Spatial Complexity and Scaling in Cities

v0.0

Phil Chodrow

May 29th, 2016

The problem of identifying natural, socio-economically defined neighborhoods arises in applied contexts including Census reporting, measurement of segregation, and dimension reduction in urban computing. This problem is also of interest for urban theory, since the difficulty of identifying neighborhoods may be viewed as a measure of socio-spatial complexity.

We develop a rigorous, information-theoretic approach to this topic, using open data on race in American cities as a case study. First, we formulate the *mean local information* $J(X, Y)$ as a localization of the mutual information between spatial and socio-economic variables. The measure $J(X, Y)$ is closely related to the Fisher information of the underlying joint distribution, and is therefore a measure of the intrinsic spatial complexity of an urban phenomenon. Unlike standard global 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 relate this algorithm's behavior to the mean local information. Questions raised by this work include the social, economic, and policy determinants of socio-spatial complexity in cities, and the potential use of spatial information measures in quantifying temporal changes in socio-economic structure, on time scales ranging from days to decades.

1 Introduction

Information theory provides one natural approach toward thinking about complex systems. It achieves this by mapping 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 disorder to make prediction challenging. In this paper, we apply information-theoretic tools to the measurement of *spatial compositional complexity*, with a focus on the spatial variation of racial trends in American cities.

Information-theoretic concepts have already found application in urban planning problems related to zoning and predicting population distributions [5, 2, 3, 4]. Substantial recent work has addressed the measurement of difference and disparity in cities through an information-theoretic lens [20, 6, 17, 18]. Attractive features of information-theoretic measures for this purpose include their deep relationship to statistical inference [7, 8], their generalizability to multiple demographic phenomena, and the fact that Theil's ([20]) original index satisfaction of many (though not all) of the invariance properties desirable for the measurement of segregation [16]. .

While closely-related to segregation, the concept of complexity has received less attention among quantitative sociologists. This is natural, since complexity is both less operational than segregation and challenging to define. On the other hand, as we demonstrate below, a spatial information-theoretic measure can be defined which measures the granularity of neighborhood structure, distinguishing cities with large, monolithic tracts of constant racial composition from those with many adjacent, smaller, neighborhoods of varying racial profiles. In doing so, it therefore captures one natural aspect of complexity in spatial compositional phenomena, and may also be of interest as a segregation measure as well.

1.1 Information Theory and Spatial Structure

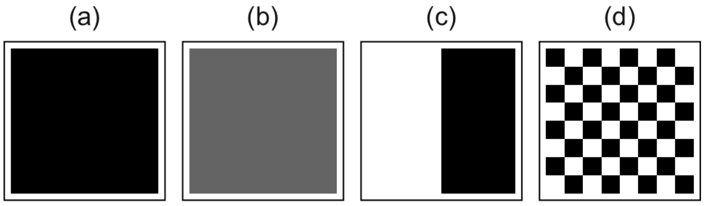
Two of the most fundamental objects of information theory are the entropy of a single random and the mutual information of two. Let X be a spatial variable, such as a set of coordinates in space or a neighborhood name. Let Y be a compositional variable that we aim to study, defined on an alphabet $\mathcal{Y}$. In the simple case in which $\mathcal{Y} = \{\text{Black}, \text{White}\}$, 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:

$$H(Y) \triangleq \sum_{y \in \mathcal{Y}} p(y) \log p(y) \ . \tag{1}$$

The entropy measures how evenly the global marginal distribution $p(Y)$ is distributed over the alphabet $\mathcal{Y}$. It thereby clearly distinguishes cities (a) and (b): since (a) is completely uniform, $H_a(Y) = 0$, while $H_b(Y) \neq 0$. 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:

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

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



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.

The divergence D is (with some caveats) 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. 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.

City (d) shares with city (c) the fact that residence and race are maximally dependent. 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 [18] 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 $p(X,Y)$.

