Published in final edited form as:

Environ Plan B Urban Anal City Sci. 2017 November ; 44(6): 1145–1167. doi:

10.1177/0265813516659286

Analyzing Walking Route Choice through Built Environments using Random Forests and Discrete Choice Techniques

Calvin P. Tribby¹, Harvey J. Miller¹, Barbara B. Brown², Carol M. Werner³, and Ken R. Smith^{2,4}

¹Department of Geography, The Ohio State University

²Department of Family and Consumer Studies, University of Utah

³Department of Psychology, University of Utah

⁴Huntsman Cancer Institute, University of Utah

Abstract

Walking is a form of active transportation with numerous benefits, including better health outcomes, lower environmental impacts and stronger communities. Understanding built environmental associations with walking behavior is a key step towards identifying design features that support walking. Human mobility data available through GPS receivers and cell phones, combined with high resolution walkability data, provide a rich source of georeferenced data for analyzing environmental associations with walking behavior. However, traditional techniques such as route choice models have difficulty with highly dimensioned data. This paper develops a novel combination of a data-driven technique with route choice modeling for leveraging walkability audits. Using data from a study in Salt Lake City, Utah, USA, we apply the data-driven technique of random forests to select variables for use in walking route choice models. We estimate data-driven route choice models and theory-driven models based on predefined walkability dimensions. Results indicate that the random forest technique selects variables that dramatically improve goodness of fit of walking route choice models relative to models based on predefined walkability dimensions. We compare the theory-driven and data-driven walking route choice models based on interpretability and policy relevance.

Keywords

walkability; route choice; random forests; built environment

1. Introduction

Understanding the relationships between built environment walkability measures and walking is a vital research question due to the numerous benefits of walking. Walking has positive health benefits, such as the prevention of weight gain and chronic diseases (Berrigan et al., 2012; Lee and Buchner, 2008). Walking has positive environmental impacts because it can substitute for short car trips or be combined with transit trips, reducing pollution and individuals' contributions to climate change (Ewing and Cervero, 2010; Maibach et al., 2009). Research suggests that walking can promote neighborhood social interaction and

improves individuals' mental health (Ettema and Smajic, 2015; Miller et al., 2011; Sugiyama et al., 2008).

Until recently, analysis of *walkability* or the quality of the walking environment with respect to safety, comfort, and convenience (Litman, 2003), was typically conducted at a spatially aggregate level, such as census block groups, tax parcels, or buffers around residences (Forsyth et al., 2009; Frank et al., 2010; Saelens et al., 2012; Sallis et al., 2011; Witten et al., 2012). These area-based measures are problematic since they mask spatial variation in walkability and do not correspond to actual walking experiences (Brown et al., 2007; Houston, 2014; Isaacs, 2001; Tribby et al., 2015; Werner et al., 2010). Disaggregate, route-level analysis is more appropriate since it more closely matches walking behavior and corresponding measurements of the built environment are more specific to actual walking experiences (Gallimore et al., 2011; Harrison et al., 2014; Kelly et al., 2011; Larsen et al., 2012; Rodríguez et al., 2015).

Facilitating disaggregate analysis of walking behavior are large amounts of georeferenced data available on mobility behavior within built environments. Mobile data collection technologies, such as global positioning system (GPS) loggers and mobile phones, enable the collection of high locational precision movement data for individuals walking within a built environment. There is a concurrent increase in the availability of highly dimensioned, high spatial resolution built environment data (Nethery et al., 2015). An example is the Irvine Minnesota Inventory, a walkability instrument that researchers use to collect 162 attributes per block face in a street network (Boarnet et al., 2006; Day et al., 2006). It is straightforward to reference and integrate these mobility and built environment data within GIS. However, appropriate methods for analyzing these data are less clear.

A consideration in walkability research is the construction of theoretically-driven built environment dimensions. In general, the long list of built environment items to collect is a major barrier to widespread use and efficient analysis (Brownson et al., 2009). Furthermore, there is little consensus on which items to include in each grouping and how to summarize these items into walkability dimensions (Brown et al., 2009; Yamada et al., 2012). This presents the problem that multiple configurations of individual items can produce the same dimension summary values, which holds for walking route-level analysis as well (Gallimore et al., 2011). A thorough comparison is needed to gauge the explanatory value for walking route choice of individual built environment items compared to theoretically constructed dimensions.

Walking route choice models based on discrete choice theory (Ben-Akiva and Lerman, 1985; Ortúzar and Willumsen, 2001) are an effective method for relating traveler demographics, trip characteristics and route-level built environmental factors to individual route choice (Guo, 2009; Guo and Loo, 2013; Rodríguez et al., 2015). However, route choice models, like many traditional statistical methods, have difficulty with highly dimensioned data, including overfitting, multicollinearity and sensitivity to outliers (Breiman, 2001b). Furthermore, it is very challenging to explore, specify and test the large number of possible route choice variables and variable interactions available in these highly dimensioned data. An additional complicating factor is that assessing walkability at the

route-level requires methods for summarizing built environment attributes for actual or potential walking routes. Different summary methods for attributes can lead to varying results (Brownson et al., 2009).

This paper meets the challenge of leveraging highly dimensioned georeferenced built environment data and individual-level mobility data in the analysis of walking route choice. We assess the influence of built environment attributes and traveler and trip interaction terms on walking route choice in a neighborhood in Salt Lake City, Utah, USA. We use the random forest technique to select important variables and types of route choice summary measures for walking route choice models and compare these data-driven models with theory-driven models based on predefined walkability dimensions. Results indicate that the random forest technique selects variables that dramatically improve goodness of fit of walking route choice models relative to theory-driven models. We also compare the data-driven and theory-driven walking route choice models based on ease of interpretation and the ability to identify actionable policy knowledge.

2. Methodology

The data implementation steps are matching GPS-derived walking trips to the street network, generating a set of alternative routes for each trip, and summarizing the built environment attributes from the street network for all routes. With these routes, we use the technique of random forests to select important variables and summary types that best distinguish between routes chosen versus not chosen. We compare walking route choice models that use theory-driven variables and data-driven variables, including interaction terms for transit trips, night trips, and female travelers.

2.1 Data Collection and Preparation

The data for this project are from the Moving Across Places Study (MAPS) in Salt Lake City, Utah, USA. This project assesses walkability, walking behavior, transit use, and physical activity before (2012) and after (2013) a complete streets intervention that includes the construction of a new light rail line (called TRAX), new bicycle lanes, improved landscaping, and widened sidewalks (Brown et al., 2015). The data for this study are from 939 participants who wore accelerometers at least three days for 10 hours per day in 2012, the first phase of data collection. Non wear hours were defined as hours with zero accelerometer counts per minute, allowing for up to 2 minutes of 100 counts per minute, following procedures used in a national study (Troiano et al., 2008). Three days of wear has been a standard used in past research (Hart et al., 2011; Zenk et al., 2011).

As part of the research project, trained research assistants collected Irvine Minnesota Inventory (IMI) data for each street block face (both sides of the street, bounded by intersections) which we join to the GIS street network. The IMI is a walkability audit that inventories characteristics of the built environment for a street block (Boarnet et al., 2006; Day et al., 2006). There are 162 IMI items that can be organized into six walkability dimensions: i) *attractiveness* (features that enhance the pleasure of walking, such as benches and flowers); ii) *crime safety* (high levels of outdoor lighting and well-maintained buildings, and the absence of incivilities or cues of danger from crime); iii) *density* (single or

multifamily dwellings); iv) *diverse destinations* (kinds of public buildings, shops, or services and excludes unpleasant destinations); v) *pedestrian access* (ease of walking and the absence of barriers to walking), and; vi) *traffic safety* (how safe it is for pedestrians to cross the street). For a complete description of the construction of these dimensions, see Werner et al. (2010). We normalize the IMI items by removing variables with no variance, converting to z-scores and Winsorising these scores (limiting the extreme values) (Howell, 2011).

We processed the raw GPS data to identify trips, trip stages, and travel modes using an algorithm based on the mean speed, standard deviation speed and acceleration/decelerations during each trip stage (see Miller et al., 2015; Tsui and Shalaby, 2006). We select only the walking trips that are at least 5 minutes in duration, so trips are long enough to have at least one alternative route, and have origins and destinations within 100m of the study area street network. We remove trips that are predominantly along non-street paths where we did not collect IMI data.

