Modeling the impacts of park access on health outcomes: a choice-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 few 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; 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 choice-based accessibility measure compares the continuous distance to all parks in the region, weighted against the sizes of the each park and 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 choice-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

Numerous previous studies have found relationships between characteristics of the built environment and aspects of physical, mental, social, and economic health. For the purposes of this study, it is clear that parks and other vegetated areas

support outdoor physical activity in ways distinct from other urban environments (Giles-Corti et al., 2005; Wells et al., 2007)

In spite of these generally positive findings, the evidence of a specific link between green space and physical activity is somewhat mixed. Wolf (2008) reports that people self-identify as more active and attain a higher quality of life when a greater portion of their environment is greenspace. Larson et al. (2016) similarly used self-reported scores on the Gallup-Healthways Wellbeing Index to evaluate the relationship between different areas of wellbeing, including 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. 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 relationships between the density of parkland, parkland per capita, and levels of physical activity and obesity for 67 metropolitan statistical areas in the U.S. The study found a significant, positive association between park density and physical activity and a significant, negative association between park density and obesity. Conversely, Richardson et al. (2012) examined the relationship between urban greenspace and selected mortality rates, and 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. 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.

While several of the aforementioned studies reported results consistent with the hypothesis that park access and use confer health benefits stemming from increased physical activity and attendant decreases in obesity, the large majority of extant research has been conducted at the level of the city or metropolitan statistical area. Metropolitan-level analyses 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.

In a study of neighborhoods in New York City, Stark et al. (2014) used hierarchical linear models to explore relationships between the proportion of zip codes that was park space, the cleanliness of parks, and the body mass index of adult residents. The authors 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.

Choice-based Accessibility

Consider that an individual is choosing a park for a recreation activity. If we apply random utility choice theory, specifically the multinomial logit model (McFadden, 1974), the expected probability of individual i choosing park j from the set of all regional parks J is

$$P(j|V_{ij}) = \frac{\exp(V_{ij})}{\sum_{p \in J} \exp(V_{ip})} \tag{1}$$

where parks are differentiated from each other by their relative measurable utilities V_{ij}. In principle, V may include any measurable attributes of either the choice maker or the park. In this study we use a linear formulation of

$$V_{ij} = size_j \lambda_s + distance_{ij} \lambda_d \tag{2}$$

incorporating the size of the park in acres and the distance of the park from the census tract i in miles. The coefficients $\boldsymbol{\lambda}$ can be estimated from household surveys, though in the absence of a survey we may assert reasonable values.

A key theoretical outcome of random utility choice models is that the consumer surplus to individual i, or the total weighted value of all alternatives in individual i's choice set, can be obtained as the logarithm of the sum in the denominator of the choice probability equation (Small and Rosen, 1981):

$$C_i = \ln \left(\sum_{p \in J} \exp(V_{ip}) \right) \tag{3}$$

There are several advantages to such a log-sum defined metric relative to buffer-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 any attribute of the park as part of its utility specification; in this case, we consider the size of the park weighted against its distance as an element of relative accessibility. This suggests that not all parks are equal, and that a large park such as Central Park in Manhattan may provide health and activity benefits over a much larger area than a smaller community square.

Note that this formulation is an extension of gravity-based accessibility statistics, in the same way that the gravity-based trip distribution model is a specific case of a destination-choice model (Daly, 1982). But the extension is meaningful and allows us to consider more attributes of the park or the individual beyond size and impedance. For example, we can aggregate the number of tweets

emanating from a park as a measure of the park's attractiveness beyond its size. The utility for the choice maker from Equation 2 then becomes

$$V_{ij} = size_j\lambda_s + distance_{ij}\lambda_d + tweets_j\lambda_t \tag{4}$$

In spite of its flexibility and basis in choice theory, accessibility measures based in log-sums have not received much application in the accessibility literature compared with distance-based or even gravity-based measures. Zhang et al. (2011) present a metric with a similar functional form, though the authors arrive at it from a motivation other than destination choice theory. Log-sums 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 holistic logsum-based accessibility to parks measure 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. Finally, we model the relationship between a modified logsum measure considering geolocated Twitter data and physical activity and obesity rates.

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`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 sub-facilities, we created variables for each 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 used (Wang and Taylor, 2016).

Model

We predict the rate of physical activity (y) in a census tract as a function of the tract's sociodemographic characteristics and choice-based accessibility to parks with the following specification:

$$y_i = f(\boldsymbol{x}_i, C_i, g(W, \boldsymbol{y}, \boldsymbol{X})) \tag{5}$$

where $\boldsymbol{x}_i$ is a vector of tract-level sociodemographic attributes, C_i is a choice-based accessibility measure for the tract computed from Equation 3 and $g(W, \boldsymbol{y}, X)$ represents the impact of spatial neighborhood effects on the outcome.