2 Methods

2.1 Measuring Socio-Spatial Complexity

To motivate our methods, we first consider an idealized problem in which we have a differentiable field $p(y|x)$ of observed probability distributions for each $x \in M \subset \mathbb{R}^n$, where M is intuitively 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 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 pattern in local variations. $I_r(x_0) = 0$ if and only if $p(y|x)$ is constant for each y in the ball $B_r(x_0)$, corresponding to a complete lack of local compositional complexity.

In this definition, 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:

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

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

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

A more 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

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

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.

2.2 Identifying Natural Neighborhoods

Suppose that we have a collection of tracts labeled with the random variable X, and suppose that each tract has an additional label C denoting the *cluster* in which it lies. Then, it is possible to show that

$$I(X, Y) = I(Y, C) + I(X, Y|C) \tag{8}$$

Equation (8) has a natural interpretation: the information I have about Y if I know X is equal to the information I have if I know the cluster C, plus the amount of information I have if I subsequently learn the exact location X. Usefully, equation (8) decomposes the system-wide mutual information $I(X, Y)$ into a *between-cluster* component $I(Y, C)$ and an *within-cluster* component $I(X, Y|C)$. It is therefore comparable to various "sum of squares" decompositions that frequently appear in classical statistics. Indeed, it is possible to show that, when the differences between locations are small, the mutual information can be interpreted as a variance, in which case (8) expresses the sum of squares composition directly.

Finally, equation (8) motivates a natural, information-based scheme for identifying natural neighborhoods. If all I had access to were the cluster labels C and not the locations X, then my information would be $I(Y, C)$. 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 with good performance. Suppose that we face the problem of choosing a collection $I = \{1, 2, \ldots\}$ of locations to cluster together. The reduction in information associated with aggregating the locations I into a single cluster is

$$d(I) \triangleq \sum_{i \in I} p(X = i) D[p(Y|X = i) \| p(Y|X \in I)] - p(X \in I) D[p(Y|X \in I) \| p(Y)] \tag{9}$$

where $p(Y) = \sum_x p(x, Y)$ is the global marginal distribution. If we choose $I = \{i, j\}$, then equation (9) defines a natural information distance between locations i and j. Like the KL divergence, this distance is strictly nonnegative; unlike the KL divergence, this distance is symmetric. 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 stage, the algorithm aggregates together two locations i and j such that

1. Locations i and j are adjacent in space. This requirement ensures that the results may be naturally understood as neighborhoods, rather than disparate tracts with similar demographics.
2. Among all adjacent pairs, $d(i,j)$ is the smallest possible.

This greedy procedure defines a form of agglomerative 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 neighborhoods.

2.3 Complexity and Neighborhood Identification

3 Results

We assembled block-group level data from the 2008-2012 American Community Survey (ACS), conducted by the U.S. Census Bureau, on race and ethnicity for counties corresponding to 28 large US cities. At this early stage, the counties used were chosen to correspond to the most densely populated urban cores; in future developments we could standardize to conduct analyses for Metropolitan Statistical Areas. We then aggregated the detailed racial and ethnic groups into five meta-categories: 'Asian', 'Black', 'Hispanic', 'Other', and 'White'. 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 1km. 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 a complete summary of cities and their associated information measures is provided in Appendix 3.

Figure 2 shows the relationship of the global spatial variability $I(X,Y)$ and mean local variability $J(X,Y)$. An overall positive trend is evident, reflecting the fact that global spatial variability is a prerequisite for 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 comparable global variability $I(X,Y)$, Philadelphia's $J(X,Y)$ 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

