Check for updates

Article

B Urban Analytics and City Science

Environment and Planning B: Urban Analytics and City Science
0(0) 1–21
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2399808318797135
journals.sagepub.com/home/epb
SAGE

Parcels, points, and proximity: Can exhaustive sources of big data improve measurement in cities?

Kevin Kane
Southern California Association of Governments, USA

Young-An Kim
University of California, USA

Abstract

While there has been no shortage of discussion of urban big data, smart cities, and cities as complex systems, there has been less discussion of the implications of big data as a source of individual data for planning and social science research. This study takes advantage of increasingly available land parcel and business establishment data to analyze how the measurement of proximity to urban services or amenities performed in many fields can be impacted by using these data—which can be considered "individual" when compared to aggregated origins or destinations. We use business establishment data across five distinctive US cities: Long Beach, Irvine, and Moreno Valley in California; Milwaukee, Wisconsin; and the New York borough of Staten Island. In these case studies, we show how aggregation error, a previously recognized concern in using census-type data, can be minimized through careful choice of distance measures. Informed by these regions, we provide recommendations for researchers evaluating the potential risks of a measurement strategy that differs from the "gold standard" of network distance from individually measured, point-based origins and destinations. We find limited support for previous hypotheses regarding measurement error based on the abundance or clustering of urban services or amenities, though further research is merited. Importantly, these new data sources reveal vast differences across cities, underscoring how accurate proximity measurement necessitates a critical understanding of the nuances of the urban landscape under investigation as measures appear heavily influenced by a city's street layouts and historical development trajectories.

Corresponding author:
Kevin Kane, Department of Research & Analysis, Southern California Association of Governments, 900 Wilshire Blvd., Suite 1700, Los Angeles, CA 90017, USA.
Emails: kevin@kevinkane.org; kane@scag.ca.gov

Keywords
Urban morphology, big data, accessibility, parcel-level analysis

Introduction and background

There is no shortage of articles discussing and defining "big data" in the spatial, and particularly the urban spatial contexts (Batty, 2013; Glaeser et al., 2015; Kitchin and McArdle, 2016; Shearmur, 2015; Ye and He, 2016). The "smart city" represents the ability, given the proliferation of timely sensor-based data, to see city planning in the span of days or minutes rather than years or decades (Batty, 2013), transformed through technology-enabled data such as Google Street View, geo-referenced tweets, and real-time captures of human or vehicular travel. Additionally, the paradigms of "computational social science" (Chang et al., 2014) or "data-driven science" (Kitchin, 2014) challenge the philosophical underpinnings of urban scholarship and consider the city as a complex adaptive system comprised of myriad entities and their interactions (Batty, 2005).

Kitchin and McArdle (2016) argue that the key definitional traits of big data are velocity and exhaustivity. While much attention has been paid to new types of high-velocity urban data, less emphasis is placed on the contribution of two exhaustive big data sources that are permeating both planning practice and social science research: tax parcels and business establishments. Government administrative records—particularly those from local and regional agencies—are increasingly available and typically feature both georeferencing and universal or near-universal coverage (Einav and Levin, 2014; Newbold and Brown, 2016). In response to calls for increased data transparency, many city agencies have placed comprehensive land, tax, and public finance records online at no cost to citizens—or researchers (Arribas-Bel, 2014). Longitudinal data are increasingly common as some counties and municipal planning authorities began digitization projects years ago. Commercially produced, georeferenced datasets of business establishments are increasingly used in urban and regional research and have been used extensively for their ability to track business relocations, interindustry linkages, and specific business types (Funderburg and Zhou, 2013; Meltzer and Capperis, 2016; Neumark et al., 2005).

While these data sources are less dynamic and less popular than Google Street View and Twitter, for example, both land parcels and business establishments are very much "big" in terms of volume, computing requirements, and programming knowledge. As Shearmur (2015) describes, such exhaustive data have the possibility of revealing different conclusions than their census analogues since the scale of measurement is much finer. In particular, they are individual data which record the "population" of structures in the built environment or of business establishments in a city. The observation in either of these data sources is an individual (land parcel or business) at a discrete point in space, at once addressing the ecological inference problem and the Modifiable Areal Unit Problem (Openshaw and Taylor, 1981)—key concerns in "small data" research.

While the implication is that socioeconomic analysis could yield different findings using these individual level data, they also raise the concern that as the sample approaches the entire population, interpretation is a-theoretical and inference approaches pure description (Newbold and Brown, 2016). While Anderson (2008) claims that big data will bring the "death of theory," the more optimistic perspective of Einav and Levin (2014) is that it will instead allow for sharper tests of existing models and theories of human behavior, and if anything, require more of an organizing theory to make research progress. Parcel-level

datasets have long been used by agricultural and natural resource economists for their ability to link observed spatial patterns of urban land use with land change processes, for example in studies of zoning's role in land conversion at the urban–rural fringe (Bell and Irwin, 2002). Additionally, the contention that such data are merely descriptive is no longer problematic when "big" data are leveraged with other sources, like the census (Glaeser et al., 2015). Glaeser also notes that "many practical problems in cities do not directly require causal inference," highlighting their utility for planners and practitioners as well as researchers.

This paper is concerned with the implications of integrating these two sources of big data—tax parcels and business establishments—into the measurement of the proximity of urban services. A varied literature discusses accessibility measurement in urban space. In planning and public health, applications often focus on healthy food access or the equity of public service provision (Apparicio et al., 2008; Talen, 2003). A broader literature in geography and planning on spatial interaction addresses accessibility and urban morphology through an explicit treatment of the relationship between origins and destinations. Spatial interaction modeling can be used to evaluate not only service accessibility but job access and transportation demand (Bhat et al., 2002; Fotheringham and O'Kelly, 1989; Kwan, 1998). However, distance measures are also relied upon by a wide array of social scientists studying neighborhood effects including economists, criminologists, sociologists, and urban planners who are concerned with how proximity in urban space affects outcomes like health, wealth, crime, or property value appreciation.

This paper compares methods of measuring accessibility between residential land parcels and neighborhood-level destinations such as restaurants, doctor's offices, and grocery stores with an empirical focus on the latter given its prominence in the literature. Using the hypothesis that the network distance between a home (origin) and a business (destination) is the optimal measure of access, we assess two spatial issues through comparison to this newly available "gold standard": the aggregation error that arises from zonal rather than individual representation of origins and destinations and the use of a street network topology versus straight-line (Euclidean) distance. The purpose of this analysis is to demonstrate how this finer-scaled, data-enabled approach toward proximity informs research on neighborhoods and urban spatial structure. Furthermore, we gauge how these foundational issues might be impacted by urban morphology—namely, the abundance and clustering present across an array of destination types. Contrasting the empirical results with the growth history and development trajectories of the five cities analyzed (Long Beach, Irvine, and Moreno Valley in California, Milwaukee, Wisconsin, and the New York borough of Staten Island) provides added insight as to how urban morphology impacts access measurement.

While the conclusions hint at some possible best practices for researchers and practitioners alike, what is most striking about our results is how these fine-grained data illustrate interurban differences in measurement error. All five cities—even the three located in Southern California—have different development histories and, when compared to the "gold standard," demonstrate the importance of a nuanced understanding of the urban landscape of the region under investigation in the measurement of proximity—an insight revealed largely through these new sources of data.

Measuring proximity in urban space

A number of studies provide a comparison of various methods for measuring the accessibility between origins and destinations in urban space. Chief concerns are aggregation error, the imprecision that arises from using areal units to represent individual origins or

destinations, and the error arising when straight-line or Euclidean distance is used rather than a more accurate street network topology. Specific results vary across studies, many of which have applications in public health, such as measuring the accessibility of health services (Apparicio et al., 2008), the proximity of tobacco retailers (Duncan et al., 2014), recreational amenities (Hewko et al., 2002), or the walkability between homes and urban destinations (Oliver et al., 2007). A particular interest has emerged in measuring healthy food access and the identification of so-called food deserts. Research here largely concludes that while measurement strategies might produce different results, most are able to identify the same key underserved areas regardless of aggregation error or use of Euclidean distances (Leete et al., 2011; Sparks et al., 2010). Spatial interaction research has also advanced to include sensor-based mobility studies and space-time analysis (Gao et al., 2013; Ye and Rey, 2011); however, the present study focuses instead on data sources and measurement.

We take into account two increasingly available fine-grained data sources: business establishments and residential land parcels. Parcel-level land-use data are widely used in studies of agricultural land conversion which seek to investigate the impact of property-level characteristics like slope, cost, and zoning designation to the likelihood of development (Bell and Irwin, 2002; Kane et al., 2014) with the advantage being vastly improved statistical identification since the scale matches that of the land-use decision maker.

