Sociospatial Complexity and Structure in Cities

Phil Chodrow

July 3rd, 2016

We consider two related problems. First, the measurement of sociospatial segregation is an ongoing concern of quantitative sociologists. Second, the identification of natural, socio-economically defined neighborhoods arises in Census reporting, applied sociological analysis, and dimension reduction in urban computing.

We unify these problems by developing a rigorous, information-theoretic approach to both, using open data on race in American cities as a case study. We formulate a suite of three information measures, including *mean local information* $J(X, Y)$. The local information is a novel measure of complexity in unevenly organized cities, and we prove that it is closely related to the Fisher information of the underlying joint distribution. We show that the local information measures both spatial exposure as defined in segregation studies and the fineness of neighborhood structure. Unlike aspatial information measures, the mean local information clearly distinguishes between cities like Detroit–which is dominated by a few huge, monoracial superclusters–and cities like Philadelphia–which is an intricate patchwork of small, racially-distinct neighborhoods. Second, we provide a practical algorithm for identifying natural neighborhoods through greedy information maximization, and show that the fineness of neighborhood structure as identified by this algorithm is closely related to these information measures. This work provides a unified methodological foundation for further research into sociospatial phenomena, the dynamics of diversity over decades or days, and into scaling relations between urban density and neighborhood size.

1 Introduction

The neighborhood in which an individual lives has profound impact on that individual's mental and physical well-being [19, 20]. Urban planners may aim to revitalize neighborhoods, and an entire cottage industry in sociology and epidemiology has grown around the idea of "neighborhood effects" (see, for example, [14, 29] for overviews of older work). Despite the persistent import of neighborhoods in social sciences, little consensus exists on how to define or demarcate neighborhoods in a non-arbitrary way. Many studies define "neighborhoods" by the boundaries supplied in their data sets, such as those determined by the U.S. Census [13]. Such an approach follows the limitations of the data provided, rather than a theoretically-principled concept of neighborhood. However, recent progress in non-arbitrary demarcations of urban form serve as proofs of concept for the theoretically-principled measurement of heterogeneity in urban form [27, 28].

We take up the question of neighborhoods in the context of segregation studies in sociology, and will argue that the measurement of segregation is intricately linked with the identification of spatial demographic structure. Many indices have been proposed in the last half-century to measure various aspects of segregation; see [22, 23] for useful synthesizing overviews. Many such approaches have also taken Census-defined boundaries as constitutive of neighborhoods, but more recent work has focused "spatialized" measures that are independent of neighborhood demarcations [24, 25, 26, 32, 33, 17]. Other work has quantified the tendency of different social groups to cluster together [18].

The approach we develop is grounded in information measures of spatial demographic structure. Information theory maps the physical concept of structure to the epistemic concept of predictability. A complex system has enough structure to be at least partially predictable, but enough variability to make prediction challenging. Information-theoretic concepts have already been introduced in the study of cities, finding application in urban planning problems, in predicting population distributions, and in quantifying difference between neighborhoods [6, 3, 4, 5, 7, 30]. Attractive features of information measures include their deep relationship to statistical inference [11, 12], their generalizability to a wide variety of phenomena, and their conceptual linkage of cities with the methods of statistical physics. In the context of the measurement of urban demographic structure, information measures provide answers to questions such as:

1. Supposing you choose an individual randomly from a city, how accurately could I guess their race?
2. Supposing you then told me where that person lived within the city, how much would this new information increase my accuracy?

These questions are intimately connected with the measurement of diversity and structure. If I can guess a random individual's race with 100% accuracy given no additional information, then it must be the case that the city is monoracial; increased diversity reduces my ability to guess. If knowing where a person lives dramatically increased my ability to guess, then there must be substantial structural dependence of demographics on space, reflecting a kind of spatial segregation.

We make two primary contributions. The first is to the theory of segregation measurement in sociology. We formulate the *mean local information* $J(X, Y)$ as a measure of spatiosocial structure. We then show that this novel measure, in concert with the mutual information $I(X, Y)$, provides a mathematically unified, operationally interpretable, and spatially-aware methodology for the measurement of diversity and segregation. These metrics possess many of the traditional virtues of segregation indices and can be easily computed from open data such as that provided by the U.S. Census. Our second contribution is to provide a simple, theoretically-motivated, and computationally-tractable method for agglomerating spatio-social data into "natural neighborhoods" across which demographic trends are approximately constant. This method has multiple applications. Sociologists can use this agglomeration procedure to study racially coherent neighborhoods, rather than the partially-arbitrary administrative boundaries such as those produced by the U.S. Census. Computational urban planners can use the method for *dimension-reduction*, in which a large data set is made more computationally tractable while preserving spatio-social relationships of interest. Finally, urban physicists may be interested in the scaling behavior of the mutual information across different levels of aggregation in analogy with coarse-graining in statistical mechanics [7]. The hierarchical character of agglomerative clustering defines a non-arbitrary way to perform this aggregation.

Structure of the Essay

In Section 2, we motivate and develop the two core components of our mathematical framework. The first of these is the mean local information $J(X, Y)$, a measure of the intrinsic spatial complexity of a compositional phenomenon. In the context of racial trends in cities, the mean local information measures the granularity of racial neighborhood structure. It is low in a city like Detroit, which consists of a few large, monolithic tracts of constant racial composition. It is high in a city like Philadelphia, in which many (spatially) smaller neighborhoods are knit together in an intricate patchwork. We develop the mathematics of the mean local information and describe its properties. The second component of our mathematical framework is a simple algorithm for identifying clusters that are both spatially and compositionally coherent through agglomerative hierarchical clustering and greedy information maximization. In the context of race in cities, this algorithm may be viewed as an automated means to identify "natural" neighborhoods.

In Section 3, we apply these techniques to the analysis of spatial trends in race across U.S. cities. We show that the mean local information $J(X, Y)$

and the mutual information $I(X, Y)$ jointly supply simple and intuitive measures of spatial complexity for various cities. We also note an intriguing scaling relationship between spatial complexity and population density. We next evaluate our information-theoretic clustering method, and show that the neighborhoods it identifies can usefully shed light on traditional concerns of quantitative sociology, such as racial affinities and spatial concentration. Next, by defining a global evaluation measure on each clustering, we show that the mean local information $J(X, Y)$ does indeed measure the "clusterability" of a spatial compositional data set.

Section 4 is a discussion of our findings and prospects for future work. We have also provided an Appendix, which includes links to software repositories, mathematical proofs, and specifications of computational procedures used.

2 Mathematical Framework

Information and the Checkerboard

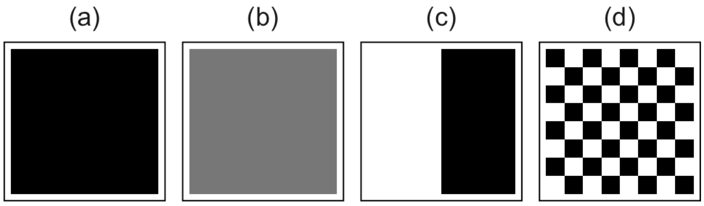
We begin by developing two fundamental measures of information theory, the entropy of one random variable and the mutual information of two. Let X be a spatial variable, such as a set of coordinates in space or a neighborhood name. Let Y be a compositional variable that we aim to study, defined on an alphabet $\mathcal{Y}$. In the simple case in which $\mathcal{Y} = \{\text{Black, White}\}$, Figure 1 illustrates a range of possible joint distributions $p(X, Y)$. Comparing the completely undiverse city (a) and the spatially uniformly diverse (b) motivates the first fundamental information measure, the entropy:

$$H(Y) \triangleq -\mathbb{E}_Y[\log p(Y)] = -\sum_{y \in \mathcal{Y}} p(y) \log p(y) \,. \tag{1}$$

The entropy measures how evenly the global marginal distribution $p(Y)$ is distributed over the alphabet $\mathcal{Y}$. Epistemically, $H(Y)$ measures the difficulty in guessing the random variable Y, given no further information. $H(Y) = 0$ when Y always takes a single value, such as in city (a) – perfect guessing is possible. On the other hand, $H(Y)$ achieves its maximum of $H(Y) = \log |\mathcal{Y}|$ when Y is uniformly distributed on $\mathcal{Y}$, as in city (b). The entropy $H(Y)$ is this sufficient to distinguish the presence of global diversity from its absence. On the other hand, the entropy H is unable to distinguish between the spatial uniformity of (b) and the spatially variability of (c). To distinguish these two cities we may use the mutual information, which is defined in terms of the Kullback-Leibler divergence D:

$$D[p(Z) \| q(Z)] \triangleq \sum_z p(z) \log \frac{p(z)}{q(z)} \tag{2}$$

$$I(X, Y) \triangleq D[p(X, Y) \| p(X)p(Y)] = \mathbb{E}_X[D[p(Y|X) \| p(Y)]] \tag{3}$$



City	$H(Y)$	$I(X,Y)$	$J(X,Y)$
(a)	0.0	0.0	0.0
(b)	0.7	0.0	0.0
(c)	0.7	0.7	0.6
(d)	0.7	0.7	2.7

Figure 1: Information theory and the checkboard problem: model cities with various kinds of spatial diversity can be distinguished through progressively more subtle spatial information measures.

