Modeling the impacts of park access on health outcomes: a utility-based accessibility approach

Gregory Macfarlane[a], Nico Boyd[b], John E. Taylor[c], Kari Watkins[c]

[a]*Civil and Environmental Engineering Department, Brigham Young University 430 Engineering Building, Provo, UT, 84602*
[b]*Los Angeles, California*
[c]*School of Civil and Environmental Engineering, Georgia Institute of Technology, 755 Atlantic Drive, Atlanta, GA, 30332*

Abstract

Recent research has underscored the potential for public green spaces to influence individual and societal health outcomes, but empirical measurements of such influences have yielded mixed results to date, with particular disagreement surrounding how access to parks ought to be defined while controlling alternate explanations. In this paper we present a comprehensive measure of park accessibility drawn from random utility choice theory, which avoids arbitrary assertions of proximity while incorporating potentially numerous amenities and attributes of both the parks and the population. We apply this choice-based accessibility measure to correlate Census tract-level obesity and physical activity rate estimates from the Centers for Disease Control and Prevention's 500 Cities project with tract-level American Community Survey socioeconomic data in New York City, paired with geographic parks data from New York City. Controlling for the socioeconomic variables and spatially correlated error terms, we show a positive and significant relationship between park access and physical activity rates, and a clear and significant negative relationship between park access and obesity rates. In doing so, this research contributes a more comprehensive modeling approach for measuring the impact of park access on health, and may improve our understanding of the role parks and access to them can serve in furthering public health objectives.

Introduction

The United States and other developed nations face an epidemic of obesity and chronic diseases, including cardiovascular diseases, chronic respiratory diseases, diabetes, stroke, joint and bone diseases, and cancer. These diseseases can severely limit the lifespan of affected individuals (World Health Organization, 2014), with substantial financial costs for treatment borne both privately and socially (Finkelstein et al., 2009). Though a moderate amount of regular physical

Email addresses: gregmacfarlane@byu.edu (Gregory Macfarlane), jetaylor@gatech.edu (John E. Taylor), kariwatkins@gatech.edu (Kari Watkins)

activity has been established as an effective strategy for reducing and managing obesity and many of the aforementioned chronic diseases (Centers for Disease Control and Prevention, 2009; Durstine et al., 2013), a large portion of U.S. adults do not participate in sufficient physical activity (Wolf, 2008).

Historically, most large-scale health promotion efforts focused on individual-level interventions intended to educate people about healthy lifestyles and behaviors, touching on topics including diet and exercise. Over the last several years a new social model of health has evolved, treating health “as an outcome of the effects of socioeconomic status, culture, environmental conditions, housing, employment and community influences.”(Duhl and Sanchez, 1999, p. 7) In this new paradigm, resources provided via civil infrastructure — in particular parks and other public green spaces — play a crucial role in promoting and sustaining public health (Bedimo-Rung et al., 2005; Wells2007; Coutts, 2008). After all, it does little to lecture individuals on the importance of exercise if their community is built in such a way that such exercise is expensive, unenjoyable, or unsafe.

In spite of the potential for green spaces to influence public health outcomes and numerous studies attempting to empirically estimate the strength of these outcomes, evidence to date has been mixed. A major challenge researchers have faced is a proliferation of techniques for measuring access to green spaces, of which many are plausible but lack robust theoretical basis. In this paper, we present a holistic and flexible measurement for park accessibility based in activity location choice theory. This utility-based accessibility measure compares the continuous distance to all parks in the region, weighted against the sizes of each park and its other assorted amenities.

We apply this measurement to study the link between park accessibility and attractiveness and Census tract-level aggregate physical activity participation and obesity rates in New York City, controlling for spatially correlated unobserved effect and tract-level socioeconomic characteristics. We find a positive relationship where the least park-accessible tracts have an expected physical activity participation rate roughly a full percentage point lower than the most accessible tracts; a traditional quarter-mile buffer analysis estimates a similar albeit marginally less significant effect. However, we also find a strong and significant correlation between the choice-based park accessibility metric and decreased obesity rates — again controlling for spatial correlation, socioeconomic attributes, and physical activity rates. In this case a traditional buffer analysis does not show significance.

We also demonstrate how the utility-based accessibility metric can potentially be expanded to account for other measures of a park’s attractiveness including its observed social media activity and the presence of various park amenities. The paper concludes with a discussion of opportunities for future research.

Literature Review

The large majority of extant research on the topic of park access and public health outcomes has been involved comparing metropolitan regions against each other. For example, a study by West et al. (2012) used park data from the

Trust for Public Land's 2010 City Park Facts and public health data from the Behavioral Risk Factor Surveillance System (BRFSS) to examine the relationship between the density of park land and physical activity and obesity rates for 67 metropolitan statistical areas in the U.S. The findings in this case conformed to expectation; that is, the study found a significant positive association between park density and physical activity and a significant negative association between park density and obesity. Larson et al. (2016) similarly used self-reported scores on the Gallup-Healthways Wellbeing Index to evaluate the relationship between physical health and park quantity, quality, and accessibility in 44 U.S. cities. While the authors found positive associations between wellbeing and park quality and accessibility, these relationships were not statistically significant. Conversely, Richardson et al. (2012) examined the relationship between total urban greenspace and mortality rates for selected maladies; though the authors did not find a statistically significant relationship between the quantity of urban greenspace and mortality caused by lung cancer, diabetes, heart disease or car accidents, they did find that all-cause mortality was oddly, *higher* in greener cities. Finally, in a meta-analysis of 20 peer reviewed journal articles exploring the relationship between parks and objectively measured physical activity, Bancroft et al. (2015) found that five studies reported a significant positive association between the two, six studies produced mixed results, and nine studies found no association at all.

Metropolitan-level analyses such as those described above do not capture the within-city variation in accessibility to parks that exist in many cities. Some cities with large amounts of total greenspace may nevertheless unequally distribute this space throughout the city, leading to areas with poor access. Conversely, a city with a smaller overall proportion of greenspace may give all of its citizens better access to sufficient greenspace to meet their needs or wants for physical activity. Considering access at the neighborhood level within a single city may also eliminate some regional or cultural fixed effects affecting metropolitan-level analyses.