Table 1 shows the means or proportions of the traveler demographics for our sample of trips modeled. Sample recruitment details are in (Brown et al., 2014). For study inclusion, adults (18) were eligible if they could speak English or Spanish, expected to live in the area for at least a year, could walk a few blocks, were not pregnant, and could give informed consent and provide survey data. Due to these eligibility requirements and to census boundaries that do not overlap precisely with the study area, it is difficult to evaluate representativeness. Proportionally fewer renters were in the study sample than in the area's underlying census statistics (48% sample, 59% area), given the requirement that participants had to remain in the neighborhood one year to participate in the follow-up study. However, the sample was more representative of the area in terms of gender (56% female sample, 48% area) and college education (35% sample, 16% area), but about the same for Hispanic ethnicity (25% sample, 26% area) and age (40 years old in sample, 44 years old in area).

Table 2 provides details about their walking trips, including distance summaries for transit trips (walking trips that are to/from light rail, commuter rail, or bus), night trips (9pm-6am), and alternative routes. The mean transit-related walking distance we observe is a little longer compared to walking distance to rail in other studies. For example, research finds a range of mean walking distance to rail stops from 1,000 m in Perth, Australia (Canepa, 2007) to 933 m in Portland, Oregon and San Francisco, California (Agrawal et al., 2008), to 564 m in Montreal, Canada (El-Geneidy et al., 2013). Furthermore, transit-related walking distance is dependent on suburban versus urban context; in the former context, walking distances are farther (O'Sullivan and Morrall, 1996). Salt Lake City is known to have wider streets and larger blocks than other comparable cities, potentially contributing to the longer observed walking distances (Zick et al., 2009). We assess trips by female travelers since females often have different travel preferences compared to males with respect to safety, particularly during trips at night time (Brown et al., 2007; Foster et al., 2004; Sallis et al., 2007; Suminski et al., 2005).

2.2 Implementation

The implementation steps are map matching, route choice set generation, route attribute summaries, random forest attribute selection, and route choice models. All methods used

Python (version 2.7.2) and ArcPy tools from ArcGIS (version 10.1), unless noted. Map matching turns the collected GPS data into GIS routes that correspond with the IMI street network data. The origins and destinations of the GPS data are the source for the route choice set generation, which creates a feasible set of alternative walking routes for each observed walking trip. The route attribute summaries calculate different types of block attribute aggregations for the routes, to see if the random forest algorithm determines if different summary types are more important for particular attributes. For the analysis, we first implement the random forest algorithm to determine the best variables for distinguishing between chosen and alternative routes and then estimate route choice models for theory-driven and data-driven built environment and walking associations.

Map matching observed routes—We match the collected GPS points for the walking trips to the street network, to attribute the trips with block-level walkability measures. First, the beginning and ending GPS points of the trips are input as origins and destinations. Next, we use the GPS trip point locations to select the street network nodes to use as intermediate waypoints. Salt Lake City has large street blocks (study area count=1026 blocks; mean length=154 m; standard deviation=92 m), so this method works well to produce no overlap or false spurs. With the origins, destinations, and intermediate stops, we solve the route with the shortest distance and export the route, the path turns, and the traversed network links for each route.

Alternative route choice set generation—An issue in walking route choice models is specification of the walking route choice set (Prato, 2009). Misspecification can result in spurious findings related to highly overlapping routes, improbable routes, or too few alternatives. Overlapping routes cause correlation amongst alternatives, violating the independence between choices that route choice models assume. An effective approach is to use a measure of route overlap to penalize highly overlapping routes in a choice set, specifically the Path Size attribute, by making them have less utility (Frejinger and Bierlaire, 2007). In addition, the use of space-time constraints reduces improbable route construction, so that very circuitous and unlikely routes are not included in the choice set (Kaplan and Prato, 2012).

The method we use to generate the alternative route choice set for each observed route follows the *calibrated labeling method* (Broach et al., 2010). This creates a collection of least cost paths by varying the impedance value of the network links using the linear combination of the length and a specified attribute: $L_k = \beta c_k + (1 - \beta) I_k$, where L_k is the link cost, c_k is the link attribute, I_k is the link distance, and β is the trade-off weight between the importance of the link attribute and distance, constrained between 0 and 1 (Broach et al., 2010). For each of the six IMI walkability dimensions, we start with a weight of 0 and step size of 0.1 to create a total of 60 data sets of least cost paths. We then calibrate these generated routes by removing routes that are infeasible and have a high ratio of five times the distance of the shortest path distance. We also remove identical routes from the sets and routes that have no alternative to the observed route.

Route attribute calculation and summary—The sets of input attributes that require processing are the street block level measures of the built environment to create route summaries. The block level attributes are the IMI street block measurements and space syntax measures. Space syntax measures assess the cognitive ease of movement through a street network using a variety of metrics.

To assess the association of the IMI street block items on the walking routes, we summarize the items to the route-level, because multiple street blocks compose each walking route. To capture the variation and potentially different relationships between block level IMI items and the routes, we create four summary measures for each of the 162 IMI items and the six dimensions for each route:

1. maximum value of the IMI item scores on the route
2. minimum value of the IMI item scores on the route
3. mean value of the IMI item scores on the route
4. length-weighted mean of the IMI item scores on the route

These four measures capture different aspects that the street block level IMI item may have on route-level walkability. The maximum and minimum measures capture the value of the street block item that has the highest or lowest score on the route. This is behaviorally meaningful because there might be a block that has a feature of the built environment that is very attractive or repulsive, such that other features on other blocks in the route are trivial in comparison. Averaging the block items over the route smooths these extremes; therefore, we calculate the maximum and minimum values in addition to the mean of block items and a modification of the mean to adjust for block length. This adjustment is related to the walking experience because it adjusts for the proportion of total walking exposure to the built environment item on each block.

We employed space syntax measures to assess the likelihood of a street block's inclusion on a route, using global and local measures. One research focus of space syntax contends that the topological influence of the street network is a primary predictor of human movement (e.g. Jiang and Jia, 2011; Koohsari et al., 2014; Lerman et al., 2014; Rybarczyk and Wu, 2014). The three measures we used are *metric choice*, *length-weighted metric choice*, and *angular connectivity* (Turner, 2007). Metric choice is a global measure of how important a street segment is within a study area. The study area for the space syntax analysis is the same set of street block for which we collected IMI data. The metric choice value is the summation of all the pair-wise shortest distances between street segments, with the impedance based on angular deviations between adjacent streets. If there is no angular deviation, i.e. the adjacent street segment continues straight through an intersection, the value is 0; if there is a 90° turn the value is 1 and a 180° turn is 2; other turn angles are scaled between those values (Hillier and Iida, 2005). The length-weighted choice measure is similar to the standard metric choice measure, with values weighted by the segment length. The angular connectivity is a local measure, since it is not assessing the importance of each segment to the overall network, but compares the geometric relationship of each street segment with its immediate neighbors. Specifically, this records the cumulative turn angle to

adjacent links, using the previous convention. We implemented the space syntax measures with the software DepthmapX (version 0.3b; Turner, 2004). We matched these measures to the street blocks, therefore we summarize these measures by the maximum and the mean of the block scores. We did not length-weight these measures because two of the three measures are length-weighted and the third is not length dependent.

Random forests—A technique that works well with highly dimensioned data is the *random forest* algorithm, an ensemble learning method (Breiman, 2001a). The basis of the random forest algorithm is *decision trees*, which is a method that successively uses input variables to arrive at a predicted outcome variable (Breiman, 2001a).

The random forest algorithm creates a large collection of decision trees using a random subset of variables from the entire independent variable set to create each tree. The random forest splits the input data into two sets: the training data used to create the trees and the test data used for determining model accuracy and variable importance. We implemented the random forest models with the randomForest (version 4.6-10) package in R (version 3.1.2). The key parameters to specify for the random forest model are: mtry, the number of variables to sample at each split point; ntree, the number of trees to grow; nodesize, the minimum size of the terminal nodes; and, sampsize, the size of the sample to draw. We performed a sensitivity analysis on these parameters and did not notice a significant effect on the ordering of the importance of the variables or the overall model classification accuracy, so we used the default values. We used the random forest to select the most important IMI and space syntax summary measures to identify which have the most influence in distinguishing observed from alternative routes.

