Optimal Transport as a Metric for Measuring Segregation

Daniel Peng, James M. Murphy
Department of Mathematics
Tufts University

***Abstract*—In this paper, we propose the *OTSeg* score, a new quantitative metric that utilizes optimal transport to measure the degree of segregation on graphs. We characterize maximizers and minimizers for the metric and consider its relationship to classical measures of segregation including those based on spectral graph theory. We validate the effectiveness of OTSeg through experiments on synthetic and real datasets derived from American political geography, highlighting its potential for applications in computational social science.**

I. Introduction

Segregation refers to the uneven distribution of groups or individuals across spatial regions, often along demographic, economic, or cultural lines. Understanding and quantifying the extent of segregation is a major step in studying its causes and consequences. The examination often involves analyzing spatial networks that model geographic units, which could be counties, Census tracts, or households [1]. A range of methods have been proposed to quantify segregation, ranging from statistical indices to spatial metrics. Two of the most widely used methods are the Gini index [2]—originally developed to quantify income inequality—and Moran's I [3]—a measure of spatial autocorrelation that captures the degree to which similar values (e.g., demographic proportions) are clustered in a spatial network.

A. Summary of Contributions & Notation

In this paper, we propose a novel approach to measuring segregation using *optimal transport on graphs*. Our index, the OTSeg cost, uses optimal transport to analyze segregation directly on graph-structured data. We establish its mathematical properties, including a characterization of maximizers and minimizers, and perform a comparison between OTSeg and Moran's I. Our analysis demonstrates that OTSeg captures patterns and disparities in population distributions that are often overlooked by Moran's I. For instance, the values of Moran's I are heavily influenced by the underlying graph topology and the spatial weight matrix. This dependency complicates the interpretation of intermediate scores (e.g., those lying near 0.5) and makes cross-locality comparisons challenging without careful normalization [4]. In contrast, OTSeg inherently incorporates transport costs and spatial distances, allowing it to capture disparities of demographic distribution directly at a finer level. Moreover, it is robust across different graph topologies as it does not depend on arbitrary choices of spatial weight matrices, making it more suitable for consistent cross-locality comparisons.

II. Background

A. Existing Segregation Metrics

One of the most commonly-used metrics to quantify inequality is the Gini index [2].

Definition II.1. *Let $p \in [0,1]$ denote the cumulative proportion of the population and let $L(p) \in [0,1]$ be the* Lorenz curve*, denoting the cumulative proportion of the quantity held by the bottom p proportion of the population. The* Gini index *is $G := 2\int_0^1 (p - L(p))\,dp$.*

One major limitation of the Gini index is that it can produce the same value for very different income distributions [5]. Indeed, since the Gini index summarizes the area between the Lorenz curve, $p \mapsto L(p)$, and the equality line, $p \mapsto p$, into a single number, it often fails to capture detailed information about the specific shape of inequality within a society.

Another common metric for measuring segregation is Moran's I, a spatial autocorrelation metric.

Definition II.2. *Let $W \in \mathbb{R}^{n\times n}$ be a nonzero spatial weight matrix. Then* Moran's I *with respect to W is defined as $I = \frac{n}{w}\frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ where $w = \sum_{i=1}^n \sum_{j=1}^n |W_{ij}|$ and x is the variable of interest.*

A checkerboard distribution, where adjacent units have alternating values for x, would yield a Moran's I value near -1, indicating dissimilarity between neighboring units as defined by W. In contrast, a distribution where neighboring units have the same value (such as a region of all 0s next to a region of all 1s) would yield a Moran's I value near 1, indicating similarity between neighboring units. Moran's I has different bounds on different graphs and can be arbitrarily large or small [4]. Finding the true maximizer and minimizer of Moran's I can be achieved by solving the generalized eigenvalue problem associated to (CWC, C)

where $C = I - \frac{1}{n}11^T$. [4]. Using these extremal values, we can normalize Morans's I by simply dividing by $\max\{\max(I), |\min(I)|\}$.

B. Transport Costs

Our new segregation score utilizes transport costs, defined below.