Though not a true metric, the divergence D is interpretable a measure of distance in the space of probability distributions, and so $I(X,Y)$ may be interpreted as the distance between the true joint distribution $p(X,Y)$ and the product of marginals $p(X)p(Y)$. Since the latter expresses statistical independence between X and Y, $I(X,Y)$ measures the extent to which X and Y are dependent. Epistemically, $I(X,Y)$ measures the extent to which knowledge of Xincreases possible accuracy in guessing Y. In city (b), X and Y are independent: knowing X (where an individual lives) conveys no information about that Y (that individual's race). In city (c), on the other hand X and Y are completely dependent: if you know where someone lives, you know their race with 100% confidence, and $I(X,Y)$ achieves its maximum.

City (d) shares with city (c) the fact that race is completely determined by residence. However, city (d) embodies the "checkerboard problem": measures that use only the joint distribution $p(X,Y)$ without additionally considering the spatial information contained within X will evaluate city (d) to be the same as city (c), despite their considerably different patterns of racial separation and potentially very different implications for planning and policy. A recent working paper [26] provides one highly operational approach to this problem, using road network topology and a weighting function that decays with distance to define a localized measure based on the Kullback-Leibler divergence. Here we pursue an alternative strategy, showing that a localization of the mutual information both measures spatial variation in race and corresponds to the estimation of a fundamental statistical property of the distribution $p(X,Y)$.

We will now derive a measure that solves the "checkerboard problem" by distinguishing the toy cities (c) and (d). To motivate our methods, we consider an idealized scenario in which we have a differentiable field $p(y|x)$ of observed probability distributions for each $x \in M \subset \mathbb{R}^n$, where the metric space M is interpretable as the "map" on which we work.

Fix a point $x_0 \in M$ and a radius $r > 0$. Let $B_r(x_0)$ denote the ball of radius r about x_0. Then, define the *local mutual information in radius r* as the mutual information between X and Y, restricted to the small ball $B_r(x_0)$ about x_0:

$$I_r(x_0) \triangleq \mathbb{E}_X[D[p(\cdot|X) \| p(\cdot|X \in B_r(x_0))]|X \in B_r(x_0)] \tag{4}$$

Intuitively, $I_r(x_0)$ measures how much knowing the "location" x adds to our information about the category Y, or, equivalently, how much the probability field $p(y|x)$ varies with x in a small neighborhood of x_0. It is therefore a natural measure of local complexity, aligned in approach with the global mutual information but designed to detect local variation. As expected, $I_r(x_0) = 0$ if and only if $p(y|x)$ is constant for each y in the ball $B_r(x_0)$; that is, if within $B_r(x_0)$ the field $p(y|x)$ resembles toy city (b) in Figure 1. In the definition (4), r may be thought of as the spatial resolution at which we conduct analysis, and $I_r(x_0)$ is highly dependent on r. However, it is possible to show that $I_r(x_0)$ is related to a fundamental statistical property of the probability field $p(y|x)$ that is resolution-independent. Under the stated conditions, the following approximation holds:

$$\frac{I_r(x)}{r^2} \cong \frac{1}{4}\text{trace } J_Y(x) \ , \tag{5}$$

where $J_Y(x)$ is the Fisher information matrix in Y about x, defined as

$$J_Y(x) \triangleq \mathbb{E}_Y \left[(\nabla_x \log p(Y|x))(\nabla_x \log p(Y|x))^T \right] \ . \tag{6}$$

A formal statement and proof of (5) are provided in Appendix 2. The Fisher information J_Y is a fundamental quantity in statistics and information theory. From a geometric perspective, J_Y provides the natural intrinsic metric in the geometric space of probability distributions parameterized by the spatial variable x. Equation (5) therefore expresses a relationship between the local mutual information $I_r(x)$ and the information geometry of the underlying probability field. Corresponding to the fact that $I_r(x_0)$ vanishes if and only if $p(y|x)$ is constant in $B_r(x_0)$, $J_Y(x_0) = 0$ if and only if $\nabla_x p(y|x_0) = 0$ for all y. This implies that x_0 is a stationary point, about which the probability field $p(y|x_0)$ exhibits only small (2nd order or smaller) changes with respect to changes in x.

Since the Fisher information is a strictly local measure of statistical variability around x, we can aggregate the Fisher information to derive a measure of average local variability. The *mean local information* is

$$J(X,Y) \triangleq \mathbb{E}_X[\text{trace } J_Y(X)] \tag{7}$$

As demonstrated in Figure 1, $J(X,Y)$ distinguishes cities (c) and (d), thereby addressing the “checkerboard problem” head on. We propose the aggregate quantity $J(X,Y)$ as a third measure–alongside the entropy $H(Y)$ and mutual information $I(X,Y)$– as a tool for the information-theoretic structure of spatial compositional complexity.

We note that, since

$$\text{trace } J_Y(x) = \mathbb{E}_Y\left[\|\nabla_x \log p(Y|x)\|^2\right] \tag{8}$$

may be viewed as the average norm of the gradient of $\log p(Y|x)$, the quantity

$$\begin{aligned} J(X,Y) &= \mathbb{E}_X[\text{trace } J_Y(X)] && (9)\\ &= \mathbb{E}_{X,Y}\left[\|\nabla_x \log p(Y|X)\|^2\right] && (10) \end{aligned}$$

may be viewed as a cousin to total variation measures often encountered in mathematical analysis.

While we have developed these methods in the context of a differential field of observed distributions $p(y|x)$, their application to a discrete data set is conceptually simple. For a given set of tracts, overlay an evenly spaced grid of radius r, and measure the mutual information $I_r(x)$ in each grid cell. Then, when data and grid resolutions are sufficiently high, $4I_r(x)/r^2$ approximates the quantity trace $J_Y(x)$, which can then be aggregated across the data set. We provide a more formal statement of this computation in the appendix.

Information Measures and Segregation Studies

Considerable attention has been paid to relating segregation indices to intuitive concepts of diversity and segregation. In this section, we relate the information measures $H(Y)$, $I(X,Y)$, and $J(X,Y)$ to three standard dimensions of diversity and segregation in sociology. In brief,

- The entropy $H(Y)$ measures global diversity.
- The mutual information $I(X,Y)$ measures *evenness*.
- The local information $J(X,Y)$ measures *spatial exposure*.

We will define the latter two terms and argue for these claims below. In doing so, we make a case for these measures as practical tools in the quantitative study of urban diversity.

One way to quantify diversity is to consider how closely the distribution of races resembles the uniform distribution. A "maximally diverse" city might be 20% each Asian, Black, Hispanic, Other, and White. The entropy measures distance from the uniform distribution, as can be formalized through the identity:

$$H(Y) = \log |\mathcal{Y}| - D[p(Y) \| u(Y)] \tag{11}$$

where u is the uniform distribution on $\mathcal{Y}$. The entropy is thus largest when the distribution $p(Y)$ is nearly uniform, and smallest when $p(Y)$ is "farthest" from the uniform distribution. Through properties of D, it is possible to show that these cases occur when $p(Y)$ is itself uniform and when Y is a constant, respectively. Since entropy measures nearness to uniformity, it is an intuitive characterization of the idea of diversity.

We now turn to the relationship between the mutual information $I(X, Y)$ and sociospatial evenness. The authors of [23] define *spatial evenness* as "the extent to which groups are similarly distributed in residential space" (page 126), and argue that this concept encapsulates those of clustering, concentration, and centralization used in [22]. To see that $I(X, Y)$ measures (un)evenness, consider the model cities (b) and (c) in Figure 1. In city (b), knowing where a person lives carries no information about their race, in the sense that, if you are attempting to guess someone's race, learning their residence will be of no help to you. In contrast, in city (c), location determines race: if you learn a person's residence, you can guess their race with 100% accuracy. This difference is cast in information terms, but it is precisely the difference between evenness and unevenness. In (c), location epistemically determines race precisely because races are unevenly distributed across locations.

