

ARTICLE



Impacts of the choice of distance measurement method on estimates of access to point-based resources

Amruta Nori-Sarma^{1,2✉}, Keith R. Spangler^{1,2}, Biqi Wang^{2,3}, Nina Cesare², Kimberly A. Dukes^{2,3} and Kevin J. Lane^{1,2}

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

BACKGROUND/OBJECTIVE: Lack of access to resources such as medical facilities and grocery stores is related to poor health outcomes and inequities, particularly in an environmental justice framework. There can be substantial differences in quantifying “access” to such resources, depending on the geospatial method used to generate distance estimates.

METHODS: We compared three methods for calculating distance to the nearest grocery store to illustrate differential access at the census block-group level in the Atlanta metropolitan area, including: Euclidean distance estimation, service areas incorporating roadways and other factors, and cost distance for every point on the map.

RESULTS: We found notable differences in access across the three estimation techniques, implying a high potential for exposure misclassification by estimation method. There was a lack of nuanced exposure in the highest- and lowest-access areas using the Euclidean distance method. We found an Intraclass Correlation Coefficient (ICC) of 0.69 (0.65, 0.73), indicating moderate agreement between estimation methods.

SIGNIFICANCE: As compared with Euclidean distance, service areas and cost distance may represent a more meaningful characterization of “access” to resources. Each method has tradeoffs in computational resources required versus potential improvement in exposure classification. Careful consideration of the method used for determining “access” will reduce subsequent misclassifications.

Keywords: Geospatial analysis; Exposure differences; Resource access; Disparities

Journal of Exposure Science & Environmental Epidemiology (2023) 33:237–243; <https://doi.org/10.1038/s41370-022-00414-z>

INTRODUCTION

Environmental factors have been investigated extensively in the environmental justice literature to understand the possibility for adverse health impacts associated with unequal access to an array of resources, including urban green spaces, grocery stores, and hospitals/medical care facilities, among others [1–3]. Additionally, examining factors that characterize individuals’ local environments can help to provide evidence on the extent to which neighborhood factors are related to behavioral choices (for example, how distance to resources may impact choices to utilize those resources), which can also lead to important health outcomes. For example, previous studies that have examined the impacts of neighborhood deprivation have reported fewer supermarkets and grocery stores in low-income and minority neighborhoods, reducing residents’ access to healthy and affordable foods [4].

One of the major challenges for estimating the potential impacts of differential access to resources on human health is the geospatial method for assessing “distance”, which often relies on a buffer of a particular straight-line distance (i.e., Euclidean or “as the crow flies”), ignoring the roads, sidewalks, and topography that substantially influence the actual distances to points of interest. Several previous studies have examined multiple alternative

methods for distance estimation within a health context; most have shown that Euclidean distance tends to underestimate the distance between two points of interest [5, 6].

The aim of this paper is to illustrate how different methods of estimating access distance can lead to meaningful differences in defining the population whose access to the place of interest is “high” or “low” based on whether they fall within a predetermined buffer distance, thereby influencing studies of the relationship between access to resources/locations of resources and health or other outcomes. Using access to grocery stores in Atlanta, Georgia (GA) as a case study, we assess three distance-estimation methods—traditional Euclidean versus two alternate methods that account for road networks and topography—for comparison of population access rates at the census block-group level. Grocery-store access is a good example of a resource to examine in this context, given the wealth of literature that has looked at the impacts of differential access to healthy foods on health outcomes, food deserts in low-socioeconomic status neighborhoods and minority race/ethnicity neighborhoods, etc [3, 7–11]. Here, we focus on exploring tradeoffs between the methodologies and discuss the opportunities for improved estimates and reduced exposure misclassification in a variety of interdisciplinary contexts.

¹Department of Environmental Health, Boston University School of Public Health, Boston, MA, USA. ²Biostatistics and Epidemiology Data Analytics Center, Boston University School of Public Health, Boston, MA, USA. ³Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. ✉email: anorisar@bu.edu

Received: 4 January 2021 Revised: 13 January 2022 Accepted: 20 January 2022
Published online: 10 February 2022

METHODS

Grocery-store location data

We identified the location of establishments that sell healthy foods using one of the components of the Modified Retail Food Environment Index, which includes identification of retail establishments that CDC considers to constitute access to healthy foods (referred to as “grocery stores” in the rest of the paper). We established the locations of grocery stores using geocoded location data through 2018 from Data Axle Reference Solutions (formerly known as Reference USA), a data aggregation service that maintains detailed records on the location of all US businesses [12]. We defined grocery stores following the criteria used by CDC in the calculation of the modified retail food environment index [13], which include: grocery stores with at least ten employees (North American Industry Classification Codes [NAICS]: 445110), produce markets (NAICS: 445230), and wholesale stores (NAICS: 452311). We imposed additional criteria and data cleaning to most accurately reflect healthy food access points and to remove misclassified locations: we included all major-chain grocery stores regardless of number of employees, additionally added in major department stores that feature grocery sections and removed locations that were classified as corporate headquarters. The final number of observations in the national data set was 38,922, of which 62 were located within the Atlanta city boundaries.

