# Spatial Proximity as a Measure of Activity-Space Segregation: Inferential Statistics and Sample Size Requirements

John R.B. Palmer

working draft - March 24, 2014

**Abstract**

This article analyzes the performance of White's index of spatial proximity as a measure of activity-space segregation using sampled data. It relies on data collected from volunteers with a mobile phone application and data generated from computer simulations to construct empirical sampling distributions of the index's estimator at a range of sample sizes and to test a bootstrap approach for evaluating the uncertainty of any individual estimate. The empirical distributions suggest that this index may be estimated with essentially no bias using coarse trajectory data with sample sizes as low as several hundred people. In addition, the uncertainty of individual estimates can be approximated well using bootstrap methods. The article concludes that these statistical properties, combined with the index's flexibility for measuring segregation at a range of scales by modifying its distance function, make it a valuable tool for research on activity-space segregation.

# Introduction

Palmer (2012) offers a framework for quantifying activity-space segregation, defined as a systematic separation in the spaces people move through as they go about their daily activities. One of the proposed measures of activity-space segregation is an extension of White's spatial proximity index, which may be used to capture information about differences in the places and the people with which different groups come into contact.

The spatial proximity index appears to be particularly well-suited for measuring activity-space segregation because it may be used with a variety of distance functions to capture segregation at different scales, and because there are reasons to think it may be estimated from sample data with effectively no bias and with low uncertainty. These statistical properties are important because activity-space segregation, unlike residential segregation, cannot be measured from census data. Activity-space segregation indexes must be constructed from samples of the trajectories followed by individuals who are themselves sampled from the population. Moreover, current methods for obtaining these samples generally require a trade-off between the number of points that may be sampled from each person's trajectory and the number of people who may be sampled from the population: Those methods that permit the sampling of fine-grained trajectory information also make it hard to generate large samples of individuals, while those methods that permit large samples of individuals are limited to coarse trajectory information.

This article explores the extent to which the spatial proximity index may be estimated from sampled data. It uses highly detailed trajectory data collected from mobile phones, combined with data drawn from large computer simulations of full city populations, to evaluate a straightforward estimator of the index. In doing this, the article contributes to the growing field of human mobility research by using unique data to test and refine a new empirical method for understanding spatio-temporal

social divisions.

# 1 Background

## 1.1 Spatial Proximity

The primary building block of the activity-space extension of White's spatial proximity index is the average proximity between individuals throughout the day, with proximity calculated using various functions depending on the specific context in which the index is employed. The index compares the average proximity among members of each group and the average proximity among individuals of both groups pooled, with each average weighted by population:

$$\mathrm{SP}^a = \frac{P_{xx} + P_{yy}}{P_{bb}} \tag{1.1}$$

where:

$$\mathrm{P}_{xx} = \sum_{t=1}^{T} \sum_{i=1}^{n_{xt}} \sum_{j=1}^{n_{xt}} \frac{n_{xt} f(d_{ijt})}{(n_{xt}^2 - n_{xt})} \tag{1.2}$$

where $n_{xt}$ is the number of members of group G1 at time $t$, $f(d_{ijt}) = e^{-d_{ijt}}$, $\forall\, i \neq j$, and $d_{ijt}$ is the geographic distance, along the spatial axes, between individuals $i$ and $j$ at time $t$, and where $\mathrm{P}_{yy}$, $\mathrm{P}_{bb}$, $n_{yt}$, and $n_{bt}$ are defined analogously for members of group G2 ($yy$) and for members of both groups pooled ($bb$).

The extent to which this index can be accurately estimated from sampled locations instead of full trajectory data and from sampled individuals instead of full population data depends largely on the estimation of Equation 1.2. This equation

can be represented as simply the weighted mean of a set of arithmetic means:

$$\mathrm{P}_{xx} = \sum_{t=1}^{T} \left[ n_{xt} \left( \sum_{i=1}^{n_{xt}} \sum_{j=1}^{n_{xt}} \frac{v_{ijt}}{n_{xt}^2 - n_{xt}} \right) \right], \forall\, i \neq j \tag{1.3}$$

where $v_{ijt} = f(d_{ijt})$. The expression in the inner parentheses is just the mean of a matrix of transformed distances, excluding the diagonal (hence the subtraction of $n_{xt}$ from the denominator). That the mean of a sample of these distances is an unbiased estimator of the full trajectory mean can be easily proved, and the Law of Large Numbers tells us that the sample mean will converge to the full trajectory mean as sample size increases. Moreover, the Central Limit Theorem shows that the sampling distribution of the mean will converge to normal, although in this case, it will be a truncated normal distribution (for most relevant distance functions) because all of the distances must be above zero (since two people cannot occupy exactly the same space).

The inclusion of $n_{xt}$ in the outer expression complicates things in that this is also a random variable—in this case, a binomially distributed one. The expected value of the product of random variables is equal to the sum of (1) the product of their expectations, and (2) their covariance ($\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y] + \mathrm{COV}[X, Y]$). In this case, however, the covariance of a sample estimate of $n_{xt}$ and the estimate of the mean in the inner expression should converge to zero as the sampling distribution converges to normal because the expectation of the mean of a normal distribution does not depend on sample size. Thus, if $P_{xx}$ is estimated using sample values for all of the distances and for $n_xt$, then the expected value of the estimate can be written as

$$\mathrm{E}[\widehat{\mathrm{P}}_{xx}] = \sum_{t=1}^{T} \left[ \mathrm{E}[\hat{n}_{xt}] \left( \sum_{i=1}^{\hat{n}_{xt}} \sum_{j=1}^{\hat{n}_{xt}} \mathrm{E} \left[ \frac{v_{ijt}}{\hat{n}_{xt}^2 - \hat{n}_{xt}} \right] \right) \right], \forall\, i \neq j \tag{1.4}$$

Because $\hat{n}_{xt}$ is a binomially distributed random variable, its expected value is equal to $n_t p_x$, where $n_t$ is the sample size at time $t$ and $p_x$ is the proportion of Group

1 members in the population at time $t$. Although this, on its own, would be a biased estimator of $n_{xt}$, dividing by sample size, $n_t$, gives an unbiased estimator of $p_x$. (This division can been done in each of the terms of the numerator and denominator of $\widehat{\mathrm{SP}}^a$, which is equivalent to multiplying the expression by $n_t/n_t = 1$.) Thus, an unbiased estimator of $P_{xx}$ is:

$$\widehat{\mathrm{P}}_{xx} = \sum_{t=1}^{T} \left[ \frac{\hat{n}_{xt}}{n_t} \left( \sum_{i=1}^{\hat{n}_{xt}} \sum_{j=1}^{\hat{n}_{xt}} \frac{v_{ijt}}{\hat{n}_{xt}^2 - \hat{n}_{xt}} \right) \right], \forall\, i \neq j \tag{1.5}$$

A final complication arises in the full expression for the estimator of $\mathrm{SP}^a$:

$$\widehat{\mathrm{SP}^a} = \frac{\hat{P}_{xx} + \hat{P}_{yy}}{\hat{P}_{bb}} \tag{1.6}$$