In addition to administratively sourced parcel data, commercially provided business establishment data are increasingly used in urban and economic research. Neumark et al. (2005) provide a thorough evaluation of Dun and Bradstreet's new National Establishment Time Series (NETS) data. This is a commercial database of business establishments by location compiled and telephone-verified by Walls & Associates with an expressed purpose for sale to other businesses. Of relevance for researchers seeking to adopt these data is an early finding by Neumark et al. (2005) of a substantial overcount of establishments and employment when compared to official US Bureau of Labor Statistics figures: 17–22% higher for employment and 38–63% higher for establishments. The overcount was highest—184%—for employment in businesses with one to four employees, substantially changing the ability to understand small business location with higher resolution data.

Proximity to businesses or other destinations in urban space is of key concern to a wide variety of social science researchers. For example, studies of crime and place have found that proximity to certain types of businesses and other urban features such as freeways or parks at the block or street segment scale can increase or decrease the risk of violent and property crime (Groff and McCord, 2012; Kim, 2016). In addition to the increasingly apparent scale-sensitivity of this line of research, many such studies frequently use Euclidean distance measures since they are less computationally demanding. The present study's comparison of individual-level measures versus aggregated origins and destinations using Euclidean and network distance can help to assess the potential benefits of increasingly "big" available data for a wide variety of researchers.

Foundational spatial issues

Distances can be calculated either by solving on a street network topology, or using a Euclidean or "straight-line" measurement. The latter measurement requires far less data, GIS processing, and specialized knowledge and might be preferred by planners, practitioners, or social scientists due to its simplicity and efficiency. Sparks et al. (2010) find that a shorter Euclidean distance threshold—roughly 35% shorter—approximates a street network distance; however, street network layout and urban morphology more generally could impact this (though few multi-city comparisons exist). While network distances are

not new to accessibility research, topological concerns remain challenging especially over short distances or when measuring pedestrian access. Van Eggermond and Erath (2015) demonstrate how differences between Euclidean and network measurements are far greater than differences between networks of different quality.

Aggregation error, a concern no matter the topology, arises when a single point is used to represent an area or zone comprised of a spatial distribution of individuals or destinations (Kwan, 1998). Since populations are often available only in aggregated areas, the entire population of a census tract, for example, is placed at a single point: the tract's centroid. While it is fairly common practice to use smaller scale census block data and area-weighted interpolation to improve tract-level accessibility estimates, even census blocks represent aggregated areas of individuals. Zoraghein et al. (2016) use parcel-level data to refine block-level population estimates; however, to our knowledge, this methodology has not yet been used for measuring proximity. Furthermore, using GIS to link a block centroid to its position on a street network is an imperfect exercise and a possible source of aggregation error even at this fairly fine scale. Van Eggermond and Erath (2015) explore these issues using pedestrian-specific street networks with varying degrees of realism, finding that the connection of buildings to the network has increased importance at a finer scale. Aggregation to blocks in network space might compound sensitivities based on which direction a building's entrance faces, or whether or not it is trivial to traverse to the other side of a block.

In addition, most prior work only considers the aggregation error arising when populations or *origins* are zonal measures. Most studies, such as those on healthy food access, focus on a fairly small count of a single type of destination. Given the computational ease of calculating Euclidean distances in a GIS or conventional statistical software, an understanding of the associated aggregation error would be of use to practitioners and researchers using proximity as a covariate in a statistical analyses of neighborhood-level well-being. For example, controlling for the region-wide accessibility between homes and liquor stores might be helpful in understanding neighborhood-level crime (Pridemore and Grubesic, 2012).

Establishment distribution and urban morphology

In addition to spatial methodological issues in accessibility measurement, we attempt to use these data and the notion of a "gold standard" to gauge how proximity measurement varies based on the abundance and spatial distribution of various types of urban destinations services, as well as provide a qualitative comparison of how results vary across cities that differ based on their development histories, street layouts, and zoning legacies. For example, the clustering of retail businesses has been shown to be an outcome of demand externalities (Nelson, 1958), yielding a very patchy spatial distribution for some business types. Shopping centers and malls are one logical outcome. Zoning as well as policies directed at enhancing commercial concentration can result in clustering as opposed to regularity (see, e.g., Boarnet et al., 2010).

Owing to an available longitudinal parcel dataset, three of this study's cities are in Southern California, a region of 19 million residents known for its urban sprawl and automobile culture. However, considerable variation exists across its more than 200 local jurisdictions in street network layout, planning history, and the resultant spatial distribution of homes and businesses. In order to expand the scope of this paper beyond one metropolitan region, we also investigate Milwaukee, Wisconsin, a legacy industrial city in the American Midwest, and Staten Island, a comparably sized borough of New York City.

Material and methods

Study area

At nearly 1/2 million residents Long Beach is the fifth-largest city in California, lying at the southern edge of Los Angeles County. The city has a variety of neighborhood types including a historic downtown, traditional suburban enclaves, disinvested and high-crime communities, and a large international port ("Intra City," 2016). Long Beach has sizeable populations of all the region's racial/ethnic groups, a variety of income levels, and a nearly even split between single-family and multi-family housing ("Profile of the City of Long Beach," 2015). Its street network is largely gridded and predates Second World War.

Irvine, in neighboring Orange County, stands out as an extreme example of master planning (Cronon, 1995). The Irvine Company, a legacy of the 19th-century landowners' expansive ranches, strategically developed the town over several decades. Irvine it is known for its high-end housing, gated communities, and carefully manicured commercial areas. The Irvine Company maintains exceptionally tight control over zoning and land-use regulations, and the street networks consist largely of winding, high-speed arterial roads with residential communities set between them. While its business subcenter nearly rivals downtown Los Angeles for regional dominance (Kane et al., 2016), it is a rare example of being entirely synthetic in that its location has no historical precedent (Forstall and Greene, 1997).

Moreno Valley, in inland Riverside County, is an archetypical urban fringe boomtown. Its history of growth coalitions, incorporation battles, and rapid, unplanned growth—particularly during the 1980s—contrasts with Irvine's more top-down planning (Jonas, 1999). Its roots as an agricultural region as well as its flatter topography contribute to a more typical 1/2-mile street grid with single-family homes developed on sections within, and retail on major N/S or E/W roads.

Milwaukee, Wisconsin is a mid-sized city along Lake Michigan's western shore. Incorporated in 1846, the city's history is characterized by industry and immigration. While it could be considered as part of the United States' "Rust Belt" and is well below its mid-century population peak, the city remains economically vibrant. It consists of both a historic core and areas annexed in the 1950s and 1960s with a suburban land-use character (Hegerty, 2015). Staten Island is by far the smallest and least dense of New York City's five boroughs. Although settled during the 17th century, the borough was not integrated into New York City until 1898 and its conversion from agricultural to largely residential use proceeded gradually (Rizzi, 2012).

Table 1 provides summary statistics of each city's street network characteristics from the US Environmental Protection Agency's (EPA) Smart Location Database (Ramsay & Bell, 2014). Total road network density roughly measures the overall presence of thoroughfares, while street intersection density may be more closely related to accessibility since it approximates the ability to traverse the street network. "Auto" and "pedestrian" designations measure streets based on speeds, intersection type, and lane count to gauge the appropriateness of auto or pedestrian travel, though no information on sidewalks is included. Long Beach scores highest on most measures, consistent with its older, more rectilinear street layout. Milwaukee and Staten Island are comparable but slightly lower in both pedestrian, total, and intersection density—expected of legacy cities. Irvine, which incorporated in 1971, has a much lower intersection and pedestrian density but a far greater density of auto-oriented links, consistent with its predominantly high-speed arterial roads. Moreno Valley is the newest city and has the lowest street network, street intersection, and pedestrian density.

Table 1. Population and street network characteristics in study areas.

City	State	Population[a] (2010 Census)	Population per square mile[b]	Year incorporated	Total road network density	Street intersection density	Network density: facility miles of auto links per sqm	Network density: facility miles of pedestrian links per sqm
Irvine	CA	212,375	3,238	1971	22.29	95.41	5.71	15.25
Long Beach	CA	462,257	9,186	1897	26.84	151.84	1.19	21.23
Moreno Valley	CA	193,365	3,772	1984	15.33	62.89	1.27	12.42
Milwaukee	WI	594,833	6,185	1846	23.85	118.93	1.81	18.64
Staten Island	NY	469,645	8,028	1898[c]	23.17	116.96	1.50	19.18

[a]*Source:* Ramsay and Bell (2014).
[b]*Source:* American Community Survey (2015), [c]Year of integration into New York City.