Park Accessibility Measures

Loosely defined, the *accessibility* of an arbitrary place describes the ease with which people can accomplish activities there. Accessibility is an abstract concept with tempting quantifiability (Handy and Niemeier, 1997), perhaps explaining the proliferation of calculation mechanisms. Dong et al. (2006) present a helpful mathematical heirarchy of common mechanisms, which we follow here. Consider a person residing at point i in a city with parks $j \in 1 \dots J$. An analyst might consider point i as "having access" to a park if the distance to the park d_ij within an "isochrone," or less than some threshold D,

$$A_i = \begin{cases} 1 & d_{ij} \leq D \\ 0 & d_{ij} > D \end{cases} \tag{1}$$

Variations of this isochrone-based framework are easily derived and relatively common. Dias et al. (2019) considered road safety perception as an extenuating factor in the effective distance to a park. ParkScore (Trust for Public Land, 2019) is a sophisticated application of this approach where d_{ij} is a carefully calculated 10-minute network-bound walk, with a demographic analysis of areas within and without this threshold. In other applications D might be more variably defined, such as the presence or percent of green space within a political or statistical boundary (Mitchell and Popham, 2008). The indicator function could also be modified into the total number of parks or percent of park space within radius D of point i (**???**). In a study of neighborhoods in New York City, Stark et al. (2014) found that a higher proportion of park space in a neighborhood was associated with a lower BMI for its residents, and lower park cleanliness scores were associated with a higher BMI, suggesting that increasing the supply of clean, well-maintained parks could positively impact adult BMI.

In spite of the flexibility of adapted isochrone techniques, the arbitrary definitions necessarily imposed by researchers in its application can inhibit holistic analysis (Logan et al., 2019). A somewhat more complete approach is the so-called "gravity" accessibility statistic. In this case the accessibility of a point is the denominator of the gravity formulation of a trip distribution model,

$$A_i = \sum_{j=1}^{J} S_j f(d_{ij}) \tag{2}$$

where S_j is the "size" of each park literally (i.e. acres) or abstractly (i.e. trip attractions) and $f(d_ij)$ is a monotonically decreasing cost function, usually a negative exponential. Dong et al. (2006) note that the isochrone framework is a special case of the gravity model where $f(d_{ij})$ is a binary function. Giles-Corti et al. (2005) compare various gravity-based accessibility scores with an isochrone specification and show that the former are more predictive of walking behavior; that is, the attributes of a park matter influence walking more than merely the distance to it. Zhang et al. (2011) present a population-weighted gravity-based accessibility to parks metric with a national rather than metropolitan scope, though they do not examine its correlation with health outcomes.

A third accessibility mechanism is a utility-based specification, termed as such for being derived from random utility choice theory. Consider that an individual living at point i is choosing a park for a recreation activity. If we apply the multinomial logit model (McFadden, 1974), the expected consumer surplus enjoyed by this individual can be shown to be

$$A_i = \ln\left(\sum_{j=1}^{J} \exp(V_{ij})\right) + C \tag{3}$$

where parks are differentiated from each other by their relative measurable *utilities* V_{ij}. C is an unknown constant, but the difference in consumer surplus between two points i and k can be quantitatively compared as $A_i - A_k$ (Bruce, 1977). In principle, V may include any measurable attributes of either the choice maker or the park, and is typically represented as a linear-in-parameters function of destination attributes X_j and the travel cost d_{ij},

$$V_{ij} = X_j\beta + \beta_d * d_{ij} \tag{4}$$

The coefficients $\boldsymbol{\beta}$ are frequently estimated from household surveys, though in the absence of a survey we may assert reasonable values.

Note that the gravity formulation is itself a special case of the utility-based specification where $\exp(X_j\beta + \beta_d d_{ij}) = X_j f(d_{ij})$ (Daly, 1982). The distinction between gravity and utility-based specification is meaningful, however. Primarily, it becomes possible to construct accessibility statistics based on revealed behavioral preferences rather than calibrated or asserted values (Handy and Niemeier, 1997).

Both gravity and utility-based specifications hold several advantages relative to isochrone-based accessibility metrics more commonly found in the literature. First, all individuals are defined as having some access to all parks, rather than an arbitrary cutoff asserted by the researcher. This allows for the fact that some people are more or less sensitive to distances, and that distance is a continuous, non-binary phenomenon. It defies reason to assume a person living 1.1 miles from a park has functionally different accessibility than someone living at 0.9 miles. Second, the random utility formulation allows the researcher to include – in principle – any attribute of the park as part of its utility specification. This suggests that not all parks are equal, and that a large park with many amenities such as Central Park in Manhattan may provide health and activity benefits over a much larger area than a smaller community square.

In spite of its flexibility and basis in choice theory, utility-based accessibility measures have not received much application in the accessibility literature compared with distance-based or even gravity-based measures. Vale et al. (2016) present a descriptive classification of active accessibility techniques, and explicitly dismiss utility-based metrics for incorporating randomness. The accessibility formula presented in 3 is the expectation of a random utility process, and is not in and of itself random any more than the gravity model is random. Utility-based accessibility metrics are commonly used, however, in alternatives analyses of transit infrastructure improvements (Jong et al., 2007). A reason for this is likely that a regional travel demand model is available to the analysts, thus making calibrated and multimodal logsums readily available (Geurs et al., 2010).

Empirical Application

In this section, we develop a model with data for New York City, where we compute a set of utility-based accessibility to parks scores for each tract and model the relationship between this measure and physical activity rates, controlling for spatial effects and socioeconomic factors. We subsequently model the effect of the accessibility metric on obesity rates, accounting for physical activity and the other controls.

Data

This study uses data available to the public from a variety of federal and state data agencies. The datasets, as well as the analysis code, are available on GitHub at https://github.com/gregmacfarlane/parks_access.

The Centers for Disease Control and Prevention makes small-area estimates on key health indicators available through its 500 Cities data program (Centers for Disease Control and Prevention, 2016). The indicators are multilevel aggregations and imputations of BRFSS responses (Wang et al., 2018, 2017), and have been recently used to study the tract-level link between gentrification and urban health (Gibbons et al., 2018). We use two indicators as our dependent variables in this study: the share of adults in a Census tract who are obese, and the share of adults who participate in no leisure-time physical activity. To improve clarity in our interpretation, we use the complement of the second variable — the share of tract adults who participate in *some* physical activity — even if the amount is not sufficient to affect overall health. Both indicators are estimated for the year 2016.