Population estimates

To estimate the population with access within each census block, we used high-resolution gridded population estimates at a 250 m resolution from the Global Human Settlements Population (GHS-POP) 2015 data set [14]. The GHS-POP disaggregates gridded population estimates from v4.10 of the Gridded Population of the World data set [15] using remotely sensed estimates of built-up density to approximate the spatial distribution of populations (i.e., the algorithm distributes populations within administrative boundaries commensurate with land-use indicators of development). To ensure consistency with the most up-to-date population estimates, we converted the gridded population estimates to points and created spatial weights by dividing the point population by the sum of the points within each census block group. We then multiplied each of these point spatial weights by the 2018 American Community Survey

block-group population totals (Fig. 1). At the census block-group level, we estimated the percentage of the population with access by summing all the population points within each block group that intersected the access buffers (described below) and then dividing by the total population of that block group (Fig. 1).

Method 1: Euclidean distance

Euclidean distance is a commonly used distance metric for environmental health studies, measured as a straight-line distance between two points “as the crow flies” [16]. This method is simple and intuitive but has very few applications in which it can yield accurate distance estimates. For this comparison, we include a Euclidean distance buffer of 1 mile—a “walkable” distance threshold as used in other studies [9–11]—around each grocery store (Fig. 2a) and calculate the percent of each census block-group population that falls within at least one of the buffers (i.e., the percent of the population with “access” to a grocery store). We compare the population proportions with access directly to the service areas (Method 3, below) and, additionally, calculate septiles of access for comparison with the cost-distance analysis (Method 2, below).

Method 2: Service area

We used the Generate Service Areas tool in ArcGIS Pro 2.7 (© Esri: Redlands, CA), which is a cloud-based network analysis service from Esri that calculates highly detailed buffers around points of interest with user-specified distances and mode of travel. The specific road network used by this tool is updated continually by Esri; our analysis was run in November 2020 using contemporaneous road networks. We calculated 1-mile walking distance buffers (Fig. 2b), for which the tool identified all the locations that can reach the grocery store within a 1-mile walk (excluding things like highways that are inaccessible to pedestrians). We then calculated the population proportions within each census block group within at least one service area, as was done for Method 1.

Method 3: Cost distance analysis

Cost-distance analysis relies on raster data to generate the distance to the nearest source (or, in our case, the nearest grocery store) for every cell in