Data sources

The primary data sources used in this study are parcel-level (cadastral) data used to identify residences (origins) and point-based business establishment data used as destinations. The Southern California Association of Governments (SCAG) maintains 2012 land-use records at the parcel-level for Los Angeles, Orange, Riverside, San Bernardino, Ventura, and Imperial Counties. A four-digit classification of land uses allows for the identification of each parcel's land use. For this study, we isolate residential parcels. Milwaukee parcel data are from the County Assessor's database (*Master Property Record Data*, 2016) while Staten Island's are from the New York Department of City Planning (*MapPLUTO*, 2015).

While these cities represent a convenience sample, the data structure, time period, and land-use classification scheme differs widely across municipal or regional data providers (Kane and Clark, 2018). Alternative approaches using sensor-based data might allow for a statistically robust sample of cities which is not currently feasible using administrative records since the lack of inter-regional harmonization necessitates deep familiarity with each dataset. For example, Gebru et al. (2017) used Google StreetView in 200 US cities to estimate socioeconomic characteristics of neighborhoods based on the types of cars parked on streets. While this approach using a single data source allows for statistically robust interregional comparison, administrative records offer more robustness than sensor data since they are official measures which have been verified and adopted by governments.

Reference USA point-based business establishment data cover 1997–2014 (Infogroup, 2015) and provide coordinate data, an employee count, and the North American Industry Classification System (NAICS) code for every business establishment nationwide. These commercial data are similar in construction and source as the commonly used NETS. Our examination of topology and aggregation error will focus on a single category—grocery stores—given both their prominence in the food access literature and that they are near the mean in abundance in our data across 31 establishment types. Street network data are derived from Environmental Systems Research Institute's (ESRI) North America Detailed Streets (ArcGIS, 2014) and distances are solved on the network using ArcGIS 10.4 software ("ArcGIS 10.4," 2016).

Measures of accessibility

Figure 1 depicts six strategies for measuring the level of accessibility between origins and destinations, each labeled with a letter. The top-left corner indicates the "gold standard" against which alternatives are compared and is the parcel-to-establishment

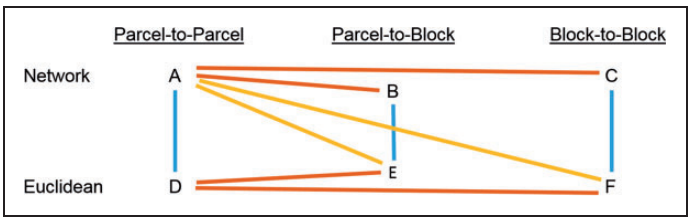

Figure 1. A typology of accessibility measurement.

(or, parcel-to-parcel) distance measure along the street network. The next column abstracts origins to census blocks, measuring distance to their centroids, similar to prior studies using zonal population measures. The third column also abstracts destinations to blocks, i.e., rather than measuring the distance to an actual destination we use the distance to the centroid of the nearest block which contains that type of establishment. This type of data structure might be used when establishments are too numerous or when due to data censoring concerns destination data are only available in zones.

Lines connecting measurement strategies indicate the comparisons undertaken. Comparisons along the verticals of Figure 1 (A–D, B–E, and C–F) investigate the Euclidean versus network tradeoff. Comparisons along the horizontals of Figure 1 (A–B, A–C, D–E, and D–F) analyze aggregation error stemming from the use of census blocks as zones to represent destinations (B and E) or both origins and destinations (C and F). Diagonals (A–E and A–F) reflect the mixed effect of aggregation error and topology. Point E is a situation that might be faced by a planner comparing the accessibility of between a set of point-based destinations and an aggregated source of population and uses Euclidean distances in order to reduce computing needs. Point F might occur when using zonal destination data (e.g., US Census ZIP Code Business Patterns) and using Euclidean distance to reduce computing needs—in particular if an accessibility study is to be compared across multiple urban regions.

This study examines three commonly used measurements of accessibility:

1. The distance between the origin and the nearest destination of that type.
2. The number of destinations within a threshold distance of the origin. A $^1/_2$-mile threshold distance is generally considered as the upper end of walkability (Congress for the New Urbanism (CNU), 2000). We also include a 1-mile threshold since this study's focus is not solely on walkability and a 3-mile threshold to better capture the retail trade areas.
3. A variety of gravity type measures which weigh nearby destinations highly and further away destinations—up to a certain threshold—less.

While a gravity type measure has the advantage of considering proximity and multiple destinations, results are sensitive to functional form. We explore two common measures: a power function and an exponential decay function (Fotheringham and O'Kelly, 1989). The power function is defined as

$$\sum_j \frac{1}{d_{ij}^{\alpha}} \forall k$$

where d_{ij} is the distance in feet from parcel i to establishment j for all establishment types k and α is a decay parameter. We examine three previously examined parameters ($\alpha = 0.5$, 1.0,

and 2.0) in order to provide an intuitive comparison for planners and researchers. While parameter choice can alter results substantially, given this paper's broad comparison of non-gravity measures, different distance cutoffs, and network versus Euclidean forms we eschew a complete simulation of all possible parameters. In instances where origins and destinations are at the same point (e.g., a small store in the lobby of a residential building), a correction of 0.02 miles was added to prevent an infinity value. Due to this requirement, the exponential decay function may be better suited for intraurban analysis and is given by

$$\sum_j \frac{1}{e^{\beta d_{ij}}} \forall k$$

where β is the distance decay parameter which we evaluate at $\beta = 0.2$ and 0.5, informed by Kwan (1998). Pearson's correlation (r) is used for comparing measurements based on nearest distance (1) and counts (2) which are scalar measures that can be directly correlated. Spearman's rank correlation (rho) is used for gravity type measures (3) since they are functions of the above equations rather than direct measurements.

Measures of abundance and clustering

Hewko et al. (2002) and Sparks et al. (2010) note that the abundance and clustering of destinations may influence aggregation error. As an additional analysis in Long Beach only, we use the simple Nearest Neighbor Index (NNI) to measure the level of clustering of each of 31 different establishment types (Lee and Wong, 2000). NNI compares the distance of each establishment to its nearest neighbor of the same type to a null hypothesis of spatial randomness. NNI for establishment type k is given by

$$NNI_k = \frac{\frac{\sum_i d_i}{n}}{\frac{0.5}{\sqrt{n/A}}}$$

where d_i is the distance between establishment i and its nearest neighbor of type k, n is the total number of establishments of type k, and A is the area of a minimum enclosing rectangle around all establishments of type k. NNI values below 1 indicate high clustering while values above 1 indicate dispersion. Inference is achieved using a z-score relating the actual and random (hypothetical) spatial distribution. Euclidean distance-based NNI rather than network distance-based NNI is used to ensure that the statistic is not dependent on street network characteristics.

Results

Aggregation error

Results for aggregation error, which compare the effect of substituting parcel-to-parcel measurement with block-to-parcel or block-to-block, are shown in Table 2.

A–B (network distances) and D–E (Euclidean distances) measure the effect of substituting zonal measures of destinations. In the older cities of Long Beach, Milwaukee, and Staten Island this has almost no effect on measurement no matter the distance threshold or functional form with all correlations above 0.96 and most above 0.99. No matter whether

Table 2. Correlations between block and parcel-level measures.