The sampling distribution of this estimator is very hard to evaluate because it results from the ratio of random variables that are themselves the products of normally and binomially distributed random variables. However, we can show that this is effectively an unbiased estimator of $\mathrm{SP}^a$ under the conditions we would expect to encounter in real data. To see this, consider the expression as:

$$\widehat{\mathrm{SP}^a} = \frac{\hat{P}_{xx}}{\hat{P}_{bb}} + \frac{\hat{P}_{yy}}{\hat{P}_{bb}} = \sum_{t=1}^{T} \left[ \hat{p}_{xt} \frac{\bar{v}_{xt}}{\bar{v}_{bt}} + \hat{p}_{yt} \frac{\bar{v}_{yt}}{\bar{v}_{bt}} \right] \tag{1.7}$$

where $\hat{p}_{xt}$ and $\hat{p}_{yt}$ are the sample estimates of the proportions of Group 1 and Group 2 members in the population at time $t$ and $\bar{v}_{xt}$, $\bar{v}_{yt}$, and $\bar{v}_{bt}$ are the sample means of the distance function values for the distances between, respectively, Group 1 members, Group 2 members, and the members of both groups pooled. Assuming zero covariance between $\hat{p}_{xt}$ and $\bar{v}_{xt}/\bar{v}_{bt}$ (a reasonable assumption, as discussed above), the expected

value of this estimator is:

$$\mathrm{E}[\widehat{\mathrm{SP}^a}] = \sum_{t=1}^{T}\left[\mathrm{E}[\hat{p}_{xt}]\mathrm{E}\left[\frac{\bar{v}_{xt}}{\bar{v}_{bt}}\right] + \mathrm{E}[\hat{p}_{yt}]\mathrm{E}\left[\frac{\bar{v}_{yt}}{\bar{v}_{bt}}\right]\right] = \sum_{t=1}^{T}\left[p_{xt}\mathrm{E}\left[\frac{\bar{v}_{xt}}{\bar{v}_{bt}}\right] + p_{yt}\mathrm{E}\left[\frac{\bar{v}_{yt}}{\bar{v}_{bt}}\right]\right] \tag{1.8}$$

The terms $\bar{v}_{xt}/\bar{v}_{bt}$ and $\bar{v}_{yt}/\bar{v}_{bt}$ are ratios of normally distributed random variables for which we would expect positive covariance between the variable in the numerator and that in the denominator (since the denominator includes all of the distances that are in the numerator). Importantly, the distributions of these variables are truncated such that they can only take values above zero. Rice (2008) derives the following formula for the expected value of just such an expression:[1]

$$\mathrm{E}\left[\frac{a}{b}|b \neq 0\right] = \frac{\mathrm{E}[a]}{\mathrm{E}[b]} + \sum_{i=1}^{\infty}(-1)^i\frac{\mathrm{E}[a]\langle^i b\rangle + \langle a,^i b\rangle}{\mathrm{E}[b]^{i+1}} \tag{1.9}$$

where $\langle^i b\rangle$ is the $i^{\text{th}}$ central moment of $b$ and $\langle a,^i b\rangle$ is the mixed central moment of $a$ and $b$, defined as:

$$\mathrm{E}\left[(a - \mathrm{E}[a])(b - \mathrm{E}[b])^i\right] \tag{1.10}$$

The summation in Equation 1.9 is a Taylor series expansion that shrinks as the variances of $a$ and $b$ shrink relative to their expected values, and as the expected value of $a$ shrinks relative to that of $b$. The variances of $\bar{v}_{xt}$, $\bar{v}_{yt}$, and $\bar{v}_{bt}$ are equal to the population variances divided by the sample sizes, so they necessarily shink as sample size is increased. Moreover, the sample size in question is the number of distances between people, which is a square function ($n^2 - n$) of the number of people in the sample ($n$), and therefore increases rapidly as additional people are added.

The relationship between the expected values of $\bar{v}_{xt}$ and $\bar{v}_{bt}$ and between $\bar{v}_{yt}$ and $\bar{v}_{bt}$ depend on the distance function used and the level of segregation. In a

[1]The derivation is shown in detail in Rice's supplementary material at `http://www.faculty.biol.ttu.edu/rice/ratio-derive.pdf`.

highly segregated city, we would generally expect $\bar{v}_{xt}$ and $\bar{v}_{yt}$ to be larger than $\bar{v}_{bt}$ when the negative exponential is used as the distance function, signifying that people tend to be more proximate to members of their own group then to members of the opposite group. We would expect the reverse to be true when the identity function is used (since proximity is captured by smaller values of the identity function). In less segregated cities, we would expect the values $\bar{v}_{xt}$ and $\bar{v}_{bt}$ and of $\bar{v}_{yt}$ and $\bar{v}_{bt}$ to be closer together, regardless of the function used. Even in the case of the negative exponential function in a highly segregated city, however, it appears that the difference between $\bar{v}_{xt}$ and $\bar{v}_{yt}$, on one hand, and $\bar{v}_{bt}$, on the other, is outweighed by the effect of the small variances noted above, which shrink this term to the point of disappearance even at moderate sample sizes.

It appears, therefore, that we can treat $\widehat{\mathrm{SP}}^a$ as an effectively unbiased estimator of $\mathrm{SP}^a$, even if it is not formally unbiased.

The precision of this estimator remains to be seen and is difficult to calculate. There is necessarily some degree of correlation between sampled distances across individuals, and if the index is calculated by sampling people and then estimating each person's trajectory, there will necessarily be correlation of distances across time. This correlation can lead to variance that is difficult to estimate using standard parametric analysis, but that may be estimated empirically.

## 1.2 Activity-Space Data Sources

Collecting data on activity-space segregation means collecting data on human movement. There are a number of methods that can be used for this purpose, and it is useful to think of these methods as falling into two basic categories (which also apply to the collection of data on any group of moving objects): Eulerian and Lagrangian. The Eulerian approach involves picking points in space and recording the people who move past these points. In contrast, the Lagrangian approach involves picking people

and recording the space through which these people move (Ōkubo & Levin, 2001). Censuses are a type of Eulerian approach, whereas time-use surveys are a type of Lagrangian approach. The latter are better suited for studying activity-space segregation, but there are limits to what can be learned from self-reported movement data, given faulty perception and memory, and the logistical problems of implementation on a large scale (Stopher, FitzGerald, & Xu, 2007; Murakami & Wagner, 1999; Golob & Meurs, 1986).

An alternative to self-reported movement data is to attach to each individual a device that automatically measures and records the individual's location or transmits signals to an external receiver that does so. Until recently, this was an expensive proposition that could be done only on a relatively small scale. The rapid and widespread adoption of mobile phones by people around the world has changed all of this. The simplest of these devices leave traces of their locations every time they transmit signals to a cell tower, and the more sophisticated ones can determine their own locations based on signals received from cell towers, satellites, and other sources. A growing body of research shows the variety of ways in which mobile phones may be used to study human movement and the quality of the data they are able to produce (Palmer et al., 2013; Ahas, 2011; Ahas & Mark, 2005; Asakura & Hato, 2004).