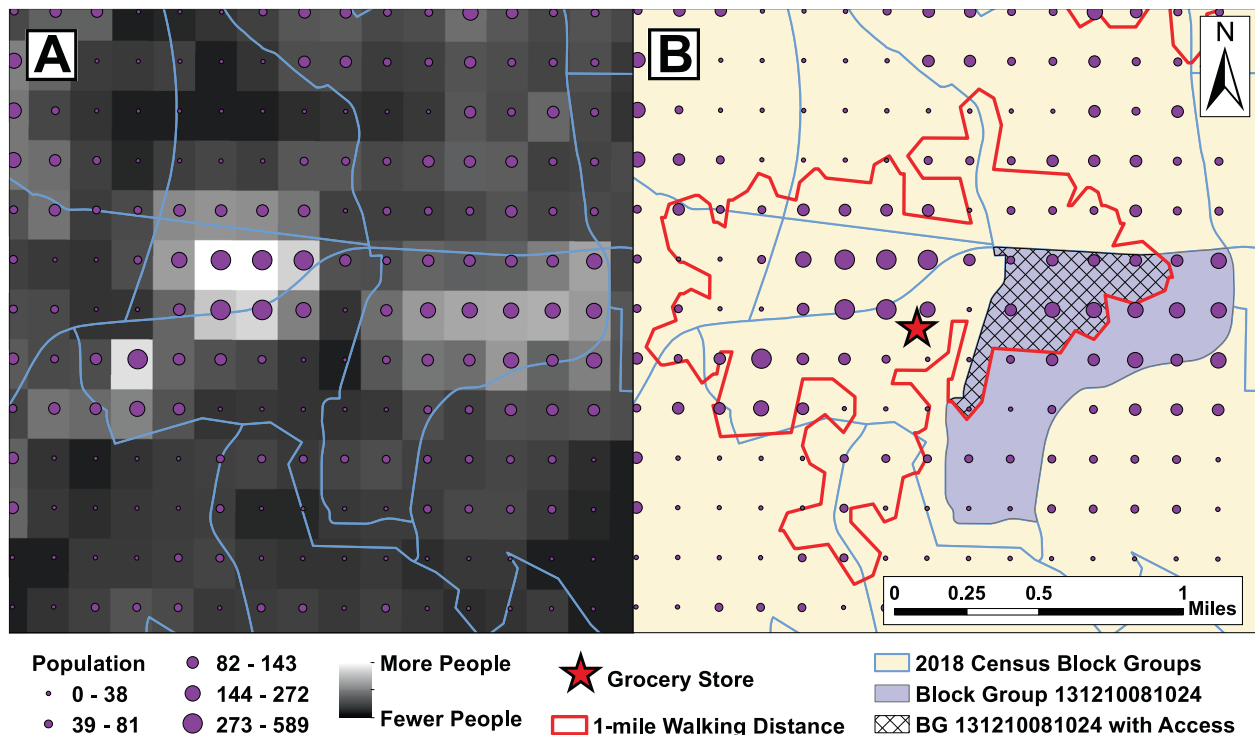


Fig. 1 Generation of gridded population estimates, by census block group. High-resolution gridded population estimates were converted to points and overlaid on polygons of 2018 census block groups (A). Percent of population with access to grocery stores was calculated by summing the population points that intersect with the 1-mile access buffers within a particular block group (B cross-hatched polygon) and dividing by the sum of all population points that intersect that same block group (B purple polygon).