City	Long Beach				Irvine				Moreno Valley				Milwaukee				Staten Island			
Analysis (see Figure 1): *Measurement*	A–B	D–E	A–C	D–F	A–B	D–E	A–C	D–F	A–B	D–E	A–C	D–F	A–B	D–E	A–C	D–F	A–B	D–E	A–C	D–F
Distance to nearest	0.988	0.997	0.863	0.970	0.916	0.989	0.836	0.841	0.962	0.996	0.887	0.949	0.970	0.999	**0.794**	0.809	0.966	0.995	**0.744**	0.917
Count within 1/2-mile	0.988	0.998	0.970	0.990	0.918	0.969	**0.336**	0.847	0.883	0.978	**0.520**	0.843	0.993	0.993	0.976	0.904	0.992	0.994	0.962	0.980
Count within 1 mile	0.997	0.999	0.992	0.998	0.936	0.990	**0.619**	0.950	0.969	0.993	0.815	0.946	0.997	0.998	0.986	0.967	0.995	0.997	0.987	0.993
Count within 3 miles	0.999	1.000	0.997	0.944	0.986	0.996	0.893	0.875	0.942	0.998	0.926	0.986	0.994	1.000	0.979	0.987	0.999	1.000	0.998	0.954
Power ($\alpha = 0.5$), 1/2-mile	0.980	0.996	0.943	0.982	0.811	0.950	**0.485**	0.818	**0.772**	0.941	**0.509**	0.811	0.990	0.990	0.952	0.896	0.987	0.984	0.932	0.958
Power ($\alpha = 0.5$), 1-mile	0.994	0.999	0.963	0.996	0.911	0.989	**0.645**	0.955	0.959	0.994	0.805	0.951	0.996	0.998	0.966	0.956	0.993	0.997	0.961	0.985
Power ($\alpha = 0.5$), 3-mile	0.997	1.000	0.969	0.964	0.980	0.997	0.809	0.923	0.932	0.999	0.897	0.992	0.995	1.000	0.969	0.982	0.997	1.000	0.975	0.984
Power ($\alpha = 1$), 1/2-mile	0.978	0.996	0.944	0.973	0.807	0.952	**0.491**	0.811	**0.773**	0.943	**0.508**	0.813	0.987	0.989	0.944	0.882	0.987	0.984	0.914	0.952
Power ($\alpha = 1$), 1-mile	0.991	0.998	0.968	0.988	0.911	0.985	**0.671**	0.924	0.961	0.994	0.811	0.947	0.992	0.997	0.961	0.938	0.990	0.996	0.951	0.970
Power ($\alpha = 1$), 3-mile	0.994	0.998	0.975	0.981	0.962	0.994	**0.734**	0.951	0.896	0.997	0.850	0.983	0.992	0.999	0.965	0.957	0.992	0.997	0.976	0.976
Power ($\alpha = 2$), 1/2-mile	0.974	0.994	0.936	0.952	0.804	0.953	**0.494**	0.804	**0.773**	0.944	**0.507**	0.814	0.983	0.979	0.907	0.834	0.986	0.976	0.853	0.931
Power ($\alpha = 2$), 1-mile	0.985	0.995	0.958	0.960	0.916	0.982	**0.692**	0.872	0.960	0.993	0.811	0.932	0.985	0.975	0.919	0.852	0.988	0.974	0.886	0.914
Power ($\alpha = 2$), 3-mile	0.986	0.995	0.965	0.962	0.941	0.983	**0.714**	0.887	**0.761**	0.994	**0.728**	0.944	0.986	0.971	0.924	0.836	0.990	0.964	0.925	0.879
Exponential ($\beta = 0.2$), 1/2-mile	0.979	0.994	0.952	0.983	0.810	0.943	**0.480**	0.820	**0.772**	0.938	**0.507**	0.810	0.990	0.989	0.969	0.899	0.987	0.982	0.938	0.959
Exponential ($\beta = 0.2$), 1-mile	0.995	0.999	0.980	0.996	0.903	0.987	**0.622**	0.957	0.955	0.992	0.800	0.951	0.997	0.997	0.987	0.958	0.994	0.996	0.976	0.987
Exponential ($\beta = 0.2$), 3-mile	0.999	1.000	0.997	0.953	0.981	0.997	0.852	0.890	0.939	0.999	0.922	0.993	0.997	1.000	0.989	0.986	0.999	1.000	0.999	0.978
Exponential ($\beta = 0.5$), 1/2-mile	0.980	0.995	0.953	0.983	0.810	0.945	**0.480**	0.821	**0.772**	0.939	**0.508**	0.811	0.991	0.989	0.969	0.899	0.987	0.983	0.940	0.960
Exponential ($\beta = 0.5$), 1-mile	0.995	0.999	0.981	0.996	0.907	0.989	**0.630**	0.960	0.958	0.993	0.804	0.953	0.997	0.998	0.987	0.958	0.994	0.997	0.977	0.988
Exponential ($\beta = 0.5$), 3-mile	0.999	1.000	0.997	0.973	0.984	0.998	0.840	0.948	0.933	0.999	0.912	0.996	0.998	1.000	0.992	0.980	0.999	1.000	0.999	0.986

Table displays aggregation error-based results using lettered codes from the accessibility typology in Figure 1. Values are correlation coefficients, using Pearson's *r* for distances and counts, and Spearman's rho for power-based measures. Values below 0.8 are in boldface.

a researcher is interested in measuring the nearest destination, a count of destinations within a threshold, or a variety of possible distance decay-based measures, so long as either origins or destinations are point-based there is minimal loss of accuracy. However, this is not the case in Irvine or Moreno Valley where several correlations are near or below 0.8 (shown in boldface). In Irvine, the A–B correlation for distance to the nearest destination is only correlated at the level of $r = 0.916$. Gravity measures using a $1/2$-mile threshold are also very low (near 0.8) while a 1-mile threshold is an improvement (near 0.91) and a 3-mile threshold shows very minimal error, with correlations between 0.94 and 0.98 depending on functional form. In Moreno Valley, gravity-type 1/2-mile measures for A–B correlation perform especially poorly with most values around 0.77. Gravity-type 1-mile measures suffer less aggregation error than 3-mile variants. A 3-mile threshold is inferior using a power law, especially with a steep decay (e.g., a 0.7614 correlation for power $\alpha = 2$, 3-mile).

Turning to A–C and D–F, which measure the impact of abstracting both origins and destinations to polygons versus a point-to-point "gold standard," we see that the distance to nearest measure is deficient across all cities, especially using network distances. In Milwaukee (0.794) and Staten Island (0.744), these are the only correlations of any kind below 0.8. The D–F (Euclidean) equivalent measures are actually higher, which supports the conclusion of Van Eggermond and Erath (2015) who suggest that topological issues become magnified at short distances.

In the legacy cities of Long Beach, Milwaukee, and Staten Island, A–C and D–F correlations are slightly lower than A–B and D–E, indicating that using zonal measures for both points is costly; however, in most cases, correlations remain above 0.9. For count or gravity measures in all cities, a wider threshold is better: correlations tend to be higher at 1-mile than $1/2$-mile and at 3-miles than 1-mile. Minimal differentiation can be seen between exponential β values of 0.2 or 0.5. The steepness of the α distance decay parameter in the power function depends on both the city and the distance threshold. With a generous 3-mile threshold, a steeper parameter typically increases aggregation error. However, it appears to have less impact at a narrow $1/2$-mile threshold, with almost no change in correlation in the California cities and a slight decrease in correlation value in Milwaukee and Staten Island.

In the newer developing cities of Irvine and Moreno Valley, aggregating both origins and destinations results in some very deficient measures: as low as 0.34 for A–C comparisons in Irvine which use a count of destinations within $1/2$-mile. This suggests that while aggregation error can be minimized, choosing an appropriate accessibility measure is very important. In Moreno Valley, the $1/2$-mile threshold which is most deficient with correlations hovering around 0.5—roughly the same as Irvine. At 1-mile thresholds in Moreno Valley, correlations improve to around 0.8, improving to around 0.9 in most cases at 3 miles (a steeply decaying power function is an exception at 0.728). However, this improvement at wider thresholds is not seen in Irvine: many 1-mile measures are around 0.6 and some 3-mile measures are below 0.8.

Case A–C demonstrates the most consistently low correlations, again highlighting Van Eggermond and Erath's (2015) point that network measures can strongly affect aggregation error, especially with narrow distance thresholds. These results suggest that this is much more problematic in newer cities. In the legacy cities of Long Beach, Milwaukee, and Staten Island, the distance to nearest measure can raise concerns, which makes sense in a denser urban context with more rectilinear streets: the point at which a destination is located on network is more important.

Based on these findings, aggregation error has the potential to be problematic but this can be mitigated through careful choice of measures such as avoiding $1/2$-mile thresholds and

the use of zonal measures for *both* origins and destinations. Consistent with Sparks et al. (2010), Apparicio et al. (2008) and others in the food desert literature, we find that aggregation error is not an overwhelming problem in identifying underserved areas.

Euclidean versus network topology

Euclidean versus street network comparisons are conducted for parcel-to-parcel relationships (AD), blocks-to-parcels (BE), and blocks-to-blocks (CF). Results are displayed in Table 3. Long Beach stands out from all other cities—even the other legacy cities of Milwaukee and Staten Island—for having substantially fewer measures below 0.8. Many gravity-type measures appear better in Milwaukee and Staten Island than the newer cities of Irvine and Moreno Valley, which both have numerous correlations dipping below 0.5.

