Spatial Dimensions of Algorithmic Transparency: A Summary

Jayant Gupta
gupta423@umn.edu
Computer Science & Engineering
University of Minnesota
USA

Alexander Long
longx552@umn.edu
Computer Science & Engineering
University of Minnesota
USA

Corey Kewei Xu
kx16@my.fsu.edu
Askew School of Public
Administration and Policy
Florida State University
USA

Tian Tang
ttang4@fsu.edu
Askew School of Public
Administration and Policy
Florida State University
USA

Shashi Shekhar
shekhar@umn.edu
Computer Science & Engineering
University of Minnesota
USA

ABSTRACT

Spatial data brings an important dimension to AI's quest for algorithmic transparency. For example, data driven computer-aided policy-decisions use measures of segregation (e.g., dissimilarity index) or income-inequality (e.g., Gini index), and these measures are affected by space partitioning choice. This may lead policymakers to underestimate the level of inequality or segregation within a region. The problem stems from the fact that many segregation based analyses use aggregated census data but do not report result sensitivity to choice of spatial partitioning (e.g., census block, tract). Beyond the well-known Modifiable Areal Unit Problem, this paper shows (via mathematical proofs as well as case studies with census data and census based synthetic micro-population data) that values of many measures (e.g., Gini index, dissimilarity index) diminish monotonically with increasing spatial-unit size in a hierarchical space partitioning (e.g., block, block-group, tract), however the ranking based on spatially aggregated measures remain sensitive to the scale of spatial partitions (e.g., block, block group). This paper highlights the need for social scientists to report how rankings of inequality are affected by the choice of spatial partitions.

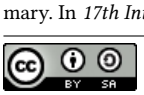
CCS CONCEPTS

• **Computing methodologies → Spatial and physical reasoning**.

KEYWORDS

Fairness, Accountability, and Transparency, Public policy, Spatial data science, Urban Planning

ACM Reference Format:
Jayant Gupta, Alexander Long, Corey Kewei Xu, Tian Tang, and Shashi Shekhar. 2021. Spatial Dimensions of Algorithmic Transparency: A Summary. In *17th International Symposium on Spatial and Temporal Databases (SSTD '21), August 23–25, 2021, virtual, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3469830.3470898



SSTD '21, August 23–25, 2021, virtual, USA

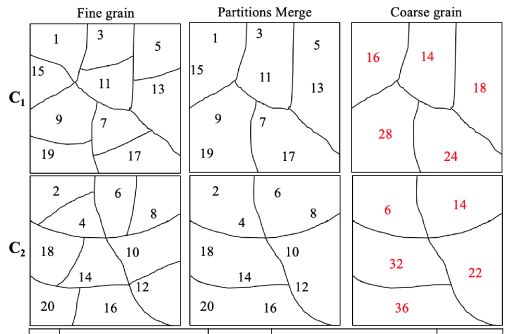
ACM ISBN 978-1-4503-8425-4/21/08.
https://doi.org/10.1145/3469830.3470898

1 INTRODUCTION

Spatial data raises special concerns for algorithmic transparency, which calls for the factors influencing algorithmic decisions to be made visible to users, regulators, policy makers, and people affected. Computational operations (e.g., partitioning, ranking) which may not affect the empirical analysis of non-spatial datasets can result in empirically inconsistent spatial analysis. These operations are often used for quantifying societal inequality, segregation, etc., to be used by policymakers to assess and develop relevant policies. Currently, issues such as reducing global inequality is part of the UN Sustainability Development Goals (SDGs) [1] and a lack of transparency in assessment may lead to inaccurate conclusions with global ramifications. For example, measures of segregation and income-inequality (e.g., dissimilarity index, Gini index) are affected by the choice of spatial partitioning.

Most studies in the US, including Richard Florida and Charlotte Millander's notable book *Segregated City* [9], are based on data aggregated from the US Census. However, these works often do not report the sensitivity of their results to the spatial partitioning (e.g., census tract, census block group). This raises questions about the validity of the findings. Underestimation (or overestimation) of income inequality can have implications for a region as policymakers rely on geographically aggregated data to assess subsidies [16], health insurance policies [19], compute spatial risk adjustment [33], etc. Therefore, knowing the sensitivity of inequality values at different scales is useful to account for possible errors and assess the need for additional data collection (e.g., random surveys) to reduce the error.

Given a space partitioning, our goal in this work is to characterize the sensitivity (e.g., change in rankings) of income inequality and segregation measures computed on spatially aggregated data to the choice of spatial unit. We illustrate the problem with the following example. Figure 1 shows two cities (say C_1 and C_2) having 10 partitions each, where each partition has some income shown within them. We can use these income values for ranking the two cities based on their income inequality (e.g., Gini Index [11]). When the income inequality is computed using fine-grained partitions, city C_1



City	Fine grain Inequality Rank	Gini Index	Coarse Grain Inequality Rank	Gini Index
C_1	1	0.33	2	0.144
C_2	2	0.3	1	0.283

Figure 1: Example of inconsistent ranking at different scales.

is ranked higher than city C_2. By contrast, when the partitions are merged and income inequality is computed using coarse-grained partitions, city C_2 is ranked higher than city C_1. The Gini index values are shown in the table at the bottom of Figure 1.

