: 10.1016/S1361-9209(98)00010-8

Nelson, D., & Niles, J. (1999). Essentials for transit-oriented development planning: Analysis of non-work activity patterns and a method for predicting success. *Proceedings of the 7th TRB Conference on the Application of Transportation Planning Methods,* Boston, Massachusetts. Retrieved from http://docs.trb.org/00939750.pdf

Nikšič, M., Campagna, M., Massa, P., Caglioni, M., & Nielsen, T. (2017). Opportunities for volunteered geographic information use in spatial planning. In *Mapping and the citizen sensor* (pp.

327–349). London: Ubiquity Press. Retrieved from https://www.ubiquitypress.com/site/chapters/10.5334/bbf.n/

Noland, R. B., Weiner, M. D., DiPetrillo, S., & Kay, A. I. (2017). Attitudes towards transit-oriented development: Resident experiences and professional perspectives. *Journal of Transport Geography, 60*, 130–140. doi: 10.1016/j.jtrangeo.2017.02.015

Olszewski, P., & Wibowo, S. (2005). Using equivalent walking distance to assess pedestrian accessibility to transit stations in Singapore. *Transportation Research Record: Journal of the Transportation Research Board, 1927*, 38–45. doi: 10.3141/1927-05

O'Sullivan, S., & Morrall, J. (1996). Walking distances to and from light-rail transit stations. *Transportation Research Record: Journal of the Transportation Research Board, 1538*, 19–26. doi: 10.3141/1538-03

Quantcast. (2017). Yelp audience insights and demographic analytics. Retrieved from https://www.quantcast.com/yelp.com/demographics/WEB?country=US

R Core Team. (2017). *R: A language and environment for statistical computing.* Retrieved from https://www.r-project.org/

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behavior: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies, 75*, 197–211. doi: 10.1016/j.trc.2016.12.008

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532–538). New York: Springer. doi: 10.1007/978-0-387-39940-9_565

Reilly, M. K., O'Mara, M. P., & Seto, K. C. (2009). From Bangalore to the Bay Area: Comparing transportation and activity accessibility as drivers of urban growth. *Landscape and Urban Planning, 92*(1), 24–33. doi: 10.1016/j.landurbplan.2009.02.001

Rissel, C., Curac, N., Greenaway, M., & Bauman, A. (2012). Physical activity associated with public transport use—a review and modelling of potential benefits. I*nternational Journal of Environmental Research and Public Health, 9*(7), 2454–2478. doi: 10.3390/ijerph9072454

Rybarczyk, G., Banerjee, S., Starking-Szymanski, M. D., & Shaker, R. R. (2018). Travel and us: The impact of mode share on sentiment using geo-social media and GIS. *Journal of Location Based Services, 12*(1), 40–62. doi: 10.1080/17489725.2018.1468039

Sedera, D., Lokuge, S., Atapattu, M., & Gretzel, U. (2017). Likes—the key to my happiness: The moderating effect of social influence on travel experience. *Information & Management, 54*(6), 825–836. doi: 10.1016/j.im.2017.04.003

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science, 31*(1), 139–167. doi: 10.1080/13658816.2016.1189556

Statistics Canada. (2016). Data products, 2016 Census. Retrieved from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics — challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168. doi: 10.1016/j.ijinfomgt.2017.12.002

Stopher, P., Jiang, Q., & FitzGerald, C. (2005). Processing GPS data from travel surveys. *Australasian Transport Research Forum (ATRF), 28th, 2005, Sydney, New South Wales, Australia*.

Stopher, P. R., & Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice, 41*(5), 367–381. doi: 10.1016/j.tra.2006.09.005

Trajdos, P., & Kurzynski, M. (2018). Weighting scheme for a pairwise multi-label classifier based on the fuzzy confusion matrix. *Pattern Recognition Letters, 103*, 60–67. doi:org/10.1016/j.patrec.2018.01.012

Tung, E. (2015, September 2). Automatically categorizing Yelp businesses. Retrieved from https://engineeringblog.yelp.com/2015/09/automatically-categorizing-yelp-businesses.html

U.S. Census Bureau. (2016). US Census Bureau: American community survey, 2016 5-year estimates. Suitland, MD: US Census Bureau.

Vinithra, S. N., Selvan, S. J. A., Kumar, M. A., & Soman, K. P. (2015). Simulated and self-sustained classification of Twitter data based on its sentiment. *Indian Journal of Science and Technology, 8*(24), 1–7. doi: 10.17485/ijst/2015/v8i24/80205

Vuchic, V. R. (2017). *Urban transit: Operations, planning and economics*. Hoboken, NJ: J. Wiley & Sons.

Walle, S., & Steenberghen, T. (2006). Space and time related determinants of public transport use in trip chains. *Transportation Research Part A: Policy and Practice, 40*(2), 151–162. doi: 10.1016/j.tra.2005.05.001

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, 26*(1), 97–107. doi: 10.1109/TKDE.2013.109

Xiang, Z., & Gretzel, U. (2010). Role of social media in online travel information search. *Tourism Management, 31*(2), 179–188. doi: 10.1016/j.tourman.2009.02.016

Yelp. (2017a). API 2.0: All category list. Yelp for developers. Retrieved from https://www.yelp.com/developers/documentation/v3/all_category_list

Yelp. (2017b, January). Yelp dataset challenge. Retrieved from https://www.yelp.com/dataset_challenge

Yelp. (2017c, March). Yelp factsheet. Retrieved from https://www.yelp.com/factsheet