A–D correlations, which measure point-to-point distance using network versus Euclidean topologies, generally are more closely correlated than B–E or C–F measures. As seen in the previous section, a general rule appears to be that a $1/2$-mile threshold distance can result in more topological error than a 1-mile threshold, with a 3-mile threshold offering the highest correlation. The difference is fairly dramatic: for exponential distance decay with $\beta = 0.5$ in Moreno Valley, correlations increase from 0.634 to 0.795 to 0.961 as the threshold increases. A notable exception to this is in the count measures in three legacy cities, which does not monotonically increase as the threshold distance increases.

While an increasingly steep distance decay parameter had an ambiguous or slightly negative effect on aggregation error, when comparing network versus Euclidean topologies, steeper α or β parameters in the gravity-type measures tend to perform better. Using a 1-mile threshold and a power law in Milwaukee, correlations increase from 0.857 ($\alpha = 0.5$) to 0.901 ($\alpha = 1$) to 0.950 ($\alpha = 2$). However, there are exceptions such as a 3-mile threshold and a power law in Long Beach, where correlations are highest at $\alpha = 1$ (0.937 vs. 0.922 for the other two α values).

On the whole, the CF case (block-to-block) shows the lowest correlations between Euclidean and network distances. For example, for the CF case in Irvine, the correlation between network and Euclidean measures for the "count of establishments within $1/2$-mile" of a residence is a meaningless $r = 0.269$. Even when using a power law ($\alpha = 1$) and a 1-mile threshold, the correlation is only 0.546.

Most striking is the difference between the three cities. In most instances in Long Beach, Euclidean distances can substitute for network distances with a correlation approaching or above 0.900 and as high as 0.977 for a power law ($\alpha = 1$) and a 1-mile threshold. Correlations in Moreno Valley and Irvine are much lower and reach a high of 0.962 in Moreno Valley (AD, exponential $\beta = 0.2$, 3-mile), but only 0.862 in Irvine (AD, exponential $\beta = 0.5$, 3-mile). In particular for power laws, Moreno Valley, 0.5-mile thresholds are particularly deficient while 1-mile and even some 3-mile thresholds have low correlations in Irvine. In other words, Euclidean approximations in Moreno Valley are not nearly as deficient when 1 mile is used as the cutoff. This may be due to the regularity of the street grid in Moreno Valley, which typically consists of $1/2$-mile spaced arterial streets. A trip from a typical home to the nearest intersection of two arterials (where businesses are often found) requires navigating out of a cul-de-sac through a feeder street before reaching the arterial, resulting in short trips that are far longer on the street grid than straight line. Once on an arterial road, there is less difference between Euclidean and street network distance measurements, and thus this problem subsides. Irvine's correlations are typically lower than Moreno Valley's and do not exhibit as much differentiation based on the threshold. This could reflect the more curvilinear nature of Irvine's arterial roads—a feature of its history of master

Table 3. Correlation between Euclidean and network accessibility measures.

City	Long Beach			Irvine			Moreno Valley			Milwaukee			Staten Island		
Analysis (see Figure 1*): Measurement*	A–D *r/rho=*	B–E *r/rho=*	C–F *r/rho=*	A–D *r/rho=*	B–E *r/rho=*	C–F *r/rho=*	A–D *r/rho=*	B–E *r/rho=*	C–F *r/rho=*	A–D *r/rho=*	B–E *r/rho=*	C–F *r/rho=*	A–D *r/rho=*	B–E *r/rho=*	C–F *r/rho=*
Count within 1/2-mile	0.939	0.934	0.922	**0.559**	**0.548**	**0.269**	**0.602**	**0.619**	**0.474**	0.821	**0.787**	0.835	**0.740**	**0.715**	**0.699**
Count within 1 mile	0.968	0.970	0.966	**0.582**	**0.541**	**0.403**	**0.786**	**0.764**	**0.689**	**0.716**	**0.687**	**0.684**	**0.704**	**0.691**	**0.655**
Count within 3 miles	0.908	0.897	0.892	0.827	**0.733**	**0.796**	0.940	0.910	0.891	**0.623**	**0.624**	**0.650**	**0.765**	**0.777**	**0.747**
Power ($\alpha = 0.5$), 1/2-mile	0.914	0.910	0.869	**0.468**	**0.569**	**0.420**	**0.635**	**0.628**	**0.499**	0.819	**0.797**	0.819	**0.715**	**0.688**	**0.684**
Power ($\alpha = 0.5$), 1-mile	0.951	0.940	0.892	**0.585**	**0.606**	**0.467**	0.809	**0.788**	**0.698**	0.857	0.844	0.860	**0.779**	**0.766**	**0.692**
Power ($\alpha = 0.5$), 3-mile	0.942	0.943	0.922	0.821	**0.795**	0.811	0.958	0.945	0.924	0.951	0.938	0.948	0.837	0.839	0.834
Power ($\alpha = 1$), 1/2-mile	0.928	0.921	0.883	**0.490**	**0.593**	**0.451**	**0.643**	**0.640**	**0.506**	0.859	0.833	0.843	**0.787**	**0.761**	**0.754**
Power ($\alpha = 1$), 1-mile	0.963	0.948	0.907	**0.640**	**0.683**	**0.546**	0.843	0.819	**0.733**	0.901	0.892	0.895	0.850	0.832	**0.779**
Power ($\alpha = 1$), 3-mile	0.977	0.965	0.937	0.860	0.831	**0.767**	0.930	0.939	0.910	0.982	0.966	0.969	0.839	0.834	0.822
Power ($\alpha = 2$), 1/2-mile	0.937	0.924	0.885	**0.524**	**0.618**	**0.488**	**0.656**	**0.655**	**0.519**	0.913	0.879	0.872	0.867	0.840	0.842
Power ($\alpha = 2$), 1-mile	0.966	0.945	0.906	**0.706**	**0.747**	**0.641**	0.876	0.850	**0.769**	0.950	0.926	0.915	0.937	0.916	0.880
Power ($\alpha = 2$), 3-mile	0.976	0.953	0.922	0.863	0.834	**0.745**	0.820	0.919	0.875	0.978	0.954	0.948	0.944	0.927	0.904
Exponential ($\beta = 0.2$), 1/2-mile	0.907	0.903	0.889	**0.462**	**0.559**	**0.413**	**0.634**	**0.626**	**0.498**	**0.795**	**0.773**	**0.794**	**0.662**	**0.638**	**0.626**
Exponential ($\beta = 0.2$), 1-mile	0.938	0.929	0.910	**0.543**	**0.559**	**0.442**	**0.793**	**0.774**	**0.695**	0.809	**0.794**	**0.800**	**0.699**	**0.695**	**0.591**
Exponential ($\beta = 0.2$), 3-mile	0.923	0.929	0.950	**0.779**	**0.780**	0.837	0.962	0.940	0.938	0.865	0.870	0.893	0.837	0.843	0.823
Exponential ($\beta = 0.5$), 1/2-mile	0.907	0.903	0.889	**0.462**	**0.559**	**0.413**	**0.634**	**0.626**	**0.498**	**0.796**	**0.773**	**0.795**	**0.663**	**0.639**	**0.627**
Exponential ($\beta = 0.5$), 1-mile	0.942	0.932	0.913	**0.555**	**0.564**	**0.445**	**0.795**	**0.775**	**0.696**	0.823	0.808	0.818	**0.719**	**0.713**	**0.615**
Exponential ($\beta = 0.5$), 3-mile	0.959	0.953	0.964	0.863	0.827	0.842	0.961	0.954	0.945	0.946	0.941	0.950	0.824	0.829	0.819

Pearson's *r* is used for distances and counts; Spearman's rho is used for gravity-based measures. Values below 0.8 are in boldface.

planning and hillier topography. These two cities do have the lowest population densities of the five cities analyzed (see Table 1) and likely have larger census blocks in terms of land area. In addition, the road network density and street intersection density are both higher in Long Beach than the other legacy cities of Milwaukee and Staten Island. In fact, most back alleys in the city have street names and some addresses. We suspect this urban morphological idiosyncrasy may account for far higher Euclidean-network correlations there.

Finally, while a number of count and exponential decay measures are poorly approximated by Euclidean distance in Milwaukee and Staten Island, most power law measures in these places have correlations above 0.8—even measures that performed poorly in Moreno Valley and Irvine. While we suspected exponential functions may perform better due to non-infinity values at zero distance, they show lower correlations than a power law in these cities.

Joint effects

Joint effects, the AE and AF cases in Figure 1, are shown in Table 4. Planners are increasingly called upon to integrate and manipulate newer, bigger urban data but they may not have advanced GIS expertise or substantial computing power. Relationship AE is one such

Table 4. Correlations considering joint effects.