To the health data, we join socioeconomic data collected through the Census Bureau API via the `tidycensus` package for R (Walker, 2019). The primary dataset is a geographic polygons layer of Census tracts in the five boroughs of New York City. We append to each Census tract relevant sociodemographic data for each tract from the American Community Survey 2013-2017 5-year estimates. For a small handful of tracts in our sample, Census supressed the median income estimate; these appear to be primarily wealthy tracts and in almost all cases the CDC estimates of obesity and physical activity are missing as well. After removing these tracts from the estimation dataset, we have 2,102 complete cases. Table 1 presents key descriptive statistics for these data.

In a destination choice framework, the tracts represent the "origins" and the "destinations" are parks and green spaces in New York City. We retrieved a polygons layer of public parks and greenspaces within New York City's municipal boundaries and checked it for accuracy and relevance (City of New York, 2018). Upon inspection, we removed several facilities that do not qualify as publicly accessible green space, such as Yankee Stadium, Citi Field, and their surrounding parking lots. We also removed parks of less than half an acre in size, as these appear to be predominantly planted medians rather than legitimate public green space. We also consolidated individual geographic polygons comprising a single facility — as is the case in Flushing Meadows — and eliminated sub-facilities such as tennis courts or baseball fields included within larger park facilities. Instead of these distinct sub-facilities, we created variables for each containing park indicating the presence of sports courts, playgrounds, and trails.

To this dataset of parks and regularized green spaces, we add a polygons layer of cemeteries in New York that are open to the public. Cemeteries are important green spaces that can be used for many types of physical activity. These operations leave us with 1,277 distinct parks. For each park and cemetery we generate calculate the size of the park in acres. Using a Twitter application programming interface (API) implemented with the Python package Tweepy, we also collected and stored tweets containing geotags of precise geographic coordinates located within the boundaries of the parks; from this information, we were able to segregate tweets that were generated within a park in September 2014. This geolocated twitter activity could provide an additional point of information on the degree to which individual parks are actually used (Wang and Taylor, 2016).

Accessibility

We calculate a set of utility-based accessibility statistics for each tract in New York City. The most basic utility specification includes only the the park size in acres and the distance from the population-weighted tract centroid to the boundary of the park in miles,

$$V_{ij} = \lambda_s * \log(size_j) + \lambda_d * \log(distance_{ij}) \tag{5}$$

The logarithmic transform allows for diminishing marginal utility of distance and park size: A 1-mile increase to a trip matters more for a 1-mile trip than a 10-mile trip. (**???**) estimated destination choice parameters for park trips in the Alameda County (Oakland), California using this utility specification, obtaining values of $\lambda_s = 0.373$ and $\lambda_d = -1.98$. The ratio of these estimates implies people are willing to travel roughly six times further to reach a park twice as large. In the absence of park trip destination choice coefficients in New York City, we apply these previous estimates.

The second utility specification is adapted from a park destination choice model estimated by Kinnell et al. (2006) in New Jersey. In this study the authors used a revealed preference survey to determine which factors of a park and its surrounding environment affected its selection. By bringing over common variables, we can create a second utility specification as

$$V_{ij} = \lambda_s{*}\log(size_j){+}\lambda_d{*}\log(distance_{ij}){+}\lambda_t{*}trails{+}\lambda_c{*}courts{+}\lambda_p{*}playgrounds \tag{6}$$

with $\lambda_t = 0.99$, $\lambda_c = 0.43$, and $\lambda_p = 0.26$. (**???**) did not transform their size and distance estimates, so we retain the previously estimated and applied values. We assume that the other covariates in the (**???**) model which we do not have available to us (e.g., boat launches, picnic areas) are orthogonal to the included parameters and leaving them out will not affect the values of the included coefficients.

The final utility specification includes the number of geolocated tweets emanating from within the park in September of 2014 in addition to the basic specification in Equation 5,

$$V_{ij} = \lambda_s * \log(size_j) + \lambda_d * \log(distance_{ij}) + \lambda_{tw} * YJ(tweets) \tag{7}$$

with $YJ(x)$ being a Yeo-Johnson transformation that implements a diminishing marginal return similar to $log(x)$, but where $YJ(0) = 0$ (**???**). In this case we have no external information describing how destination choice to parks may be affected by the twitter activity in the park, so we assert a value of $\lambda_{tw} = 0.1$

Table 1: Descriptive Statistics of Tract and Park Variables

	Description	Minimum	Median (IQR)	Maximum	Source
Tract Variables, N = 2102					
Obesity	Share of adults over 18 who are obese	10.20	24.90 (19.83, 30.80)	45.40	CDC 500 Cities
Physical Activity	Share of adults over 18 who engage in some leisure-time physical activity	45.90	72.10 (66.60, 76.77)	90.70	CDC 500 Cities
Income	Median tract income	9053.00	59,592.50 (41,928.25, 79,092.75)	250001.00	ACS
Density	Households per square kilometer	9.20	5,848.46 (3,173.62, 9,674.36)	43621.52	ACS
Fulltime	Share of adults over 18 with full-time work	8.80	49.48 (44.45, 55.29)	100.00	ACS
College	Share of adults over 24 with a college degree	0.61	16.30 (12.37, 20.11)	44.94	ACS
Single	Share of adults over 18 living alone or in a non-partnership household	16.38	59.39 (50.30, 68.65)	100.00	ACS
Youth	Share of population under 18	0.00	20.54 (16.79, 24.92)	64.07	ACS
Young adults	Share of population between 18 and 34	0.00	25.73 (21.66, 29.98)	86.75	ACS
Seniors	Share of population who are 65 and over	0.00	12.83 (9.47, 16.88)	89.88	ACS
Black	Share of population who is black	0.00	10.03 (2.13, 44.62)	220.65	ACS
Asian	Share of population who is Asian	0.00	7.66 (2.40, 20.78)	88.07	ACS
Hispanic	Share of population who is Hispanic	0.00	19.07 (9.39, 41.07)	96.27	ACS
Other	Share of population who belong to other minority groups	0.00	0.00 (0.00, 0.53)	19.47	ACS
Park Variables, N = 1277					
Size	Park size in acres	0.50	1.66 (0.95, 5.62)	1961.00	NYC
Courts	Presence of sport courts / ball fields	0.00	0.00 (0.00, 0.00)	1.00	NYC
Playgrounds	Presence of playgrounds	0.00	0.00 (0.00, 1.00)	1.00	NYC
Trails	Presence of trails	0.00	0.00 (0.00, 0.00)	1.00	NYC
Tweets	Tweets emanating from park in September 2014	0.00	0.00 (0.00, 2.00)	1593.00	Twitter API