Methods that rely on locations estimated by each research subject's mobile phone are generally referred to as "active" mobile phone positioning (Ahas, 2011), since they require the active participation (and consent) of the subject. These methods generally involve the distribution to participants of some sort of tracking software, and sometimes also the phones on which it will run. As a result, there are practical limits to the sample sizes that may be achieved from active positioning. At the same time, active positioning methods offer the highest resolution information on participants' movement trajectories.

In contrast, methods that rely on the traces of mobile phone locations that

are detected by external receivers are referred to as "passive" mobile phone positioning (Ahas, 2011), since these methods do not require any effort by the people being tracked. These methods generally involve the researcher obtaining massive volumes of anonymized call detail records (CDRs) from a mobile network provider. Each CDR includes an identifier of the cell tower with which the user was in closest proximity when placing a call, receiving a call, or transmitting data (Isaacman et al., 2010; Wesolowski & Eagle, 2010). Although the identity of the individual in each record is hidden for privacy reasons, researchers are sometimes able to obtain basic demographic information that can be used in studies of segregation (Toomet, Silm, Saluveer, Tammaru, & Ahas, 2012). Although CDRs have much lower resolution than active positioning data—both because each location is estimated with lower precision and because each person's trajectory is sampled much less frequently—they have the advantage of very large sample sizes.

On one hand, both active and passive positioning can be thought of as Lagrangian approaches to human location since they both track the trajectories of individuals through time. Passive CDR methods, however, produce samples that are large enough to be treated in a more Eulerian manner: One can select areal units in a city, for example, and observe the people who move through these areal units using sub-sampled CDR data—something that would not be practically possible with the small sample sizes obtained with active positioning.

# 2 Data and Methods

This article relies on two data sources: (1) high-resolution active positioning data collected from volunteers using an open source Android app called Space Mapper, and (2) data drawn from computer simulations of full census populations moving along the street networks of Buffalo and Utica, New York.

## 2.1 Space Mapper Data

Space Mapper is an open source Android application that was released on Google Play in August 2012 and available until December 2013 for volunteers who wished to share their location data with the author.[2] The app presented users with a detailed demographic survey, and it then estimated and recorded their phones' locations at regular intervals. It did this without interfering with other phone operations, with minimal drain on the battery, and in a manner that protected users' data from being disclosed to third parties. The only requirements were that volunteers be at least 18 years old, have an Android device, and consent to be research subjects.[3]

The Space Mapper data analyzed in this article was downloaded on May 26, 2013 and it includes all data shared by users between that date and the app's August 2012 release (except for information from 67 users who withdrew from the study and requested that their data be deleted). In total, there are 739,224 location estimates from 900 users in 80 countries.

The analysis focuses on the highest resolution daily work-day trajectories from the Space Mapper data: daily trajectories made during work days (Monday through Friday) from which at least 288 location estimates were obtained (an average of at least 1 estimate every 5 minutes) falling in at least 12 different hours. There are 419 such daily trajectories, representing 90 different participants. These participants' ages and sexes are shown in Figure 1. Among these trajectories, the combined total number of locations observed is 326,381.

The analysis of this data follows a Monte Carlo approach in the sense that it uses repeated random sampling of locations from the high resolution trajectories to

[2]The application is still available at `https://play.google.com/store/apps/details?id=edu.princeton.jrpalmer.asm`, but the data-sharing functions have been shut off, so it is now entirely for users who wish to track their own activity-spaces. The source code is available at `http://activityspaceproject.com` and `http://github.com/JohnPalmer/SpaceMapper`.

[3]The study was approved in advance by Princeton University's Institutional Review Board for Human Subjects, Office of Research Integrity and Assurance (Protocol 5310).

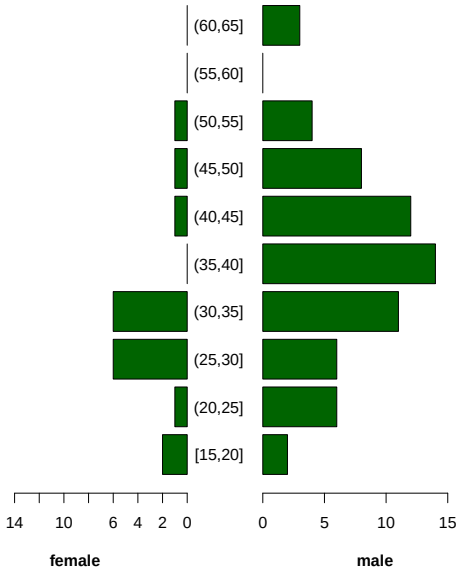



**Figure 1** – Age and sex of Space Mapper participants included in the high resolution work-day trajectories analyzed in this article.

construct empirical sampling distributions of the estimator of the spatial proximity index (Herzog & Lord, 2002). This provides a way to resolve questions of bias and uncertainty that would be very difficult, if not impossible, to solve directly.

## 2.2 Simulation Data

The other source of data used in this article is two large computer simulations of movement in the cities of Buffalo and Utica, New York. The primary purpose of these simulations was to generate population-level movement data from which sampling distributions for the index estimator could be constructed empirically. The basic idea is this: If we have full population data, then even if there is no way to algebraically calculate the expected value or distribution of the sample estimates of a given index, we can estimate both of these things empirically by drawing repeated samples from the full population. In so doing, we can assess the bias and uncertainty of the sample estimates, as well as the effect of any procedures that may be used to correct the bias

| | Buffalo | Utica |
|---|---|---|
| Land Area (miles$^2$) | 40 | 16 |
| Density (persons per mile$^2$) | 7,206 | 3,710 |
| Population | 292,648 | 60,651 |
| Black (% ) | 37.2 | 2.9 |
| Asian (%) | 1.4 | 2.2 |
| Hispanic (%) | 7.5 | 5.8 |
| Foreign Born (%) | 4.4 | 11.9 |
| White-Black D | 73.9 | 48.5 |

**Table 1** – Selected geographic and demographic characteristics of Buffalo and Utica, New York, from 2000 census data.

or reduce uncertainty. All of this information can then be used to tailor and assess the estimators we use when full population data is not available (i.e., in most research settings).

For this purpose, real-life trajectories would be preferable, but obtaining them for the full population of a city is simply not feasible. Simulated population data is a useful alternative to the extent that the simulations are able to capture the properties of real-life populations that influence the sampling distributions of the index. These properties are likely to include: (1) the size of the total population, (2) the area of the city and its population density, (3) the ratio of the populations of the two social groups whose segregation is being measured, (4) the level of segregation in the city, and (5) the geometry of the individual movement trajectories.

Buffalo and Utica were selected as the sites of the simulations because these cities exhibit important differences along the first four of the above properties and, therefore, provide a useful comparison. Buffalo is the second largest city in New York State, with a population of 292,648, and over 7,000 people per square mile. It also has a large and highly segregated black population: over 37% of Buffalo's population is black and, with a residential dissimilarity index score of 73.9, the city has one of the highest levels of black-white segregation in the United States. Utica is smaller and less dense, with a total population of 61,651 spread over 16 square miles (3,700 people

per square mile). It also has a much smaller black population (2.9%) and lower levels of residential segregation ($D = 48.5$). A summary of the geographic and demographic characteristics of each city is shown in Table 1. Figure 2 shows the proportion of black residents by census block group, making the high level of segregation in Buffalo, and the lower level in Utica immediately apparent.