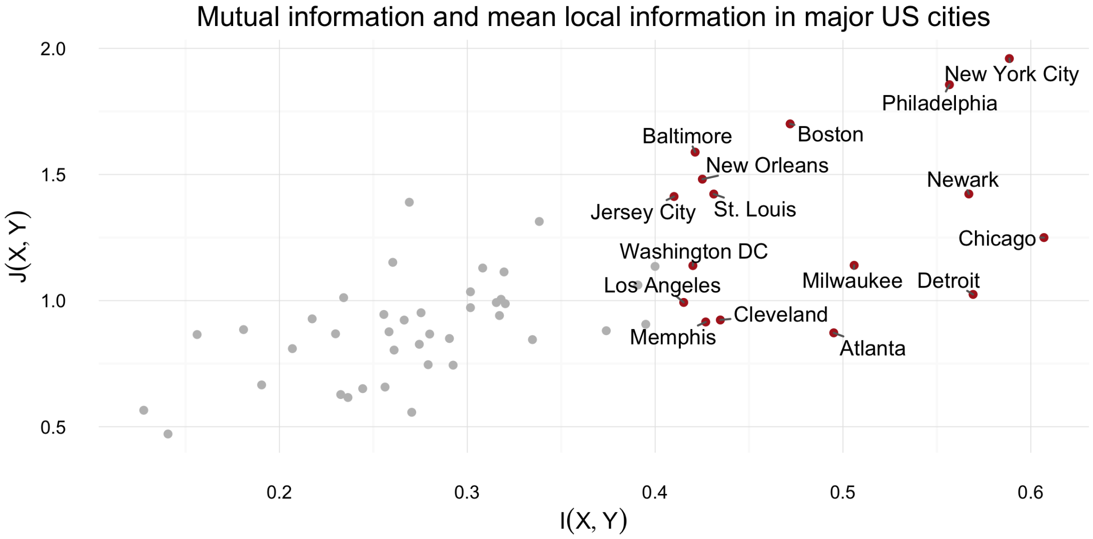


Figure 2: Relationship of global mutual information $I(X,Y)$ and mean local information $J(X,Y)$.

information measures $H(Y)$, $I(X,Y)$, and $J(X,Y)$ can be used to construct taxonomies diversity for American cities.

Intriguingly, the mean local variability $J(X,Y)$ appears obey a scaling relation with respect to urban density. In Figure 3, we plot $J(X,Y)$ against the population density ρ of our studied cities. A clear trend is evident, with $J(X,Y)$ growing 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 much less distance in order to reach a neighborhood with substantially different racial trends than one's own.

4 Discussion

We have formulated a novel measure of urban spatial complexity, and applied it to an analysis of spatial distributions of race in American cities. We emphasize that the mean local spatial variability $J(X,Y)$ is novel piece of a more complete information-theoretic characterization that also includes the entropy $H(Y)$ and the global variability $I(X,Y)$. Our results suggest two major directions of further exploration:

One major physical question raised by our results is the origin of the scaling relation seen in 3. This scaling relation may suggest a dynamical process of neighborhood formation and growth common to many American cities. If so, a physical understanding of this process in terms of individual-level behavior