We then calculate the utility-based accessibility of each tract A_i as defined in Equation 3 under each utility specification. Recall that the total value of the accessibility is relative to an unknown scalar C; for this reason we standardize the accessibility values for all tracts within each utility specification,

$$A'_i = \frac{A_i - \bar{A}}{sd(A)} \tag{8}$$

Figure 1 shows a map of New York City with each tract highlighted based on its relative basic utility-based accessibility score (Equation 5). The most continuous region of good park access is in upper Manhattan and the Bronx, bracketed by Central Park and the Bronx River. Conversely, some of the poorest-accessibility areas are in Brooklyn tracts not immediately adjacent to Prospect Park. Because the accessibility statistic is normalized, the worst values are slightly below -3, the best somewhere above 3.

For comparative purposes, we also employ an isochrone analysis using the 10-minute walk threshold calculated for ParkScore (Trust for Public Land, 2019). If the population-weighted centroid of a tract is within a 10-minute walk of a green space as defined by ParkScore, the tract is considered as having "access" to a park. The centroid is necessary as all tracts in New York City have at least *some* intersection with the 10-minute walk buffer; as it is, only 24 of the 2,102 tract centroids are not located in this buffer.

Model

We predict either the physical activity or the obesity rate (y) in a census tract as a linear function of the tract's sociodemographic characteristics X and accessibility to parks (A),

$$y = X\beta + \beta_a A + \varepsilon \tag{9}$$

As the observations are related to each other spatially, it is likely that this process involves spatial spillovers of one kind or another. A complete treatment of spatial econometrics is not warranted here; the reader is directed to LeSage and Pace (2009) as well as LeSage and Pace (2014). Suffice it to say that the presence of spatial data generating processes can negatively affect econometric interpretation in a variety of ways. Each of these process relies on spatial autoregression, where elements of some random variable x are spatially dependent on other elements,

$$x = \rho W x + \varepsilon \tag{10}$$

with the strength of dependence an estimated parameter ρ, the structure of dependence governed by the asserted matrix W (details of W follow below), and some random uncorrelated residual ε. A collection of models is available to represent spatial autodependence in the model residuals, the independent variables, and the dependent variable.

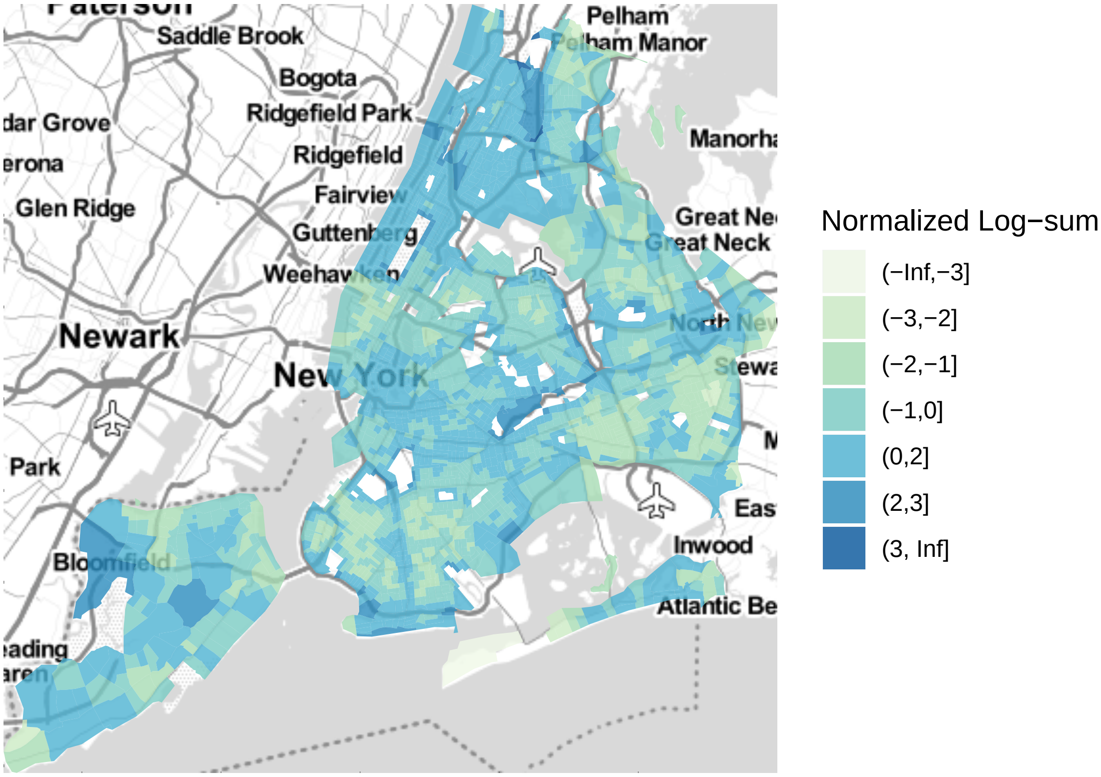

Figure 1: Normalized choice-based park accessibility log-sum values in New York City.

The specific spatial model applied is both an econometric choice — as improperly specified spatial models can lead to inconsistent estimates of model parameters and standard errors — as well as a philosophical choice about the likely structure of spillovers in the problem at hand. In case of physical activity and obesity rates, we believe that these outcomes are locally determined: That is, a particular individual decides whether or not to participate in physical activity independently of whether his or her neighbors participate in physical activity. Mathematically, we assert that the autodependence relationship on the dependent variable $y = \rho W y + \varepsilon$ has $\rho = 0$ in the current case.