To capture these demographic and geographic properties, the simulations rely on the full census population of non-Hispanic blacks and non-Hispanic whites in each city, with each person initially placed within their reported census tract of residence. As exact addresses are not included in the public census data, each person was placed randomly along a road within his or her census tract. Thus, the simulations directly incorporate the real-life demographic and geographic structure of each city to a large extent.

Of the properties that likely influence sampling distributions, the fifth, trajectory geometry, is the hardest to capture. The simulations rely on a type of random walk known as the Lévy walk as a way to approximate the trajectory geometry of human movement that has been observed in other studies (Jiang, Yin, & Zhao, 2009; Rhee et al., 2011). Each simulated person independently performs road-network-constrained, truncated Lévy walks for 8 hours of simulated time, moving at 4 km per hour. This 8-hour period is followed by 8 hours during which everyone retraces their steps back to their homes. For the remainder of the 24 hour period over which the index is measured, the simulated people are kept at their home locations.

The Lévy walk is an extension of a simpler random walk, Brownian motion, that was first developed to model the movement of particles suspended in a fluid and later applied to the movement of living organisms and a wide range of other dynamic processes (Bartumeus, 2007). Brownian motion involves a series of steps in which the direction and distance of each step is selected at random. The direction of each step is chosen from a uniform distribution (i.e., all directions are equally likely), whereas

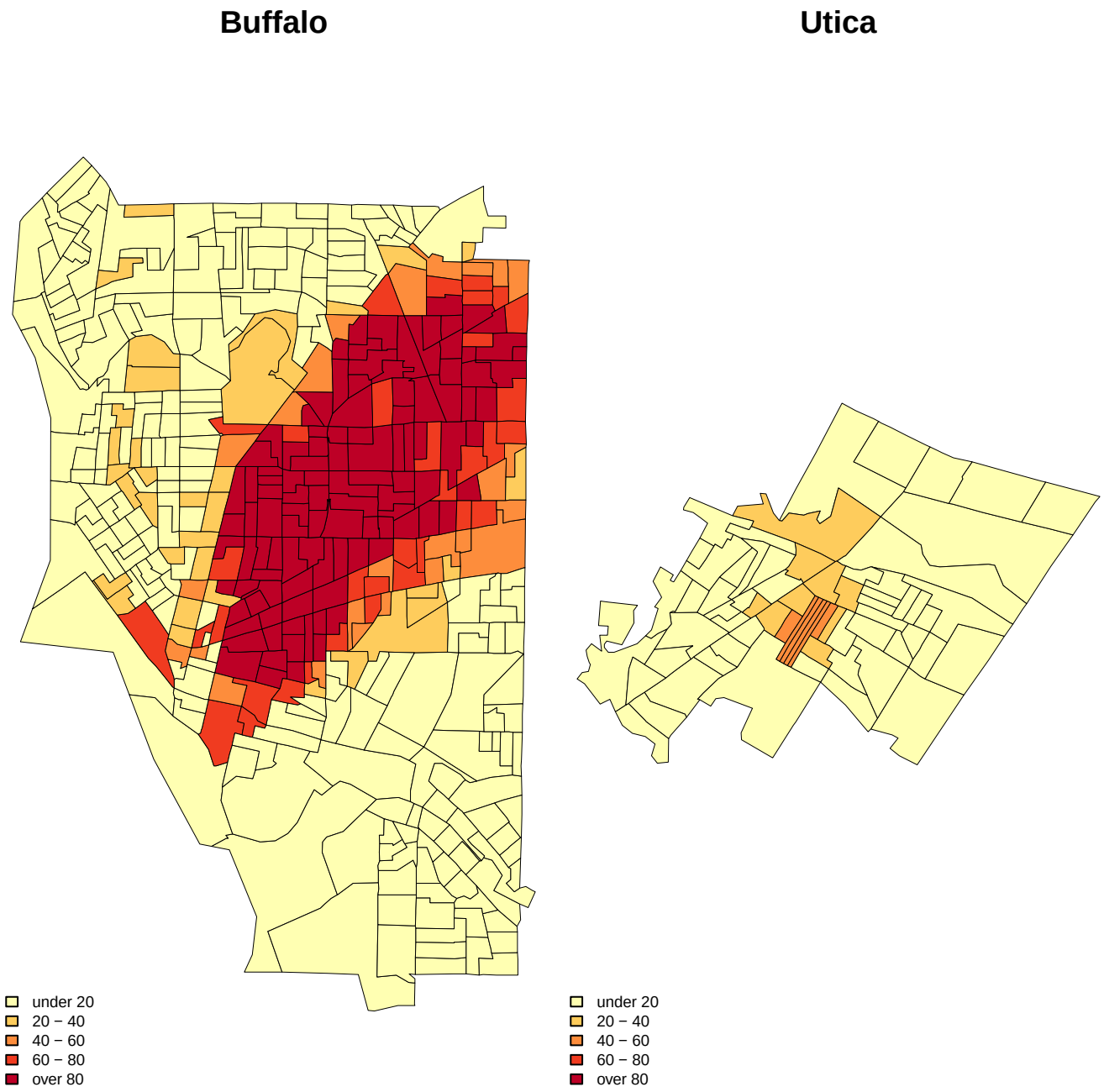



**Figure 2** – Proportion of black residents in Buffalo and Utica census block groups, based on 2000 census data.

the distances are drawn from a normal distribution. In Lévy walks, the direction of each step is also drawn from a uniform distribution, but the distances, $\ell$, are drawn from a power-law distribution, such that $P(\ell) = \ell^{-\mu}$ and $1 < \mu < 3$. If $\mu$ is set to 3 or above, this distribution converges to the normal and the movement becomes Brownian (Bartumeus, Catalan, Fulco, Lyra, & Viswanathan, 2002; Viswanathan et al., 2002; Bartumeus, da Luz, Viswanathan, & Catalan, 2005; Bartumeus, 2007; Jiang et al., 2009).

During the past three decades, ecologists have begun turning to the Lévy walk as a model of animal movement because the Lévy walk is able to capture the observed directional persistence of this movement, whereas such persistence is ignored by Brownian models (Bartumeus et al., 2005). Lévy walks have been proposed as a general model of animal movement at large spatial and temporal scales, and as a specific model for the strategies that animals employ to search for food and other targets in nature when they lack information about locations (Viswanathan et al., 2002). For this purpose, Lévy walks outperform other search strategies under certain conditions, and they may well be driven by internal biological mechanisms that result from evolution (Bartumeus, 2007; Bartumeus & Levin, 2008).

One might expect the movement of humans within modern urban areas to differ from that of animals in natural environments, and yet the Lévy walk appears to do a very good job of characterizing the statistical properties of human movement, even in cities. This was suggested for large scale movement by Brockmann, Hufnagel, and Geisel (2006), on the basis of their analysis of the movement of bank notes throughout the United States. (They used bank note circulation as a way to measure large-scale movement patterns that would otherwise be hard to detect.) More recently, Jiang et al. (2009) analyzed high resolution GPS traces from 50 taxicabs in four Swedish towns over a 6-month period, and found that the statistical properties of the cab rides were closely approximated with Lévy statistics. Rhee et al. (2011) found the same

with respect to the movement of people going about regular daily activities in a set of urban and suburban locations.