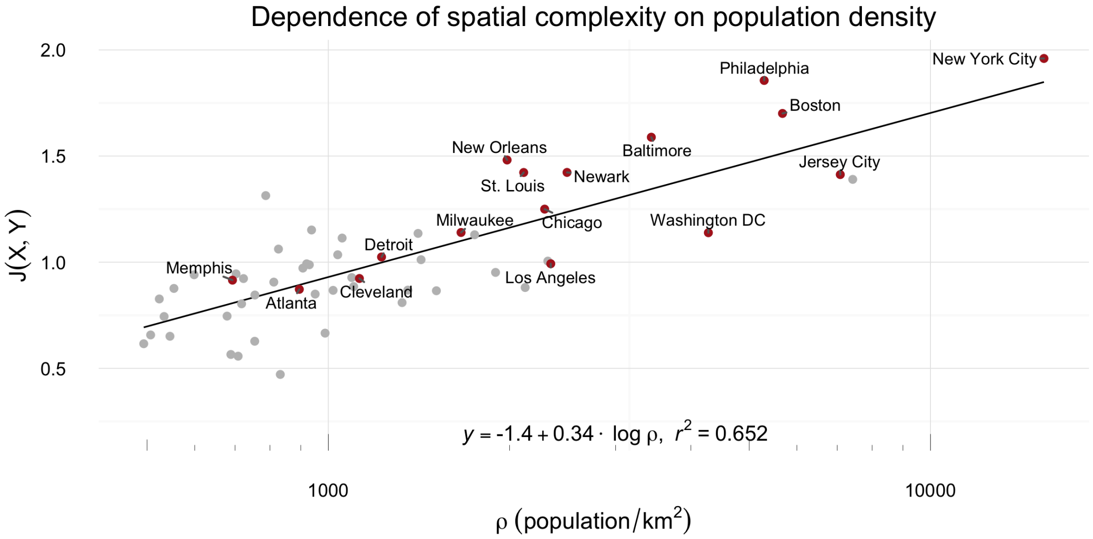


Figure 3: The mean local information $J(X, Y)$ scales with the logarithm of population density.

is called for. It is tempting to view the scaling relationship as an interplay of spatial preferential attachment and intrinsic limits to neighborhood density due to housing availability, but this view is speculative at this point. A theory to explain this behavior would be most welcome.

Another question raised is more operational. In many urban planning contexts, the amount of available data may make direct computations prohibitive. In such contexts, it may be useful to construct a simplifying model of the data the data (such as a clustering) in order to reduce dimensionality and make "the big picture" more visible. When a particular demographic phenomenon (such as race) is under consideration, it may be important to ensure that the model preserves the major patterns of variation in the granular data. We conjecture that the local variability $J(X, Y)$ measures the "modelability" of such data sets. Intuitively, a city like Detroit with relatively low $J(X, Y)$ and large, monoracial neighborhoods should be easily representable with relatively few modeled clusters, one for each of the relatively few major neighborhoods. On the other hand, relatively speaking, a city like Philadelphia with very intricate neighborhood structure may require substantially greater model complexity to represent without large loss of information. A promising course of further study is to design an information-theoretic clustering algorithm designed to model urban patterns, and then compare this algorithm's performance to $J(X, Y)$.

A third question and final question relates to the time-dependence of spatial structure. On one time scale, daily movement around a city for work or leisure activities has the impact of "mixing" separated residents in public spaces, potentially leading to very different spatial patterns of racial difference. On another time scale, these measures may track how the spatial structure of cities evolves over decades, potentially shedding further light on the dynamics of city formation.

A final question is....

References

[1] Alex Anas, Richard Arnott, and Kenneth Small. Urban spatial structure. *The transportation center*, page 63, 1997.

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

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

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

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

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

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

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

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

[10] Steven R. Holloway, Richard Wright, and Mark Ellis. The Racially Fragmented City? Neighborhood Racial Segregation and Diversity Jointly Considered. *The Professional Geographer*, 64(1):63–82, 2012.

[11] YM Ioannides and LD Loury. Job Information Networks, Neighborhood Effects. XLII(December):1056–1093, 2004.

[12] M J Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings. Biological sciences / The Royal Society*, 266(1421):859–67, 1999.

[13] 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.

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

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

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

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

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

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

[20] Henri Theil and Anthony J Finezza. A note on the measurement of racial integration of schools by means of informational concepts. *Taylor and Francis*, 1971.

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

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

5 Appendix

Relation 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:

$$\begin{aligned} I_r(x_0) &\triangleq \mathbb{E}_X[D[p(\cdot|X)\|p(\cdot|X \in B_r)]|X \in B_r] & (10) \\ &= \int_{B_r} p(x|X \in B_r) D[p(\cdot|x)\|p(\cdot|X \in B_r)] d^n x \ . & (11) \end{aligned}$$