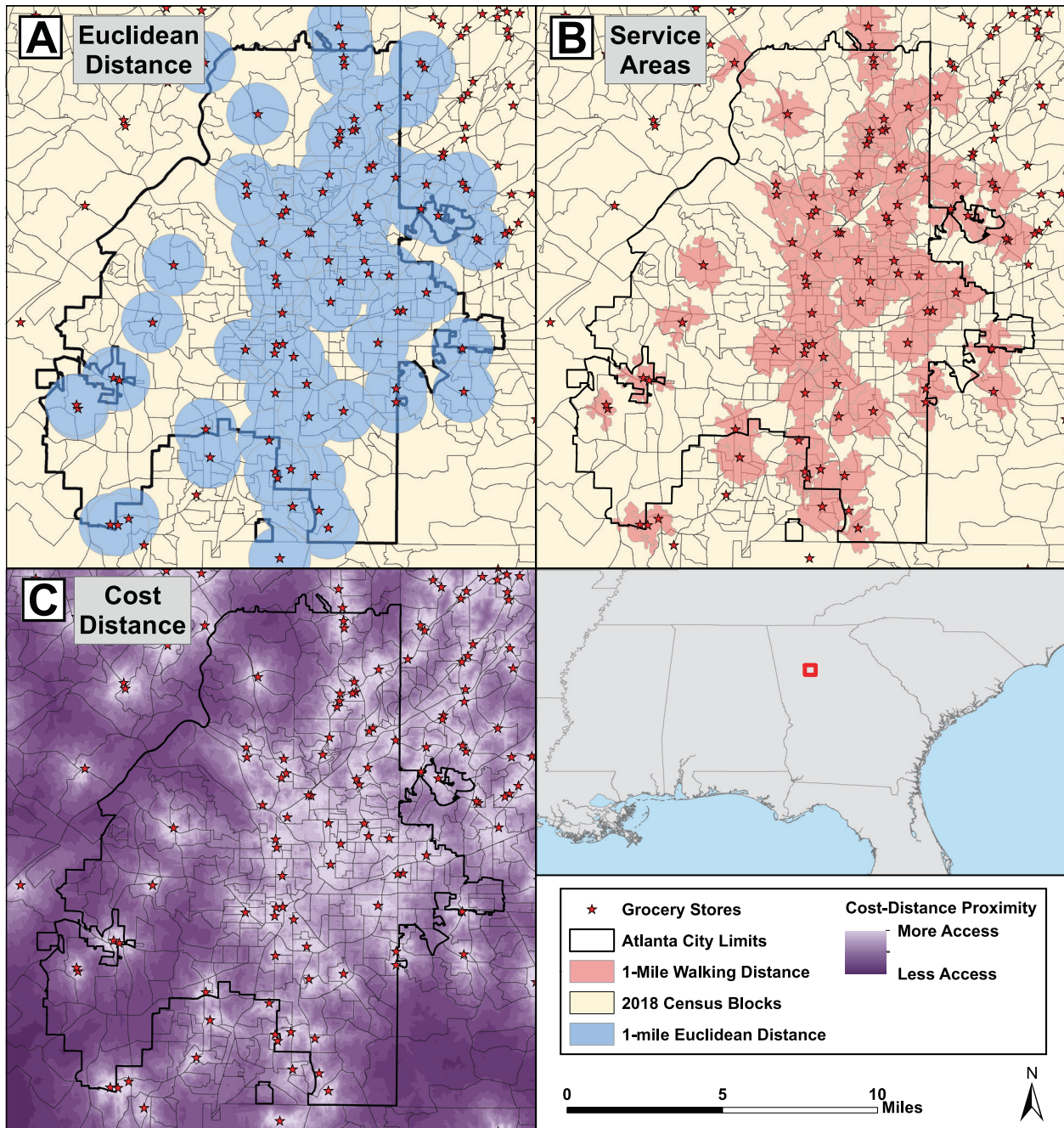


Fig. 2 Illustration of methods for estimating access to the nearest grocery store. Euclidean distance (A), service area (B), and cost distance (C).

the raster, based on the least-accumulative “cost” estimated by a cost surface; using road networks from OpenStreetMap [17] as inputs, this method essentially quantifies the path of least resistance between grid cells and points of interest to provide a proxy value of relative distances. For this analysis, the cost surface was estimated in ArcGIS Pro using the “Cost Distance” tool relying on major roads, minor roads, and non-road areas, to generate a map of the relative “cost” for every point on the map of Atlanta to the nearest grocery store (Fig. 2c) at a 10 meter grid resolution. In contrast to Methods 1 and 2, the cost-distance approach estimates relative distances between each population point and the nearest point of interest, rather than designating a population point as within a particular threshold of access. To make comparisons between these different methods, these relative distances were averaged to the census block group and results were categorized into septiles to map areas of lower and higher access to grocery stores.

Comparison across access measures

To ensure the cost-distance estimate would be comparable with the other access estimation methods in the raw data, since the cost-distance approach does not make use of the population points (as described previously for the other two approaches) we used min–max normalization of the estimates of cost-distance generated in the analysis. Additionally, to compare both the Euclidean distance and the service area methods with the cost-distance method in terms of access to the nearest grocery store, we generated septiles of the block group-level population estimates within a 1-mile buffer of the nearest grocery store. By comparing across septiles, we can provide a comparison of potential differences in exposure classification at the neighborhood level using all 3 proposed methods. We further estimated the intraclass correlation coefficient (ICC) across the three methods, to provide a statistical comparison of the differences in estimation between the three methods.

All analyses were conducted in the ESRI suite of GIS software (ArcMap, ArcGIS Pro 2.7, and ArcGIS Online; © Esri: Redlands, CA) and R version 3.6.2 [18].

RESULTS

Figure 3 shows the distribution of the actual estimates generated using each of the methods. In Fig. 3, lower estimates on the y axis indicate census block groups with “better” access (lower % of low access/outside of the indicated buffers in the Euclidean distance and service area methods, or less time to the nearest resource in the cost-distance method). Higher estimates indicate worse access for the census block group. We found differences in the distributions of grocery store access at the census block-group level between the three distance-estimation measures assessed. In particular, the Euclidean method identified substantially fewer block groups as being completely without access (i.e., block groups where 0% of the population falls within an access buffer): the service areas method identified 106 block groups compared to 188 using the Euclidean method. Moreover, the distributions of block-group-level percent population within access buffers differed as well: the median for Euclidean was 0% (inter-quartile range [IQR]: 0% to 45.9%), compared to 21.9% for the service areas (IQR: 0% to 85.4%) and 0.032 units for the cost distance (IQR: 0.003, 0.118). Spread also differs starkly between the methods, with cost-distance having the least variability and service areas having the most (Fig. 3).

We found differences in spatial patterns of access to grocery stores based on the distance-estimation method used (Fig. 4). In particular, we observe substantial homogeneity of access in the areas where more of the population has access to grocery stores using the Euclidean method, a current standard in the literature, owing to the high degree of overlap between the buffers in the most-densely populated parts of the city. By contrast, both the service areas and cost-distance methods reveal more variation in the amount of grocery store access, particularly in the city center.

In Fig. 5, the number of census block groups falling into each septile is demonstrated by the thickness of the septiles by access

method. Under the Euclidean distance method, Fig. 5 illustrates that many block groups had 100% access to grocery stores, resulting in less differentiation of septiles by comparison with the Cost-Distance method which has differentiation for all septiles, so some of the categories (6, 5, 4—indicating relatively high access) have been dropped from the figure. Change in septile of access to grocery stores between different access methods is illustrated as the line shifts between access method, and the number of block groups that are changing between methods is shown by the thickness of the connecting lines. Supplementary Table S1 shows the number of block groups within each septile, by access method.

