15 *Explainer: Measuring clustering and segregation*

MOON DUCHIN AND JAMES M. MURPHY

BACKSTORY

This explainer will focus on a statistic—"Moran's I"—that is so ubiquitously used in geography to measure spatial structure in a dataset that it has become almost interchangeable with the *concept* of spatial structure.

Though P.A.P. Moran developed it slightly earlier, Moran's I was brought into geography during the rise of *spatial analysis*, a subdiscipline that emerged during the late 1940s. Before that turn, geography as a university discipline had been framed as a study of places and regions, with an emphasis on description and characterization. After World War II, universities became increasingly entangled with what Eisenhower had famously dubbed the "military-industrial complex," which led to increased research emphasis in areas connected to defense, planning, and decision science. This brought a so-called "quantitative revolution" to geography, among many other domains.

By the 1950s, with a boost from Red Scare politics, this new muscularly mathematized toolset had pushed cultural and Marxist geography to the margins, sometimes seeing geography departments entirely eliminated in the course of postwar modernization. In the 1970s, the pendulum had begun to swing back, and critiques of the spatial analysis framework—as positivist, politically, and culturally disconnected, and too far from the more descriptive geography of the early twentieth century—became more audible. But by then, the rise of computing meant that the calculational spirit of spatial analysis was fairly entrenched. Metrics like I, which were developed to measure the degree of spatial patterning in data, could now become instantly accessible in spatial software. Now one could load a dataset with population demographics for Chicago and, at the push of a button, learn that I = .884 for Black population and I = .828 for Latino population, both very high numbers in a city where random distributions would yield scores closer to zero. With this ease of use, the straight-up comparison of one score to another, across different localities and time periods and contexts, became unavoidably tempting.

What is a score like this trying to measure? Arthur Getis, one of the standard-bearers of the spatial analysis school, cites the following definition:

> Given a set S containing n geographical units, *spatial autocorrelation* refers to the relationship between some variable observed in each of

> the n localities and a measure of geographical proximity defined for all $n(n-1)$ pairs chosen from n [1].

This is essentially just a quantification of the common-sense maxim called Tobler's First Law of Geography: *Everything is related to everything else, but near things are more related than distant things.* As we will see, the `I` metric attempts to take this literally by measuring how much the values in a unit are like the values at the neighbors. When the score was built to focus on literally adjacent units, as in the early work of Moran and his contemporaries, it was sometimes called a *contiguity ratio*. In its more general form, it was given the lasting name of spatial autocorrelation in an influential conference paper by Cliff and Ord in 1968, followed by a monograph in the early 1970s [2, 3].

So the story of `I` is a story of a chalkboard-math intervention in spatial statistics that became possible because of the academic politics of its era. Just as the need to prove geography's mathematical bona fides was starting to fade, the rise of computers made it easy to crunch numbers on larger and larger datasets. This also made it possible to stop thinking about the formulas.

Below, we'll turn back the clock and look at measurements of clustering (a.k.a. segregation), poking them mathematically to see what we find.

WHAT IS SEGREGATION?

A classical problem in quantitative social science is to define a measure of segregation that matches up with the ways that people talk about their communities. More precisely, given a geography with some information about a demographic subgroup, the problem is to quantify how much the group is separated rather than undifferentiated from the rest of the population in terms of residential patterns.

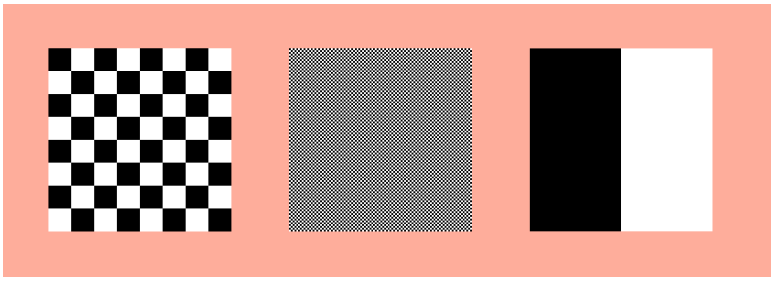

Figure 1: How interspersed or separated are the black and white colors—i.e., how segregated are these cities?