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,*

$$\lim_{r \to 0} \frac{I_r(x_0)}{r^2} = \frac{n}{2(n+2)} \text{trace } J_Y(x_0) \ . \tag{12}$$

where the Fisher information matrix J_Y is given by

$$J_Y(x) \triangleq \mathbb{E}_Y \left[\nabla_x S_Y(x) \nabla_x S_Y(x)^T\right] \tag{13}$$

$$S_y(x) \triangleq \log p(y|x) \ . \tag{14}$$

The proof of Theorem 1 proceeds by the application of a number of Taylor approximations, in tandem with a fundamental relationship of information geometry. We first expand out $I_r(x_0)$ explicitly as

$$I_r(x_0) = \int_{B_r} p(x|X \in B_r) D[p(\cdot|x) \| p(\cdot|X \in B_r)] d^n x \ . \tag{15}$$

Lemma 1. *The following approximation relationships hold for the components of* (15)*:*

(a) $p(X \in B_r, Y) = p(x_0, Y)v(B_r) + O(r^{n+2})$

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

(c) $p(x|X \in B_r) = \frac{1+O(r)}{v(B_r)}$

Proof. For each approximation, we directly apply Taylor expansions about $X = x_0$.

(a) We have

$$p(X \in B_r, Y) = \int_{B_r} p(x, Y) \ d^n x \tag{16}$$

$$= \int_{B_r} p(x_0, Y) + \frac{\partial p(x_0, Y)}{\partial x}(x - x_0) + O(\|x - x_0\|^2) \ d^n x \tag{17}$$

$$= p(x_0, Y)v(B_r) + \frac{\partial p(x_0, Y)}{\partial x} \int_{B_r} (x - x_0) \ d^n x \tag{18}$$

$$+ O\left(\int_{B_r} \|x - x_0\|^2 \ d^n x\right) \tag{19}$$

$$= p(x_0, Y)v(B_r) + O(r^{n+2}) \ , \tag{20}$$

where the middle term vanishes due to spherical symmetry.

(b) The fact that the error terms e_y must satisfy $\sum_{y \in \mathcal{Y}} e_y = 0$ follows from the fact that $p(Y|X \in B_r)$ must be a valid probability distribution over $\mathcal{Y}$. We'll now show that $e_y \in O(r^2)$. First,

$$p(X \in B_r) = \sum_{y \in \mathcal{Y}} p(X \in B_r, y) \tag{21}$$

$$= \sum_{y \in \mathcal{Y}} \left[p(x_0, y)v(B_r) + O(r^{n+2})\right] \tag{22}$$

$$= p(x_0)v(B_r) + O(r^2); \tag{23}$$

from part (a). Next,

$$
\begin{aligned}
p(Y|X \in B_r) &= \frac{p(X \in B_r, Y)}{p(X \in B_r)} && (24)\\
&= \frac{p(x_0, Y)v(B_r) + O(r^{n+2})}{p(x_0)v(B_r) + O(r^{n+2})} && (25)\\
&= p(Y|x_0) + O(r^2)\ , && (26)
\end{aligned}
$$

which completes this part of the argument.

(c) First,

$$
\begin{aligned}
p(X \in B_r) &= \int_{B_r} p(x)\ d^n x && (27)\\
&= \int_{B_r} \left[p(x_0) + \nabla p(x_0)(x - x_0) + O(r^2)\right]\ d^n x && (28)\\
&= p(x_0)v(B_r) + O(r^{n+2})\ , && (29)
\end{aligned}
$$

where the middle term again vanishes through spherical symmetry. Thus, for $x \in B_r$, we have

$$
\begin{aligned}
p(x|X \in B_r) &= \frac{p(x)}{p(X \in B_r)} && (30)\\
&= \frac{p(x_0) + \nabla p(x_0)(x - x_0) + O(r^2)}{p(x_0)v(B_r) + O(r^{n+2})} && (31)\\
&= \frac{1 + O(r)}{v(B_r)}\ . && (32)
\end{aligned}
$$

□

Lemma 2. *The following approximation holds for the divergence factor in the integral* (15)

$$
D[p(\cdot|x) \| p(\cdot|X \in B_r)] = D[p(\cdot|x) \| p(\cdot|x_0)] + O(r^3) \qquad (33)
$$