Table 1 shows the estimation of ICC, a statistical measure of the reliability of the estimates across the three methods. For this study, we chose the one-way random effect model with a single rater and without bias, notated as in Table 1 as the ICC1. The ICC1 may be regarded as an estimate of the population intraclass correlation ρ_1 . ICC1 values higher than 0.8 indicate good agreement, while ICC values between 0.5 and 0.8 indicate moderate agreement; our ICC of 0.69 (0.65, 0.73) shows that the three methods have only moderate agreement, likely being driven at least in part by the relative agreement between the cost distance and service area methods shown in Fig. 5.

DISCUSSION AND CONCLUSIONS

We found meaningful differences in the quantiles of access as defined by the Euclidean distance as compared with service area and cost distance (Figs. 4, 5). The cost-distance approach, which is a continuous measure of relative distances rather than proportions with access within a predefined buffer, can easily be distributed into discrete septiles. By contrast, the Euclidean and service area approaches resulted in many areas with either entirely no access or full access: the Euclidean method found more than half of block groups to have 100% of populations with access (i.e., everyone in half of the block groups in Atlanta live within a 1-mile Euclidean buffer of a grocery store), while the service area method found

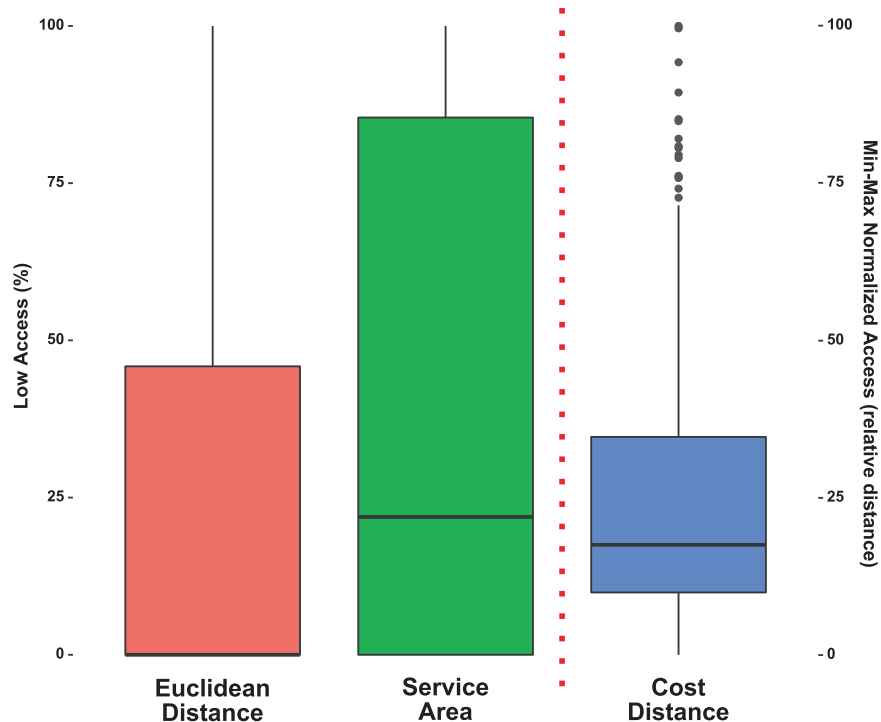


Fig. 3 Distribution of access measures—% with low access (for Euclidean distance and service area methods)/min–max normalized access (for cost-distance method); lower “access measure” indicates block groups with better access.