Finally, we consider the local information $J(X, Y)$ and its connection to spatial exposure. The authors of [23] define *spatial exposure* as "the extent that members of one group encounter members of another group...in their local spatial environments" (page 126). Our contention is that $J(X, Y)$ measures exposure conditional on a fixed level of evenness (as measured by $I(X, Y)$). To see this, consider cities (c) and (d) in Figure 1. These cities are equally diverse as measured by $H(Y)$ and equally uneven as measured by $I(X, Y)$. The difference between them is that (c) consists of two large, racially-distinct neighborhoods, of which only residents near the center are exposed to residents of the other. On the other hand, in (d), all residents are just one "block" away from residents who do not share their race. $J(X, Y)$ detects this difference because the localized mutual information I_r in terms of which $J(X, Y)$ is defined precisely as the unevenness in a small area. We should emphasize, however, that $J(X, Y)$ should always be considered in the context of $I(X, Y)$. City (d) clearly has more exposure than (c), but less than the perfectly even (b). The difference is that (b) has no neighborhood structure at all.

This point highlights that, though the concepts of diversity, evenness, and exposure are conceptually distinct, they are not fully independent. There exists a hierarchical relationship among them. Complete lack of diversity ($H(Y) = 0$) implies perfect evenness ($I(X, Y) = 0$), as in city (a). In turn, perfect evenness implies maximal exposure, but also a lack of neighborhood structure $J(X, Y)$, as in city (b). While this point may appear theoretical, it is necessary to keep in mind in the context of practical data analysis. Interpreting whether or not $I(X, Y)$ is "high" in a given city requires contextualizing its value against $H(Y)$. Similarly, evaluating a value of $J(X, Y)$ requires keeping $I(X, Y)$ in mind.

Properties of Information Diversity Measures

There has been much work showing the desirable properties of various segregation indices. To give a brief sampling, many scholars agree that such indices should be invariant to changes in overall population size; they should decrease when populations "even themselves out," and they should behave predictably under aggregation. How do the measures $I(X, Y)$ and $J(X, Y)$ compare? The author of [25] shows that the mutual information $I(X, Y)$ (which she calls the "Divergence Index") satisfies the core desiderata of segregation indices, generally as well or better than existing alternatives. We highlight one property of $I(X, Y)$ for special note, as it will be central to our development of natural neighborhood identification. The principle of "additive decomposability" [23] stipulates that, when the data is grouped along either racial or spatial axes, a good segregation index should split into "within group" and "between group" components. In the context of the mutual information $I(X, Y)$, additive decomposability is simply the familiar chain rule of mutual information. For concreteness, let C be a random variable giving the cluster label of location X; importantly, C is completely determined by X. Then, the chain rule expresses additive decomposability as

$$I(X, Y) = I(C, Y) + I(X, Y|C) \; ; \tag{12}$$

This information identity has a simple epistemic interpretation: he information I have about someone's race Y given that I know their residence X is equal to the information I have if I first learn their broad neighborhood C, plus the amount of additional information I gain if I subsequently learn the exact residence X as well. Mathematically, this expression decomposes $I(X, Y)$ into two useful components. The first term is interpretable as the between-group information, while the second is interpretable as the within-group information. It is therefore comparable to the familiar between-group/within-group "sum of squares" decompositions that frequently appear in classical statistical analysis. A similar version of the chain rule expresses additive decomposability for aggregation of racial groups rather than spatial locations. In the next subsection, we use (12) to define a clustering procedure to identify natural neighborhoods in cities.

We now turn to the properties of $J(X,Y)$. Our core argument is that $J(X,Y)$ possesses a variety of intuitive properties, and that those it does not either underscore the importance of reading it in conjunction with $I(X,Y)$ or need not apply to spatialized measures. See [23, 24] for detailed discussion of these criteria.

Organizational Equivalence: Both the theoretical definition (7) and the procedure to compute it from tract data remain unchanged when a tract is subdivided into smaller tracts, each of which with identical demographic structure.

Size and Density Invariance: Since $J(X,Y)$ is completely determined by the marginal distribution $p(X,Y)$, it is invariant under changes in population density.

Additive Group Decomposability: When demographic groups are aggregated into super-groups, the chain rule of mutual information applied to the demographic variable Y provides an an additive decomposition of the form $J(X,Y) = J(X,G) + J(X,Y|G)$, which can be interpreted as the sum of a between-groups term and a within-groups term as needed.

Scale Interpretability: We have $J(X,Y) = 0$ if and only if $p(Y|X)$ is constant on each connected component of the metric space M. When only one connected component exists, this implies that every locale has the same demographic structure as the global environment. When multiple connected components exist (e.g. the city is divided by a river), demographics must be constant in space on each side of the division. $J(X,Y)$ does not satisfy the additional condition of achieving its maximum value when all locales are monoracial, but this point only emphasizes that $J(X,Y)$ and $I(X,Y)$ should be read jointly. When all locales are monoracial, $I(X,Y)$ achieves its maximum value and $J(X,Y) = 0$.

Boundary Independence: $J(X,Y)$ is defined in terms of a continuous underlying probability distribution $p(X,Y)$, rather than arbitrarily-defined tracts. In computational practice, $J(X,Y)$ is indeed sensitive to the boundaries supplied with the data; however, (5) guarantees that this sensitivity vanishes as the resolution grows sufficiently small. No alternative measure can overcome limited data resolution, and, to our knowledge, no other measure comes with the same assurance of resolution-independence in the limit.

Transfers and Exchanges: The core idea of the principles of transfers and exchanges is that population permutations that tend to "smooth out" differences should reduce segregation measures. However, the concrete principles as stated in [23] do not hold of $J(X,Y)$. The core point

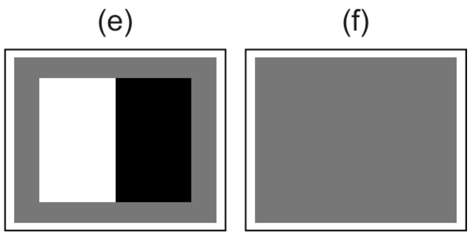


Figure 2: A problem with additive spatial decomposability: aggregating the inner two regions transforms (e) into (f), smoothing out all spatial variability. A "between groups" term must therefore vanish, while a "within-groups" term considers only the inner boundary; not the outer ones.

is that $J(X, Y)$ compares distributions to those nearby, so whether or not a transfer or exchange reduces $J(X, Y)$depends not only on the two regions involved in the transfer or exchange, but also the regions surrounding them. $I(X, Y)$ does satisfy the two principles.

Additive Spatial Decomposability: This criterion states that "If X spatial subareas are aggregated into Y larger spatial areas, then a segregation measure should be decomposable into a sum of within- and between-area components" [24], page 136. However, considering that we have imposed additional spatial structure on our model, this criterion does not appear to be motivated. Consider, for example, Figure 2. Aggregating the two central regions transforms (e) into (f), erasing all spatial variability. Any measure should therefore have a between-groups component equal to 0. On the other hand, the within-groups component can only consider the middle white/black boundary between the central two regions. The sum of the between-groups and within-groups components must therefore consist only in information included in the middle boundary. However, the spatial variability in city (e) is not exhausted by this middle boundary; there are a total of six other frontiers of racial difference that should be considered in any spatial segregation measure. We therefore content that additive decomposability is a desideratum of nonspatial measures (it is satisfied by $I(X, Y)$), but not of explicitly spatial ones.

In summary, $J(X, Y)$ possesses a variety of useful, intuitive properties. The ones it fails to possess reflect the fact that these desiderata are typically formulated to apply to a single measure. While we have not exhausted the full complement of desiderata for segregation measures. However, many of the remaining ones – such as composition invariance – are either difficult to define mathematically or controversial within the sociological community, and we will not them consider further.

We close with a discussion of some of the theoretical and operational virtues of $I(X, Y)$ and $J(X, Y)$ not considered above.

Interpretability: In contrast to many previous works, we make no attempt to encapsulate every aspect of diversity in a single measure. The result of maintaining distinct measures is that each of our two separate measures are highly interpretable. The mutual information $I(X, Y)$has a standard interpretation in information theory: it is the epistemic value of knowing where an individual lives for the purposes of predicting their race, as measured using the logarithmic loss function; see [11] for a detailed discussion of information measures and prediction. The local information $J(X, Y)$ is then interpretable as the epistemic value of knowing where an individual lives when restricting one's attention to a small area. The direct operational meaning of these metrics is the source of their desirable mathematical properties, and reflects the virtue of starting from a unified mathematical framework.

Kernel Independence: Many existing spatial information measures depend on a choice of spatial kernel function which serves as a functional definition of proximity. While providing flexibility, dependence on a spatial kernel function increases researcher degrees of freedom and makes it more difficult to compare studies. There is no theoretical consensus on the appropriate kernel. The mutual information $I(X, Y)$ is aspatial and requires no kernel. In practice, computing $J(X, Y)$ does require a choice of a grid radius (see Appendix for details), but does not require the specification of a functional kernel.