Definition II.3. *Let P and Q be distributions on a set of elements X and Y, respectively, each with n elements. Let $\Pi(P,Q)$ be the set of couplings between P and Q, namely matrices $\pi \in \mathbb{R}^{n\times n}_{\geq 0}$ such that:*

$$\pi \mathbb{1} = Q, \quad \pi^T \mathbb{1} = P,$$

where $\mathbb{1}$ denotes the n-dimensional column vector of ones. Let $D \in \mathbb{R}^{n\times n}$ denote a (symmetric) cost matrix where each entry D_{ij} represents the cost of transferring mass from the i-th element in the support of P to the j-th element in the support of Q. Then, the transport cost between P and Q is

$$\mathcal{C}(P,Q) := \min_{\pi \in \Pi(P,Q)} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}\pi_{ij}.$$

III. OTSeg

We introduce a new measurement, the OTSeg score, which uses the transport cost to quantify the extent of segregation within a population. The idea is grounded in comparing an arbitrary distribution P with a reference that represents the absence of segregation—the uniform distribution. In this framework, the farther P is from the uniform distribution, the higher the OTSeg score, signifying greater segregation.

Definition III.1. *Let X be a set of n elements. Let $P, U \in \mathbb{R}^{n\times 1}$ be distributions on X with U the uniform distribution. Then,*

$$OTSeg(P) := \frac{\mathcal{C}(P,U)}{d_{max}},$$

where

$$d_{max} := \max_{1\leq i\leq n} \sum_{j=1}^{n} \frac{D_{ij}}{n}.$$

A. Extrema of OTSeg

It is useful to characterize the maximizers and minimizers of segregation metrics in order to make meaningful comparisons and inferences.

The minimizer is trivial: if P is the uniform distribution on X, there is no need to move any mass from P to U resulting in an OTSeg score of 0.

Intuitively, a maximizer corresponds to a distribution P that concentrates all of its mass at a single element of X that is on average farthest from all other nodes. We formalize this intuition by first considering the case where P is supported on a single element in the following lemma.

Lemma III.2. *Let X be a set of n elements. Let $P, U \in \mathbb{R}^{n\times 1}$, where U is the uniform distribution on X and P is a distribution on X that has support size 1. Then, $\mathcal{C}(P,U) \leq d_{max}$.*

Proof. Let P be a Dirac measure centered at x_i. The only coupling from U to P is $\pi^* \in \mathbb{R}^{n\times n}$ with:

$$\pi^*_{jk} = \begin{cases} \frac{1}{n}, & \text{if } j = i, \\ 0, & \text{if } j \neq i. \end{cases}$$

The transport cost is then:

$$\begin{aligned} \mathcal{C}(P,U) &= \min_{\pi \in \Pi(P,U)} \sum_{j=1}^{n} \sum_{k=1}^{n} \pi_{jk} D_{jk}. \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} \pi^*_{jk} D_{jk} \\ &= \sum_{k=1}^{n} \pi^*_{ik} D_{ik} = \sum_{k=1}^{n} \frac{D_{ik}}{n}. \end{aligned}$$

So, $\mathcal{C}(P,U)$ is the average distance of x_i to all other elements in X. Since d_{max} is defined as

$$\max_{1\leq i\leq n} \sum_{k=1}^{n} \frac{D_{ik}}{n},$$

it follows directly that $\mathcal{C}(P,U) \leq d_{max}$. □

Theorem III.3. *Let X be a set with n elements and let U denote the uniform distribution on X. A distribution P_{max} which maximizes $\mathcal{C}(P,U)$ is given by:*

$$(P_{max})_i = \begin{cases} 1, & \text{if } i \text{ maximizes } \frac{1}{n}\sum_{j=1}^{n} D_{ij}, \\ 0, & \text{otherwise}, \end{cases}$$

where if there are multiple indices i that achieve the maximum value, then any such index works. The cost associated to a maximizer distribution is d_{max}.

Proof. Let $P, U \in \mathbb{R}^{n\times 1}$, where U is the uniform distribution on X and P is a distribution on X. We claim $\mathcal{C}(P,U) \leq d_{max}$. We proceed by induction over $k \in \{1, \ldots, n\}$, the support size of P.