City	Long Beach		Irvine		Moreno Valley		Milwaukee		Staten Island	
Analysis (see Figure 1*)* *Measurement*	A–E	A–F	A–E	A–F	A–E	A–F	A–E	A–F	A–E	A–F
Distance to nearest	0.954	0.939	0.875	**0.755**	0.944	0.897	0.978	0.800	0.948	**0.689**
Count within 1/2-mile	0.948	0.947	**0.608**	**0.558**	**0.706**	**0.653**	0.810	0.832	**0.717**	**0.673**
Count within 1 mile	0.977	0.977	**0.562**	**0.573**	0.813	**0.799**	**0.725**	**0.717**	**0.686**	**0.630**
Count within 3 miles	0.915	0.923	0.819	**0.786**	0.943	0.931	**0.613**	**0.615**	**0.768**	**0.744**
Power ($\alpha = 0.5$), 1/2-mile	0.949	0.937	**0.562**	**0.446**	**0.768**	**0.619**	0.834	0.824	**0.695**	**0.595**
Power ($\alpha = 0.5$), 1-mile	0.966	0.963	**0.599**	**0.590**	0.857	0.833	0.865	0.856	**0.766**	**0.640**
Power ($\alpha = 0.5$), 3-mile	0.954	0.975	0.810	**0.792**	0.962	0.960	0.953	0.933	0.834	0.802
Power ($\alpha = 1$), 1/2-mile	0.957	0.937	**0.576**	**0.449**	**0.775**	**0.620**	0.873	0.843	**0.773**	**0.645**
Power ($\alpha = 1$), 1-mile	0.971	0.961	**0.672**	**0.622**	0.879	0.840	0.913	0.893	0.838	**0.708**
Power ($\alpha = 1$), 3-mile	0.977	0.978	0.836	0.803	0.929	0.923	0.983	0.954	0.837	**0.792**
Power ($\alpha = 2$), 1/2-mile	0.960	0.926	**0.598**	**0.461**	**0.781**	**0.620**	0.924	0.845	0.862	**0.690**
Power ($\alpha = 2$), 1-mile	0.970	0.942	**0.762**	**0.663**	0.906	0.842	0.957	0.892	0.934	**0.763**
Power ($\alpha = 2$), 3-mile	0.975	0.951	0.828	**0.745**	**0.791**	**0.779**	0.983	0.917	0.943	0.834
Exponential ($\beta = 0.2$), 1/2-mile	0.946	0.937	**0.560**	**0.450**	**0.768**	**0.621**	**0.799**	0.812	**0.624**	**0.547**
Exponential ($\beta = 0.2$), 1-mile	0.959	0.959	**0.571**	**0.584**	0.854	0.841	0.816	0.823	**0.687**	**0.565**
Exponential ($\beta = 0.2$), 3-mile	0.941	0.970	**0.789**	**0.764**	0.967	0.966	0.869	0.871	0.838	0.818
Exponential ($\beta = 0.5$), 1/2-mile	0.946	0.937	**0.560**	**0.450**	**0.768**	**0.621**	0.804	0.815	**0.631**	**0.554**
Exponential ($\beta = 0.5$), 1-mile	0.961	0.960	**0.573**	**0.584**	0.850	0.836	0.830	0.836	**0.707**	**0.586**
Exponential ($\beta = 0.5$), 3-mile	0.967	0.982	0.868	0.852	0.964	0.961	0.947	0.943	0.824	0.814

Table shows correlations between measures A–E and A–F for joint effects. Pearson's *r* is used for distances and counts, and Spearman's rho is used for gravity-based measures. Values below 0.8 are in boldface.

Table 5. Abundance and clustering in Long Beach.

Establishment type	Count	NNI
Home products retailing	4415	0.120
Full-service restaurants	1275	0.324
Auto services	849	0.446
Specialty retailing	837	0.463
Healthcare provider offices	815	0.331
Personal financial	669	0.362
Social service organizations	655	0.490
Hospitals	483	0.607
Apparel retailing	443	0.440
Religious organizations	439	0.593
Elementary and secondary schools	314	0.571
Other personal services	312	0.576
Repair services	311	0.627
Personal products retailing	298	0.632
Groceries	268	0.635
General merchandise retailing	267	0.645
Specialty food	247	0.582
Limited-service food and beverage	183	0.763
Medical laboratories	180	0.644
Child care services	155	0.709
Deposit-taking institutions	152	0.494
Beer, wine, and liquor stores	151	0.859
Drug stores	150	0.587
Gas stations	149	0.772
Recreational facilities and instruction	145	0.683
Drinking places (alcoholic beverages)	93	0.722
Hair care services	92	0.809
Movie theaters	69	0.643
Laundry	68	0.552
Convenience stores	63	0.909
Other learning	11	1.230

Note: NNI refers to nearest neighbor index. Counts include the city boundary plus a 1-mile (Euclidean) buffer to capture establishments accessible to residents living near the city limits.

instance for a planner interested in measuring the distance from census blocks to a set of points, e.g., a single business type. Relationship AF may be encountered in multi-city research when distance to a neighborhood service is being used as a covariate explaining some socioeconomic outcome and destinations are represented zonally.

Overall patterns look similar to previous analyses, with substantial differences across cities. No correlations in Long Beach are below 0.9, while only a select few correlations in Milwaukee are below 0.8—mostly count measures. Every measure using a $1/2$-mile threshold in Moreno Valley has a correlation below 0.8 while in Irvine the only measures above 0.8 are distance-to-nearest and 3-mile thresholds. Staten Island exhibits a number of low correlations, particularly the A–F measures which compare the network gold standard against the Euclidean measure using zonal origins and destinations. A peculiarity emerges in that some A–F measures are better than measures in Table 3 which compare Euclidean versus network distance. For example, the count of establishments within a $1/2$-mile threshold is less deficient here (0.755) than in Table 3 (0.269–0.559), suggesting that a component of the

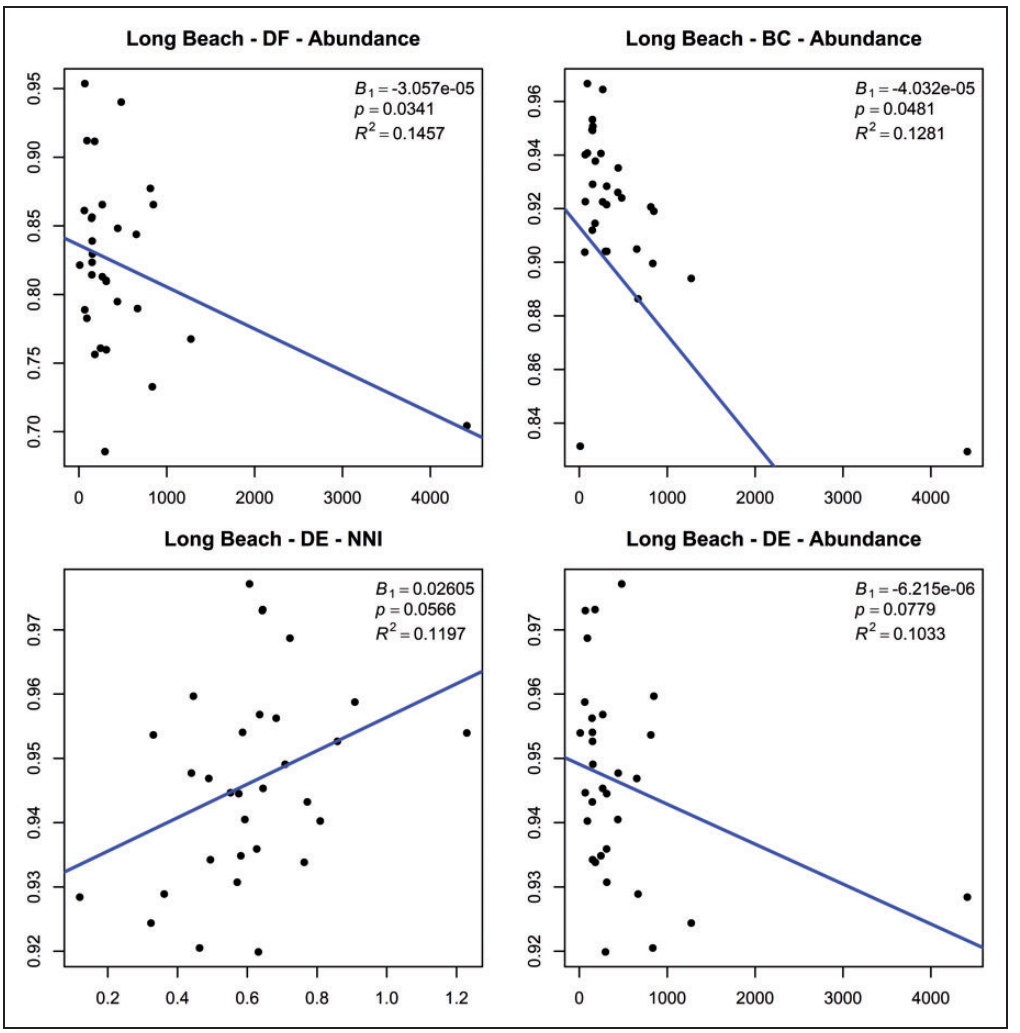

Figure 2. Scatterplots of abundance, clustering, and error in Long Beach.

high error involving B and C measurements may be network-based, consistent with Van Eggermond and Erath (2015).