With this relationship ruled out, it remains a possibility that the model residuals are spatially dependent,

$$y = X\beta + \beta_a A + u; u = \lambda W u + \varepsilon \tag{11}$$

that the outcome is partially dependent on the socioeconomic variables in *neighboring* tracts,

$$y = X\beta + \beta_a A + W X \lambda + \varepsilon \tag{12}$$

or a linear combination of both.

$$y = X\beta + \beta_a A + W X \gamma + u; u = \lambda W u + \varepsilon \tag{13}$$

These models are referred as the spatial error model (SEM), the spatial lag of X model (SLX), and the spatial Durbin error model (SDEM). Pace and LeSage (2008) suggest that a Hausman-style test of the estimates of β derived from the OLS (Equation 9) and SEM (Equation 11) can identify whether the estimates are consistent. If the β estimates are dissimilar, both specifications are inappropriate as they contain untreated omitted variables. Similarly, if the SLX and SDEM estimates are dissimilar it is an indication of missing variables in that specification. On the other hand if the estimates of both pairs of models are similar and the residual correlation parameter λ is non-zero, the model accounting for spatially dependent residuals will produce proper estimates of the model standard errors.

Note that in the lag-X specifications (SLX and SDEM) we exclude the accessiblity component A, that is we do *not* consider that the accessibility in a neighboring zone will have an effect on physical activity or obesity rates. In a practical sense, this is because the accessibility term A is itself spatially determined.

Spatial Weights

The spatial weights matrix W is constructed of individual elements where $w_{ij} = 0$ if observations i and j are spatially independent of each other, and $w_{ij} > 0$ if i and j are spatially related in some way. Dubin (1998) presents details on constructing W, but in this study we assume that tracts sharing at least one common border point are neighbors of each other[1]. The elements of W

[1] We also considered a distance-discounted weights matrix, which produced similar results. This matrix is selected for simplicity.

are then row-standardized so that each observation has an equal total influence from all its neighbors.

Results

We estimated SDM and SEM models regressing the physical activity rate against the base covariates (without any accessibility component). A likelihood ratio test failed rejected that the SDM is different from the SEM, so we use the SDM specification only going forward. Recall that in an SDM the effect of a covariate on the dependent variable is *not* the estimated coefficient. As a consequence of this Table **??** presents the estimated coefficients for the base model with the socioecononomic controls and no representation of park accessibility. For the most part the coefficients are highly significant and of the expected sign. Tracts with higher shares of full-time workers, college-educated adults, young adults, and high-income households all show a greater share of individuals engaging in regular physical activity. Conversely, tracts with greater population density and a greater share of single adults, children, seniors, minorities of all types, and low income households all have lower modeled rates of physical activity.

To this base model we add a measure of park accessibility considered in two ways: a logsum-based statistic as described earlier and a more typical quarter-mile binary access buffer. The estimated coefficients for these models are also given in Table **??**. In both accessibility specifications, the access measure is positively correlated with physical activity rates; all else equal, people in tracts with higher accessibility to parks have higher estimated rates of physical activity. Though the estimated coefficient on the quarter-mile buffer appears to have a larger absolute value than the logsum-based statistic, this masks a key difference in the specification worth further comment. The logsum allows for variation in access unavailable in the buffer-based method. In our dataset almost every tract (0%) is within a quarter-mile of *some* kind of park, but this park may not be large enough to support physical activities all individuals would like to participate in. Because tracts with only adequate park access are grouped with tracts having excellent park access, the overall effect is comparatively unclear. The logsum-based coefficient, on the other hand, suggests going from the least-accessible (-3.0) to the most-accessible (3.0) tract implies roughly a 1.1 percentage point rise in estimated physical activity rates.

\end{table*}

We now consider the impacts of a model where the dependent variable is the *obesity* rate, and physical activity becomes an independent covariate alongside the controlling variables and the accessibility metric. Table **??** again presents the base model without any accessiblity measure, and models with accessibility captured through a logsum and through a 1/4 mile buffer. As in the physical activity models, the coefficients are typically significant and of the expected sign, with a few exceptions: in this case an increasing hispanic population share and an increasing share of children have no significant relationship on obesity. Unlike in the physical activity models however, the inference on the accessibility statistic *does* change between specifications. The 1/4 mile buffer

Table 2: SEM Coefficients Predicting Physical Activity Rates

	Base	Logsum	10-min walk
(Intercept)	-0.1737	-0.3979	-0.6864
	(3.3363)	(3.3287)	(3.3549)
log(density)	0.0555	0.0713	0.0485
	(0.0803)	(0.0804)	(0.0804)
log(income)	6.7541^{***}	6.7649^{***}	6.7516^{***}
	(0.2495)	(0.2492)	(0.2494)
fulltime	0.1453^{***}	0.1451^{***}	0.1455^{***}
	(0.0103)	(0.0103)	(0.0103)
college	0.0378^{**}	0.0366^{**}	0.0389^{**}
	(0.0137)	(0.0137)	(0.0137)
single	-0.0394^{***}	-0.0394^{***}	-0.0397^{***}
	(0.0093)	(0.0092)	(0.0093)
youth	-0.1401^{***}	-0.1397^{***}	-0.1405^{***}
	(0.0147)	(0.0147)	(0.0147)
young_adults	0.0382^{***}	0.0398^{***}	0.0379^{***}
	(0.0115)	(0.0115)	(0.0115)
seniors	0.0401^{**}	0.0373^{*}	0.0398^{**}
	(0.0154)	(0.0154)	(0.0154)
black	-0.0583^{***}	-0.0577^{***}	-0.0581^{***}
	(0.0047)	(0.0047)	(0.0047)
asian	-0.1131^{***}	-0.1122^{***}	-0.1132^{***}
	(0.0053)	(0.0053)	(0.0053)
hispanic	-0.1151^{***}	-0.1160^{***}	-0.1151^{***}
	(0.0050)	(0.0050)	(0.0050)
other	-0.0319	-0.0294	-0.0331
	(0.0490)	(0.0490)	(0.0490)
λ	0.9268^{***}	0.9238^{***}	0.9266^{***}
	(0.0164)	(0.0169)	(0.0164)
access_ls		0.1689^{*}	
		(0.0692)	
walk_10TRUE			0.6740
			(0.4865)
Num. obs.	2102	2102	2102
Parameters	15	16	16
Log Likelihood	-4810.8483	-4807.8887	-4809.8890
AIC (Linear model)	10405.0179	10330.6080	10401.2340
AIC (Spatial model)	9651.6966	9647.7773	9651.7781
LR test: statistic	755.3213	684.8307	751.4560
LR test: p-value	0.0000	0.0000	0.0000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Standard errors in parentheses.

neither substantially improves the model fit nor is its term significant at any threshold, but the accessibility logsum indicates a significant negative correlation: moving from the least to most accessible tract decreases the expected obesity rate by approximately half a percentage point.