On the other hand, Gonzalez, Hidalgo, and Barabasi (2008) suggest that the Lévy-like statistical patterns observed in the aggregate in these studies do a poor job of capturing the directed (non-random) and very regular patterns of movement that individuals engage in on a daily basis as they shuttle between home, work, and a small set of additional places. That finding is supported by evidence that individual movement trajectories can be characterized by a relatively small number of "motifs" (Schneider, Belik, Couronné, Smoreda, & González, 2013).

In spite of these doubts, however, Lévy walks may well be a good model for population data needed to estimate sampling distributions of the spatial proximity index because this index depends more on the aggregate relationships between people than on the shapes of individual trajectories.

The specific rules of the simulations used here are as follows: For each step of the Lévy walk every simulated person selects a direction and a distance at random. The direction is constrained to whatever road they were on (forward or backward) and it is selected from a uniform distribution. The distance is drawn from a truncated power law distribution with $\mu$ set to 2 and the minimum and maximum set at 100 m and 100 km. If the simulated person encounters an intersection during one of the steps, their decision to change roads is also based on a random draw from a uniform distribution, while their direction (forward or backward) remains the same. If the person reaches the end of a road or the simulation boundary, they automatically reverse direction and continue whatever step they are on.

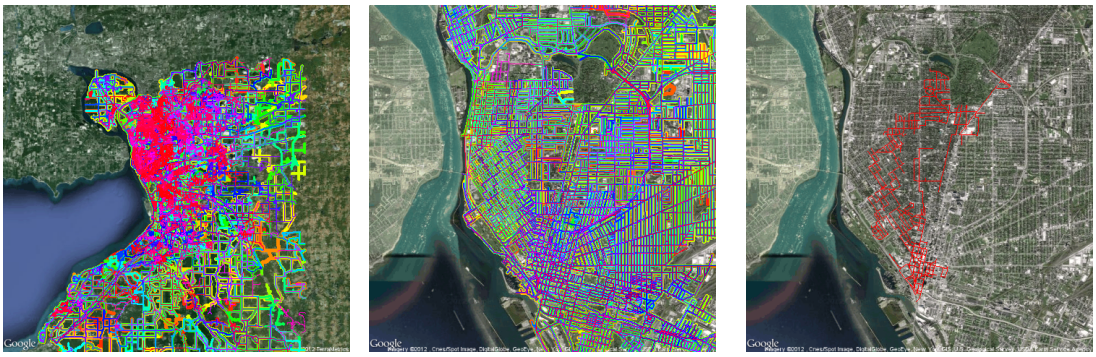
Several weaknesses of these rules should be addressed: (1) The simulated people moved at a constant speed, whereas in reality, people move at a range of speeds, particularly in an urban environment with multiple transportation modes. (2) The constant speed was a walking speed rather than a vehicle speed, whereas in

reality, prevalent vehicle use would be expected. (3) The simulated people were in constant motion for 16 hours (8 hours out, 8 hours back) and then motionless for the rest of the day, and the movement hours were identical for all simulated people, conditions that obviously do not hold in reality. (4) A boundary was placed around the simulation space and a maximum distance placed on the distribution from which distances were drawn, whereas in reality, maximum distances are set by a number of constraints and preferences that are likely to vary among people. (5) Finally, the simulation boundary was set as the county boundary for Buffalo and as the city boundary for Utica, whereas the index was calculated only for locations within the city boundaries in both cases. This means that the Buffalo simulation allowed for people exiting and entering the city during the course of the day, whereas the Utica simulation did not.

The simulations might well be made more realistic by modifying each of these characteristics. At the same time, however, doing so would make them increasingly complex. These simulations were designed with a goal of simplicity, as initial evaluations of the spatial proximity index with full population data.

One further concern about the simulations is whether the road network or the truncation changed the trajectory geometries to such an extent that they cannot be considered the product of Lévy walks. Although more testing should be done in this regard, initial analysis of the trajectories from the Buffalo simulation showed that the pattern of mean squared displacements over time (i.e. the mean squared distance each person was from their starting point at each time slice) were in line with the expectation of the Lévy walk.

The simulations were programmed in R and run on a set of high-memory, multi-core servers using parallel processing techniques. As a general visualization of the results, Figures 3 and 4 show simulated trajectories on a satellite map of each city, with each person assigned a unique color (and with progressively decreasing



**(a)** Full street network with all simulated paths

**(b)** Close-up with all simulated paths

**(c)** Close-up with one simulated path

**Figure 3** – Buffalo satellite image with simulated paths. Each person is given a unique color and lines are drawn with varying thicknesses to aid in visualization.

**(a)** Full street network with sampled paths

**(b)** Close-up with sampled simulated paths

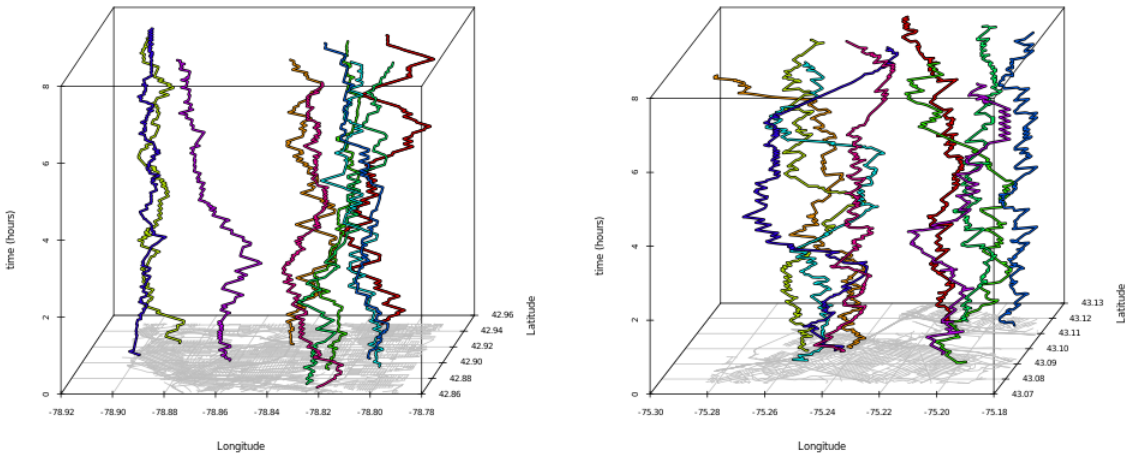
**(c)** Close-up with one simulated path

**Figure 4** – Utica satellite image with simulated paths. Each person is given a unique color and lines are drawn with varying thicknesses to aid in visualization.

line thicknesses to aid in distinguishing overlapping lines). Figure 5 shows the three-dimensional space-time paths of ten people from each simulation.

# 3 Sampling Locations from Trajectories

To better evaluate the magnitude of the variability in spatial proximity estimates made with coarse location samples, the index was calculated using replicate subsamples from the Space Mapper data. This was done using only high resolution trajectories that fell substantially within the Barcelona city limits (72 daily trajecto-




**(a)** Buffalo **(b)** Utica

**Figure 5** – Space-time paths of 10 people's paths in Buffalo and Utica simulations. Time is the vertical axes; space is horizontal, with each city's street network drawn in grey at the base.

ries in total, made by 13 people). For each person, the highest resolution trajectory was selected, and 500 sub-samples were drawn from this trajectory at each of 9 sample sizes. The spatial proximity index calculated from these subsamples was compared with that calculated from the full trajectory data.