In Chapter 10, Chris Fowler talked about this question, and noted that geographers think of this as a multiscale measurement problem—as Figure 1 makes clear, the answer might depend on how the pattern falls against a chosen set of units. We'll develop that idea mathematically here.

(DIS)SIMILARITY ACROSS GEOGRAPHICAL UNITS

One approach is to demarcate the geography into smaller units and analyze the demographics on these units. For instance, we can divide up a city into its census tracts and look at demographic population proportion by tract. Suppose the units are numbered 1 through n, and we want to study the population of a group B relative to the total population. We can denote the number of B-type people and the total population of unit i as b_i and p_i, respectively, and write the totals for the large geography as $B = \sum_{i=1}^{n} b_i$ and $P = \sum_{i=1}^{n} p_i$. Let's write $\rho_i = b_i/p_i$ for the share of B population in unit i and $\rho = B/P$ for the global ratio. Then, if the local population shares ρ_i are the same for all i, and therefore equal to the global ratio ρ, we would declare the city completely unsegregated at the scale of the units we have chosen. But if there is one part of the city where the local ratio is far higher and another part of the city where it's far lower, that sounds segregated.

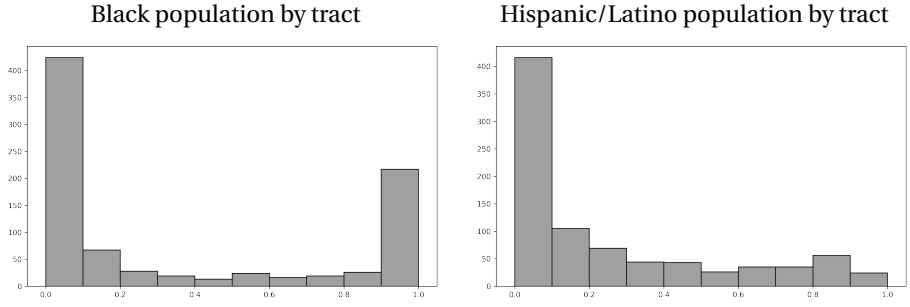

Figure 2: These histograms show how the 853 census tracts in Chicago are distributed by Black population (left) and Hispanic/Latino population (right). There are over 200 tracts that are more than 90% Black, but no comparably large number of heavily Latino tracts.

Reasoning this way, we can define the *dissimilarity index* by comparing b_i/p_i to the global ratio B/P in each unit, which turns out to be equivalent to comparing b_i/B to p_i/P. We define:

$$D = \frac{1}{2}\sum_{i=1}^{n}\left|\frac{b_i}{B} - \frac{p_i}{P}\right| = \frac{1}{2\rho}\sum_{i=1}^{n}\frac{p_i}{P}\cdot\left|\rho_i - \rho\right|.$$

That means a geographical unit makes no contribution to dissimilarity if the population share in that unit, ρ_i, is the same as the share in the whole geography, ρ. But if one unit has very different population proportions than the region overall, it contributes significantly to the overall dissimilarity.[1,2]

[1] There is a large body of literature on the dissimilarity index. This expression for D matches the one used in Frey and Myers [4]; an expression with a different normalization coefficient is cited in Massey and Denton [5], where it is noted that the formulations have varied in the literature.

[2] More generally, the dissimilarity index allows us to compare any two populations B and C with an exactly similar formula; this is the special case that C is the total population. For this case, let's

We can record the population shares in an ordered list, or vector; we can then subtract off the average to get the deviations.

$$(\rho_1, \rho_2, \ldots, \rho_n) \longrightarrow (\rho_1 - \rho,\ \rho_2 - \rho, \ldots,\ \rho_n - \rho).$$

If we call this deviation vector $x = (x_1, \ldots, x_n)$, then we see that dissimilarity D is based on the average magnitude of deviation—it begins with the average of the $|x_i|$, weighted by the population of each unit. (Then there's a normalization by a factor out front that drops out when the group has half of the total population, but scales up the final answer when the size of the subgroup is small.)