Additional Amenities

As discussed when presenting the choice-based accessibility metric above, a key benefit it offers is the natural way in which park amenities beyond simply size can be accommodated. For example, the number of tweets emanating from a park might be seen as a measure of the park's popularity, or a proxy for its use. It may be construed that parks with high popularity or use contain other amenities that encourage residents to use the park space, raising physical activity rates and lowering obesity rates beyond what would be expected with the park's size and proximity alone. It is reasonable to question whether park size and Twitter activity are highly correlated, and thus would result in a double-counting of park size in the accessibility statistic. In our data the number of tweets is positively but only loosely correlated with park size ($\rho = $), indicating that the Twitter data in fact could provide new information or proxy for the desired amenities.

To test this hypothesis, we can use the twitter data described earlier and suggested in Equation **??**. In this application, we again log-transform all the components of utility:

$$V_{ij} = \log(size_j)\lambda_s + \log(distance_{ij})\lambda_d + g(tweets_j)\lambda_t \tag{14}$$

where $g(x)$ is a Yeo-Johnson transformation to preserve cases where $log(0)$ would be otherwise undefined [**?**]. As before, there is little information by which to determine the values of λ in this utility specification; we use the previous values with the addition of $\lambda_t = 0.1$. The ratio between the utility coefficients implies residents would travel almost twenty times as far or use a park three times smaller if it had twice the Twitter activity, all else equal.

Figure **??** shows the percent change in the normalized log sum statistic after including the twitter information. There is not a recognizable overarching pattern to the changes, though tracts immediately surrounding Prospect Park and Central Park gain even more relative accessibility, and tracts in midtown Manhattan, Queens, and southern Bronx tend to lose some.

Table **??** presents the estimated coefficients relating accessibility to obesity and physical activity rates, with and without the inclusion of Twitter data. In both cases, including twitter information in the accessibility statistic tempers the strength of the relationship with the dependent variable and modestly widens the standard error without substantively affecting the significance of the test statistics or the interpretation. From this limited example, we cannot say whether Twitter activity enhances the ability of parks to increase physical activity rates or lower obesity rates, though further investigation of this and other park amenities would certainly be warranted.

Table 3: SEM Coefficients Predicting Obesity Rates

	Base	Logsum	Tweets	Attributes	10-min walk
(Intercept)	59.2637***	59.2516***	59.2232***	59.2565***	59.3463***
	(1.8141)	(1.8118)	(1.8114)	(1.8125)	(1.8171)
log(density)	−0.0492	−0.0543	−0.0538	−0.0539	−0.0475
	(0.0406)	(0.0408)	(0.0407)	(0.0407)	(0.0407)
log(income)	−0.3310*	−0.3386*	−0.3364*	−0.3391*	−0.3320*
	(0.1462)	(0.1462)	(0.1461)	(0.1462)	(0.1461)
fulltime	−0.0164**	−0.0165**	−0.0165**	−0.0164**	−0.0165**
	(0.0054)	(0.0054)	(0.0054)	(0.0054)	(0.0054)
college	0.0306***	0.0309***	0.0309***	0.0309***	0.0303***
	(0.0069)	(0.0069)	(0.0069)	(0.0069)	(0.0069)
single	0.0118*	0.0118*	0.0119*	0.0119*	0.0119*
	(0.0047)	(0.0047)	(0.0047)	(0.0047)	(0.0047)
youth	0.0128	0.0128	0.0129	0.0128	0.0130
	(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0076)
young_adults	−0.0139*	−0.0144*	−0.0144*	−0.0145*	−0.0138*
	(0.0058)	(0.0058)	(0.0058)	(0.0058)	(0.0058)
seniors	−0.0974***	−0.0966***	−0.0964***	−0.0965***	−0.0973***
	(0.0077)	(0.0078)	(0.0078)	(0.0078)	(0.0077)
black	0.0671***	0.0670***	0.0670***	0.0669***	0.0671***
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
asian	−0.1200***	−0.1202***	−0.1202***	−0.1202***	−0.1199***
	(0.0030)	(0.0030)	(0.0030)	(0.0030)	(0.0030)
hispanic	−0.0005	−0.0001	−0.0001	−0.0001	−0.0004
	(0.0028)	(0.0028)	(0.0028)	(0.0028)	(0.0028)
other	−0.0534*	−0.0541*	−0.0541*	−0.0543*	−0.0530*
	(0.0246)	(0.0246)	(0.0246)	(0.0246)	(0.0246)
physact	−0.4062***	−0.4054***	−0.4054***	−0.4055***	−0.4059***
	(0.0110)	(0.0110)	(0.0110)	(0.0110)	(0.0110)
λ	0.9755***	0.9753***	0.9752***	0.9754***	0.9755***
	(0.0069)	(0.0070)	(0.0070)	(0.0069)	(0.0069)
access_ls		−0.0504			
		(0.0352)			
tweets_ls			−0.0539		
			(0.0360)		
multi_ls				−0.0567	
				(0.0356)	
walk_10TRUE					−0.1997
					(0.2464)
Num. obs.	2102	2102	2102	2102	2102
Parameters	16	17	17	17	17
Log Likelihood	−3379.5886	−3378.5650	−3378.4731	−3378.3221	−3379.2603
AIC (Linear model)	8401.3863	8391.0914	8388.2862	8403.1101	8399.6563
AIC (Spatial model)	6791.1771	6791.1299	6790.9463	6790.6441	6792.5206
LR test: statistic	1612.2092	1601.9614	1599.3400	1614.4660	1609.1357
LR test: p-value	0.0000	0.0000	0.0000	0.0000	0.0000

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$. Standard errors in parentheses.