Using tracts as the unit of observation in the model allows us to consider the availability of parks at a sub-metropolitan level, but it also introduces the problem of spatial correlation of unobserved errors and spatial dependence in the data generating process (Anselin, 1980; Cliff and Ord, 1970). Spatial correlation

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

occurs when spatially-distributed unobservable or missing variables (school quality, neighborhood prestige, etc.) influence the modeled outcomes. Spatial dependence occurs when the outcome *depends* on the outcome in neighboring areas; for example, seeing neighbors exercising may encourage exercise. The relaxed spatial Durbin model (SDM) proposed by Burridge (1981) controls for both processes:

$$\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{6}$$

where ρ is the estimated spatial correlation between the outcome variables $y_i, y_j, \ldots \in \boldsymbol{y}$, W is a matrix specifying the spatial relationship between observations i and j in the dataset, $\boldsymbol{\beta}$ is a vector of estimable coefficients relating the attributes X of an observation to its outcome, and $\boldsymbol{\gamma}$ is a vector of estimated coefficients relating the attributes of an observation's *neighbors* to its outcome. If only correlation and not dependence exists in the data generating process, it may be more efficient to estimate a restricted spatial autoregressive model and therefore save degrees of freedom in the estimation. Macfarlane et al. (2015) illustrate a decision algorithm originally proposed by Florax et al. (2003) which uses a likelihood ratio test of the SDM against the restricted spatial error model (SEM):

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} = \gamma W\boldsymbol{u} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= X\boldsymbol{\beta} + (I - \gamma W)^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{7}$$

If the likelihood ratio test rejects that the models are equivalent, then the SDM should be used because its estimates are unbiased. One consequence of using the SDM is that the estimated coefficients do not represent the

The researcher must assert W *a priori*, though there are several common specifications [**?**]. In this study we specify that tracts with population-weighted centroids located within 1.8 Euclidean miles of each other are considered neighbors, and the strength of the relationship is the row-standardized (each row sums to 1) inverse of the distance between them:

$$W_{ij} = \begin{cases} \frac{d_{ij}^{-1}}{\sum_{k=1}^{n} d_{ik}^{-1}} & \text{for } d_{ij} \leq 1.8 \text{ miles} \\ 0 & \text{for } d_{ij} > 1.8 \text{ miles} \end{cases} \tag{8}$$

This is a similar specification to that used by **?**] in a study of home prices in Atlanta. We selected this specification after comparing overall model likelihoods between it and a specification where tracts sharing a boundary are considered neighbors.

Accessibility

Extant destination choice models used in practice typically use travel time in minutes as a travel impedance term, and employment by sector as a size term or attraction component. Typically such models handle recreational trips together with other non-work and non-school trips, and we therefore can find no previously estimated coefficients using size terms relevant to park acreage or

amenities. Given this gap in the literature, we instead assert plausible values for β to generate the choice-based accessibility statistic used in this study.

The functional form of the accessibility statistic is adapted from Equation 2 by log-transforming first the distance (Euclidean, in miles) between the park boundaries and population-weighted tract centroids and second the park size (in acres).

$$V_{ij} = \log(size_j)\lambda_s + \log(distance_{ij})\lambda_d \tag{9}$$

We then calculate the consumer surplus C (log-sum of the exponentiated utilities) for each tract and standardize the result,

$$C_i' = \frac{C_i - \bar{\boldsymbol{C}}}{sd(\boldsymbol{C})} \tag{10}$$

We initially estimated potential values of the λ coefficients by applying a limited-memory, box-constrained optimization algorithm maximizing the model log-likehood.[1] The constraint required that the coefficient on size be positive and that on distance negative; all else equal, we assert people will prefer to use larger parks and parks that are nearer to their residence. This resulted in $\lambda_d = -1.98$ and $\lambda_s = 0.000$. Though the "most likely" parameter value, in this case it was unsatisfactory for two reasons: first, our prior assertion that park size is a valuable attribute of a park, and second the observation that with these maximum likelihood values tracts bordering on the corners of Central Park were measured with poor overall accessibility, an unacceptably unintuitive result. We subsequently adjusted the λ_s value to 0.3; the relative ratio of the coefficients implies that all else equal, an individual would travel 6.6 times further to access a park twice as large.

Figure 1 shows a map of New York City with each tract highlighted based on its relative accessibility score. 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.