We used the out-of-bag (OOB) performance estimates for ranking variable importance (Breiman, 2001a). The OOB estimates use a subset of the input data, also called the test data, which is not used in generating the random forest. The attribute importance, or utility, is calculated by comparing the change in classification accuracy of each variable computed from the training data set with the classification accuracy of the test data using a random permutation of the same variable. A larger change in this accuracy indicates the variable is more important to classifying the data. We used the rankings to select the top 20 attributes for distinguishing observed from alternative routes, which corresponds to a noticeable gap in the attribute importance, to test subsequently in a route choice model. We chose 20 because of the expected correlation between some variables will reduce our final set.

Route choice models—The route choice models used the conditional logistic regression model in Stata (version 12.1). We used this model specification to cluster robust standard errors at the participant level, to account for multiple trips per participant (Rodríguez et al., 2015). To account for route overlap in the choice sets, we modified the conditional logistic model with a Path Size factor, following (Frejinger and Bierlaire, 2007):

$U_{in} = V_{in} + \beta_{PS} \ln PS_{in} + \varepsilon_{in}$ where V_{in} is the deterministic part of the utility and ε_{in} is the random part. We used the basic formulation of the Path Size factor (PS):

$$PS_{in} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \frac{1}{\sum_{j \in C_n} \delta_{aj}}$$

where Γ_i is the set of all links in path i , l_a is the length of link a and L_i is the length of path i .

δ_{aj} is an indicator which equals 1 if link a is on path j , 0 if not. $\sum_{j \in C_n}$ is the summation of paths in choice set C_n which share a link a .

3. Results

We present two sets of route choice models to compare the model fit and interpretation between theory-driven and data-driven models. The first set of models uses the predefined six walkability dimensions to explain observed walking routes. The second set uses the random forest, data-driven built environment variables. Within both sets of models, we examine three separate interaction cases to investigate whether the effects of the built environment change with context: transit-related trips, night trips, and trips made by female travelers. We check the model specifications with a link test and assess the collinearity by the variance inflation factor and find no substantial issues for any of the models.

Table 3 shows the theory-driven six walkability dimensions base model (Model 1a); Model 2a adds an interaction term between distance and transit trip; and, Model 3a adds interaction terms between distance and the six dimensions with transit trip. In Table 4, we add interactions between night and distance (Model 4a) and between night and the base model variables (Model 5a). In Table 5, the interaction is between female traveler and distance (Model 6a) and between female and all variables (Model 7a). These models are significantly different than the intercept only null model and the restricted model of only distance and Path Size variables, based on the log likelihood ratio test and the resulting in χ^2 p values.

Across all models, the significant variables are density, traffic safety, Path Size, and distance, except Model 3a and Model 7a. Interpretation of positive coefficients is that an increase in the walkability feature, such as density, increases the probability of choosing a route and the positive coefficient for distance, such as in Model 1a (Table 3), means that slightly longer routes, in general, are more likely to be chosen. We were unable to screen recreational trips from the data, which may contribute to the preference of slightly longer routes, on average. However, in Model 2a (Table 3), the coefficient of the transit trip and distance interaction term is negative, which means that shorter routes are preferred for transit trips, compared to non-transit related trips. The interpretation of the interaction term is such that the base attribute effect plus interaction effect (when true) equals the total attribute effect. In Model 1a (Table 3), the negative sign of the traffic safety coefficient means that higher traffic safety is associated with a lower probability of choosing the route. In Model 3a (Table 3), the interaction term between diverse destinations and transit is positive and significant, meaning there is an increased likelihood of the chosen route having more diverse destinations for transit-related trips. Also in Model 3a, the coefficients of traffic safety and the traffic safety by transit interaction are negative and significant. The interaction term enhances the negative effect of traffic safety on chosen routes for transit-related trips; that is, the higher the traffic

safety the less likely the route will be chosen for a transit-related trip. Model 4a (Table 4) has the same relationships and similar coefficients as the base model and the distance with night interaction term is insignificant. Model 5a (Table 4), for night interacted with all variables, shows a significant positive interaction effect between traffic safety and night; in contrast, the relationship is consistently negative in the other models. Models 6a and 7a (Table 5), which include the female interaction terms has only two significant variable relationships, namely, lower traffic safety and positive effect of distance related to route choice for trips by female travelers. The path size variable is associated with the chosen route and is negative for all models. This implies that routes that have overlaps with alternative routes were preferred. This is not observed in automobile or bicycle route choice, but has been observed in other models of pedestrian route choice (Rodríguez et al., 2015) and transit route choice (Hoogendoorn-Lanser et al., 2005).

We retain insignificant variables because we are not using the models for predictive purposes. Rather, we retain these variables to compare their relationship to route choice between models. Furthermore, we want to test the theoretically constructed walkability dimension variables with walking route choice among the transit, night, and female traveler models, thus we keep all of the dimensions. We report Bayesian information criterion (BIC) for the models, because it provides a measure to aid in the selection of a model. Specifically, it provides a relative value to compare between models with the same functional form, on the same set of data, with the same sample size, while adjusting for the number of predictor variables (Raftery, 1995) A smaller value represents a more suitable model, compared to alternatives.

The second set of models use the random forest-derived (data-driven) built environment attributes for the independent variables. There are sixteen built environment variables, which is less than the initial twenty selected, due to the removal of some variables after screening for correlations. The removed variables are commercial land uses present, any office/service, any public space, and yellow pedestrian sign at end of block. Some of the variables identified as important by the random forest are nested, which is a result of their organization in the IMI collection form. For instance, commercial land uses present (yes/no) category had sub-questions about specific types, such as retail stores/restaurants, gas/service stations, or other commercial. In these nested cases, we select the variable identified as having the higher importance by the random forest algorithm. Other variables that the random forest algorithm did not select as important to distinguish between observed and alternative routes are the space syntax variables and the number of turns for each route.

Tables 6 and 7 provide the results from the data-driven models and all variables are selected by the random forest, except the distance, Path Size, and interaction terms, which are included in the model for conceptual reasons. In Table 6, the base model (Model 1b), has the random forest selected attributes and Model 2b includes an interaction term between distance and transit. Model 3b and 4b (Table 7) additionally report the interaction terms of distance with night and female, respectively. All of the variables chosen by the random forest algorithm are length-weighted, except how noisy is the block, which is the mean, and angled/on-street parking, which is the minimum. All these models fit better compared to the

null model intercept only and the restricted model of distance and Path Size, based on the log likelihood test and the resulting χ^2 p values.

The base model (Model 1b, Table 6) shows positive and significant relationships between offices, graffiti, and angled/on-street parking and observed walking routes. It indicates negative relationships for choosing a route for yellow pedestrian sign (at beginning of block), retail stores/restaurant, any industrial/manufacturing, other sidewalks/green belts/trails/paths, curb bulb-out, people waiting for/using services, and how noisy is the block. Model 2b (Table 6) has the same significant variables as Model 1b, except any industrial/manufacturing and curb bulb-out are insignificant; the interaction term between distance and transit is negative and significant. Model 3b and 4b (Table 7) have the same significant variables with the same signs as Model 1b, except in Model 4b, the distance and the distance by female interaction term are both insignificant.

4. Discussion

The models with the six walkability dimensions consistently indicate a positive effect of density on taking a route and a negative effect of traffic safety. The density result suggests that routes with greater housing density are preferable to routes with less density, similar to previous research on the association of density on walking route choice (Sun et al., 2014). The explanation for traffic safety may relate to the type of walking trip. If many trips are to destinations that are on higher traffic volume streets, this may lead to a lower traffic safety score of the observed route compared to alternatives; this is similar to an effect that Rodriguez et al. found in certain contexts (Rodriguez et al. 2015). We see the negative effect of traffic safety on route choice magnified when interacted with transit. This suggests that routes with lower traffic safety than alternative routes are especially likely for transit trips. The other interesting interaction term from the theory-driven models is diverse destinations and transit, which is a positive value. This suggests that the observed walking routes to transit are associated with more diverse land uses compared to the alternative routes. Lastly, we see an effect of time of day on traffic safety, where the night interaction with traffic safety is positive, the opposite sign we see in other models. This suggests participants chose streets at night that provided more safety from traffic compared to alternatives in the choice set (Lovasi et al., 2014).