Base case: When $k = 1$, the result follows from Lemma III.2.

Inductive step: Assume that any distribution $R \in \mathbb{R}^{n\times 1}$ on X with support size $k \geq 2$ satisfies $\mathcal{C}(R,U) \leq d_{max}$. Let P be an arbitrary distribution on X with support size $k+1$. Let x_i be an element in the support of P, and let $p_i > 0$ denote the probability mass assigned to x_i by P. Define the Dirac measure δ_{x_i} centered at x_i, and let P' be the distribution obtained by removing x_i from the support of P and rescaling the remaining masses to ensure $\|P'\|_1 = 1$, i.e.,

$$P = p_i \delta_{x_i} + (1 - p_i)P'.$$

By convexity of the transport cost [6], we have:

$$\mathcal{C}(P,U) = C\big(p_i\delta_{x_i} + (1-p_i)P', U\big) \\ \le p_i\mathcal{C}(\delta_{x_i}, U) + (1-p_i)\mathcal{C}(P', U).$$

By Lemma III.2, we know that $\mathcal{C}(\delta_{x_i}, U) \le d_{max}$. By the inductive hypothesis, $\mathcal{C}(P', U) \le d_{max}$. Therefore:

$$\mathcal{C}(P,U) \le p_i d_{max} + (1-p_i)d_{max} = d_{max}.$$

Thus, any distribution P with support size $k \in \{1, \dots, n\}$ satisfies $\mathcal{C}(P,U) \le d_{max}$. Selecting x_i to be the node which realizes d_{max} is such that δ_{x_i} realizes this upper bound. □

IV. Numerical Experiments

A. Experiment Setup

To see and evaluate how OTSeg captures segregation trends, we compare this score with Moran's I using the 2010 National Census data [7]. Every state was separated into four major demographic categories according to the Census: Black, White, Asian, and Hispanic. Each Census tract was treated as a node and a demographic proportion was the distribution on that node. Every state was scaled to fit within a 1×1 box around the origin. Moran's I for each demographic was calculated using Definition II.2, and then scaled by $\max\{\max(I), |\min(I)|\}$. The spatial weight matrix W was constructed to reflect adjacency between Census tracts. Specifically, $W_{ij} = 1$ if tracts i and j share a boundary, and $W_{ij} = 0$ otherwise. OTSeg was calculated using Definition III.1 in a similar fashion, using the Euclidean distance between Census tracts as the distance function.

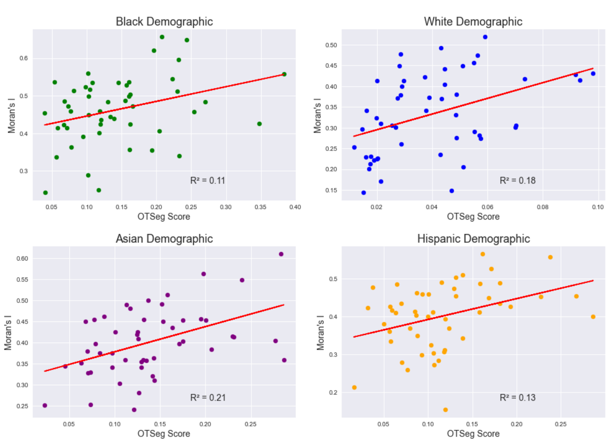

Fig. 1: Moran's I vs. OTSeg Scores by Demographic (Euclidean Distance)

These results are shown in Figure 1, which shows a relatively positive correlation between these two segregation metrics, with an r-squared value of around 0.2 for the four demographics. Varying the cost matrix to use the squared Euclidean distance instead gives a similar result, as shown in Figure 2, with an even higher r-squared value for the White demographic across all fifty States. Figure 3 shows the correlation between Moran's I and OTSeg for the Hispanic population within the U.S. An R-squared value of 0.16 indicates that both Moran's I and OTSeg capture some overlapping trends in the distribution of the Hispanic population. What is more interesting is the discrepancy between these two metrics.