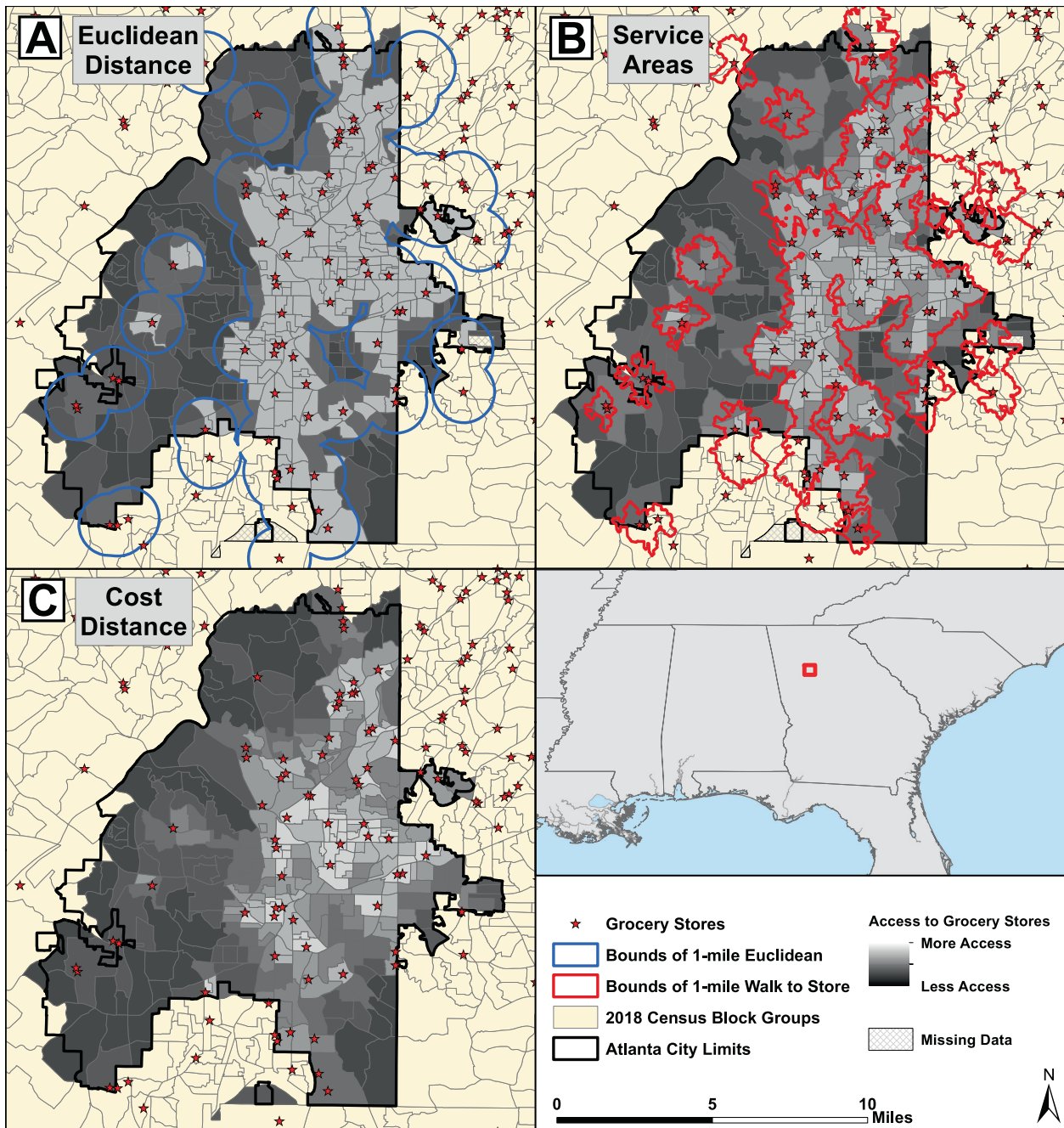


Fig. 4 Septiles of access to grocery stores in Atlanta at the census block-group level, by access method. Euclidean distance (A), service area (B), and cost distance (C).

about one-fourth of block groups each with 0% or 100% of populations with access. As a result, we observed substantial amounts of septile differentiation between the methods, particularly when comparing to the Euclidean approach (Fig. 4b, c). We also found potential for high levels of possible misclassification moving from the cost-distance method to the service area method and onto the Euclidean distance method, as portions of very high-access census block groups in one move into the lowest-access septile in another method, and vice versa. If only the ICC were taken into consideration, the differences between these measures may not be fully appreciated, indicating the need for alternative methods such as the geospatial plots (Fig. 4) and the alluvial plot (Fig. 5) to more fully evaluate the agreement between each of these different measures of access.

Many previous studies have relied heavily on Euclidean distance to assess access to resources, including healthy food access [7, 8]. Relatively fewer previous studies have aimed to characterize the potential for exposure misclassification by measurement method and have found mixed results regarding the potential for misclassification. One study focused on accessibility of urban health resources and noted important variability in the correlation between Cartesian measurements (including Euclidean distance) and network analysis, especially concentrated in suburban areas where the population density as well as the frequency of resources may diminish [19]. In another study focused on food access measures in Portland, OR, results suggested the same relative patterns of food access regardless of street network or Euclidean distance measures, although estimates varied with