Isotropy: As a result of its kernel independence, $J(X, Y)$ is *isotropic*, taking no consideration of natural boundaries or infrastructure. This property has both virtues and drawbacks. On the one hand, an isotropic measure is simple to compute, simple to interpret, and easily generalized. On the other, boundaries and infrastructure unquestionable shape interactions between people, and are clearly relevant in segregation studies. We note, however, that most urban anisotropies exist on the scale of a few hundred meters at most. Therefore, we expect that anisotropic measures will only be relevant at scales of analysis substantially smaller than the 500m used in this study. A detailed numerical comparison of isotropic and anisotropic measures would be most welcome.

Connection to Data Analysis: As we will see below, the mutual information $I(X, Y)$ can be used to define a practical algorithm for the identification of natural, demographically coherent neighborhoods, and the local information $J(X, Y)$ predicts its performance. These two measures are thus not only of theoretical interest, but also provide tools for the practical analysis of urban diversity. We are not aware of any other measures that share this helpful property.

Identifying Natural Neighborhoods

Equation (12) motivates a simple scheme for identifying natural neighborhoods based on maximizing the between-group term $I(C, Y)$. If all I had access to were the cluster labels C and not the locations X, then the information I had about Ywould be $I(C, Y)$. A "good" clustering therefore maximizes the between-cluster information $I(Y, C)$, which entails minimizing the within-cluster information $I(X, Y|C)$. Solving this problem exactly is a challenging discrete optimization problem, and may not be computationally tractable. We can, however, construct a greedy algorithm which leads to satisfactory results. At each stage of the algorithm, we consider the problem of choosing a pair of tracts $\{i*, j*\}$ to cluster together. The reduction in information associated with aggregating the locations I into a single cluster is

$$d(i, j) \triangleq \sum_{k \in \{i,j\}} p(X = k) D[p(Y|X = k) \| p(Y|X \in \{i, j\})] \tag{13}$$

$$- p(X \in \{i, j\}) D[p(Y|X \in \{i, j\}) \| p(Y)] \tag{14}$$

where $p(Y) = \sum_x p(x, Y)$ is the global marginal distribution. the first term is the information associated with the two separate tracts, while the second is the information associated with a merged tract. Equation (14) defines a natural information distance between locations i and j. Like the KL divergence, this distance is strictly nonnegative; unlike the KL divergence, it is symmetric, and defines an axiomatic metric on the space of tracts. Importantly, we can therefore use the distance $d(i, j)$ as a dissimilarity measure for the purposes of clustering. Our greedy procedure is simple: at each iteration, determine

$$(i^*, j^*) = \underset{i \text{ neighbors } j}{\text{argmin}} \; d(i, j), \tag{15}$$

and then combine i^* amd j^* into a single tract, repeating until only one cluster remains. This procedure defines a form of agglomerative hierarchical clustering distinguished by two characteristics: its spatial constraints and its pursuit of minimal information loss at each step. As a greedy algorith, it possesses no guarantees for optimal solutions, but in practice its performance leads to intuitive, racially-coherent regions. It thereby enables a study of spatial difference using non-arbitrarily-defined regions.

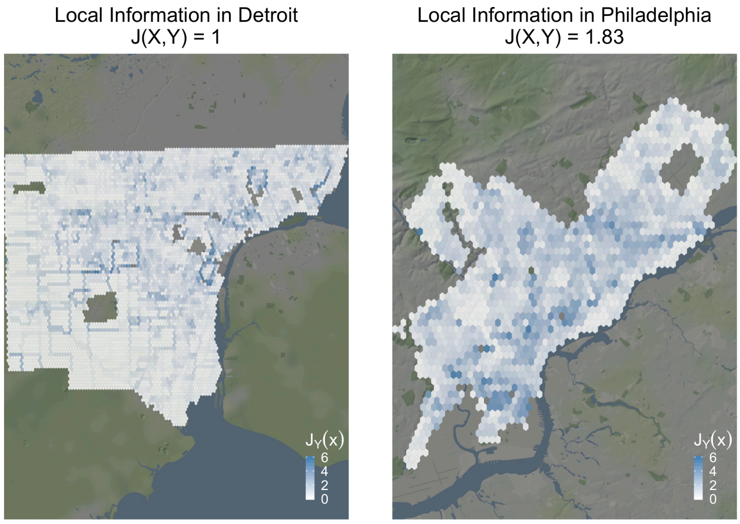


Figure 3: Illustration of methodology for estimating local information in Suffolk County, MA (Boston). We first cover the map in a hexgrid, and compute the mutual information within each hex from the Census tracts that overlap the hex, weighted by density.

3 Findings

Data Used

We assembled block-group level data from the 2010-2014 American Community Survey (ACS), conducted by the U.S. Census Bureau, on race and ethnicity for counties housing 51 large US cities. We then aggregated the detailed racial and ethnic groups into five meta-categories: 'Asian', 'Black', 'Hispanic', 'Other', and 'White'.

Information Measures

For each city, we computed the entropy $H(Y)$ and the mutual information $I(X,Y)$. To compute the estimated aggregate Fisher information $J(X,Y)$, we tiled the map with a hexagonal grid of cell radius 0.5km. We then computed the estimated mutual information within each grid cell, and averaged the results weighted by population population. A technical specification of this approach is provided in Appendix 1, and an illustration of it in Figure 3.

Figure 4 shows the relationship of the mutual information (or evenness) $I(X,Y)$ and mean local information (or exposure) $J(X,Y)$. The global positive trend reflects the fact that global spatial variability is a prerequisite for

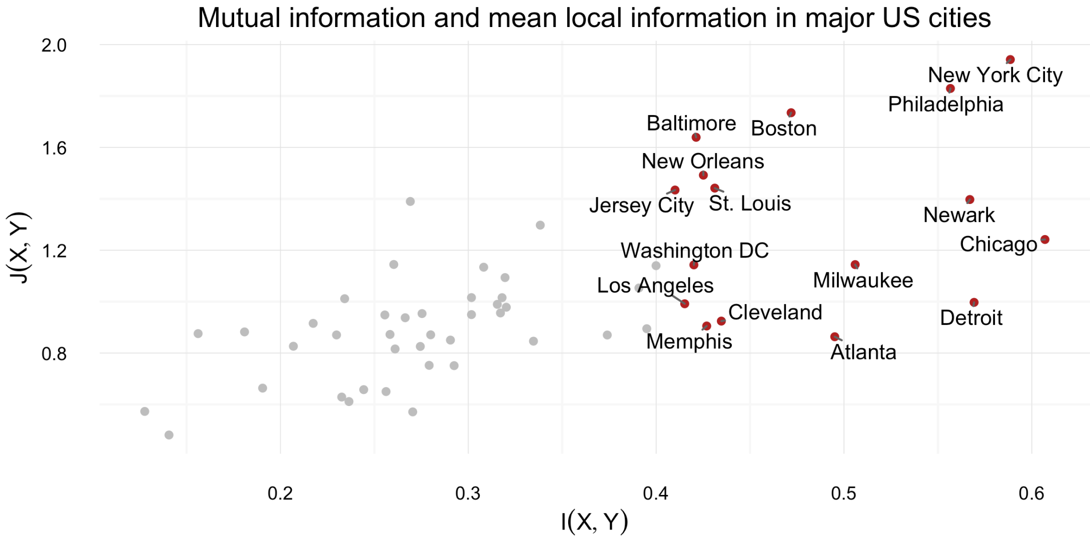


Figure 4: Relationship of global mutual information $I(X, Y)$ and mean local information $J(X, Y)$.

local variability: if the city is uniform (like Figure 1(a)), then no local differences either exist. On the other hand, there are substantial variations in $J(X, Y)$ even in cities with comparable global variability $I(X, Y)$. For example, the cities of Detroit and Philadelphia provide a striking contrast. While they have mutual information $I(X, Y)$, Philadelphia's mean local information $J(X, Y)$ is substantially higher. This reflects the fact that Detroit is composed of large, highly-segregated, monoracial neighborhoods, whereas Philadelphia has a much more fine-grained, intricate neighborhood structure. These patterns illustrate how combinations of the information measures $H(Y)$, $I(X, Y)$, and $J(X, Y)$ can be used to construct taxonomies diversity for American cities. It may be useful summarise the two measures by calling cities close to the bottom-right "most segregated", but we recommend reporting both $I(X, Y)$ and $J(X, Y)$. It is best to compare $J(X, Y)$ only between cities with similar $I(X, Y)$; thus, while it may be right to say that Detroit has lower levels of exposure than Phildelphia, when comparing Detroit to Baltimore it should be kept in mind that Baltimore is more even in its overall sociospatial structure.