The distance to nearest measure does not appear to suffer from as much "joint" error as the count or gravity approaches, particularly in the A–E case, where correlations for all cities are at least 0.875. Staten Island's A–F correlation is far lower (0.689) indicating that zonal origins and destinations may be problematic there; however, we can fairly confidently state across a variety of urban morphologies that sticking with minimum distance likely results in lower errors when A–E approximation is necessary.

Abundance and clustering

Our final analysis tests whether the level of abundance or clustering of establishments impacts the levels of error when measuring accessibility to them. This analysis only investigates one city—Long Beach—and uses a power ($\alpha = 1$) and a 1-mile threshold since this

yielded some of the most consistently high correlations earlier. Abundance is measured as the count of each type of establishment in each city, while the NNI measures its level of clustering. Table 5 shows the count and NNI for 31 establishment types in Long Beach.

We analyzed relationships AB, AC, DE, and DF—all four aggregation error cases. Plots were created with the level of correlation on the vertical axis and either abundance or NNI on the horizontal axis (Figure 2). Only four had an R-squared over 0.10. DF abundance, AC abundance, and DE abundance show a negative slope, suggesting that measurement error is higher when there are more destinations in a city. However, these three plots have a far outlier in the highly abundant home products retailing category, removal of which would weaken the relationship. In contrast, the nearest neighbor index plot for relationship D–E shows a positive slope, suggesting that more clustering leads to higher correlation. However, this is only weakly significant ($p = 0.056$) and is not present using network distances.

This could provide some weak evidence that measuring accessibility to more abundant destinations may increase aggregation error and measuring accessibility to more clustered destinations may decrease aggregation error. However, these relationships are weak and further research either across more cities or more measures is merited.

Discussion and conclusions

The evolving city data landscape provides many new opportunities for planners, practitioners, and researchers. This paper illustrates an integrative approach combining GIS and urban theory using two sources of individual level data: residential land parcels and business establishments. The goal is to inform social science research that involves a component of urban proximity in understanding outcomes, in addition to improving the ability of planners and practitioners to leverage these data sources in order to better understand their cities and improve public service delivery—even without advanced computing resources. In particular, these data sources allow for a "gold standard" of proximity measurement against which alternatives can be evaluated. Most prior research on proximity has focused on a single type of destination—often grocery stores or public services—and has aggregated at least origins or destinations to polygon areas.

First, while this investigation covers only five cities and does not simulate all parameter values, results hint at possible best practices. Aggregation error has the potential to be problematic but results suggest it might be mitigated by carefully choosing an accessibility measure. When only origins are measured zonally, using a shorter threshold distance such as $1/2$-mile should be avoided especially with network distances—consistent with Van Eggermond and Erath (2015). A zonal representation of both origins (homes) and destinations (businesses) can be more problematic and even impact the distance-to-nearest measure in legacy cities, whose measurement errors tended to be lower. If origins, destinations, or both are to be aggregated, a steeper distance decay parameter using either a power or exponential function typically increases error: $\alpha = 0.5$ or $\beta = 0.2$ generally performed better when compared to the "gold standard" but there are exceptions. In contrast, a steeper distance decay parameter such as $\alpha = 2$ or $\beta = 0.5$ typically performs better when a researcher is restricted to using Euclidean distance measures. A narrow, $1/2$-mile threshold can result in more topologically-induced error while a 3-mile threshold performed best; however, 3 miles may be too far a threshold when investigating walkability. When zonal origins (but not destinations) are used in conjunction with a Euclidean topology, a "distance to nearest" measure appears fairly robust. The selection of these well-trodden decay parameters and threshold distances for investigation provides an intuitive comparison for planners and researchers; however, a known shortcoming is that the entire parameter space is not

simulated. Such an approach might shed more light on idiosyncrasies or non-monotonic changes and is one we recommend for future research whose goal is to calibrate a travel or accessibility model.

Second, while cities were selected for this study based on qualitative differences, results varied more than expected. A set of five cities does not provide a statistically robust sample; however, these differences suggest that it would be challenging to confidently substitute measurement strategies across cities: a distribution of values of "the proximity to a grocery," for example, cannot simply be compared between one city and another. Some guidance might be had by considering urban morphology and development trajectory. The older, "legacy" cities of Long Beach, Milwaukee, and the New York borough of Staten Island performed consistently better on most measures and suffered from far less aggregation error than newly built Irvine and Moreno Valley. Moreno Valley's robust performance using 1-mile rather than 0.5-mile measures makes sense in light of its "cul-de-sacs within $^1/_2$-mile arterials" morphology, while Irvine's hilly topography and curvilinear roads make measurement of proximity more challenging. Long Beach performed strongly on most measures, likely a result of its especially high street intersection density. Myriad additional considerations—including controlling for city-level differences—are outside the scope of this work; however, a critical understanding of the nuances of the urban landscape under investigation appears to be an important component of accurate measurement. A promising area of future research might be to harmonize parcel-level land-use data from a statistically robust, random sample of cities or regions. Alternatively, sensor-based data such as Google StreetView could be used to identify urban features in the same manner across all cities. Nonetheless, this study's results caution that not every intraurban landscape feature can be codified, while many practical policy applications are in a single region.

Finally, this study provides some limited evidence to suggest that more abundant or less clustered destinations could be subject to more aggregation error. Despite the high-resolution of the data, residential and business location decisions remain idiosyncratic and these relationships are fairly weak and inconsistent. It is possible that a more systematic simulation or machine-learning approach that pools measurement strategies or synthetically distributes establishments might be able to establish a more generalizable rule, while consolidation of urban parcel data through a common platform similar to the General Transit Feed Specification (GTFS) system for standardizing transit data (Hadas, 2013) could increase the applicability of these conclusions to other US and international cities. However, our results stress that while big data can offer increasingly robust, individual-level descriptive statistics involving accessibility and other topics, not every intraurban landscape feature can be codified and many policy applications are regionally-specific nonetheless.

Authors' note

Young-An Kim is now affiliated with Florida State University, USA.

Acknowledgements

This paper would not have been possible without the help of MFI Director John R. Hipp as well as Simon Choi and Frank Wen of the Southern California Association of Governments and Scott B. Kelley. Findings reflect the views of the authors and not those of the Southern California Association of Governments, by whom this research was not funded.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We are grateful for the support of the Metropolitan Futures Initiative (MFI) in UC-Irvine's School of Social Ecology for funding support and data.

ORCID iD

Kevin Kane http://orcid.org/0000-0002-6109-7358

References

Anderson C (2008) The end of theory: Will the data deluge make the scientific method obsolete? *Wired, July 16*. Online publication. URL https://www.wired.com/2008/06/pb-theory/. Date accessed unknown, however the currently available version is stable/similar to the version accessed during the writing of this paper.

Apparicio P, Abdelmajid M, Riva M, et al. (2008) Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. *International Journal of Health Geographics* 7: 7.

ArcGIS 10.4 (2016) Redlands, CA: Environmental Systems Research Institute (ESRI).

ArcGIS (2014) *North America Detailed Streets*. Available at: https://www.arcgis.com/home/item.html?id=f38b87cc295541fb88513d1ed7cec9fd (accessed 24 May 2016).

Arribas-Bel D (2014) Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography* 49: 45–53.

Batty M (2005) *Cities and Complexity: Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals*. Cambridge, MA: MIT Press.

Batty M (2013) Big data, smart cities and city planning. *Dialogues in Human Geography* 3(3): 274–279.

Bell KP and Irwin EG (2002) Spatially explicit micro-level modelling of land use change at the rural-urban interface. *Agricultural Economics* 27(3): 217–232.