Limitations and Future Research Direction

We readily acknowledge limitations in this study. As in any study conducted with areal data, we are at risk of falling victim to the ecological inference fallacy, where aggregate statistics mask or contradict disaggregated or individual-level trends. A large-sample survey of individuals in New York City, including measured physical activities and obesity would always be preferable to the tract-level data used in this study. It would also be preferrable to have obtained the $\boldsymbol{\lambda}$ accessibility utility coefficients from a high-quality survey of leisure destination choice rather than asserting them to match our prior expectations of what constitutes quality park access. An ideal survey to address the question would incorporate both sets of questions: physical activity and health data on one hand and park use (including which parks were used and how frequently) on the other. As no such dataset exists to our knowledge, this tract-level aggregate analysis with asserted utility coefficients is the possibility that remains.

This paper presents a holistic accessibility statistic that could, in theory, accommodate many attributes of the destination parks as well as the people who might use them. As an illustration: the park-going population could be separated into at least four delineated clusters, each preferring different amenities of a park:

- runners and cyclists: long, interesting trail systems
- sports players: soccer fields, basketball courts, or baseball diamonds
- families with small children: water features and playgrounds
- casual users: water features, gardens, performances, etc.

An analyst could then compute the accessibility logsum for each cluster with different utility values for each park's amenities, and obtain a measure of a neighborhood's accessibility to park features that its residents most care about. In this paper, we proceed only incrementally beyond this theory by adding Twitter activity as an element of a park's attractiveness above and beyond its size. Exploring additional amenities or market segmentation strategies could provide a more definitive understanding of the relationship between park accessibility and health outcomes.

As a travel impedance measure, we used the Euclidean distance between each tract's population-weighted centroid and the nearest point on the edge of a park's border. Euclidean distances have well-rehearsed limitations regarding their fidelity with the underlying infrastructure network, etc. Network-based distances can also suffer from challenges when applied to multimodal problems; these challenges are exacerbated when the non-highway mode share is high, as is likely when considering access to parks. A better metric of travel impedance may be a mode choice model logsum, which weights all travel alternatives against each other.

Finally, this study is primarily focused on the hypothesis that accessibility to parks encourages physical activity, which in turn reduces obesity. There are a multitude of other hypotheses that might be proposed and tested with the basic methodology we have employed here, or competing explanations for the outcomes we have observed. It is distinctly possible, for instance, that individuals who wish to exercise regularly in parks choose to live near them. Given that the CDC models generating the obesity and physical activity estimates presumably include variables likely to influence such preferences (income, etc.), our investigation cannot isolate the preferences from the effect. Controlling for such a self-selection effect would be necessary to isolate the exogenous impacts of park access on obesity or other health outcomes. And regarding these other variables: this study did not consider potential relationships between park access and hospitalization rates, life expectancy, respiratory disease, mental health, or any number of potential beneficial outcomes hypothesized or explored in the existing literature. Exploring these connections and their underlying mechanisms should be a priority as city planners and urban architects attempt to improve the quality of life of urban residents in the future.

Conclusion

Increased physical activity and decreased obesity rates are critical measures of improvement in public health. Although many have theorized the link between park accessibility and these metrics, previous literature has produced mixed findings, perhaps owing to the range of variables modeled and the coarse spatial scale of the analyses. Using New York City as a case study, we presented a holistic and flexible measurement for park accessibility that compares the continuous distance to all parks in the region, weighted against the size of the park and its other amenities. In terms of physical activity, we found a positive relationship where the least park-accessible tracts have an expected physical activity participation rate roughly one percentage point lower than the most accessible tracts. In terms of obesity, we also found a strong and significant correlation between increased park accessibility and decreased obesity rates, controlling for spatial correlation, socioeconomic attributes, and physical activity rates.

In both cases a traditional, buffer-based analysis estimated a weaker relationship with greater uncertainty. Buffer analyses are relatively common in the literature, perhaps because of the widespread availability of GIS software. In spite of their widespread use, they are relatively limited in terms of both their theoretical underpinnings and their flexibility to accommodate attributes of parks beyond their proximity. And even proximity may not be adequately handled, as the buffer distance may be arbitrarily asserted by the researcher. The model we develop and apply in this paper extends buffer driven models to contribute a more comprehensive and flexible approach for measuring the impact of park access on health outcomes. Adopting choice-based accessibilities of the kind used in this study will allow researchers to encompass the full range of park amenities in their accessibility analyses. This will in turn enable planners to consider how multiple attributes of a park — from its location to its size to its amenities and beyond — benefit the health of the community the park serves.

Acknowledgements

This project was funded by the Speedwell Foundation. We are grateful to Rachel Samuels for helping to assemble the geolocated Twitter data.

References

Bancroft, C., Joshi, S., Rundle, A., Hutson, M., Chong, C., Weiss, C.C., Genkinger, J., Neckerman, K., Lovasi, G., 2015. Association of proximity and density of parks and objectively measured physical activity in the United States: A systematic review. Social Science & Medicine 138, 22–30. doi:10.1016/j.socscimed.2015.05.034

Bedimo-Rung, A.L., Mowen, A.J., Cohen, D.A., 2005. The significance of parks to physical activity and public health. American Journal of Preventive Medicine 28, 159–168. doi:10.1016/j.amepre.2004.10.024

Bruce, N., 1977. A Note on Consumer's Surplus, the Divisia Index, and the Measurement of Welfare Changes. Econometrica 45, 1033. doi:10.2307/1912692

Centers for Disease Control and Prevention, 2009. CDC - Healthy Places - Physical Activity.

Centers for Disease Control and Prevention, 2016. 500 Cities Project: Local data for better health.

City of New York, 2018. NYC Open Data – Open Space (Parks).

Coutts, C., 2008. Greenway Accessibility and Physical-Activity Behavior. Environment and Planning B: Planning and Design 35, 552–563. doi:10.1068/b3406

Daly, A., 1982. Estimating choice models containing attraction variables. Transportation Research Part B: Methodological 16, 5–15. doi:10.1016/0191-2615(82)90037-6

Dias, A., Gaya, A., Brand, C., Pizarro, A., Fochesatto, C., Mendes, T., Mota, J., Maia Santos, M., Gaya, A., 2019. Distance from home to the nearest park and the use of the parks for physical activity: the mediator role of road safety perception in adolescents. Public Health 168, 9–16. doi:10.1016/J.PUHE.2018.11.021

Dong, X., Ben-Akiva, M.E., Bowman, J.L., Walker, J.L., 2006. Moving from trip-based to activity-based measures of accessibility. Transportation Research Part A: Policy and Practice 40, 163–180. doi:10.1016/J.TRA.2005.05.002

Dubin, R.A., 1998. Spatial autocorrelation: a primer. Journal of Housing Economics 7, 304–327. doi:10.1006/jhec.1998.0236

Duhl, L.J., Sanchez, A.K., 1999. Healthy Cities and the City Planning Process: a background document on links between health and urban planning, European health 21. World Health Organization, Copenhagen.