The results are shown in Figure 6. This plot shows the distribution of the estimates from each sample, with the central 95% of estimates marked in grey, the central 50% marked in black, the median marked in orange, and the mean marked in yellow. The index value computed with full trajectory data is shown by the horizontal red line.

That the full trajectory index value is nearly identical to the sample mean supports the analytical conclusion, above, that this is an effectively unbiased estimator. In addition, the results show a relatively small amount of variability in the sample estimates, with 95% of the data falling within a 0.1 point range on the index scale. This is important because variability was the major concern with this index. At the

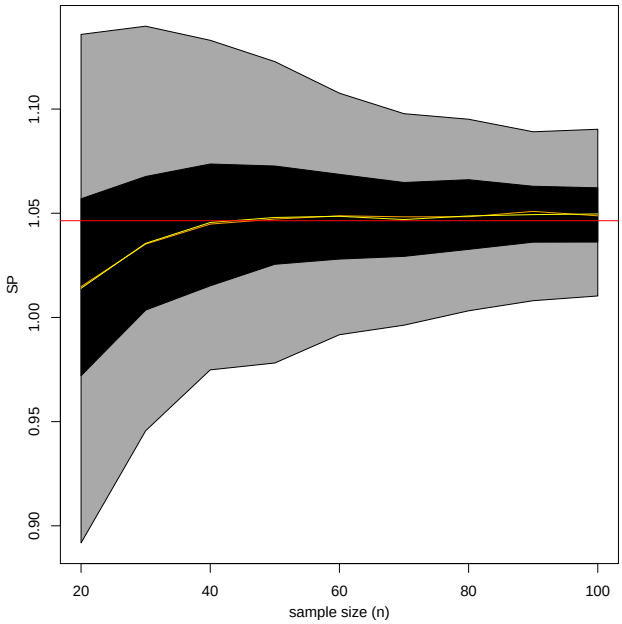



**Figure 6** – Male-female activity-space spatial proximity index (SP) using identity as distance function. Indexes are calculated using 500 replicate location samples drawn at each of 9 sample sizes from high resolution trajectories of 9 men and 4 women in Barcelona. Plot shows central 95% (grey), central 50% (black), median (orange), and mean (yellow) of the sample estimates. The index value computed with full trajectory data is shown by the horizontal red line.

same time, however, we should be cautious in reading too much into this result because it includes only variability introduced by sampling locations from trajectories, not the additional variability that can be expected from sampling people from the population.

# 4 Sampling People from the Population

Estimating the uncertainty of sample inferences of the spatial proximity index using samples of people from the population is complicated. At the core of the problem is the fact that the indexes are calculated from observations of pairs of people, whereas sampling is most likely to be done on individuals. If $n$ individuals are sampled, then each sampled individual will be used in the measurement of $n - 1$ distances, and the probabilities of those distances will not be independent of one another. Without

knowing the covariance among the distances, it will be impossible to estimate the uncertainty of the inferences drawn from these samples algebraically.

One solution, of course, would be to sample pairs instead of individuals. If pairs were drawn without replacement of either member of each pair—such that no individual was sampled in any more than one pair—then we could assume independence and estimate uncertainty algebraically using standard methods. The drawback of this approach, however, is that the recruitment of $n$ volunteers would be necessary to achieve a sample size of $n/2$ pairs. In contrast, sampling $n$ individuals using the first approach would yield $n^2 - n$ pairs. It is worthwhile, therefore, to explore how the covariation among sampled individuals affects the uncertainty of sample estimates of the index in practice.

Figures 7 and 8 show population values and estimates of the black-white activity-space spatial proximity index for each simulation, calculated from individuals sampled at sample sizes ranging from 20 to 1000. The index is calculated using both the negative exponential (Figure 7) and the identity function (Figure 8).

The population value for the negative exponential version of the index is 1.3 in Buffalo and 1.04 in Utica, marked on the plot with the dotted and dashed horizontal lines (Figure 7). These values indicate that whites were more proximate to other whites than they were to blacks and that blacks more proximate to other blacks than they were to whites in both simulations, but in Utica the differences were very small (this index takes the value 1 when there is equal mean proximity among same and opposite group members).

For the identity function version (Figure 8), the population value is approximately 0.96 for Buffalo and 0.99 for Utica, marked again with the with the dotted and dashed lines. These values are consistent with the negative exponential values: They also indicate that people in each simulation tended to be more proximate to members of their own race than to members of the opposite race. For the negative

exponential, this is shown by values above 1, whereas for the identity function, it is shown by values below 1. The negative exponential version is farther from 1, showing greater segregation, than the identity version because it places more weight on close distances, dropping off quickly as distance increases and $e^{-d}$ approaches 0. In contrast, the identity function places equal weight on all distances. The difference can be understood best as a question of the spatial scale on which the segregation is measured, with the identity function measuring segregation at a much larger scale—essentially zooming out and viewing the city from a distance. This has substantive importance, given, for example, changing contours of suburbanization.

The light red (Buffalo) and light orange (Utica) areas show the central 95% of estimates calculated with each function. The dark red and orange areas show the central 50% and the solid red and orange curves show the means. Finally, the red and yellow circles with vertical confidence bars show the mean and 95% confidence intervals estimated from one sample with 500 bootstrap replicates at each sample size for each city and function (Efron, 1979; Davison, 1997).

Three important insights may be drawn from these plots:

First, our estimator is effectively unbiased as predicted in Part 1. For both versions of the index and both cities, the expected value of the estimator (the solid red and orange curves)—at all sample sizes—is almost identical to the population value (the dashed and dotted lines).

Second, the uncertainty in the estimates is relatively small. For sample sizes of at least 100, 95% of the sample estimates of the identity function version fall within approximately 0.03 units of the true value in Buffalo and 0.01 units of the true value in Utica. This margin of error drops to 0.01 in Buffalo and 0.005 in Utica for sample sizes of 400. For the negative exponential version, these margins of error are 0.16 for Buffalo and 0.05 for Utica at sample sizes of 100, and 0.08 and 0.02 at sample sizes of 400. Even without conducting a formal analysis of statistical power, it is clear