Bhat CR, Handy SL, Kockelman K, et al. (2002) Development of an urban accessibility index: Formulations, aggregation, and application. In: T. D. o. Transportation (ed) (Vol. Working Paper Report 7-4938-4). Austin: University of Texas at Austin. http://ctr.utexas.edu/wp-content/uploads/pubs/4938_4.pdf

Boarnet MG, Joh K, Siembab W, et al. (2010) Retrofitting the suburbs to increase walking: Evidence from a land-use-travel study. *Urban Studies* 48(1): 129–159.

Chang RM, Kauffman RJ and Kwon Y (2014) Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems* 63: 67–80.

Congress for the New Urbanism (CNU) (2000) *Charter of the New Urbanism*. New York: McGraw-Hill.

Cronon W (1995) In search of nature. In: Cronon W (ed) *Uncommon Ground: Rethinking the Human Place in Nature*. New York: W.W. Norton & Company, pp. 23–56.

Duncan DT, Kawachi I, Subramanian SV, et al. (2014) Examination of how neighborhood definition influences measurements of youths' access to tobacco retailers: a methodological note on spatial misclassification. *American Journal of Epidemiology* 179(3): 373–381.

Einav L and Levin J (2014) Economics in the age of big data. *Science* 346(6210): 1243089.

Forstall RL and Greene RP (1997) Defining job concentrations: The Los Angeles case. *Urban Geography* 18(8): 705–739.

Fotheringham AS and O'Kelly ME (1989) *Spatial Interaction Models: Formulations and Applications*. London: Kluwer.

Funderburg RG and Zhou X (2013) Trading industry clusters amid the legacy of industrial land-use planning in southern California. *Environment and Planning A* 45(11): 2752–2770.

Gao S, Liu Y, Wang Y, et al. (2013) Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17(3): 463–481.

Gebru T, Krause J, Wang Y, et al. (2017) Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences of the United States of America* 114(50): 13108–13113.

Glaeser EL, Kominers SD, Luca M, et al. (2015) Big data and big cities: The promises and limitations of improved measures of urban life *National Bureau of Economic Research Working Paper, 16-605*.. http://www.nber.org/papers/w21778

Groff E and McCord ES (2012) The role of neighborhood parks as crime generators. *Security Journal* 25(1): 1–24.

Hadas Y (2013) Assessing public transport systems connectivity based on Google Transit data. *Journal of Transport Geography* 33: 105–116.

Hegerty SW (2015) Commercial bank locations and "banking deserts": A statistical analysis of Milwaukee and Buffalo. *The Annals of Regional Science* 56(1): 253–271.

Hewko J, Smoyer-Tomic KE and Hodgson MJ (2002) Measuring neighbourhood spatial accessibility to urban amenities: Does aggregation error matter? *Environment and Planning A* 34(7): 1185–1206.

Infogroup (2015) *Reference USA Historical Business Dataset*. firm's location in Papillon, Nebraska, USA.

Intra City (2016, April 20, 2016) Special Issue, *The Daily 49er*. Available at: http://neighborhoods.d49erspecial.com/ (accessed 19 December 2016).

Jonas AE (1999) Making edge city: Post-suburban development and life on the frontier in Southern California. In: Harris R and Larkham P (eds) *Changing suburbs: Foundation, Form and Function*. London: E & FN Spon, pp.202–222.

Kane K and Clark WAV (2018) Mapping the landscape of urban work: Home-based businesses and the built environment. *Environment and Planning A*. (accessed 20 March 2018).

Kane K, Hipp JR and Kim JH (2016) Los Angeles employment concentration in the 21st century. *Urban Studies* 55(4): 844–869. DOI: 10.1177/0042098016678341.

Kane K, York AM, Tuccillo J, et al. (2014) Residential development during the Great Recession: A shifting focus in Phoenix, Arizona. *Urban Geography* 35(4): 486–507.

Kim Y-A (2016) Examining the relationship between the structural characteristics of place and crime by imputing census block data in street segments: Is the pain worth the gain? *Journal of Quantitative Criminology* 34(1): 67–110. DOI: 10.1007/s10940-016-9323-8.

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 1–12. DOI: 10.1177/2053951714528481.

Kitchin R and McArdle G (2016) What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3(1): 205395171663113.

Kwan MP (1998) Space-time and integral measures of individual accessibility: A comparative analysis using a point-based framework. *Geographical Analysis* 30(3): 191–216.

Lee J and Wong D (2000) *Statistical Analysis with ArcView GIS*. New York: Wiley.

Leete L, Bania N and Sparks-Ibanga A (2011) Congruence and coverage: Alternative approaches to identifying urban food desserts and food hinterlands. *Journal of Planning Education and Research* 32(2): 204–218.

MapPLUTO (2015). New York City Department of Planning, New York, NY.

Master Property Record Data (2016) Available at: http://city.milwaukee.gov/HistoricalMPROP.htm#.WcBONdVSzIU (accessed 27 June 2017).

Meltzer R and Capperis S (2016) Neighbourhood differences in retail turnover: Evidence from New York City. *Urban Studies* 54(13): 3022–3057.

Nelson R (1958) *The Selection of Retail Locations*. New York, NY: FW Dodge Corporation.

Neumark D, Zhang J and Wall B (2005) Business establishment dynamics and employment growth. *SSRN Electronic Journal*. Neumark: Regional Research Frontiers-Vol. 1. Springer, Cham, 2017, 317–328. DOI: 10.2139/ssrn.909249.

Newbold KB and Brown WM (2016) Human capital research in an era of big data: Linking people with firms, cities, and regions. *Regional Research Frontiers*.

Oliver LN, Schuurman N and Hall AW (2007) Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *International Journal of Health Geographics* 6: 41.

Openshaw S and Taylor PJ (1981) The modifiable areal unit problem. In: Wrigley N and Bennett R (eds) *Quantitative Geography: A British View*. London: Routledge and Kegan Paul, pp. 60–69.

Pridemore WA and Grubesic TH (2012) Community organization moderates the effect of alcohol outlet density on violence. *The British Journal of Sociology* 63(4): 680–703.

Profile of the City of Long Beach (2015) *Local Profiles Report*. Southern California Association of Governments. http://www.scag.ca.gov/DataAndTools/Pages/LocalProfiles.aspx (accessed 19 December 2016).

Ramsay K and Bell A (2014) *Smart Location Database*. Available at: https://www.epa.gov/sites/production/files/2014-03/documents/sld_userguide.pdf (accessed 7 July 2016).

Rizzi N (2012) New exhibit tracks the history of Staten Island. *DNA Info*., New York, NY.

Shearmur R (2015) Dazzled by data: Big Data, the census and urban geography. *Urban Geography* 36(7): 965–968.

Sparks AL, Bania N and Leete L (2010) Comparative approaches to measuring food access in urban areas: The case of Portland, Oregon. *Urban Studies* 48(8): 1715–1737.

Talen E (2003) Neighborhoods as service providers: A methodology for evaluating pedestrian access. *Environment and Planning B: Planning and Design* 30(2): 181–200.

Van Eggermond MAB and Erath A (2015) Pedestrian and transit accessibility on a micro level: Results and challenges. *Journal of Transport and Land Use* 9(3): 127–143. DOI: 10.5198/jtlu.2015.677.

Ye X and He C (2016) The new data landscape for regional and urban analysis. *GeoJournal* 81(6): 811–815.

Ye X and Rey S (2011) A framework for exploratory space-time analysis of economic data. *The Annals of Regional Science* 50(1): 315–339.

Zoraghein H, Leyk S, Ruther M, et al. (2016) Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems* 58: 19–28.

Kevin Kane is a regional planner in the Research & Analysis Department at the Southern California Association of Governments. His focus is regional demographic forecasting and research and analytical support of the land use and transportation elements of the agency's quadrennial Regional Transportation Plan and Sustainable Communities Strategy. Dr. Kane's PhD in Geography was completed at Arizona State University followed by a post-doctoral fellowship at the University of California-Irvine's Metropolitan Futures Initiative. His research interests include urban land-use change, neighborhood change, regional economic development, demographic forecasting, spatial analysis and modeling, sustainable cities, and urban transportation.

Young-An Kim is an assistant professor in the College of Criminology and Criminal Justice at Florida State University. His research interests focus on crime patterns at micro places, effects of structural characteristics of street segments on crime, and immigration and crime. Besides criminology, he is interested in sociology of health, urban sociology, and quantitative research methods. He approaches these research topics using quantitative methods as well as geospatial analysis.