Intriguingly, the mean local variability $J(X, Y)$ appears obey a scaling relation with respect to urban density. In Figure 5, we plot $J(X, Y)$ against the population density ρ of our studied cities. It appears that $J(X, Y)$ grows linearly with the logarithm of density. We interpret this trend as reflecting a compression of social space in dense urban areas: the same structure of racial variability fits into much less geographical area in New York than in Phoenix. One aspect of diversity in large, dense cities is that one need walk

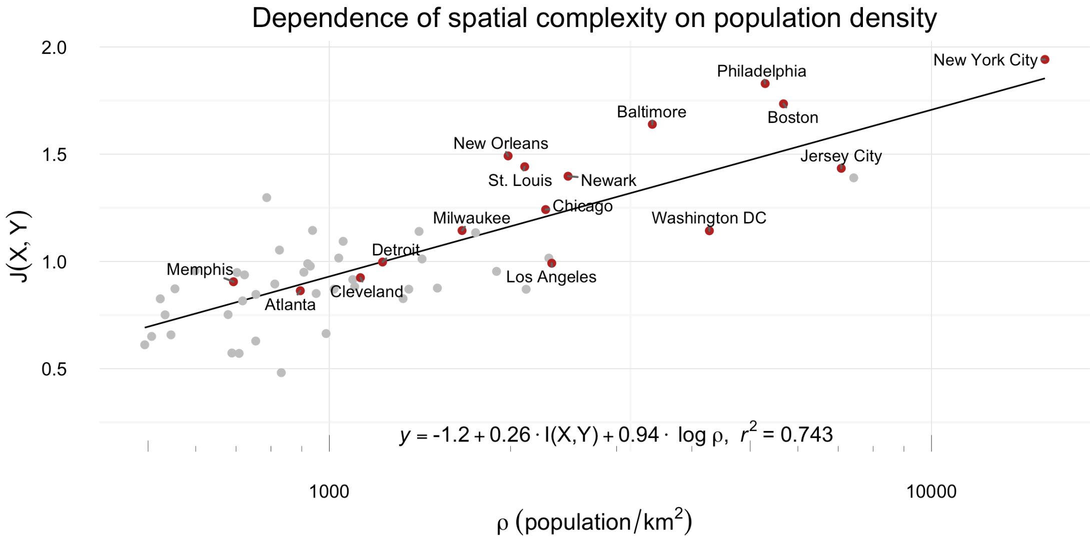


Figure 5: The mean local information $J(X,Y)$ scales with the logarithm of population density.

much less distance in order to reach a neighborhood with substantially different racial trends than one's own. It may be of interest for later studies to consider why this trend might arise through dynamical processes of urban formation.

Learning From Natural Neighborhoods

Figure 6 shows an illustrative clustering of Wayne County (housing the city of Detroit) into six regions using the greedy information maximization procedure defined by equation (14). The procedure cleanly divides the city into racially coherent zones. Cluster A is predominantly white and B predominantly black, demarcating the major racial divide in the city. Cluster C is a small transitional zone in which the two races overlap. Clusters D and E are the demographically distinct independent cities of Hamtramck and Grosse Pointe, respectively, while Cluster F is a Hispanic community known as Mexicantown. The existence of clusters like C reflects that not every racially coherent tract is interpretable as a "neighborhood"; some may be more naturally thought as transitions between neighborhoods.

Cities vary sharply in the level of cluster detail needed to convey equal amounts of demographic information. Figure 7 shows example cluster divisions for four major cities. Each clustering conveys approximately equal demographic information, as measured by the mutual information $I(C,Y)$ between cluster labels and racial distribution. In sharply segregated Atlanta, just two clusters–one predominantly white, the other predominantly

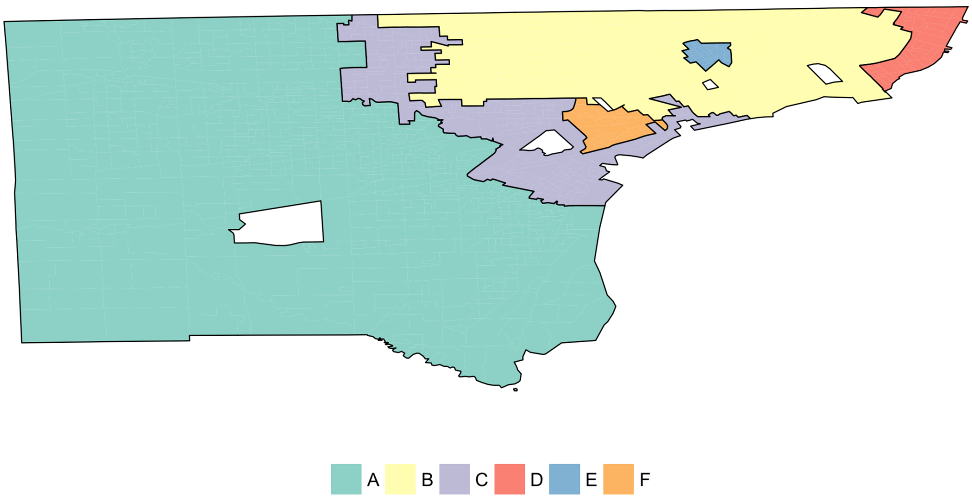


Figure 6: Clustering of Detroit into 6 neighborhoods based on race.

black–suffice to convey as much information as five in Boston, where the clusters are substantially more nuanced. Boston's Cluster A is predominantly white, cluster D majority Hispanic, and cluster E majority black. Cluster B is again majority white, but is distinct in that its minority citizens are almost all Hispanic. Cluster C is approximately equally black, Hispanic, and white.

Clusterings such as those shown in Figure 7 shed useful light on traditional concerns in the study of segregation. We have already argued that $I(X,Y)$ and $J(X,Y)$ naturally encode the high-level ideas of evenness and exposure endorsed by [24]. We now consider the further concepts of clustering and concentration. According to [22], "*clustering* measures the degree to which minority areas are located adjacent to one another" and "*concentration* refers to the degree of a group's agglomeration in urban space" (pages 309-310). When areas with similar racial composition are located close to one another, fewer spatial clusters are necessary to convey equivalent information about racial trends. Thus, 7 suggests that Atlanta is more clustered than Boston, a suggestion that we can quantify using the AUC methodology developed below. The concept of concentration invites us to consider the composition and density of each cluster. In Detroit, for example, whites make up 70% of "their" Cluster A, while black residents make up a full 86% of "their" Cluster B. Cluster B is also more concentrated in that it is more densely populated than Cluster A: it contains 34% of the Detroit's population, but accounts for just 19% of land in the analyzed area. This makes Cluster B more than twice as dense as Cluster A. Thus, in Detroit, "black" neighborhoods tend to be more racially homoge-

Figure 7: Example clusters for different cities, and the associated mutual information contained in each. The number in each box reflects the percentage of the population in the labeled racial group and cluster.

neous and more densely packed in urban space than "white" ones. Finally, we note that these clusterings allow us to examine the somewhat abstract concept of "exposure" in considerable detail. Figure 7 indicates that across all cities, Asians are much more likely to be found in predominantly white clusters than they are in predominantly black or Hispanic ones, indicating significant spatial integration between these two groups. In Chicago, some groups of Hispanics are exposed primarily to Hispanics and other whites (Cluster B), while other groups are exposed primarily to black residents (Cluster D).

Complexity and Neighborhood Structure

We expect that spatially complex cities with high $J(X, Y)$ would require more complicated models in order to capture similar levels of spatiosocial structure. Testing this expectation requires a quality measure defined on cluster models for each city. The natural measure of loss for a clustering is the mutual information $I(C, Y)$ between cluster and racial labels.

Figure 8: Evaluation of clustering using an Area Under the Curve (AUC) metric. The AUC is the fraction of the bounding box lying under the information gain curve. Boston's AUC is smaller than Detroit's, showing that Boston's spatial complexity (measured by $J(X,Y)$) requires more complex models than Detroit's.

To develop a global evaluation of the hierarchical clustering of a city according to the method defined by (14), we borrow the methodology of the "area under the curve" (AUC) used extensively in statistical learning. By plotting the information $I(C,Y)$ against the number of clusters n, we obtain a curve reflecting how the information evolves at varying scales of aggregation. Figure 8 illustrates these curves for Boston and Detroit, with N plotted on a logarithmic scale.

Figure 8 also shows the bounding rectangle, whose top right corner is defined by the number N of tracts in the census data set and the full mutual information $I(X,Y)$. The AUC is defined as the ratio of the shaded area in Figure 8 to the bounding rectangle, and is computed via the formula

$$A = \frac{1}{I(X,Y)\log N}\sum_{n=1}^{N} I(C_n,Y)\log\frac{n+1}{n}, \tag{16}$$

where N is the total number of Census blockgroups and C_n is the cluster label at the level in which n total clusters are defined.