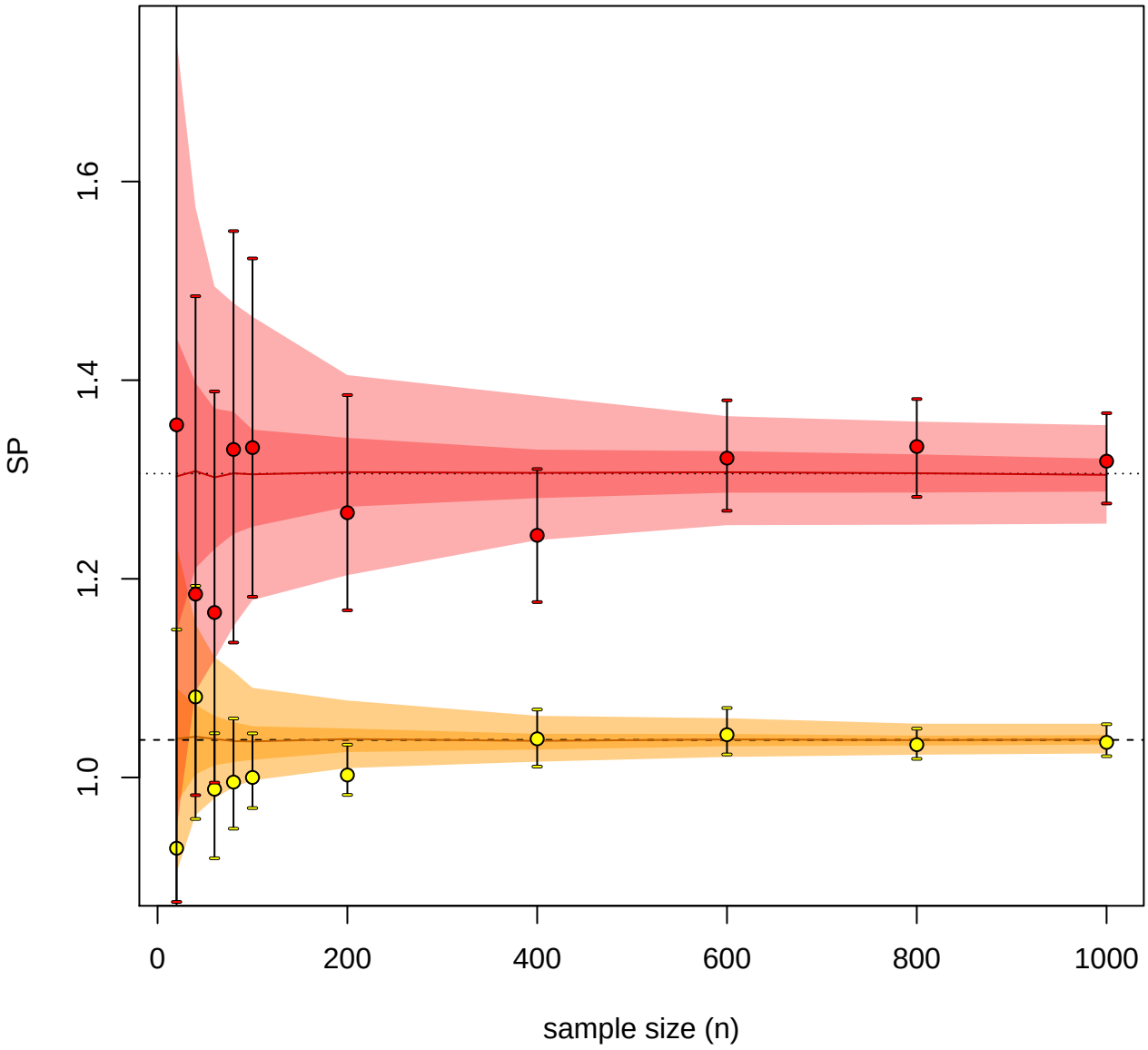



**Figure 7** – Black-white activity-space spatial proximity index (SP) for Utica (yellow) and Buffalo (red) simulation, using negative exponential distance function. Index values were calculated using 500 population samples drawn at each of 10 sample sizes from each city's census population and each individual was simulated performing Lévy walks along the street network. Plot shows central 95% (light areas), central 50% (dark areas), and means (solid curves) of the sample estimates. The index values computed with full population data are shown by horizontal dotted (Buffalo) and dashed (Utica) black lines. The colored circles with confidence bars show bootstrap estimates from individual samples drawn at each sample size.

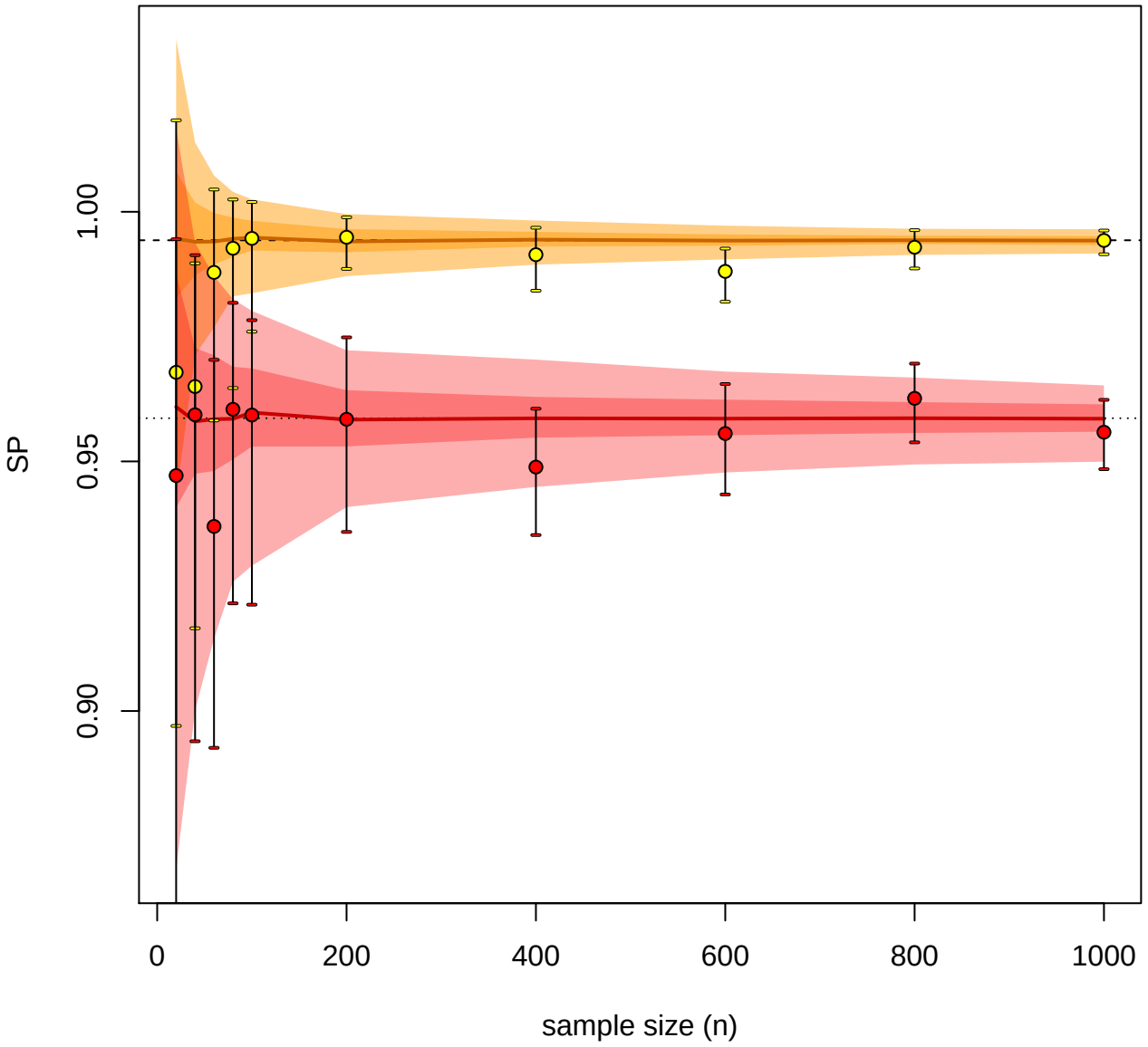



**Figure 8** – Black-white activity-space spatial proximity index (SP) for Utica (yellow) and Buffalo (red) simulation, using the identity function as the distance function. Index values were calculated using 500 population samples drawn at each of 10 sample sizes from each city's census population and each individual was simulated performing Lévy walks along the street network. Plot shows central 95% (light areas), central 50% (dark areas), and means (solid curves) of the sample estimates. The index values computed with full population data are shown by horizontal dotted (Buffalo) and dashed (Utica) black lines. The colored circles with confidence bars show bootstrap estimates from individual samples drawn at each sample size.