The random forest attributes show that routes have higher probability of being chosen if they have higher values of offices, graffiti, and on-street parking. On first pass, these variables appear to be proxies for unmeasured variables. With further reflection, offices are a measure of land use mix, which a greater mix is associated with a change in walking route behavior (Sun et al., 2014). The presence of graffiti is similar to the positive relationship found for adolescent girls' walks and abandoned buildings (Rodríguez et al., 2015), although, graffiti, and incivilities in general, has mixed associations with walking (Foster et al., 2011; Slater et al., 2013). The positive association of on-street parking and route selection agrees with previous research on the walking benefit of on-street parking, where it provides a buffer from automobile traffic (Marshall et al., 2008). So, these specific variables are not necessarily accounting for other, unmeasured attributes since they agree with previous findings, but instead are important associations worthy of further investigation.

The negative effects of the built environment on route choice are much more numerous: yellow pedestrian sign (at beginning of block), retail stores/restaurant, any industrial/manufacturing, other sidewalks/green belts/trails/paths, curb bulb-out, people waiting for/using services, and how noisy is the block. The pedestrian environment variables, such as yellow pedestrian sign and curb bulb-out, are contrary to what we would expect since a greater presence of these items reduces the likelihood that a pedestrian chooses a route. However, some previous research finds a similar negative coefficient for pedestrian safety infrastructure, with the explanation that these items are only found adjacent to streets with a lot of traffic volume and high speeds, therefore pedestrians may be avoiding traffic instead of these items (Guo and Loo, 2013). The negative effect of noise is understandable, since people tend to avoid noisy blocks and industrial land uses (Dyck et al., 2011).

The theory-driven and data-driven route choice models tell similar stories for route choice and the built environment for some attributes. We see that overall an increase in the traffic safety dimension has a negative relationship with route choice (except for night trips) and this relationship is also in the individual variables that comprise traffic safety, such as yellow pedestrian signs or crosswalks. It is most likely that such elements are on streets with higher speeds and traffic volumes (Rodríguez et al., 2015). One dimension where the two methods diverge is the diverse destinations measure, which is generally not significant in the theory-driven models, except for the transit trip interaction term. However, in the data-driven models, the disaggregated individual measures of diverse destinations, such as plaza, park, landscaped open (in the variable interaction models); retail stores/restaurant; offices; and, any industrial/manufacturing are significant, though again, there are opposing directions to the expected relationship with route choice. These are just a few of the many destinations measured, so the insignificant ones may outweigh the significant ones. This compares with the significance of diverse destinations found in other studies, though not necessarily the same type of relationship (Frank et al., 2010; Hirsch et al., 2014; Lin and Moudon, 2010).

The data-driven models better fit this dataset than the theory-driven models (by comparing the BIC scores), although the data-driven models generate some counter-intuitive results, pointing to unexpected effects worth exploring in future research. The trade-off between interpretability and model fit is not necessarily bad, because the random forest variables agree with previous research and identify significant variables that are usually aggregated into a composite dimension. This is an advantage to the data-driven model: it provides specific features of the built environment to modify to encourage walking route selection, some of which may be surprising. For example, from the data-driven models, we see that on-street parking consistently improves the probability of choosing a route. This is an actionable item that city planning departments can implement, especially on busier streets, to encourage walking route selection. Other items however are not apparently actionable, such as graffiti: we would hardly suggest that city planners tag buildings to encourage walking. Also, there is a need to consider the context of the trip and the traveler characteristics since some of these associations may modify walking route selection. This suggests the need for more targeted research into associations revealed by the data-driven approach.

Another noteworthy result from the data-driven approach is the random forest algorithm selects length-weighted route summary types as the most influential, with only two variables

that have different summaries: the mean of the block noisiness item and the minimum of the angled/on-street parking item. This is important because it indicates that for certain built environment attributes, different route summary types have more influential associations with walking route choice than the length-weighted mean. More research into summary types for walking route attributes is needed, however.

5. Conclusion

We show that the use of random forests provide detailed results that reduce the need for predefined theoretical dimension constructs and the assumed directions of their relationships on walking activity; these assumptions are, so far, an unaddressed criticism of the field (Andrews et al., 2012). While this research is a specific application that yields interesting empirical relationships for this study area and the participants, it also provides a methodology for constructing large attribute data sets, choosing important variables, and assessing them in a confirmatory statistical model. The six dimensions of the IMI performed reasonably well and are easily understandable. However, when we compare it to the attributes identified by the random forest, we see a decrease in the BIC indicating a better model fit. The data-driven variables are slightly less interpretable than theory-driven dimensions, but in the context of existing research and background knowledge they can be understood. The data-driven modeling approach also identifies some associations that are counter-intuitive and point to additional research. Additionally, the selection of important random forest variables is made with minimal assumptions. These variables are still theoretically associated with walking because they are included in the IMI audit. However, the selection of important attributes by the random forest algorithm does not assume a direction of their relationship with route choice or the functional form of the relationship. Further research needs to compare the important variables identified by the random forest with other data mining algorithms to compare the variations in importance due to algorithm selection methods. These techniques and findings are a strong argument for this mixed methodology.

Although this research covers an expansive set of data and methods, drawing from many research fields, there are some shortcomings. First, we do not explore the interactions between built environment variables themselves. It may be that some variables are not correlated, but that their co-presence is more influential than an additive linear model captures. We think this may be overcome by some of the important variables from the random forest, but is not captured in the subsequent route choice models. In addition, the validity of dimension construction and attribute measures of the built environment, such as content validity, construct validity, and criterion validity are an ongoing concern (Brownson et al., 2009), which we address with the four route attribute summary types. Also, we do not use explicit trip purposes, though we provide interactions with transit trips and night trips to capture varying built environment effects on qualitatively different trips. Trip purpose is a difficult concept to implement as walking trips often have overlapping purposes (Handy et al., 2002). Finally, there is ongoing discussion as to the most appropriate way to operationalize measures of the built environment to assess walking. For example, our method of standardization of the individual IMI items and the six walkability dimensions may obscure the range of variation of these variables and also creates relative values that are

study area dependent. This may lead to a finding of insignificance for variables that are actually significant.

Implications for urban planning, with regards to walking, are that theoretically-driven walkability dimensions are useful for explaining walking route choice. Also, data-driven measures are useful to select important measures of the built environment from a large attribute dataset for explaining route choice. Further research is needed into its utility for explaining mode choice. One trade-off between the two approaches is that the theory-driven dimensions require less data (saving time and money) compared to the data-driven approach. Also, the theory-driven route choice models may be more generalizable than the data-driven models. Another trade-off is that theory-driven models do not fit the route choice data as well as the data-driven models. While the data-driven models increase model fit, they are slightly more difficult to interpret and require expertise in data science techniques. Planning agencies may not currently have this expertise, but the use of open source software and the support of online knowledge communities can ease the financial and technical burden of implementation. We suggest that practitioners collect the most detailed built environment and walking data given cost constraints, because the more detailed data yield better results.

Insights from modeling the associations between the built environment and walking routes provide a solid foundation for advocating for modification of the built environment to increase the utilization of little used walking routes. However, cross-sectional data are the predominant source for route choice modeling, but there are ongoing calls for longitudinal research studies (Coevery et al., 2015; Næss, 2015). Determining causality of the built environment features on route choice is a more data-intensive endeavor, requiring longitudinal data collection and a natural experiment or quasi-experimental research framework to measure how changes in the built environment change walking behavior among the same sample of people (McCormack and Shiell, 2011). However, research is starting to be published in this area with promising results (Brown et al., 2015; Hirsch et al., 2014; Sun et al., 2014), and using random forests and choice models may provide even deeper insights for modeling the causality of the built environment on walking route choice.