When $A = 1$, it must hold that $I(C_1,Y) = I(X,Y)$. This implies that just one region carries full information; this can only occur in a city with no spatial unevenness, such as city (b) in 1. More generally, a large AUC indicates that simple models with few distinct regions capture much of the information about spatial variation in the city. In Atlanta, there exists a

	Dependent variable:
	AUC
$H(Y)$	-0.184^{***}
	$(-0.292, -0.075)$
$I(X,Y)$	0.450^{***}
	$(0.374, 0.525)$
$J(X,Y)$	-0.147^{***}
	$(-0.239, -0.055)$
Intercept	0.811^{***}
	$(0.733, 0.889)$
Observations	56
R^2	0.764
Adjusted R^2	0.750
Residual Std. Error	0.040 (df = 52)
F Statistic	55.959^{***} (df = 3; 52) (p = 0.000)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Figure 9: Summary of regression of the clustering AUC on information measures $H(Y)$, $I(X,Y)$, $J(X,Y)$. The AUC has been geographically adjusted to account for cities like New York, whose Census blockgroups form eight disconnected components, which would mildly distort results without adjustment. Table produced using [21]

.

clear dividing line between a predominantly white region in the north and a predominantly black region in the south. A two-cluster model therefore captures much of the information in the city, giving a high AUC of 0.71. In Boston, on the other hand, no such clear divide exists, and more complex models are necessary to capture similar amounts of information. Thus, Boston has a lower AUC of 0.65.

We argued previously that $H(Y)$, $I(X,Y)$, and $J(X,Y)$ provide a multidimensional characterization of spatial diversity in a city. Based on this discussion, we would expect that the performance of our clustering algorithm, as measured by the AUC, is closely related to these measures. Figure 9 confirms this expectation using simple linear regression of the AUC on the three measures. Overall, these three information measures explain approximately an adjusted 75% of the "clusterability" of cities. The signs associated with $I(X,Y)$ and $J(X,Y)$are suggestive: it is easiest to identify natural neighborhoods when there is substantial spatial variation (high $I(X,Y)$) but relatively simple local structure (low $J(X,Y)$).

4 Discussion

We have presented an information-theoretic framework for conceptualizing three core dimensions of urban diversity. Entropy measures global diversity; mutual information measures spatial evenness; and local information measures spatial exposure. Using information-based agglomerative clustering, we have also shown that these three measures are intimately related to the fineness of neighborhood structure. As 9 indicates, cities whose neighborhood structure can be easily characterized with just a few clusters tend to be those with either low diversity $H(Y)$ or low spatial exposure $J(X,Y)$, like Detroit. Cities with finer-grained spatial structure tend to have higher diversity $H(Y)$ and greater spatial exposure $J(X,Y)$, like Philadelphia. These tools are emminently practical; allowing both for the quantitative comparison of diversity profiles between cities (Figure 4) and more detailed analysis within a single city using naturally-defined neighborhoods (Figures 6 and 7). We conclude by noting some areas for future work, and some intriguing system properties illuminated by our findings.

While we have presented our information-theoretic framework in the context of racial residential segregation, these methods are substantially more general. Generalizing from race, information-theoretic methods applicable to any categorical variable or combinations thereof. It would be simple to conduct similar analysis for e.g. occupation type, or even to for a combination of race and occupation type. Information methods are somewhat less applicable in the case of continuous indicators [11]. However, even in this case it would be possible to define interpretable qualitative categories such as {"working class", "middle class", "upper class"} and proceed with analysis.

Generalizing from space, the remaining dimension of sociological interest is time. Many adults spend less than half of their waking hours at home [1], indicating that residential segregation is only a partial characterization of city-wide diversity. Fortunately, our ability to learn about daily patterns of human mobility is progressing at a rapid pace. Recent years have seen an enormous increase in the use of mobile devices, allowing the passive collection of digital traces. These traces can be processed, analyzed, and validated to derive insight into daily activity patterns [31, 34, 16, 15]. In the context of these developments, the information-theoretic suite of tools is attractive in its generality. Given appropriate data, only minimal mathematical changes are necessary:

- The entropy $H(Y)$ remains unchanged.
- Spatiotemporal evenness is the degree to which spatiosocial distributions change throughout the day. It can be measured by the global spatiotemporal mutual information $I([T,X];Y)$ between time T, space X, and demographics Y. Also of interest is the quantity $I(X,Y|T=t)$,

which measures spatial unevenness at a given time of day t.

It is also possible to define a greedily information-maximizing spatiotemporal clustering algorithm. Such an algorithm may be of special interest in temporal analysis on longer time-scales. For example, given many years of Census data, spatiotemporal clustering would allow the identification of neighborhoods and a characterization of their demographic and spatial evolutions.

The scaling relationship between the local information $J(X,Y)$ and the population density as shown in Figure 5 also deserves further investigation. Keeping in mind that $J(X,Y)$ measures the fineness of neighborhood structure, the upshot of Figure 5 is that, controlling for evenness, denser cities tend to have more racially-distinct neighborhoods per area. While further analysis is necessary to make this observation precise, it is intriguing to postulate a dynamical process of racial neighborhood formation and growth common to many American cities. It may be, for example, that an explanation is possible in terms of the interplay between spatial preferential attachment and infrastructure constraints.

References

[1] Occupational Employment Statistics.

[2] Shun-Ichi Amari and Hiroshi Nagaoka. Methods of Information Geometry, 2000.

[3] Michael Batty. Spatial Entropy. *Geographical Analysis*, 6(1):1–31, 1974.

[4] Michael Batty. Entropy in spatial aggregation. *Geographical Analysis*, 8(1):1–21, 1976.

[5] Michael Batty. Space, Scale, and Scaling in Entropy-Maximizing. 44(0):0–28, 2010.

[6] Michael Batty. Entropy and Spatial Geometry. *Royal Geographical Society*, 25(5):269–283, 2014.

[7] Luis M A Bettencourt, Joe Hand, and José Lobo. Spatial Selection and the Statistics of Neighborhoods. 2015.

[8] Roger Bivand, Tim Keitt, and Barry Rowlingson. rgdal: bindings for the Geospatial Data Abstraction Library, 2014.

[9] Roger Bivand and Nicholas Lewin-Koh. maptools: Tools for reading and handling spatial objects, 2014.

[10] Roger Bivand and Colin Rundel. rgeos: Interface to Geometry Engine - Open Source (GEOS), 2014.

[11] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, 1991.

[12] Imre Csiszar and Paul C Shields. Information Theory and Statistics: A Tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004.

[13] Robert D. Dietz. The estimation of neighborhood effects in the social sciences: An interdisciplinary approach. *Social Science Research*, 31(4):539–575, 2002.

[14] A V Diez Roux. Investigating neighbourhood and area effects on health. *Am J Public Health*, 91(11):1783–1789, 2001.

[15] Shan Jiang, Joseph Ferreira, and Marta C. González. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3):478–510, 2012.

[16] Shan Jiang, Y; Yang, G; Fiore, J; Ferreira, E; Frazzoli, and Marta C. González. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing,*. Association for Computing Machinery, 2013.

[17] Barret A Lee, Sean F Reardon, Glenn Firebaugh, Chard R Farrell, Stephen A Matthews, and David O'Sullivan. Beyond the Census Tract: Patterns and Determinants of Racial Segregation at Multiple Geographic Scales. *American Sociological Review*, 73(5):766–791, 2008.

[18] Rémi Louf and Marc Barthélemy. Patterns of residential segregation. 2015.

[19] Jens Ludwig, Greg J Duncan, Lisa A Gennetian, Lawrence F Katz, Ronald C Kessler, Jeffrey R Kling, and Lisa Sanbonmatsu. Neighborhood Effects on the Long-Term Well-Being of Low-Income Adults. *Science*, 337(September):1505–1510, 2012.

[20] Jens Ludwig, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, and Lisa Sanbonmatsu. Long-term neighborhood effects on low-income families: Evidence from moving to opportunity. *American Economic Review*, 103(3):226–231, 2013.

[21] Hlavac Marek. stargazer: Well-Formatted Regression and Summary Statistics Tables., 2015.

[22] Douglas S Massey and Nancy A Denton. The Dimensions of Residential Segregation. *Social Forces*, 67(2):281–315, 1988.

[23] Sean F Reardon and G Firebaugh. Measures of multigroup segregation. *Sociological Methodology*, 32:33–67, 2002.

[24] Sean F. Reardon and David O'Sullivan. Measures of Spatial Segregation. *Sociological Methodology*, 34(1):121–162, 2004.