that the differences between Buffalo's and Utica's values for each index would be very likely detected in samples as low as 100.

Third, bootstrap methods do a good job of capturing the uncertainty of the estimates. At each sample size, the bootstrapped 95% confidence interval is approximately the same length as the interval of the central 95% of the sample estimates (which can be taken as the true uncertainty of the estimator) and nearly all of the bootstrapped confidence intervals contain the true population value.[4]

All of this suggests that the activity-space spatial proximity index can be estimated quite well with a relatively small sample, so long as that sample is representative. If faced with the choice between spending money on making a sample more representative (e.g., but offering sufficient compensation and equipment to minimize non-response among participants who are selected randomly) or making it larger (e.g., by acquiring CDR data or developing a large citizen-science project), for this index it is probably better to spend the money on representativeness, even if this means the sample will include only 100 or 200 people.

An additional issue raised by Figures 7 and 8 is that the estimates made with the negative exponential distance function have much more variability than those made with the identity function. This might be taken as a reason to prefer the identity function version of the index, but that would probably be a mistake: The two versions measure segregation on different scales and there may be important substantive reasons to prefer the negative exponential. Most notably, the negative exponential version places more weight on the distances within which in-person social interaction is actually possible, whereas the identity version does not differentiate between the difference between cities' activity-space spatial proximity values are probably larger

[4]Although the bootstrapped confidence intervals are not totally aligned with the central 95% intervals, this is expected because each bootstrap interval is created from only one sample. This lack of alignment would also exist with parametric estimates of the confidence interval—indeed, the proper frequentist interpretation of a confidence interval is that it is expected to encompass the true value 95% of the time under repeated sampling.

when the index is measured using the negative exponential than when it is measured using the identity function, so the greater variability may not make it any harder to detect differences. Indeed, it may well be easier to detect them.

# 5 Conclusions

The activity-space extension of White's spatial proximity index is well-suited to capture differences in contact with places and people at different scales, depending on the distance function used. It can be estimated effectively without bias using sampled trajectory points from samples of people and it performs well even at sample sizes in the hundreds range or lower. Although trajectory resolution is not vital for this index's estimation, its robustness to small samples of people makes it ideal for a sampling scheme in which a small, representative sample of people is drawn and offered compensation, as well as any necessary equipment, to share several days of data on their movements using a mobile phone tracking application.

# References


Ahas, R. (2011). Mobile positioning. In M. Büscher, J. Urry, & K. Witchger (Eds.), *Mobile methods.* Routledge. 8, 9

Ahas, R., & Mark, Ü. (2005). Location based services: New challenges for planning and public administration? *Futures*, *37*(6), 547–561. 8

Asakura, Y., & Hato, E. (2004). Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, *12*(3-4), 273–291. 8

Bartumeus, F. (2007). Lévy processes in animal movement: An evolutionary hypothesis. *Fractals*, *15*(02), 151–162. 13, 15

Bartumeus, F., Catalan, J., Fulco, U., Lyra, M., & Viswanathan, G. (2002). Optimizing the encounter rate in biological interactions: Lévy versus Brownian strategies. *Physical Review Letters*, *88*(9). 15

Bartumeus, F., da Luz, M. G. E., Viswanathan, G., & Catalan, J. (2005). Animal search strategies: A quantitative random-walk analysis. *Ecology*, *86*(11), 3078–3087. 15

Bartumeus, F., & Levin, S. A. (2008). Fractal reorientation clocks: Linking animal behavior to statistical patterns of search. *Proceedings of the National Academy of Sciences*, *105*(49), 19072–19077. 15

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*(7075), 462–465. 15

Davison, A. C. (1997). *Bootstrap methods and their application*. Cambridge university press. 22

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 1–26. 22

Golob, T. T., & Meurs, H. (1986). Biases in response over time in a seven-day travel diary. *Transportation*, *13*(2), 163–181. 8

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. 16

Herzog, T. N., & Lord, G. (2002). *Applications of Monte Carlo methods to finance and insurance*. Actex Publications. 11

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., & Varshavsky, A. (2010). A tale of two cities. In *Proceedings of the eleventh workshop on mobile computing systems & applications* (pp. 19–24). 9

Jiang, B., Yin, J., & Zhao, S. (2009). Characterizing the human mobility pattern in a large street network. *Physical Review E*, *80*(2), 021136. 13, 15

Murakami, E., & Wagner, D. P. (1999). Can using global positioning system (GPS)

improve trip reporting? *Transportation Research Part C: Emerging Technologies*, *7*(2-3), 149–165. 8

Ōkubo, A., & Levin, S. (2001). *Diffusion and ecological problems: Modern perspectives*. Springer Verlag. 8

Palmer, J. R. B. (2012). *Activity-space segregation: Understanding social divisions in space and time.* Retrieved from `http://paa2012.princeton.edu/papers/121063` (Presented at the 2012 annual meeting of Population Association of America, San Francisco) 2

Palmer, J. R. B., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E., & Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, *50*(3). 8

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J., & Chong, S. (2011). On the Levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking*, *19*(3), 630–643. 13, 15

Rice, S. H. (2008). A stochastic version of the price equation reveals the interplay of deterministic and stochastic processes in evolution. *BMC Evolutionary Biology*, *8*(1), 262. 6

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, *10*(84). 16

Stopher, P., FitzGerald, C., & Xu, M. (2007). Assessing the accuracy of the Sydney household travel survey with GPS. *Transportation*, *34*(6), 723–741. 8

Toomet, O., Silm, S., Saluveer, E., Tammaru, T., & Ahas, R. (2012). *Where do ethnic groups meet? copresence at places of residence, work, and free-time.* Retrieved from `http://www.obs.ee/~siim/Segregation_domains.pdf` 9

Viswanathan, G., Bartumeus, F., V. Buldyrev, S., Catalan, J., Fulco, U., Havlin, S., ... Eugene Stanley, H. (2002). Lévy flight random searches in biological

phenomena. *Physica A: Statistical Mechanics and Its Applications*, *314*(1), 208–213. 15

Wesolowski, A. P., & Eagle, N. (2010). Parameterizing the dynamics of slums. In *Proceedings of the AAAI artificial intelligence for development symposium.* 9