Proof. We compute directly:

$$
\begin{aligned}
D[p(\cdot|x) \| p(\cdot|X \in B_r)] &= \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{p(y|X \in B_r)} && (34)\\
&= -H[Y|X=x] - \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|X \in B_r) && (35)\\
&= -H[Y|X=x] - \sum_{y\in\mathcal{Y}} p(y|x) \log \left(p(y|x_0) + e_y\right) \\
& && \text{(from Lemma 1)}\\
&= -H[Y|X=x] - \sum_{y\in\mathcal{Y}} p(y|x) \left[\log p(y|x_0) + \frac{e_y}{p(y|x_0)} + O(e_y^2)\right] \\
& && (36)\\
&= D[p(\cdot|x)\|p(\cdot|x_0)] + \sum_{y\in\mathcal{Y}} \frac{p(y|x)}{p(y|x_0)} e_y \\
& && \text{(quadratic terms negligible)}\\
&= D[p(\cdot|x)\|p(\cdot|x_0)] && (37)\\
&\quad + \sum_{y\in\mathcal{Y}} \left(1 + \frac{1}{p(y|x_0)} \nabla p(y|x_0)(x-x_0) + O(r^2)\right) e_y \\
& && (38)\\
&= D[p(\cdot|x)\|p(\cdot|x_0)] + \sum_{y\in\mathcal{Y}} \left[e_y + O(r^3)\right] && (e_y \in O(r^2))\\
&= D[p(\cdot|x)\|p(\cdot|x_0)] + O(r^3) && (\textstyle\sum_{y\in\mathcal{Y}} e_y = 0)
\end{aligned}
$$

□

Lemma 3. *For any positive-semidefinite matrix $A \in \mathbb{R}^{n\times n}$,*

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

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,

$$
\begin{aligned}
\int_{B_r} \langle x - x_0, A(x-x_0)\rangle \, d^n x &= \int_{B_r} \left\langle x - x_0, P^T D P(x-x_0)\right\rangle d^n x && (39)\\
&= \int_{B_r} \langle P(x-x_0), DP(x-x_0)\rangle \, d^n x \; . && (40)
\end{aligned}
$$

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

$$\begin{aligned}\int_{B_r} \langle P(x-x_0), DP(x-x_0)\rangle \, d^n x &= \int_{B_r} \langle x-x_0, D(x-x_0)\rangle \, d^n x && (41)\\ &= r^n \int_{B_n} \langle rx, rDx\rangle \, d^n x && (42)\\ &= r^{n+2} \int_{B_n} \langle x, Dx\rangle \, d^n x \ , && (43)\end{aligned}$$

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

$$\begin{aligned}r^{n+2} \int_{B_n} \langle x, Dx\rangle \, d^n x &= r^{n+2} \int_{B_n} \sum_{i=1}^{n} x_i^2 \lambda_i \; d^n x && (44)\\ &= r^{n+2} \sum_{i=1}^{n} \lambda_i \int_{B_n} x_i^2 \; d^n x && (45)\\ &= \frac{r^{n+2}}{n} \sum_{i=1}^{n} \lambda_i \int_{B_n} \|x\|^2 \; d^n x && (46)\\ & && \text{(spherical symmetry)}\\ &= \frac{r^{n+2}}{n} \text{trace}(A) \int_{B_n} \|x\|^2 \; d^n x && (47)\\ & && \text{(spherical symmetry)}\\ &= \frac{r^{n+2}}{n} \text{trace}(A) \int_{\rho \in [0,1]} \rho^2 S_{n-1}(\rho) d\rho && (48)\\ &= \frac{r^{n+2}}{n} \text{trace}(A) \int_{\rho \in [0,1]} \rho^{n+1} S_{n-1}(1) d\rho && (49)\\ &= \frac{r^{n+2}}{n} \text{trace}(A) \frac{1}{n+2} S_{n-1}(1) && (50)\\ &= \frac{r^2}{n+2} \text{trace}(A) n r^n v(B_n(1)) && (51)\\ &= \frac{n}{n+2} r^2 v(B_r) \text{trace}(A) \ , && (52)\end{aligned}$$

as was to be shown. □

Fact. *The Kullback-Leibler divergence and the Fisher information J_Y are related according to the approximation*

$$D[p(\cdot|x) \| p(\cdot|x_0)] = \frac{1}{2} \langle x - x_0, J_Y(x_0)(x-x_0)\rangle + O(\|x-x_0\|^3) \qquad (53)$$