6. Literature Cited

- Agrawal A, Schlossberg M, Irvin K. How Far, by Which Route and Why? A Spatial Analysis of Pedestrian Preference. *Journal of Urban Design*. 2008; 13(1):81–98.
- Andrews GJ, Hall E, Evans B, Colls R. Moving beyond walkability: On the potential of health geography. *Social Science & Medicine*. 2012; 75(11):1925–1932. [PubMed: 22954467]
- Ben-Akiva, ME., Lerman, SR. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press; Cambridge, MA: 1985.
- Berrigan D, Carroll DD, Fulton JE, Galuska DA, Brown DR, Dorn JM, Armour B, Paul P. Vital signs: walking among adults - United States, 2005 and 2010. *Morbidity and Mortality Weekly Report*. 2012; 61(31):595–601. [PubMed: 22874838]
- Boarnet MG, Day K, Alfonzo M, Forsyth A, Oakes M. The Irvine–Minnesota Inventory to Measure Built Environments. *American Journal of Preventive Medicine*. 2006; 30(2):153–159.e43. [PubMed: 16459214]
- Breiman L. Random Forests. *Machine Learning*. 2001a; 45(1):5–32.
- Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*. 2001b; 16(3):199–231.

- Broach J, Gliebe J, Dill J. Calibrated Labeling Method for Generating Bicyclist Route Choice Sets Incorporating Unbiased Attribute Variation. *Transportation Research Record: Journal of the Transportation Research Board*. 2010; 2197:89–97.
- Brown BB, Werner CM, Amburgey JW, Szalay C. Walkable Route Perceptions and Physical Features Converging Evidence for En Route Walking Experiences. *Environment and Behavior*. 2007a; 39(1): 34–61.
- Brown BB, Werner CM, Amburgey JW, Szalay C. Walkable Route Perceptions and Physical Features Converging Evidence for En Route Walking Experiences. *Environment and Behavior*. 2007b; 39(1):34–61.
- Brown BB, Werner CM, Tribby CP, Miller HJ, Smith KR. Transit Use, Physical Activity, and Body Mass Index Changes: Objective Measures Associated With Complete Street Light-Rail Construction. *American Journal of Public Health*. 2015; 105(7):1468–1474. [PubMed: 25973829]
- Brown BB, Wilson L, Tribby CP, Werner CM, Wolf J, Miller HJ, Smith KR. Adding maps (GPS) to accelerometry data to improve study participants' recall of physical activity: a methodological advance in physical activity research. *British Journal of Sports Medicine*. 2014; 48(13):1054–1058. [PubMed: 24815545]
- Brown BB, Yamada I, Smith KR, Zick CD, Kowaleski-Jones L, Fan JX. Mixed land use and walkability: Variations in land use measures and relationships with BMI, overweight, and obesity. *Health and Place*. 2009; 15(4):1130–1141. [PubMed: 19632875]
- Brownson RC, Hoehner CM, Day K, Forsyth A, Sallis JF. Measuring the Built Environment for Physical Activity: State of the Science. *American Journal of Preventive Medicine*. 2009; 36(4, Supplement):S99–S123.e12. [PubMed: 19285216]
- Canepa B. Bursting the Bubble: Determining the Transit-Oriented Development's Walkable Limits. *Transportation Research Record: Journal of the Transportation Research Board*. 2007; 1992:28–34.
- van de Coevering P, Maat K, van Wee B. Multi-period Research Designs for Identifying Causal Effects of Built Environment Characteristics on Travel Behaviour. *Transport Reviews*. 2015; 35(4):512–532.
- Day K, Boarnet M, Alfonzo M, Forsyth A. The Irvine–Minnesota Inventory to Measure Built Environments: Development. *American Journal of Preventive Medicine*. 2006; 30(2):144–152. [PubMed: 16459213]
- Dyck DV, Cardon G, Deforche B, De Bourdeaudhuij I. Do adults like living in high-walkable neighborhoods? Associations of walkability parameters with neighborhood satisfaction and possible mediators. *Health & Place*. 2011; 17(4):971–977. [PubMed: 21570333]
- El-Geneidy A, Grimsrud M, Wasfi R, Tétreault P, Surprenant-Legault J. New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. *Transportation*. 2013; 41(1):193–210.
- Ettema D, Smajic I. Walking, places and wellbeing. *The Geographical Journal*. 2015; 181(2):102–109.
- Ewing R, Cervero R. Travel and the Built Environment. *Journal of the American Planning Association*. 2010; 76(3):265–294.
- Forsyth A, Michael Oakes J, Lee B, Schmitz KH. The built environment, walking, and physical activity: Is the environment more important to some people than others? *Transportation Research Part D: Transport and Environment*. 2009; 14(1):42–49.
- Foster C, Hillsdon M, Thorogood M. Environmental perceptions and walking in English adults. *Journal of Epidemiology and Community Health*. 2004; 58(11):924–928. [PubMed: 15483308]
- Foster S, Giles-Corti B, Knuijman M. Creating safe walkable streetscapes: Does house design and upkeep discourage incivilities in suburban neighbourhoods? *Journal of Environmental Psychology*. 2011; 31(1):79–88.
- Frank LD, Sallis JF, Saelens BE, Leary L, Cain K, Conway TL, Hess PM. The development of a walkability index: application to the Neighborhood Quality of Life Study. *British Journal of Sports Medicine*. 2010; 44(13):924–933. [PubMed: 19406732]
- Frejinger E, Bierlaire M. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological*. 2007; 41(3):363–378.

- Gallimore JM, Brown BB, Werner CM. Walking routes to school in new urban and suburban neighborhoods: An environmental walkability analysis of blocks and routes. *Journal of Environmental Psychology*. 2011a; 31(2):184–191.
- Gallimore JM, Brown BB, Werner CM. Walking routes to school in new urban and suburban neighborhoods: An environmental walkability analysis of blocks and routes. *Journal of Environmental Psychology*. 2011b; 31(2):184–191.
- Guo Z. Does the pedestrian environment affect the utility of walking? A case of path choice in downtown Boston. *Transportation Research Part D: Transport and Environment*. 2009; 14(5):343–352.
- Guo Z, Loo BPY. Pedestrian environment and route choice: evidence from New York City and Hong Kong. *Journal of Transport Geography*. 2013; 28:124–136.
- Handy SL, Boarnet MG, Ewing R, Killingsworth RE. How the built environment affects physical activity. *American Journal of Preventive Medicine*. 2002; 23(2):64–73. [PubMed: 12133739]
- Harrison F, Burgoine T, Corder K, van Sluijs EM, Jones A. How well do modelled routes to school record the environments children are exposed to?: a cross-sectional comparison of GIS-modelled and GPS-measured routes to school. *International Journal of Health Geographics*. 2014; 13(1):5. [PubMed: 24529075]
- Hart TL, Swartz AM, Cashin SE, Strath SJ. How many days of monitoring predict physical activity and sedentary behaviour in older adults? *International Journal of Behavioral Nutrition and Physical Activity*. 2011; 8:62. [PubMed: 21679426]
- Hillier, B., Iida, S. Network and Psychological Effects in Urban Movement. In: Cohn, AG., Mark, DM., editors. *Spatial Information Theory*. Springer; Berlin Heidelberg: 2005. p. 475–490. *Lecture Notes in Computer Science*
- Hirsch JA, Moore KA, Clarke PJ, Rodriguez DA, Evenson KR, Brines SJ, Zagorski MA, Roux AVD. Changes in the Built Environment and Changes in the Amount of Walking Over Time: Longitudinal Results From the Multi-Ethnic Study of Atherosclerosis. *American Journal of Epidemiology*. 2014; 180(8):799–809. [PubMed: 25234431]
- Hoogendoorn-Lanser S, van Nes R, Bovy P. Path Size Modeling in Multimodal Route Choice Analysis. *Transportation Research Record: Journal of the Transportation Research Board*. 2005; 1921:27–34.
- Houston D. Implications of the modifiable areal unit problem for assessing built environment correlates of moderate and vigorous physical activity. *Applied Geography*. 2014; 50:40–47.
- Howell, DC. *Fundamental statistics for the behavioral sciences*. Wadsworth Cengage Learning; Belmont, CA: 2011.
- Isaacs R. The Subjective Duration of Time in the Experience of Urban Places. *Journal of Urban Design*. 2001; 6(2):109–127.
- Jiang B, Jia T. Agent-based simulation of human movement shaped by the underlying street structure. *International Journal of Geographical Information Science*. 2011; 25(1):51–64.
- Kelly CE, Tight MR, Hodgson FC, Page MW. A comparison of three methods for assessing the walkability of the pedestrian environment. *Journal of Transport Geography*. 2011; 19(6):1500–1508.
- Koohsari MJ, Kaczynski AT, McCormack GR, Sugiyama T. Using Space Syntax to Assess the Built Environment for Physical Activity: Applications to Research on Parks and Public Open Spaces. *Leisure Sciences*. 2014; 36(2):206–216.
- Larsen K, Gilliland J, Hess PM. Route-Based Analysis to Capture the Environmental Influences on a Child's Mode of Travel between Home and School. *Annals of the Association of American Geographers*. 2012; 102(6):1348–1365.
- Lee I-M, Buchner DM. The Importance of Walking to Public Health. *Medicine & Science in Sports & Exercise*. 2008; 40(Supplement):S512–S518. [PubMed: 18562968]
- Lerman Y, Rofè Y, Omer I. Using Space Syntax to Model Pedestrian Movement in Urban Transportation Planning. *Geographical Analysis*. 2014; 46(4):392–410.
- Lin L, Moudon AV. Objective versus subjective measures of the built environment, which are most effective in capturing associations with walking? *Health & Place*. 2010; 16(2):339–348. [PubMed: 20004130]