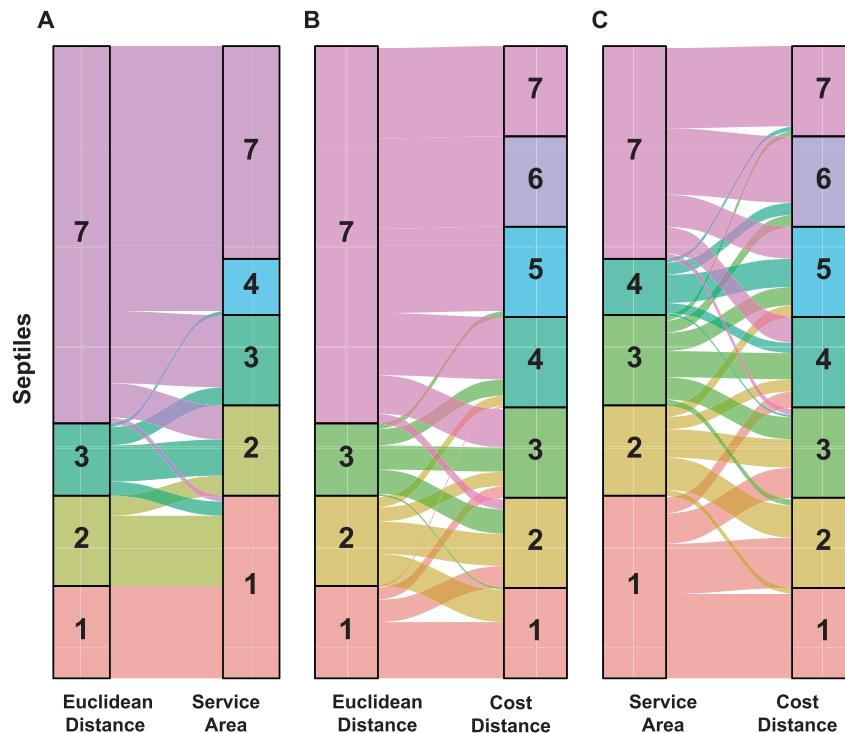


Fig. 5 Change in septic of access to grocery stores in the Atlanta metropolitan area, by access method (1—lowest access, 7—highest access).). Comparing Euclidean distance with service area (A), Euclidean distance with cost distance (B), and service area with cost distance (C).

Table 1. Results from calculation of intraclass correlation coefficient (ICC), for differences in septiles between access estimation method.

Intraclass correlation coefficient								
	Type	ICC	Lower bound	Upper bound	F	df1	df2	<i>p</i>
Single raters absolute	ICC1	0.69	0.65	0.73	7.7	314	630	3.70E–103

Number of subjects = 315; Number of Judges = 3

population density (again suggesting the importance of the type of urban environment) [8]. Importantly, these previous studies focused on estimating the distance from the centroid of the census tract or census block group to the nearest grocery store, while our analysis builds buffers around the resource itself and estimates access using highly spatially resolved population data, representing an improvement in the estimation of access using population density, which has been noted as a limitation in the previous work.

There are pros and cons of each of these different access estimation measures, that may be relevant to future research assessing issues of access to resources across different contexts and communities. The Euclidean distance method requires low computational effort but may introduce exposure misclassification by assuming that geometric area matches with surface streets. The service area method, which we consider to be more accurate because it takes into consideration pedestrian footpaths and nonlinear transportation routes, is more computationally intensive and requires either specialized software and/or access to data that is not freely accessible to all researchers. Investigators should take these computation and financial costs into consideration when proceeding with data analysis using the service area method.

By comparison, the cost-distance method is the most computationally intensive, relying on raster data and generating unique distance estimates for every pixel within a map. However, our results found that cost-distance provided more conservative estimates of low and high-access areas. Cost-distance analysis

may be run on ArcGIS relying on open-source data and does not contain embedded traffic or other information which may require funding. In this way, Cost-Distance analysis is limited to relative comparisons as opposed to providing absolute distances. The results may be highly informative at the aggregate or neighborhood level for subsequent ecological analysis, but investigators should be cautious in attempting to apply this measure to individual level studies, to avoid potential biases in exposure classification.

Although we did not evaluate the potential impact of exposure misclassification as part of a health study, we hypothesize that bias stemming from spatial error in assigning exposure could impact point estimates and standard errors in regression modeling. Our results illustrate that the Euclidean distance method overestimates the percentage of people who do have access to a given resource, in our example to grocery stores, compared to using a network or cost-distance approach, which would be more restrictive and approximate real-world access. This has the potential to overstate the accessibility of resources leading to more people incorrectly assigned to the “exposed” or “having access” group. We hypothesize that the result would be non-differential exposure misclassification which would bias results towards the null with regard to the standard error. Utilizing the Euclidean distance approach, the majority of the bias would occur in the unexposed (non-access) group being misassigned as exposed (with access); therefore, the bias would not be equally distributed as the true exposed (with access) population would be

less likely to be misassigned to the nonexposed group. This could result in bias impacting the standard errors of a health analysis but also biasing the point estimate as well and potentially leading to inversion of the associations. This could additionally impact population health studies and interventions if areas that are believed to have access are left out of plans to improve access. As one application relevant to our study, healthy food access may be impacted, for example through farmers markets or incentives for grocers to expand operations to fill the demonstrated gaps in access. Future health studies applying Euclidian distance to measure access should consider investigating the magnitude and directionality of this error in a formal health study.