For example, let's consider Black population in Chicago, according to the 2010 decennial census. There are 853 (populated) census tracts in the city, and we can make a vector of length 853 recording the Black population share by tract. The citywide Black population was 32.2%, so we can subtract off .322 from each coordinate to get our deviation vector. We compute $D = .54$ for Black population. The Hispanic population share citywide is .288 and $D = .45$. If you rescale these to get dissimilarity on a zero-one scale, you'd see that the Black population is scored by D as having roughly 80% of the maximum possible dissimilarity for a population of that size, while the Hispanic/Latino dissimilarity is 63% of its maximum.

The limitation of dissimilarity for understanding segregation is probably clear at this point: each unit is treated separately, with no spatiality taken into account, so the score makes no distinction between a left/right split and a checkerboard (see Figure 1). That's not a great fit for how we talk about segregation, where the former is clearly more segregated than the latter.

SPATIALIZING SEGREGATION

Next, we can treat the geography as a *spatial network*, recording the spatial relationships by placing edges between the nodes when the units they represent are adjacent. Possibly the simplest network model is an undirected graph $G = (V, E)$ where the vertices correspond to geographical units (like the census tracts of Chicago) and the edge set encodes spatial adjacency of units. We show Chicago as a *dual graph* (a graph dual to the tracts) in the right column of Figure 3.[3]

Now quantifying how much a subgroup is segregated is a question not only of statistics but also of the graph structure (sometimes called the *network topology*). The connectivity of the underlying graph determines to a large extent what kinds of patterns will count as segregated. In this sense, meaningful measures of segregation

examine how this is normalized. Consider a group with population share $\rho = B/P$. It's clear that the lowest possible dissimilarity would be $D = 0$ when every unit has ρ share. The highest possible would occur if some units (roughly ρn of them) have share approaching 1 and the rest have share 0. With the normalization shown here, this yields $D = 1 - \rho$. That is, small populations can register as very segregated, but large populations can't get a very high D score. This is not crazy, as a reflection of how we talk about segregation! But if you want to rescale to get D ranging from zero to one, you would divide by $1 - \rho$.

[3] Just to fix terminology: we use "graph" and "network" interchangeably, and we use "node" and "vertex" interchangeably, as is common in the literature.