- Litman T. Economic Value of Walkability. *Transportation Research Record: Journal of the Transportation Research Board*. 2003; 1828:3–11.
- Lovasi GS, Goh CE, Pearson AL, Breetzke G. The independent associations of recorded crime and perceived safety with physical health in a nationally representative cross-sectional survey of men and women in New Zealand. *BMJ Open*. 2014; 4(3):e004058.
- Maibach E, Steg L, Anable J. Promoting physical activity and reducing climate change: Opportunities to replace short car trips with active transportation. *Preventive Medicine*. 2009; 49(4):326–327. [PubMed: 19591861]
- Marshall W, Garrick N, Hansen G. Reassessing On-Street Parking. *Transportation Research Record: Journal of the Transportation Research Board*. 2008; 2046:45–52.
- McCormack GR, Shiell A. In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. 2011
- Miller HJ, Tribby CP, Brown BB, Smith KR, Werner CM, Wolf J, Wilson L, Oliveira MGS. Public transit generates new physical activity: Evidence from individual GPS and accelerometer data before and after light rail construction in a neighborhood of Salt Lake City, Utah, USA. *Health & Place*. 2015; 36:8–17. [PubMed: 26340643]
- Miller WD, Pollack CE, Williams DR. Healthy homes and communities: Putting the pieces together. *American Journal of Preventive Medicine*. 2011; 40(1 SUPPL. 1):S48–S57. [PubMed: 21146779]
- Næss P. Built Environment, Causality and Travel. *Transport Reviews*. 2015; 35(3):275–291.
- Nethery RC, Warren JL, Herring AH, Moore KAB, Evenson KR, Diez-Roux AV. A common spatial factor analysis model for measured neighborhood-level characteristics: The Multi-Ethnic Study of Atherosclerosis. *Health & Place*. 2015; 36:35–46. [PubMed: 26372887]
- de Ortúzar, JD., Willumsen, LG. *Modelling transport*. 3rd. J Wiley; Chichester ; New York: 2001.
- O’Sullivan S, Morrall J. Walking Distances to and from Light-Rail Transit Stations. *Transportation Research Record: Journal of the Transportation Research Board*. 1996; 1538:19–26.
- Raftery AE. Bayesian model selection in social research. *Sociological methodology*. 1995; 25:111–164.
- Rodríguez DA, Merlin L, Prato CG, Conway TL, Cohen D, Elder JP, Evenson KR, McKenzie TL, Pickrel JL, Veblen-Mortenson S. Influence of the Built Environment on Pedestrian Route Choices of Adolescent Girls. *Environment and Behavior*. 2015; 47(4):359–394. [PubMed: 25969591]
- Rybaczky G, Wu C. Examining the impact of urban morphology on bicycle mode choice. *Environment and Planning B: Planning and Design*. 2014; 41(2):272–288.
- Saelens BE, Sallis JF, Frank LD, Cain KL, Conway TL, Chapman JE, Slymen DJ, Kerr J. Neighborhood environment and psychosocial correlates of adults’ physical activity. *Medicine and Science in Sports and Exercise*. 2012; 44(4):637–646. [PubMed: 21946156]
- Sallis JF, King AC, Sirard JR, Albright CL. Perceived environmental predictors of physical activity over 6 months in adults: Activity Counseling Trial. *Health Psychology*. 2007; 26(6):701–709. [PubMed: 18020842]
- Sallis JF, Slymen DJ, Conway TL, Frank LD, Saelens BE, Cain K, Chapman JE. Income disparities in perceived neighborhood built and social environment attributes. *Health & Place*. 2011; 17(6): 1274–1283. [PubMed: 21885324]
- Slater SJ, Nicholson L, Chiqui J, Barker DC, Chaloupka FJ, Johnston LD. Walkable Communities and Adolescent Weight. *American Journal of Preventive Medicine*. 2013; 44(2):164–168. [PubMed: 23332334]
- Sugiyama T, Leslie E, Giles-Corti B, Owen N. Associations of neighbourhood greenness with physical and mental health: do walking, social coherence and local social interaction explain the relationships? *Journal of Epidemiology and Community Health*. 2008; 62(5):e9–e9. [PubMed: 18431834]
- Suminski RR, Poston WSC, Petosa RL, Stevens E, Katzenmoyer LM. Features of the neighborhood environment and walking by U.S. adults. *American Journal of Preventive Medicine*. 2005; 28(2): 149–155. [PubMed: 15710269]
- Sun G, Oreskovic NM, Lin H. How do changes to the built environment influence walking behaviors? a longitudinal study within a university campus in Hong Kong. *International Journal of Health Geographics*. 2014; 13(1):28. [PubMed: 25069949]

- Tribby CP, Miller HJ, Brown BB, Werner CM, Smith KR. Assessing built environment walkability using activity-space summary measures. *Journal of Transport and Land Use*. 2015
- Troiano RP, Berrigan D, Dodd KW, Mâsse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*. 2008; 40(1):181–188. [PubMed: 18091006]
- Tsui S, Shalaby A. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. *Transportation Research Record: Journal of the Transportation Research Board*. 2006; 1972:38–45.
- Turner, A. Bartlett School of Graduate Studies. University College London; London, UK: 2004. Depthmap 4: a researcher's handbook. <http://eprints.ucl.ac.uk/2651/>
- Turner A. From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. *Environment and Planning B: Planning and Design*. 2007; 34(3):539–555.
- Werner CM, Brown BB, Gallimore J. Light rail use is more likely on 'walkable' blocks: Further support for using micro-level environmental audit measures. *Journal of Environmental Psychology*. 2010; 30(2):206–214.
- Witten K, Blakely T, Bagheri N, Badland H, Ivory V, Pearce J, Mavoa S, Hinckson E, Schofield G. Neighborhood Built Environment and Transport and Leisure Physical Activity: Findings Using Objective Exposure and Outcome Measures in New Zealand. *Environmental Health Perspectives*. 2012; 120(7):971–977. [PubMed: 22456536]
- Yamada I, Brown BB, Smith KR, Zick CD, Kowaleski-Jones L, Fan JX. Mixed Land Use and Obesity: An Empirical Comparison of Alternative Land Use Measures and Geographic Scales. *Professional Geographer*. 2012; 64(2):157–177. [PubMed: 22665941]
- Zenk SN, Schulz AJ, Matthews SA, Odoms-Young A, Wilbur J, Wegrzyn L, Gibbs K, Braunschweig C, Stokes C. Activity space environment and dietary and physical activity behaviors: A pilot study. *Health & Place*. 2011; 17(5):1150–1161. [PubMed: 21696995]
- Zick CD, Smith KR, Fan JX, Brown BB, Yamada I, Kowaleski-Jones L. Running to the Store? The relationship between neighborhood environments and the risk of obesity. *Social Science & Medicine*. 2009; 69(10):1493–1500. [PubMed: 19766372]