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 socioeconononomic 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,

[1] Using the `nlm` function in R [@RCoreTeam2018]

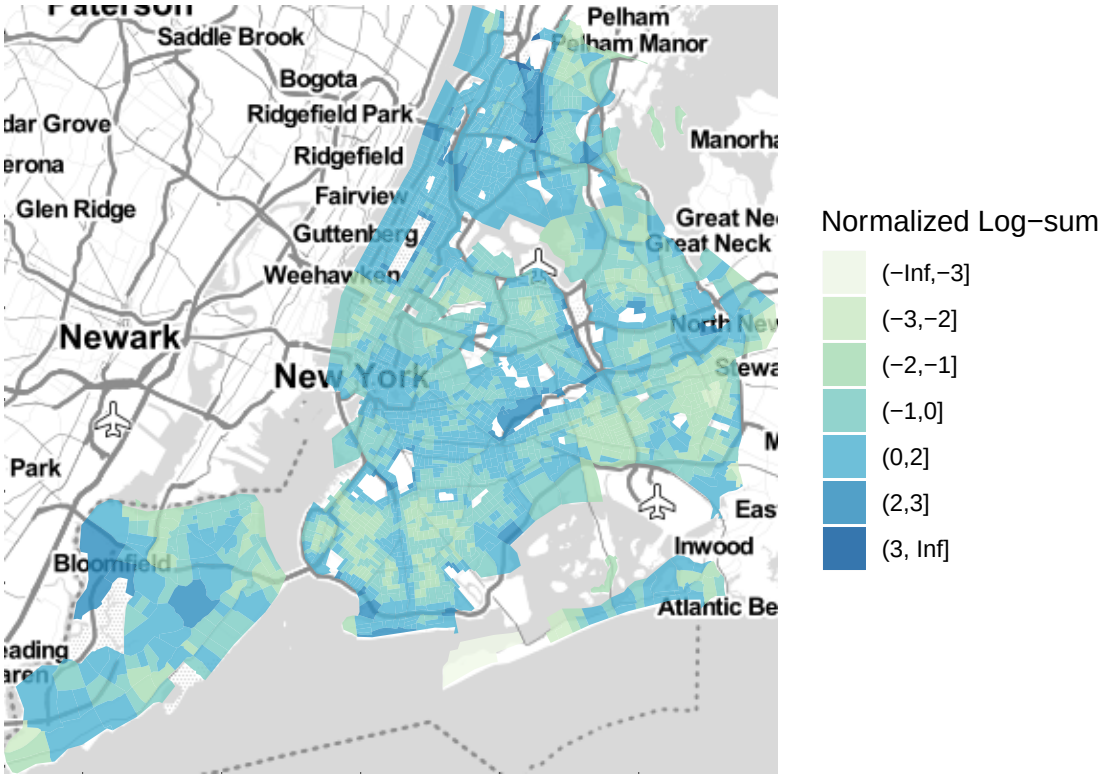

Figure 1: Normalized choice-based park accessibility log-sum values in New York City.

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 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

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.

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.

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 4. In this application, we again log-transform all the components of utility:

$$V_{ij} = \log(size_j)\lambda_s + \log(distance_{ij})\lambda_d + g(tweets_j)\lambda_t \tag{11}$$

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.

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

Anselin, L., 1980. Estimation methods for spatial autoregressive structures. Regional Science Dissertation & Monograph Series, Program in Urban and Regional Studies, Cornell University.

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

Burridge, P., 1981. Testing for a common factor in a spatial autoregression model. Environment and Planning A 13, 795–800.

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).

Cliff, A.D., Ord, J.K., 1970. Spatial autocorrelation: a review of existing and new measures with applications. Economic Geography 46, 269–292.

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

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

Florax, R.J.G.M., Folmer, H., Rey, S.J., 2003. Specification searches in spatial econometrics: The relevance of Hendry's methodology. Regional Science and Urban Economics 33, 557–579.

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. American Journal of Preventive Medicine 28, 169–176. doi:10.1016/j.amepre.2004.10.018

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

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

Macfarlane, G.S., Garrow, L.A., Moreno-Cruz, J., 2015. Do Atlanta residents value MARTA? Selecting an autoregressive model to recover willingness to pay. Transportation Research Part A: Policy and Practice 78, 214–230.

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.

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

Small, K.A., Rosen, H.S., 1981. Applied Welfare Economics with Discrete Choice Models. Econometrica 49, 105. doi:10.2307/1911129

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

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

Wells, N.M., Ashdown, S.P., Davies, E.H.S., Cowett, F.D., Yang, Y., 2007. Environment, Design, and Obesity. Environment and Behavior 39, 6–33. doi:10.1177/0013916506295570

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