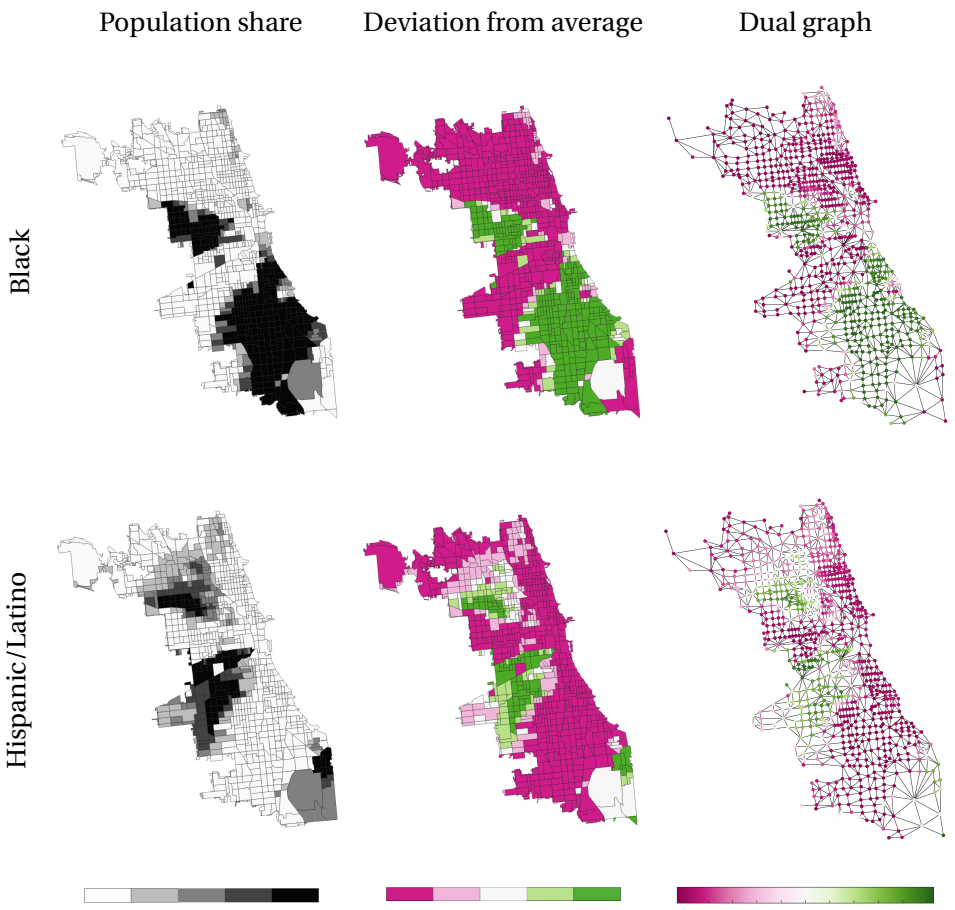

Figure 3: Demographics in spatial context: population share of minority group, deviation from average, and dual graph. Green is above average, purple below average.

must account not only for fluctuations of ρ_i around its mean, but must relate to the structure of the underlying network that records spatiality.

Moran's I is a classical quantitative measure of segregation [6, 7]. If we start with any numerical values associated with the nodes, such as the population shares ρ_i, Moran's I returns a real number, usually between -1 and 1. The standard interpretation is that values near 1 indicate extreme segregation, values near zero indicate no pattern, and negative values can be achieved by checkerboards, where the units alternate between one population subgroup and the other. Just as before, we're going to start with a vector of population shares by unit and subtract off the average to get a deviation vector. We'll do one thing differently this time: we'll assume that all we know is the *share* at each node, and not the total population. So when we take an average, we'll do so weighting all nodes equally. Thus, if $(\rho_1, \ldots, \rho_n)$ is the share by unit, and $\beta = \frac{1}{n}\sum_{i=1}^{n} \rho_i$ is the average of these values, then we get the deviation vector by $x_i = \rho_i - \beta$. For census tracts in Chicago, for instance, this way of averag-

ing makes four percentage points of difference: the average Black population in a census tract is $\beta = .362$ rather than the citywide $\rho = .322$ Black population share, indicating that tracts with high Black population are relatively underpopulated.

Now suppose the graph/network G has m pairs of adjacent units (i.e., the graph has m edges) in all. Then we can define

$$\mathtt{I}(x_1, \ldots, x_n) = \frac{n}{m} \frac{\sum_{i \sim j} x_i x_j}{\sum_i {x_i}^2}$$

where $i \sim j$ if spatial units i and j are adjacent. This is asking how the average product of neighboring values compares to the average product of a value with itself. Computing for the Black and Hispanic population in Chicago, we get $\mathtt{I} = .881$ and $\mathtt{I} = .825$, respectively.

We can interpret I as looking for patterns in the locations where the x values are positive or negative. To see this, think again about the left/right pattern versus the checkerboard. If the units are chosen so that they have solid, alternating colors (the checkerboard situation), each term in the numerator will be negative (because x_i and x_j have different signs), making I negative overall. In the left/right division, most terms will contribute positively to the numerator because they will have negative next to negative or positive next to positive, making the expression positive overall. And if there is no pattern, we will tend to see a lot of cancellation.

I AS A SLOPE

There are two intuitive interpretations of I that bear mentioning. First, I is the slope of the best-fit line relating the value at a node to the value at the neighbors (which is called the *lagged* value).[4]

So if tracts are just like their neighbors except in a small transitional area, $\mathtt{I} \approx 1$. If everything is the opposite of its neighbor, then there will be clusters in the northwest and southeast corners of the scatterplot, and $\mathtt{I} \approx -1$. And if there are no patterns at all, then the average of your neighbors will tend to be the same as the overall average, no matter what your value is, which makes the fit line flat, or $\mathtt{I} \approx 0$.