Table 1

Traveler Demographics for trips modeled

Traveler Demographics (n=292)	Mean or proportion	Standard Deviation
Age (years)	40.46	15.32
Body Mass Index	28.16	6.78
College Graduate (Yes=1)	0.35	0.48
Employment (Yes=1)	0.56	0.50
Gender (Female=1)	0.56	0.50
Hispanic (Yes=1)	0.25	0.44
Obesity Status (Yes=1)	0.33	0.47
Retired (Yes=1)	0.08	0.27
White (Yes=1)	0.68	0.47
Years Residence	5.53	9.14

Table 2

Summary of observed neighborhood trip characteristics and alternative routes

Trip characteristics (n=1048 observed; n=2020 alternative routes; total route sample=3068)	n	Mean or proportion	Standard Deviation
Trips (per participant)		3.59	3.52
Trip length (m)		1005.28	706.93
Trip duration (for >5min trips)		15.29	11.20
Transit trip length (m)	101	894.83	583.54
Night trip length (m)	56	1203.47	753.43
Female trip length (m)	613	936.47	650.33
Alternative routes (per trip)		1.93	1.63
Alternative route lengths (m)		1138.02	739.46

Table 3
Estimated route choice model based on walkability dimensions with transit trip interactions

Variable Name	Model 1a (n=3068)				Model 2a (n=3068)				Model 3a (n=3068)			
	Coefficient (95% CI)	SE	p value		Coefficient (95% CI)	SE	p value		Coefficient (95% CI)	SE	p value	
Attractiveness	1.08 (-1.31, 3.47)	1.22	0.38		1.29 (-1.02, 3.60)	1.18	0.27		1.90 (-0.69, 4.49)	1.32	0.15	
Crime Safety	-1.28 (-3.33, 0.76)	1.04	0.22		-1.23 (-3.29, 0.82)	1.05	0.24		-1.41 (-3.51, 0.68)	1.07	0.19	
Density	1.32 (0.09, 2.56)	0.63	0.04		1.40 (0.16, 2.64)	0.63	0.03		1.14 (-0.13, 2.40)	0.64	0.08	
Diverse Destinations	1.43 (-1.90, 4.76)	1.70	0.40		1.61 (-1.72, 4.93)	1.70	0.34		1.00 (-2.59, 4.59)	1.83	0.59	
Pedestrian Access	-1.53 (-3.84, 0.78)	1.18	0.20		-1.56 (-3.90, 0.77)	1.19	0.19		-1.32 (-3.75, 1.11)	1.24	0.29	
Traffic Safety	-4.77 (-7.08, -2.47)	1.18	0.001		-4.51 (-6.84, -2.18)	1.19	0.001		-4.01 (-6.51, -1.51)	1.28	0.001	
Distance	0.006 (0.004, 0.007)	0.0006	0.001		0.006 (0.005, 0.008)	0.0008	0.001		0.006 (0.004, 0.007)	0.0008	0.001	
Path Size	-0.21 (-0.29, -0.13)	0.04	0.001		-0.21 (-0.30, -0.13)	0.04	0.001		-0.22 (-0.30, -0.14)	0.04	0.001	
Inter-actions												
Attractiveness* transit									-3.85 (-10.15, 2.46)	3.22	0.23	
Crime Safety* transit									0.26 (-5.63, 6.15)	3.00	0.93	
Density*transit									2.91 (-0.08, 5.90)	1.53	0.06	
Diverse Destinations* transit									12.03 (-0.20, 24.25)	6.24	0.05	
Pedestrian Access*transit									-7.57 (-17.08, 1.93)	4.85	0.12	
Traffic Safety*transit									-10.07 (-18.06, -2.08)	4.08	0.01	
Distance*transit					-0.005 (-0.007, -0.002)	0.001	0.001		-0.001 (-0.005, 0.003)	0.002	0.53	
Log likelihood null	-1024.22				-1024.22				-1024.22			
Log likelihood restricted	-863.18				-863.18				-863.18			

Variable Name	Model 1a (n=3068)			Model 2a (n=3068)			Model 3a (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Log likelihood full	-821.49			-811.03			-801.09		
Goodness of fit χ^2	<0.0001			<0.0001			<0.0001		
BIC	-5589.85			-5603.81			-5581.96		

All variables are length-weighted means, except distance. Robust standard errors (SE) are clustered at the individual level. χ^2 p value is the statistical significance of the model compared to a null model. Statistically significant variables at the 5% level are bold. Interactions are set such that condition is true (e.g., transit related=1)

Table 4

Estimated route choice model based on walkability dimensions with night trip interactions

Variable Name	Model 4a (n=3068)			Model 5a (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Attractiveness	1.08 (-1.32, 3.47)	1.22	0.38	0.87 (-1.49, 3.23)	1.20	0.47
Crime Safety	-1.28 (-3.32, 0.77)	1.04	0.22	-1.14 (-3.30, 1.01)	1.10	0.30
Density	1.33 (0.08, 2.58)	0.64	0.04	1.59 (0.35, 2.83)	0.63	0.01
Diverse Destinations	1.43 (-1.91, 4.76)	1.70	0.40	1.45 (-1.97, 4.87)	1.74	0.41
Pedestrian Access	-1.52 (-3.82, 0.78)	1.17	0.20	-1.59 (-4.1, 0.92)	1.28	0.21
Traffic Safety	-4.78 (-7.07, -2.49)	1.17	0.001	-5.10 (-7.46, -2.74)	1.21	0.001
Distance	0.006 (0.004, 0.007)	0.04	0.001	0.006 (0.004, 0.007)	0.0007	0.001
Path Size	-0.21 (-0.29, -0.13)	0.04	0.001	-0.21 (-0.29, -0.13)	0.04	0.001
Attractiveness*night	1.03 (-6.49, 8.55)	3.84	0.79			
Crime Safety*night	-1.77 (-7.23, 3.69)	2.79	0.53			
Density*night	-2.05 (-4.86, 0.76)	1.43	0.15			
Diverse Destinations*night	2.55 (-12.98, 18.08)	7.93	0.75			
Pedestrian Access*night	-0.65 (-7.06, 5.75)	3.27	0.84			
Traffic Safety*night	5.67 (0.27, 11.06)	2.75	0.04			
Distance*night	-0.0004 (-0.004, 0.003)	0.001	0.82	-0.001 (-0.005, 0.003)	0.002	0.54
Log likelihood null	-1024.22			-1024.22		
Log likelihood restricted	-863.18			-863.18		

		Model 4a (n=3068)		Model 5a (n=3068)	
		Coefficient (95% CI)	SE	Coefficient (95% CI)	SE
			p value		p value
Model fit	Variable Name				
	Log likelihood full	-821.44		-816.43	
	Goodness of fit χ^2	<0.0001		<0.0001	
	BIC	-5582.99		-5551.27	

Table 5
Estimated route choice model based on walkability dimensions with female traveler interactions