[25] Elizabeth Roberto. Measuring Inequality and Segregation. 2015.

[26] Elizabeth Roberto. The Spatial Context of Residential Segregation. *arXiv.org*, pages 1–27, 2015.

[27] Hernán D Rozenfeld, Diego Rybski, José S Andrade, Michael Batty, H Eugene Stanley, and Hernán a Makse. Laws of population growth. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18702–18707, 2008.

[28] Hernán D Rozenfeld, Diego Rybski, Xavier Gabaix, and Hernán a Makse. The area and population of cities: New insights from a different perspective on cities. *American Economic Review*, 101(5):2205–2225, 2011.

[29] Robert J. Sampson, Jeffrey D. Morenoff, and Thomas Gannon-Rowley. Assessing "Neighborhoods Effects": Social Processes and New Directions in Research. *Annual Review of Sociology*, 28(2002):443–478, 2002.

[30] Michael J. Webber. *Information Theory and Urban Spatial Structure*. Croon Heml Ltd, London, 1979.

[31] Peter Widhalm, Yingxiang Yang, Michael Ulm, Shounak Athavale, and Marta C. González. Discovering urban activity patterns in cell phone data. pages 1–27, 2015.

[32] David Wong. Comparing Traditional and Spatial Segregation Measures: A Spatial Scale Perspective. *Urban Geography*, 25(1):66–82, 2004.

[33] David W. S. Wong. Geostatistics As Measures of Spatial Segregation. *Urban Geography*, 20(7):635–647, 1999.

[34] Yingxiang Yang, Daniele Veneziano, Chaoming Song, and Marta C. González. Modeling daily trajectories : Universal mechanisms , regional efects and individual differences.

5 Appendix

Software

We are pleased to make available two software repositories accompanying this analysis.

- Package `compx` for the `R` programming language implements computation of the information measures $H(Y)$, $I(X,Y)$, and $J(X,Y)$, as well as a method for information-theoretic clustering. Access `compx` at

  `https://github.com/PhilChodrow/compx`.

  To install in `R`, install the package `devtools` at

  `https://cran.r-project.org/web/packages/devtools/index.html`

  and run the command

  ```
  devtools::install_github(‘‘PhilChodrow/compx’’)
  ```

  in the `R` console.

- The analysis repository for this project, including data acquisition and processing; core computations; and figure generation. We hope that this repository will provide useful examples of how to use `compx` for others aiming to replicate and extend our results. Download the project files at

  `https://github.com/PhilChodrow/spatial_complexity`.

Relationship of Mutual and Fisher Informations

Let X be a continuous random variable taking values in $\mathbb{R}^n$, and let Y be a discrete random variable define on finite alphabet $\mathcal{Y}$. Suppose further that $p(y|x) > 0$ and that $p(y|x)$ is differentiable as a function of x for all $x, y \in \mathbb{R}^n \times \mathcal{Y}$. Fix $x_0 \in \mathbb{R}^n$, and define $B_r \triangleq B_r(x_0) = \{x \in R^n \mid \|x - x_0\| \leq r\}$. Additionally, define the *local mutual information* in B_r as the mutual information between X and Y where X is restricted to B_r:

$$I_r(x_0) \triangleq \mathbb{E}_X[D[p(\cdot|X)\|p(\cdot|X \in B_r)]|X \in B_r] \tag{17}$$

$$= \int_{B_r} p(x|X \in B_r) D[p(\cdot|x)\|p(\cdot|X \in B_r)] d^n x \,. \tag{18}$$

where $D[p\|q] \triangleq \sum_y p(y) \log \frac{p(y)}{q(y)}$ is the Kullback-Leibler divergence of q from p.

Theorem 1. *Under the stated conditions,*

$$\lim_{r \to 0} \frac{I_r(x_0)}{r^2} = \frac{n}{2(n+2)} \text{trace } J_Y(x_0) \,. \tag{19}$$

where the Fisher information matrix J_Y is given by

$$J_Y(x) \triangleq \mathbb{E}_Y \left[\nabla_x S_Y(x) \nabla_x S_Y(x)^T \right] \tag{20}$$

$$S_y(x) \triangleq \log p(y|x) \,. \tag{21}$$

We first define the following functions. For set A, let

$$f_A(x) \triangleq D[p(y|x)\|p(y|x \in A)] \,. \tag{22}$$

Additionally, let

$$q_r(x) \triangleq p(x|X \in B_r) \,. \tag{23}$$

We can then write the local mutual information as

$$I_r(x_0) = \int_{B_r} q_r f_{B_r} \, d\lambda \,, \tag{24}$$

where λ is the Lebesgue measure in $\mathbb{R}^n$. We also let $V_r \triangleq \int_{B_r} d\lambda$ be the volume of B_r. Finally, define

$$a_r(y) \triangleq p(X \in B_r, Y). \tag{25}$$

Lemma 1. *For all $x \in B_r$,*

$$q_r(x) = \frac{1 + O(r)}{V_r} \; . \tag{26}$$

Proof. First, Taylor's theorem implies that there exists a linear map T such that

$$\begin{aligned} p(x) &= p(x_0) + T(x - x_0) + O(\|x - x_0\|^2) && (27) \\ &= p(x_0) + T(x - x_0) + O(r^2) \; . && (28) \end{aligned}$$

Then,

$$\begin{aligned} p(X \in B_r) &= \int_{B_r} p \; d\lambda && (29) \\ &= \int_{B_r} p(x_0) + T(x - x_0) + O(r^2) d^n x \; . && (30) \end{aligned}$$

The middle term vanishes by spherical symmetry, leaving

$$p(X \in B_r) = p(x_0)V_r + O(r^{n+2}) \; . \tag{31}$$

For $x \in B_r$, we therefore have

$$\begin{aligned} p(x|X \in B_r) &= \frac{p(x, X \in B_r)}{p(X \in B_r)} && (32) \\ &= \frac{p(x)}{p(X \in B_r)} && (33) \\ &= \frac{p(x_0) + T(x - x_0) + O(r^2)}{p(x_0)V_r + O(r^{n+2})} && (34) \\ &= \frac{p(x_0) + O(r) + O(r^2)}{p(x_0)V_r} && (35) \\ &= \frac{1 + O(r)}{V_r} \; , && (36) \end{aligned}$$

as was to be shown. □

Lemma 2. *For all $y \in Y$,*

$$a_r(y) = p(x_0, y)V_r + O(r^{n+2}) \; . \tag{37}$$

Proof. Taylor's Theorem again implies that, for any $(x, y) \in B_r \times \mathcal{Y}$, there is a linear map T such that

$$\begin{aligned} p(x, y) &= p(x_0, y) + T(x - x_0) + O(\|x - x_0\|^2) && (38) \\ &= p(x_0, y) + T(x - x_0) + O(r^2). && (39) \end{aligned}$$

We then have

$$p(X \in B_r, y) = \int_{B_r} p(x,y)\; d^n x \tag{40}$$

$$= \int_{B_r} p(x_0,y) + T(x-x_0) + O(r^2)\; d^n x \tag{41}$$

$$= p(x_0,y)V_r + O(r^{n+2})\; d^n x\;, \tag{42}$$

where the middle term has again vanished due to spherical symmetry. □

Lemma 3. *For any $y \in \mathcal{Y}$, we have $p(y|X \in B_r) = p(y|x_0) + e_y$, where the error terms e_y satisfy $e_y \in O(r^2)$ and $\sum_{y\in\mathcal{Y}} e_y = 0$.*

Proof. That the errors must satisfy $\sum_{y\in\mathcal{Y}} e_y = 0$ follows from the fact that $p(\cdot|X \in B_r)$ must be a valid probability distribution on $\mathcal{Y}$. We'll now show that $e_y \in O(r^2)$. We then have

$$p(y|X \in B_r) = \frac{p(X \in B_r, Y)}{p(X \in B_r)} \tag{43}$$

$$= \frac{p(x_0,y)V_r + O(r^{n+2})}{p(x_0)V_r + O(r^{n+2})} \tag{44}$$

$$= p(y|x_0)O(r^2)\;, \qquad (V_r \propto r^n)$$

as needed. □

Lemma 4. *For any $x \in B_r$,*

$$f_{B_r}(x) = f_{x_0}(x) + O(r^3)\;. \tag{45}$$

Proof. We compute directly:

$$
\begin{aligned}
f_{B_r}(x) &= D[p(\cdot|x)\|p(\cdot|X \in B_r)] && (46)\\
&= \sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{p(y|X\in B_r)} && (47)\\
&= \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) - \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|X\in B_r) && (48)\\
&= \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) - \sum_{y\in\mathcal{Y}} p(y|x)\log(p(y|x_0)+e_y) && (49)\\
&= \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) - \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x_0) + \frac{e_y}{p(y|x_0)} + O(e_y^2) && (50)\\
&= \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) - \sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x_0) + \sum_{y\in\mathcal{Y}}\frac{p(y|x)}{p(y|x_0)}e_y + O(e_y^2) && (51)\\
&= D[p(\cdot|x)\|p(\cdot|x_0)] + \sum_{y\in\mathcal{Y}}\frac{p(y|x)}{p(y|x_0)}e_y + O(e_y^2) && (52)\\
&= f_{x_0}(x) + \sum_{y\in\mathcal{Y}}\frac{p(y|x)}{p(y|x_0)}e_y + O(e_y^2)\ . && (53)
\end{aligned}
$$