Durstine, J.L., Gordon, B., Wang, Z., Luo, X., 2013. Chronic disease and the link to physical activity. Journal of Sport and Health Science 2, 3–11. doi:10.1016/J.JSHS.2012.07.009

Finkelstein, E.A., Trogdon, J.G., Cohen, J.W., Dietz, W., 2009. Annual Medical Spending Attributable To Obesity: Payer-And Service-Specific Estimates. Health Affairs 28, w822–w831. doi:10.1377/hlthaff.28.5.w822

Geurs, K., Zondag, B., Jong, G. de, Bok, M. de, 2010. Accessibility appraisal of land-use/transport policy strategies: More than just adding up travel-time savings. Transportation Research Part D: Transport and Environment 15, 382–393. doi:10.1016/J.TRD.2010.04.006

Gibbons, J., Barton, M., Brault, E., 2018. Evaluating gentrification's relation to neighborhood and city health. PLOS ONE 13, e0207432. doi:10.1371/journal.pone.0207432

Giles-Corti, B., Broomhall, M.H., Knuiman, M., Collins, C., Douglas, K., Ng, K., Lange, A., Donovan, R.J., 2005. Increasing walking: How Important Is Distance To, Attractiveness, and Size of Public Open Space? American Journal of Preventive Medicine 28, 169–176. doi:10.1016/j.amepre.2004.10.018

Handy, S.L., Niemeier, D.A., 1997. Measuring Accessibility: An Exploration of Issues and Alternatives. Environment and Planning A 29, 1175–1194. doi:10.1068/a291175

Jong, G. de, Daly, A., Pieters, M., Hoorn, T. van der, 2007. The logsum as an evaluation measure: Review of the literature and new results. Transportation Research Part A: Policy and Practice 41, 874–889. doi:10.1016/J.TRA.2006.10.002

Kinnell, J.C., Bingham, M.F., Mohamed, A.F., Desvousges, W.H., Kiler, T.B., Hastings, E.K., Kuhns, K.T., 2006. Estimating Site Choice Decisions for Urban Recreators. Land Economics 82, 257–272. doi:10.3368/le.82.2.257

Larson, L.R., Jennings, V., Cloutier, S.A., 2016. Public Parks and Wellbeing in Urban Areas of the United States. PLOS ONE 11, e0153211. doi:10.1371/journal.pone.0153211

LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. Chapman; Hall/CRC.

LeSage, J.P., Pace, R.K., 2014. Interpreting Spatial Econometric Models, in: Fischer, M., Nijkamp, P. (Eds.), Handbook of Regional Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1535–1552. doi:10.1007/978-3-642-23430-9_91

Logan, T., Williams, T., Nisbet, A., Liberman, K., Zuo, C., Guikema, S., 2019. Evaluating urban accessibility: leveraging open-source data and analytics to overcome existing limitations. Environment and Planning B: Urban Analytics and City Science 46, 897–913. doi:10.1177/2399808317736528

McFadden, D.L., 1974. Conditional Logit Analysis of Qualitative Choice Behavior, in: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Mitchell, R., Popham, F., 2008. Effect of exposure to natural environment on health inequalities: an observational population study. The Lancet 372, 1655–1660. doi:10.1016/S0140-6736(08)61689-X

Pace, R.K., LeSage, J.P., 2008. A spatial Hausman test. Economics Letters 101, 282–284. doi:10.1016/j.econlet.2008.09.003

Richardson, E.A., Mitchell, R., Hartig, T., Vries, S. de, Astell-Burt, T., Frumkin, H., 2012. Green cities and health: a question of scale? Journal of Epidemiology and Community Health 66, 160 LP–165. doi:10.1136/jech.2011.137240

Stark, J.H., Neckerman, K., Lovasi, G.S., Quinn, J., Weiss, C.C., Bader, M.D., Konty, K., Harris, T.G., Rundle, A., 2014. The impact of neighborhood park access and quality on body mass index among adults in New York City. Preventive Medicine 64, 63–68. doi:10.1016/J.YPMED.2014.03.026

Trust for Public Land, 2019. 2019 ParkScore Index | The Trust for Public Land.

Vale, D.S., Saraiva, M., Pereira, M., 2016. Active accessibility. Journal of Transport and Land Use 9, 209–235.

Walker, K., 2019. tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames.

Wang, Q., Taylor, J.E., 2016. Process Map for Urban-Human Mobility and Civil Infrastructure Data Collection Using Geosocial Networking Platforms. Journal of Computing in Civil Engineering 30, 04015004. doi:10.1061/(ASCE)CP.1943-5487.0000469

Wang, Y., Holt, J.B., Xu, F., Zhang, X., Dooley, D.P., Lu, H., Croft, J.B., 2018. Using 3 Health Surveys to Compare Multilevel Models for Small Area Estimation for Chronic Diseases and Health Behaviors. Preventing Chronic Disease 15, 180313. doi:10.5888/pcd15.180313

Wang, Y., Holt, J.B., Zhang, X., Lu, H., Shah, S.N., Dooley, D.P., Matthews, K.A., Croft, J.B., 2017. Comparison of Methods for Estimating Prevalence of Chronic Diseases and Health Behaviors for Small Geographic Areas: Boston Validation Study, 2013. Preventing Chronic Disease 14, 170281. doi:10.5888/pcd14.170281

West, S.T., Shores, K.A., Mudd, L.M., 2012. Association of Available Parkland, Physical Activity, and Overweight in America's Largest Cities. Journal of Public Health Management and Practice 18, 423–430. doi:10.1097/PHH.0b013e318238ea27

Wolf, K.L., 2008. City trees, nature and physical activity: A research review. Arborist News 17, 22–24.

World Health Organization, 2014. Noncommunicable diseases country profiles 2014. World Health Organization; World Health Organization.

Zhang, X., Lu, H., Holt, J.B., 2011. Modeling spatial accessibility to parks: a national study. International Journal of Health Geographics 10, 31. doi:10.1186/1476-072X-10-31