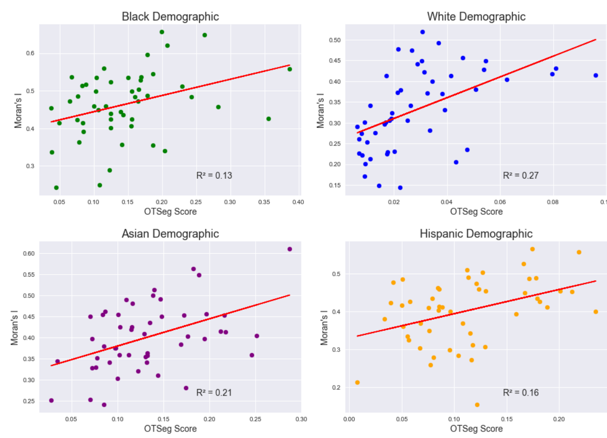

Fig. 2: Moran's I vs. OTSeg Scores by Demographic (Squared Euclidean Distance)

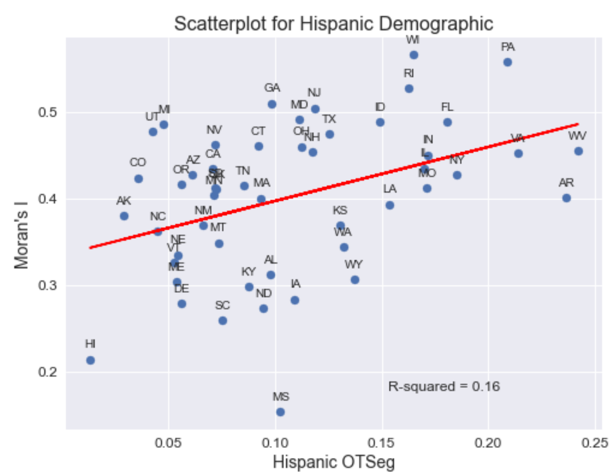

Fig. 3: Scatterplot of Moran's I vs. OTSeg Value of Hispanic Population in U.S. States

B. Discrepancies Between Moran's I and OTSeg

Figure 3 shows a clear outlier in Mississippi, whose Moran's I value is much lower with respect to its OTSeg score. This discrepancy might indicate that OTSeg captures different information regarding the geographical shape and spatial distribution of the graph as opposed to Moran's I. As shown in Figure 4, there are many high density nodes surrounded by nodes with lower Hispanic population proportion. It would make sense for Moran's

I to return a relatively low value for this distribution. Indeed, Moran's I is a measure of spatial autocorrelation, which evaluates the similarity of values at neighboring nodes. A distribution where nodes with a high Hispanic population proportion are surrounded by nodes with a low Hispanic population proportion indicates a pattern of negative spatial autocorrelation—since high and low values are clustered in opposing proximity, Moran's I will return a low or negative value. Figure 4 shows OTSeg's score on the same graph of the Mississippi Hispanic population. Notably, we can see large amounts of mass being moved from high-density clusters to low-density nodes across the state, resulting in the relatively high OTSeg score in Figure 3.

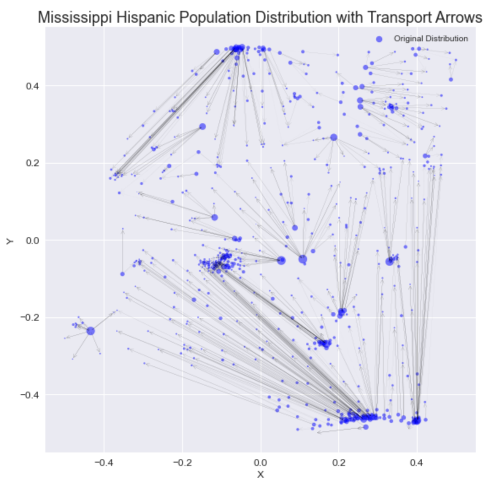

Fig. 4: OTSeg Mapping For the Hispanic Population in Mississippi

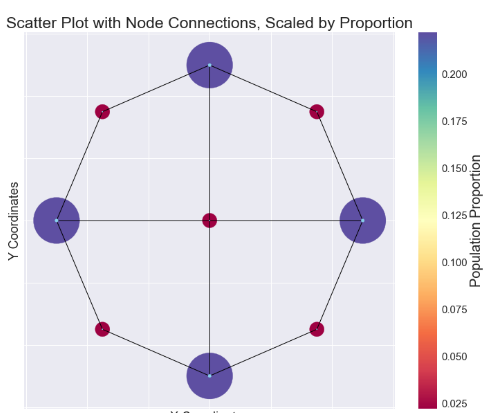

Fig. 5: 9-Node Graph With High-proportion Nodes Surrounded by Low-proportion Nodes