Perhaps a few limitations of this approach are now visible. By just reporting the *slope* of the fit line, it loses a great deal of information from the scatterplot (see Figure 4 for scatterplots and lines of best fit for the Black and Hispanic population of Chicago). It fails to adequately distinguish the bimodal distribution of Black population in Chicago, which is much more characteristic of the way we think about segregation, from the less concentrated Latino population. Another drawback is made clear by this way of visualizing the score: it's going to give meaningless answers for a very uniformly distributed population, because all the data points will be in one small area of the scatterplot. When you fit a line through a small

[4] This would be perfectly accurate for regular graphs. In general, we have to slightly modify the I formula for this interpretation to be exact. In terms of the linear algebra formulation in the next section, we should replace the adjacency matrix A with the row-standardized adjacency matrix, i.e., each rowsum normalized to one. The difference is fairly slight; I changes by less than .01 when introducing row-standardization into these Chicago examples.

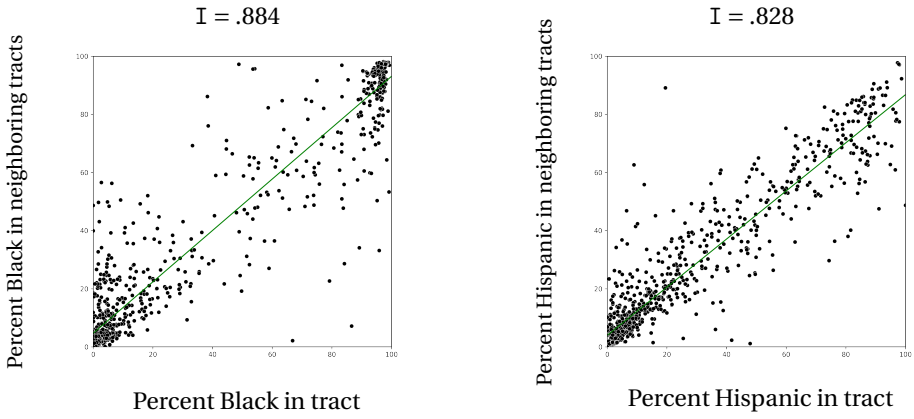

Figure 4: Moran scatterplots for the Black and Hispanic population of Chicago—note that the projection of these plots to the x-axis would give back the histograms in Figure 2. Segregation in Chicago is captured by the fit line being nearly diagonal: tracts tend to have neighboring tracts with similar demographics.

.

ball of points, its slope does not have much meaning! So in the case of a very unsegregated population, the score is more noise than signal, and in the limit when the population is exactly even over the units, the score is undefined.

I VIA LINEAR ALGEBRA

There's another ready interpretation of the I formula that sheds a lot of light on what it's doing, from a mathematician's perspective. Let A be the adjacency matrix of the graph: an $n \times n$ graph that has a zero in position i, j if those units are not adjacent, and a one if they are. Let x be the vector $(x_1, \ldots, x_n)$ of deviations, as above. Then there's a neat way to write the calculation in matrix notation: $\mathrm{I}(\mathsf{x}) = \frac{n}{2m}\left(\frac{\mathsf{x}A\mathsf{x}^\mathsf{T}}{\mathsf{x}\mathsf{x}^\mathsf{T}}\right)$.

Those who have some linear algebra background will recognize this expression in parentheses as a *Rayleigh quotient*: it is exactly what you maximize or minimize to get the largest and smallest eigenvalues of A, and the values of x where these occur are the corresponding eigenvectors.[5] The study of eigenvalues for matrices coming from graphs, like our adjacency matrix A, belongs to the kind of math called *spectral graph theory*.

A major theme in spectral graph theory is to relate the connection structure of the graph G to the eigenvalue spectrum of associated matrices. In particular, when the graph is *regular* (same number of edges incident to every node), the eigenvectors of A associated with the largest eigenvalues are known to capture latent cluster structure in the data [8].[6]

[5] We call v an eigenvector of A with associated eigenvalue λ if $A\mathsf{v} = \lambda\mathsf{v}$, and the list of eigenvalues is called the *spectrum*. Eigenvalues come up all over pure and applied mathematics.