It remains to show that the error term is of order $O(r^3)$. Invoking Taylor's Theorem again, there exists a linear map T such that

$$
\begin{aligned}
\sum_{y\in\mathcal{Y}}\frac{p(y|x)}{p(y|x_0)}e_y + O(e_y^2) &= \sum_{y\in\mathcal{Y}}\left(1+\frac{1}{p(y|x_0)}T(x-x_0)+O(\|x-x_0\|^2)\right)e_y + O(e_y^2) && (54)\\
&= \sum_{y\in\mathcal{Y}}\left(1+O(r)+O(r^2)\right)e_y + O(e_y^2) && (55)\\
&= \sum_{y\in\mathcal{Y}}\left[e_y + O(r^3)\right] && (56)\\
&= O(r^3)\ , && (57)
\end{aligned}
$$

since $\sum_{y\in\mathcal{Y}} e_y = 0$ by the previous lemma. □

We now invoke the following well-known theorem of information geometry; for reference see [2].

Theorem 2. *The Kullback-Leibler divergence and Fisher information J_Y are related according to the following approximation:*

$$D[p(\cdot|x)\|p(\cdot|x_0)] = \frac{1}{2}\langle x-x_0, J_Y(x_0)(x-x_0)\rangle + O(\|x-x_0\|^3)\ . \qquad (58)$$

Theorem 2 allows us to write $I_r(x_0)$ as the following spherical integral:

$$
\begin{aligned}
I_r(x_0) &= \int_{B_r} q_r f_{B_r} \, d\lambda && (59)\\
&= \int_{B_r} q_r(x) f_{B_r}(x_0) \, d^n x && (60)\\
&= \int_{B_r} \frac{1+O(r)}{V_r} \left(f_{x_0}(x) + O(r^3) \right) \, d^n x && (61)\\
&= \frac{1+O(r)}{V_r} \int_{B_r} \left(f_{x_0}(x) + O(r^3) \right) \, d^n x && (62)\\
&= \frac{1}{2} \frac{1+O(r)}{V_r} \int_{B_r} \left(\langle x - x_0, J_Y(x_0)(x - x_0) \rangle + O(r^3) \right) \, d^n x && (63)\\
&= \frac{1}{2} \frac{1+O(r)}{V_r} \left(\int_{B_r} \langle x - x_0, J_Y(x_0)(x - x_0) \rangle \, d^n x + V_r O(r^3) \right) . && (64)
\end{aligned}
$$

It remains to compute the integral, which we do using the following lemma:

Lemma 5. *For any positive-semidefinite matrix $A \in \mathbb{R}^{n \times n}$,*

$$\int_{B_r} \langle x - x_0, A(x - x_0) \rangle \, d^n x = \frac{n}{n+2} r^2 V_r \mathrm{trace}(A)$$

Proof. Since A is positive-semidefinite, there exist an orthonormal matrix P and a diagonal matrix D such that $A = P^T D P$. Furthermore, the entries of D are the eigenvalues $\{\lambda_i\}$ of A. Then,

$$
\begin{aligned}
\int_{B_r} \langle x - x_0, A(x - x_0) \rangle \, d^n x &= \int_{B_r} \left\langle x - x_0, P^T D P(x - x_0) \right\rangle d^n x && (65)\\
&= \int_{B_r} \langle P(x - x_0), D P(x - x_0) \rangle \, d^n x \, . && (66)
\end{aligned}
$$

We can regard P as a reparameterization of B_r; since $\det P = 1$, we have

$$
\begin{aligned}
\int_{B_r} \langle P(x - x_0), D P(x - x_0) \rangle \, d^n x &= \int_{B_r} \langle x - x_0, D(x - x_0) \rangle \, d^n x && (67)\\
&= r^n \int_{B_n} \langle r x, r D x \rangle \, d^n x && (68)\\
&= r^{n+2} \int_{B_n} \langle x, D x \rangle \, d^n x \, , && (69)
\end{aligned}
$$

where B_n is the unit n-ball. We also let $S_n(r)$ be the n-sphere of radius r.

Continuing,

$$
\begin{aligned}
r^{n+2}\int_{B_n}\langle x, Dx\rangle\, d^n x &= r^{n+2}\int_{B_n}\sum_{i=1}^{n} x_i^2\lambda_i\, d^n x && (70)\\
&= r^{n+2}\sum_{i=1}^{n}\lambda_i\int_{B_n} x_i^2\, d^n x && (71)\\
&= \frac{r^{n+2}}{n}\sum_{i=1}^{n}\lambda_i\int_{B_n}\|x\|^2\ d^n x && \text{(spherical symmetry)}\\
&= \frac{r^{n+2}}{n}\mathrm{trace}(A)\int_{B_n}\|x\|^2\ d^n x && (72)\\
&= \frac{r^{n+2}}{n}\mathrm{trace}(A)\int_{\rho\in[0,1]}\rho^2 S_{n-1}(\rho)d\rho && (73)\\
&= \frac{r^{n+2}}{n}\mathrm{trace}(A)\int_{\rho\in[0,1]}\rho^{n+1} S_{n-1}(1)d\rho && (74)\\
&= \frac{r^{n+2}}{n}\mathrm{trace}(A)\frac{1}{n+2}S_{n-1}(1) && (75)\\
&= \frac{r^2}{n+2}\mathrm{trace}(A)nr^n v(B_n(1)) && (76)\\
&= \frac{n}{n+2}r^2 V_r\mathrm{trace}(A)\ , && (77)
\end{aligned}
$$

as was to be shown. □

We may therefore complete the proof:

$$
\begin{aligned}
I_r(x_0) &= \frac{1}{2}\frac{1+O(r)}{V_r}\left(\int_{B_r}\langle x-x_0, J_Y(x_0)(x-x_0)\rangle\, d^n x + V_r O(r^3)\right) && (78)\\
&= \frac{1}{2}\frac{1+O(r)}{V_r}\left[\frac{n}{n+2}V_r\mathrm{trace}\ J_Y(x_0) + V_r O(r^3)\right] && (79)\\
&= r^2\frac{n}{2(n+2)}\left(1+O(r)\right)\left[\mathrm{trace}\ J_Y(x_0) + O(r^3)\right]\ . && (80)
\end{aligned}
$$

Dividing through by r^2 and taking the limit as $r \to 0$ proves the theorem.

Computation of Local Information

In this section, we provide a specification of the computational procedure used to estimate $J(X,Y) = \mathbb{E}_x[\mathrm{trace}\ J_Y(X)]$ using blockgroup level data from the U.S. Census. For each fixed Census blockgroup i, let P_i be the population, A_i be the area, $\rho_i = P_i/A_i$ be the population density, and $p^i(y)$ be the observed proportion of racial group y.

Over the map we impose a hexagonal grid of radius r (see 3 for an example). For each hex k in the grid, let N_k be the set of Census blockgroups overlapping k. Define $p^k(i) = \rho_i / \sum_{i \in N_k} \rho_i$ as the estimated proportion of population within hex k residing in blockgroup i. Defining $p^k(i)$ this way reflects a simplifying assumption that each blockgroup in N_k overlaps hex k with equal area. Finally, define the estimated overall racial composition of hex k through the formula $p^k(y) = \sum_{i \in N_k} p^k(i) p^i(y)$. The mutual information in hex k is then:

$$I(k) = \sum_{i \in N_k} p^k(i) D[p^i \| p^k] \,. \tag{81}$$

Using (5), the estimated Fisher information at the centroid of hex k is

$$\text{trace } J_Y(k) \approx \frac{4I(k)}{r^2} \triangleq J(k) \,. \tag{82}$$

The estimated population in hex k is $P_k = A^k \sum_{i \in N_k} \rho_i$, where A^k is the cell area. We finally compute $\mathbb{E}_X[\text{trace} J_Y(X)]$ as

$$J(X,Y) = \mathbb{E}_X[\text{trace } J_Y(X)] \approx \frac{1}{\sum_k P_k} \sum_k P_k J(k) \tag{83}$$