C. *Benefits of OTSeg over Moran's I*

Some further exploration into this discrepancy gives us Figure 5. Similar to the Figure 4, this 9-node graph has four nodes of high distribution surrounded by five nodes of low distribution. The normalized Moran's I value of this graph is -0.944. The OTSeg value, on the other hand, is 0.294, which is relatively large compared to the low Moran's I value. From a spatial autocorrelation view, it may seem that this distribution is relatively 'anti-segregated', which agrees with the large negative Moran's I value. However, from a geographical standpoint, this graph could also be seen as highly segregated, where a specific demographic is highly concentrated within specific nodes and nearly nonexistent elsewhere — this point of view is echoed by OTSeg's large value of 0.294. These nuances illustrate the complexity of spatial analysis, where these different metrics capture and offer contrasting perspectives on the same data set.

D. *Computational Complexity*

It is worth noting that finding the transport cost is computationally expensive, even when approximating with Sinkhorn's algorithm [8], which has essentially quadratic complexity in n. In practice, this becomes a serious concern as the granularity of the data is increasing. Although calculating OTSeg on every state at the Census tract level (approximately $84,414$ tracts) is relatively quick, moving to the Census block level (over 8 million blocks) significantly increases computational burden. To address this, it may be worthwhile to explore accelerated methods such as *meta optimal transport*, as we are repeatedly solving very similar transport problems within each U.S. state [9].

V. Conclusion and Future Research

In this work, we introduced OTSeg, a novel metric for quantifying segregation, leveraging optimal transport to evaluate population distributions directly on graph-structured data. We demonstrated the theoretical properties of OTSeg, including its ability to define maximizers and minimizers that correspond to extreme configurations of segregation. Furthermore, our experiments validated the utility of OTSeg in capturing subtle segregation trends that may go undetected by classical methods. These results highlight its potential to provide insights ranging from social science to urban planning and network analysis.

Future work could explore extending OTSeg to dynamic settings, where population distributions evolve over time, or to multiscale analyses that examine segregation patterns across varying spatial resolutions. Additionally, normalizing the population vector nodewise rather than globally may allow for a different perspective on measuring segregation with OTSeg.

VI. Acknowledgments

The authors acknowledge generous support from a Tufts Springboard grant.

References

[1] D. S. Massey, J. Rothwell, and T. Domina, "The changing bases of segregation in the United States," *Annals of the American Academy of Political and Social Science*, vol. 626, pp. 74–90, Nov 2009.

[2] C. Gini, *Memorie di metodologia statistica*. Rome: E.V. Veschi, 1955.

[3] P. A. P. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.

[4] M. Duchin, J. M. Murphy, and T. Weighill, "Measuring segregation via analysis on graphs," *SIAM Journal on Matrix Analysis and Applications*, vol. 44, no. 1, pp. 80–105, 2023.

[5] Y. Liu and J. L. Gastwirth, "On the capacity of the gini index to represent income distributions," *METRON*, vol. 78, no. 1, pp. 61–69, 2020.

[6] C. Villani, *Optimal Transport: Old and New*, vol. 338 of *Grundlehren der mathematischen Wissenschaften*. Berlin, Heidelberg: Springer-Verlag, 2009.

[7] D. R. Deford, "Census dual graphs," https://people.csail.mit.edu/ddeford/dual_graphs.html.

[8] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[9] B. Amos, G. Luise, S. Cohen, and I. Redko, "Meta optimal transport," in *International Conference on Machine Learning*, pp. 791–813, 2023.