[6] In spectral graph theory, it's more common to study eigenvalues of a related matrix called the graph *Laplacian* rather than the adjacency matrix; when the graph is regular of degree d, the Laplacian is $L = dI - A$, so the spectrum of L is related to the spectrum of A by translation and reflection.

This suggests that a population deviation vector x will give a large, positive Moran's I score precisely when it puts its positive and negative values in these graph clusters, which are often called "communities."[7] Back to plain English: this segregation score maxes out when you can find two parts of the graph that are relatively well connected internally but relatively separate from each other, and your population group is concentrated on one of those clusters.

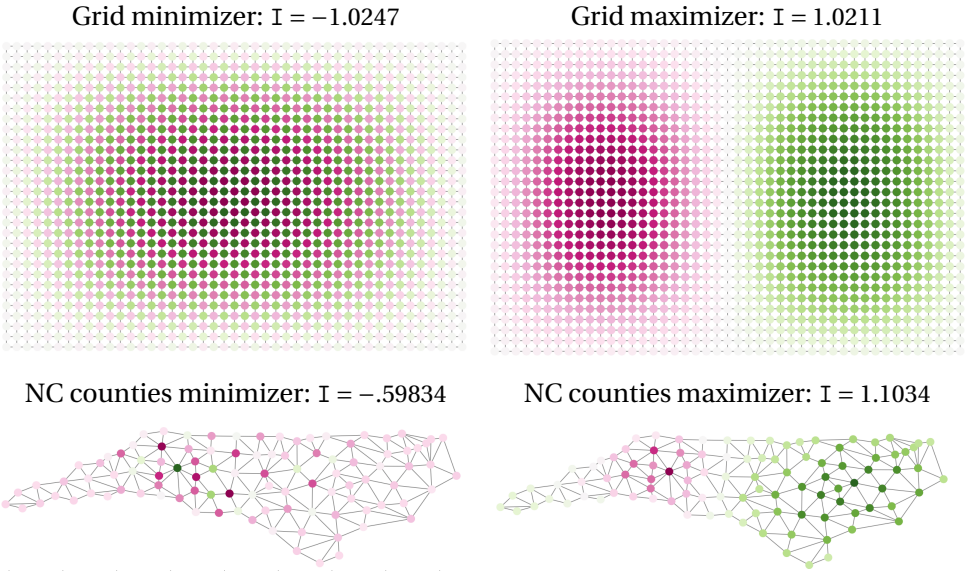

Figure 5: Top: a 45×30 grid-graph ($n = 1350$ vertices, $m = 2625$ edges). Bottom: the North Carolina counties dual graph ($n = 100$, $m = 244$). By showing which x vectors have the highest and the lowest I values, we are exploring the meaning of Moran's I. High I values detect clustering, but minimal I values are less interpretable once we depart from the world of grids.

As we've seen, I maximizers tend to concentrate the group's population in a cluster, and this remains true on a grid or a real-world graph like the counties of North Carolina (Figure 5). On grid-graphs, for example, rectangular grids, the pattern minimizing I forms a checkerboard pattern. However, most dual graphs are not bipartite (i.e., admitting an alternating pattern). The eigenvectors of A with smallest eigenvalues may be hard to interpret, both on abstract graphs [9] and on the irregular graph structures that come up in practice.[8]

For our part, we conclude that it's not advisable to read too much into negative I values—which in any case are very rare in practice for real-world demographic data. And, more problematically for overall usability, we have no good plain-English interpretation for intermediate values of I across graphs. That is, what does an

[7] Translating this into the related language of Fourier theory, the patterns that correspond to large eigenvalues have low frequency, so they may just have one negative area and one positive area, while the ones for the low eigenvalues have a high frequency, which may correspond to fast oscillation from positive to negative. Think of this as being like a sinusoid function that takes a long time to complete a period (low frequency) versus another that oscillates rapidly (high frequency).