Overall, our analysis demonstrated the high potential for misclassification in health studies that rely on Euclidean distance to estimate access to grocery stores in the Atlanta area. The over-estimation of high-access areas using Euclidian distance has the potential to leave out populations in need when designing public health interventions aiming to increase food access. Even measures that incorporate more detailed information regarding available routes vary in the extent to which they identify areas as low or high access. Future studies should aim to improve exposure estimation using alternative methods, such as calculating service areas or cost distance, depending on availability of data resources and computational power to assess these measures.

REFERENCES

- Sexton K. Socioeconomic and racial disparities in environmental health: is risk assessment part of the problem or part of the solution? *Hum Ecol Risk Assess Int J*. 2000;6:561–74. <http://www.tandfonline.com/doi/abs/10.1080/10807030008951330>.
- Lee C. Environmental justice: building a unified vision of health and the environment. *Environ Health Perspect*. 2002;110 SUPPL. 2:141–4.
- Walker RE, Keane CR, Burke JG. Disparities and access to healthy food in the united states: a review of food deserts literature. *Heal Place*. 2010;16:876–84. <https://doi.org/10.1016/j.healthplace.2010.04.013>.
- Morland K, Wing S, Diez Roux A, Poole C. Neighborhood characteristics associated with the location of food stores and food service places. *Am J Prev Med*. 2002;22:23–9.
- Phibbs CS, Luft HS. Correlation of travel time on roads versus straight line distance. *Med Care Res Rev*. 1995;52:532–42. <http://journals.sagepub.com/doi/10.1177/107755879505200406>.
- Shahid R, Bertazzon S, Knudtson ML, Ghali WA. Comparison of distance measures in spatial analytical modeling for health service planning. *BMC Health Serv Res*. 2009;9:200. <https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-9-200>.
- Li Y, Luo M, Wu X, Xiao Q, Luo J, Jia P. Grocery store access and childhood obesity: a systematic review and meta-analysis. *Obes Rev*. 2019;obr.12945. <https://onlinelibrary.wiley.com/doi/abs/10.1111/obr.12945>.
- Sparks AL, Bania N, Leete L. Comparative approaches to measuring food access in urban areas: the case of Portland, Oregon. *Urban Stud*. 2011;48:1715–37.
- Costa BVL, Menezes MC, Oliveira CDL, Mingoti SA, Jaime PC, Caiaffa WT, et al. Does access to healthy food vary according to socioeconomic status and to food store type? an ecologic study. *BMC Public Health*. 2019;19:1–7.
- Caspi CE, Kawachi I, Subramanian SV, Adamkiewicz G, Sorensen G. The relationship between diet and perceived and objective access to supermarkets among low-income housing residents. *Soc Sci Med*. 2012;75:1254–62. <https://doi.org/10.1016/j.socscimed.2012.05.014>.
- Apparicio P, Cloutier MS, Shearmur R. The case of Montréal's missing food deserts: Evaluation of accessibility to food supermarkets. *Int J Health Geogr*. 2007;6:1–13.
- DataAxle. Leading business and consumer data for libraries and their patrons. 2020. <https://www.data-axle.com/what-we-do/reference-solutions/>.
- Centers for Disease Control and Prevention. Census tract level state maps of the modified retail food environment index. *Children's Food Environ State Indic Rep*. 2011;2011:53. <https://perma.cc/79QP-QTWF>.
- Schiavina M, Freire S, MacManus K. GHS-POP R2019A—GHS population grid multitemporal (1975-1990-2000-2015). *Jt Res Centre Eur Comm*. 2019. <https://data.europa.eu/euodp/en/data/dataset/0c6b9751-a71f-4062-830b-43c9f432370f>.
- Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 10. Palisades, NY: 2018.
- Laurini R, Thompson D. Fundamentals of spatial information system. The APIC S. London: Academic Press; 1992.
- OpenStreetMap Contributors. Planet dump retrieved from <https://planet.osm.org>. OpenStreetMap. 2020.
- R Core Team. R: a language and environment for statistical computing. Vienna, Austria: 2020.
- Fisher R. Intraclass correlations and the analysis of variance. In: *Statistical methods for research workers*. Edinburgh: Tweeddale Court: Oliver and Boyd; 1934. p. 156. http://www.haghigh.com/resources/materials/Statistical_Methods_for_Research_Workers.pdf.

AUTHOR CONTRIBUTIONS

ANS and KS had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: ANS, KS, and KJL. Acquisition, analysis, or interpretation of data: ANS, KS, and BW. Drafting of the paper: ANS, KS, and KJL. Critical revision of the paper for important intellectual content: All authors. Final approval of version to be published: All authors. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: All authors.

FUNDING

Funding for this work was provided by Sharecare Inc (<https://wellbeingindex.sharecare.com/>).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41370-022-00414-z>.

Correspondence and requests for materials should be addressed to Amruta Nori-Sarma.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.