We are finally ready to prove Theorem 1. Computing directly, we have

$$
\begin{aligned}
I_r(x_0) &\triangleq \mathbb{E}_X[D[p(\cdot|X)\|p(\cdot|X \in B_r)]|X \in B_r] \ . && (54)\\
&= \int_{B_r} p(x|X \in B_r) D[p(\cdot|x)\|p(\cdot|X \in B_r)] d^n x && (55)\\
&= \int_{B_r} \left[\frac{1+O(r)}{v(B_r)}\right] D[p(\cdot|x)\|p(\cdot|X \in B_r)] d^n x && \text{(Lemma 1(c))}\\
&= \left[\frac{1+O(r)}{v(B_r)}\right] \int_{B_r} \left(D[p(\cdot|x)\|p(\cdot|x_0)] + O(r^3)\right) d^n x && \text{(Lemma 2)}\\
&= \left[\frac{1+O(r)}{v(B_r)}\right] \int_{B_r} \left(\frac{1}{2}\left\langle x - x_0, J_Y(x_0)(x-x_0)\right\rangle + O(\|x-x_0\|^3) + O(r^3)\right) d^n x\\
& && (56)\\
&= \frac{1}{2}\left[\frac{1+O(r)}{v(B_r)}\right] \int_{B_r} \left(\langle x - x_0, J_Y(x_0)(x-x_0)\rangle + O(r^3)\right) d^n x && (57)\\
&= \frac{1}{2}\left[\frac{1+O(r)}{v(B_r)}\right] \left(\frac{n}{n+2} r^2 v(B_r) \mathrm{trace}(J_Y(x_0)) + v(B_r) O(r^3)\right) && (58)\\
&= r^2 \frac{n}{2(n+2)} \left[1 + O(r)\right] \left(\mathrm{trace}(J_Y(x_0)) + O(r^3)\right) \ . && (59)
\end{aligned}
$$

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

Computational Methods and Assumptions

In this section, we provide a specification of the computational procedure used to estimate $J(X, Y) = \mathbb{E}_x[J_Y(X)]$ using blockgroup level data from the U.S. Census.

For fixed Census blockgroup i, let P_i be the population, let A_i be the area, let $\rho_i = P_i/A_i$ be the population density, and let $p^i_Y(y)$ be the observed proportion of racial group y. For hex k in our hexagonal grid, let N_k be the set of overlapping Census blockgroups. We also define $p^k_I(i) = \rho_i / \sum_{i \in N_k} \rho_i$ as the estimated proportion of population within hex k residing in blockgroup i. This definition embodies a computationally-simplifying assumption that each blockgroup in N_k overlaps hex k with equal area. Finally, $p^k_Y(y) = \sum_{i \in N_k} p^k_I(i) p^i_Y(y)$ is the estimated overall racial composition of hex k. Then, we estimate the mutual information in hex k as

$$I(k) = \sum_{i \in N_k} p^k_I(i) D[p^i_Y(\cdot)\|p^k_Y(\cdot)] \ . \tag{60}$$

Using (5), the estimated Fisher information is

$$J(k) \approx \frac{4I(k)}{r^2} \tag{61}$$

where r is the grid radius. 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 estimate $\mathbb{E}_X[J(X)]$ as

$$J(X,Y) = \mathbb{E}_X[J_Y(X)] \approx \frac{1}{\sum_k P_k} \sum_k P_k J(k) \tag{62}$$