[8] We discuss the extremization problem further, and consider replacing A with the row-standardized adjacency matrix P or with the Laplacian L, in [10].

I = .6 residential pattern in one city or state have in common with an I = .6 pattern in another?

SO... WHAT IS SEGREGATION?

This explainer has explored perhaps the two most prominent metrics in the social science literature for measuring clustering/segregation. Both are very widely used, with *D* appearing much more commonly in cross-city comparisons in the popular press[9] and I in technical work in GIS and in fields as diverse as epidemiology, urban planning, and environmental studies [1].

As for the latter, we should look at how people actually use Moran's I in the social science literature. In the examples we have found, authors usually apply a kind of significance testing for I to see if the answer is larger than you should expect [11, 12, 13, 14]. That is, does the observed demographic data score as being more segregated than would be expected under a "null model," where values are distributed at random according to a normal or some other distribution? By comparing to randomized values on the same fixed graph, this kind of inference controls for the role of graph connectivity. We explore these themes further in Duchin et al. [10].

Are there other ways to measure segregation? Of course! Many mathy readers are probably itching to play around with or replace the definition entirely, such as by using an idea like the probability that your neighbor belongs to your own group—this family of ideas is called *assortativity* in network science, and it plays out a bit differently than the two we saw here [15]. But we hope that this brief intro models a few good practices: First, when it comes to metrics, math might give you insight into how best to use the scores (and what to avoid!) and whether their meaning is stable across contexts. Also, and crucially, to do good interdisciplinary work you must engage the literature in other fields than your home discipline, rather than thinking you're painting on a blank canvas.

As for our motivating question, "what is segregation?", we think that looking hard at the notions picked out by different metrics shows us that our shared intuitions don't fully specify an answer. So there is still both conceptual and measurement work to be done!

ACKNOWLEDGMENTS

Thanks to Heather Rosenfeld and Thomas Weighill for enormously helpful research collaborations on this material, and to Gabe Schoenbach for creating notebooks—available in our GitHub—to explore the scores and data.

[9] For just a few recent examples, check out two posts from the data blog 538 on diversity vs. segregation and partisan dissimilarity.

REFERENCES

[1] A. Getis. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis*, 40(3):297–309, 2008.

[2] A.D. Cliff and J.K. Ord. The problem of spatial autocorrelation. *Studies in Regional Science London Papers in Regional Science*, page 25–55, 1969.

[3] A. D. Cliff and Ord. J.K. *Spatial Autocorrelation*. 1973.

[4] William H. Frey and Dowell Myers. Racial segregation in U.S. metropolitan areas and cities, 1990–2000: Patterns, trends, and explanations. *University of Michigan Population Studies Center Research Report*, 2005.

[5] Douglas S. Massey and Nancy A. Denton. The dimensions of residential segregation. *Social Forces*, 67(2):281–315, 1988.

[6] P.A.P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

[7] M. Tiefelsdorf and B. Boots. The exact distribution of Moran's I. *Environment and Planning A*, 27(6):985–999, 1995.

[8] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[9] M. Desai and V. Rao. A characterization of the smallest eigenvalue of a graph. *Journal of Graph Theory*, 18(2):181–194, 1994.

[10] M. Duchin, J.M. Murphy, and T. Weighill. Measures of segregation and analysis on graphs. *Preprint*, 2021.

[11] P. De Jong, C. Sprenger, and F. Van Veen. On extreme values of Moran's I and Geary's c. *Geographical Analysis*, 16(1):17–24, 1984.

[12] L. Anselin. The Moran scatterplot as an ESDA tool to assess local instability in spatial association. volume 4, page 111. CRC Press, 1996.

[13] T. J. Barnes. Spatial analysis. *The Sage Handbook of Geographical Knowledge edited by John Agnew and David Livingstone*, pages 380–391, 2011.

[14] B. J. L. Berry and D. F. Marble. *Spatial analysis: a reader in statistical geography*. Prentice-Hall, 1968.

[15] E. Alvarez, M. Duchin, E. Meike, and M. Mueller. Clustering propensity: A mathematical framework for measuring segregation. *Preprint*, 2018.