Variable Name	Model 6a (n=3068)			Model 7a (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Attractiveness	1.11 (-1.28, 3.50)	1.22	0.36	1.65 (-2.65, 5.95)	2.19	0.45
Crime Safety	-1.30 (-3.35, 0.75)	1.05	0.21	-1.42 (-4.16, 1.32)	1.40	0.31
Density	1.33 (0.09, 2.57)	0.63	0.04	1.22 (-0.69, 3.14)	0.98	0.21
Diverse Destinations	1.53 (-1.82, 4.89)	1.71	0.37	0.43 (-3.64, 4.50)	2.08	0.84
Pedestrian Access	-1.49 (-3.81, 0.82)	1.18	0.21	-1.42 (-4.56, 1.71)	1.60	0.37
Traffic Safety	-4.87 (-7.12, -2.62)	1.15	0.001	-5.45 (-8.07, -2.83)	1.34	0.001
Distance	0.005 (0.004, 0.006)	0.0008	0.001	0.005 (0.004, 0.007)	0.0008	0.001
Path Size	-0.21 (-0.29, -0.13)	0.04	0.001	-0.21 (-0.29, -0.13)	0.04	0.001
Inter-actions						
Attractiveness* female				-0.89 (-6.01, 4.22)	2.61	0.73
Crime Safety*female				0.13 (-3.96, 4.22)	2.09	0.95
Density*female				0.29 (-2.2, 2.78)	1.27	0.82
Diverse Destinations* female				2.05 (-4.35, 8.44)	3.26	0.53
Pedestrian Access*female				-0.20 (-4.83, 4.43)	2.36	0.93
Traffic Safety*female				1.01 (-3.36, 5.37)	2.23	0.65
Distance*female	0.001 (-0.002, 0.004)	0.001	0.44	0.0009 (-0.002, 0.003)	0.001	0.50
Log likelihood null	-1024.22			-1024.22		
Log likelihood restricted	-863.18			-863.18		

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Model 6a (n=3068)		Model 7a (n=3068)	
Variable Name	Coefficient (95% CI)	SE	p value
Model fit	Log likelihood full	-819.75	
	Goodness of fit χ^2	<0.0001	
	BIC	-5544.63	

Table 6
Estimated route choice model based on random forest selected walkability attributes with transit trip interactions

Variable Name	Model 1b (n=3068)			Model 2b (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Zebra striping (at end of block)	0.39 (-1.35, 2.13)	0.89	0.66	0.39 (-1.16, 1.95)	0.79	0.62
Yellow pedestrian sign (at beginning of block)	-2.81 (-4.68, -0.95)	0.95	0.001	-2.91 (-4.73, -1.10)	0.93	0.001
Plaza, park, landscaped open	-1.50 (-3.79, 0.78)	1.16	0.20	-1.63 (-4.00, 0.74)	1.21	0.18
Retail stores/restaurant	-4.14 (-6.04, -2.24)	0.97	0.001	-4.10 (-5.97, -2.22)	0.96	0.001
Commercial other	0.77 (-1.01, 2.54)	0.91	0.40	0.66 (-0.87, 2.19)	0.78	0.40
Gas/service station	-1.28 (-3.91, 1.35)	1.34	0.34	-1.17 (-3.82, 1.48)	1.35	0.39
Offices	3.22 (0.91, 5.54)	1.18	0.01	3.21 (1.04, 5.37)	1.11	0.001
Any industrial/manufacturing	-1.92 (-3.73, -0.10)	0.93	0.04	-1.86 (-3.76, 0.04)	0.97	0.06
River (barrier)	-1.87 (-6.17, 2.42)	2.19	0.39	-2.02 (-6.65, 2.61)	2.36	0.39
Other sidewalks/green belts/trails/paths	-1.17 (-2.32, -0.02)	0.59	0.05	-1.22 (-2.37, -0.08)	0.58	0.04
Mid-block crossing	0.89 (-0.24, 2.01)	0.58	0.12	0.84 (-0.24, 1.92)	0.55	0.13
Graffiti	1.17 (0.33, 2.00)	0.43	0.01	1.15 (0.32, 1.98)	0.42	0.01
Curb bulb-out	-3.75 (-7.56, 0.06)	1.94	0.05	-3.59 (-7.29, 0.11)	1.89	0.06
People waiting for/using services on block	-3.83 (-5.58, -2.08)	0.89	0.001	-3.73 (-5.34, -2.11)	0.82	0.001
How noisy is block	-6.71 (-9.02, -4.41)	1.18	0.001	-6.75 (-8.94, -4.56)	1.12	0.001
Angled/on-street parking	4.82 (2.87, 6.77)	1.00	0.001	4.99 (2.90, 7.08)	1.07	0.001

Random Forest selected variables

Variable Name	Model 1b (n=3068)			Model 2b (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Distance	0.004 (0.0007, 0.006)	0.002	0.02	0.004 (0.0006, 0.007)	0.002	0.02
Path Size	-0.32 (-0.50, -0.13)	0.10	0.001	-0.32 (-0.50, -0.14)	0.09	0.001
Inter-action						
Distance*transit				-0.008 (-0.02, 0.007)	0.00002	0.05
Model fit						
Log likelihood null	-1024.22			-1024.22		
Log likelihood restricted	-863.18			-863.18		
Log likelihood full	-95.21			-93.83		
Goodness of fit χ^2	<0.0001			<0.0001		
BIC	-6972.85			-6968.66		

All variables are length-weighted means, except distance, how noisy is block (mean) and angled/on-street parking (minimum). Robust standard errors (SE) are clustered at the individual level. χ^2 p value is the statistical significance of the model compared to a null model. Statistically significant variables at the 5% level are bold. Interactions are set such that condition is true (i.e., transit related, night, and female traveler=1).

Table 7

Estimated route choice model based on random forest selected walkability variables with night time and female interactions

Variable Name	Model 3b (n=3068)			Model 4b (n=3068)		
	Coefficient (95% CI)	SE	p value	Coefficient (95% CI)	SE	p value
Zebra striping (at end of block)	0.36 (-1.39, 2.11)	0.89	0.69	0.38 (-1.35, 2.12)	0.89	0.66
Yellow pedestrian sign (at beginning of block)	-2.86 (-4.78, -0.93)	0.98	0.00	-2.81 (-4.67, -0.95)	0.95	0.00
Plaza, park, landscaped open	-1.51 (-3.81, 0.79)	1.17	0.20	-1.49 (-3.75, 0.77)	1.15	0.20
Retail stores/restaurant	-4.18 (-6.09, -2.27)	0.97	0.00	-4.14 (-6.03, -2.25)	0.97	0.00
Commercial other	0.79 (-1, 2.57)	0.91	0.39	0.76 (-1.02, 2.54)	0.91	0.40
Gas/service station	-1.29 (-3.93, 1.35)	1.35	0.34	-1.27 (-3.88, 1.33)	1.33	0.34
Offices	3.25 (0.90, 5.60)	1.20	0.01	3.21 (0.90, 5.53)	1.18	0.01
Any industrial/ manufacturing	-1.93 (-3.76, -0.10)	0.93	0.04	-1.92 (-3.73, -0.11)	0.92	0.04
River (barrier)	-1.88 (-6.24, 2.47)	2.22	0.40	-1.89 (-6.21, 2.44)	2.21	0.39
Other sidewalks/green belts/trails/paths	-1.14 (-2.3, 0.01)	0.59	0.05	-1.17 (-2.33, -0.02)	0.59	0.05
Mid-block crossing	0.90 (-0.24, 2.04)	0.58	0.12	0.88 (-0.25, 2.02)	0.58	0.13
Graffiti	1.13 (0.31, 1.95)	0.42	0.01	1.17 (0.33, 2.02)	0.43	0.01
Curb bulb-out	-3.81 (-7.62, -0.001)	1.94	0.05	-3.81 (-7.7, 0.07)	1.98	0.05
Waiting for/using services	-3.82 (-5.59, -2.06)	0.90	0.00	-3.82 (-5.57, -2.07)	0.89	0.00
How noisy is block	-6.80 (-9.19, -4.41)	1.22	0.00	-6.72 (-9.01, -4.42)	1.17	0.00
Angled/on-street parking	4.84 (2.87, 6.81)	1.01	0.00	4.82 (2.88, 6.75)	0.99	0.00

Variable Name	Model 3b (n=3068)		Model 4b (n=3068)	
	Coefficient (95% CI)	SE	Coefficient (95% CI)	SE
Distance	0.003 (0.0006, 0.006)	0.001	0.003 (-0.001, 0.008)	0.002
Path Size	-0.31 (-0.50, -0.13)	0.10	-0.32 (-0.50, -0.13)	0.10
Distance*night	0.009 (0.002, 0.02)	0.004	0.01	
Distance*female			0.004 (-0.005, 0.006)	0.003
Log likelihood null	-1024.22		-1024.22	
Log likelihood restricted	-863.18		-863.18	
Log likelihood full	-94.54		-95.19	
Goodness of fit χ^2	<0.0001		<0.0001	
BIC	-6967.24		-6965.94	