This problem is challenging for the following reasons. First, it is computationally expensive as the number of possible partitions is exponential in the number of base units (e.g., households). Second, we lack mathematical characterization to determine the sensitivity of various inequality or segregation measures to the unit of analysis. Third, experimental validation of a mathematical characterization is difficult as collecting household data is expensive and may undermine confidentiality.

The change in values due to scale and zone (or shape of the partition) is also known as the modifiable areal unit problem (MAUP). First discussed in geography [20], MAUP, has also been applied to socioeconomic models. Two studies found that MAUP significantly affected multivariate parameter estimates of census tract employment [23] and mean family income models [10]. Studies on inequality metric sensitivity to MAUP have also been done. In [4], the authors showed that Gini moves in opposition to scale with significant variation attributable to partition shape. Other inequality measures have been analyzed to assess the zone effect [21]. Potential solutions to MAUP have included the use of T-communities [12], which are realistic, highly segregated homogeneous communities and can be a good proxy for the unit of analysis. However, to date no studies have provided theoretical tools for managing the effect of space partitioning on measures of inequality.

This work goes beyond MAUP in the following way. We study the theoretical sensitivity and behavior of two income inequality measures (Gini index and the income quintile share ratio (IQSR)) and one segregation measure (index of dissimilarity). Our mathematical proofs as well as case studies with census data and census based synthetic micro-population data show that values of all the measures in this study diminish as the scale of partitioning increases. We also provide theoretical bounds on IQSR. Through our results we find that distortion is reduced by smaller spatial units, however the ranking remains sensitive to the scale of spatial partitions (e.g., block, block group).

Contributions:
Our main contributions in this paper are as follows:

- We highlight the spatial dimension of algorithmic transparency with findings that reinforce the need to report the sensitivity of results to the choice of spatial partitioning.
- Beyond MAUP, we show theoretically that some measures (e.g., gini index, index of dissimilarity) of inequality decrease monotonically with increasing scale of analysis. For example, values calculated at the census tract level are always higher than the values calculated at the census block level. Distortion is reduced by using the smallest possible units.
- We provide formal proofs on the upper and lower bound of the IQSR.
- A case study on various income inequality measures (Gini Index, IQSR, Theil, and Atkinson) using a synthetic household level dataset supports the theoretical results on the Gini index and IQSR.
- A case study on multi-scale ranking based on the index of dissimilarity using a 2010 American Community Survey dataset. It supports the theoretical results on the index of dissimilarity and re-enforces the need to report the sensitivity of results to the choice of spatial partitioning.
- Our findings have broad implications for equity related policymaking (e.g. tax policies, gentrification, education programs). Policymakers should be cautious about the unit of analysis when using existing equity indexes or conduct sensitivity analysis with different spatial partitioning choices to provide more comprehensive evidence to inform policies that address equity issues.

Symposium Relevance: The work in this paper aligns with the call for contributions, particularly to the topic of Fairness, Accountability, and Transparency. The paper provides a perspective on the transparency of spatial analysis showing its implications for policymakers. The case studies on fine-scale census and census-like dataset show how space partitioning can affect the understanding of societal inequality and segregation.
Scope: We limit our study to mathematical analysis of an entropy based inequality measure (Gini index), a ratio-based inequality measure (IQSR), and a ratio-based segregation measure (Index of Dissimilarity) and the case studies for validation. No computational methods are proposed to measure transparency. Our analysis is limited to the effect of changes in scale and zone; other relevant issues related to spatial aggregation such as boundary effects are not considered. Further, there may be situations where the base data is given but we can consider some form of data clustering and then form the partitions. However, here we assume the partitioning is given and cannot be altered, which is the case for most studies based on census-type data.
Organization: The paper is organized in the following manner. Section 2 gives describes the application context. Section 3 presents the problem formulation with illustration. Section 4 provides the mathematical results of monotonic trend on the Gini Index, IQSR, and Index of dissimilarity and bounds on IQSR. We provide a case study on ranking US Metropolitan regions using ACS 2010 data in Section 5. In Section 6, we describe the analytical evaluation procedure and results. Section 7 discusses the spatial dimensions of algorithmic transparency in policy-making and provides a brief

review of other related work. Finally, we conclude and consider future work in Section 8.

2 APPLICATION CONTEXT

Inequality has been found to be associated with a variety of social problems. In *Economic Growth and Income Inequality* [15], Kuznets suggested that a higher level of inequality is correlated with a lower level of democracy or property rights, less redistribution, and a higher level of ethnic heterogeneity. For developing countries, high inequality threatens to stall future progress against poverty by attenuating growth prospects, and the gap of living standards between the rich and the poor is still increasing [24].

The study on inequality is typically about resource distribution among different groups of population. The Gini index and the dissimilarity index have been widely applied in policy analysis to understand the extent of income inequality. A major contribution of this paper is to illustrate the sensitivity of the dissimilarity index, to spatial partitioning and its policy implications. We applied the index of dissimilarity to study income inequality and racial segregation. The index of dissimilarity has also been applied to study inequalities in many other types. For example, Kangkang Tong et al. (2021) [32] use disparity ratio to study energy consumption inequality between high and low income group (bottom 20% income households and top 20% income households). Song et al. (2013) [30] uses index of dissimilarity to study land mix.

As individual level data is rarely available to either preserve household confidentiality or due to lack of resources [13, 31], scholarly work primarily relies on several levels of aggregated data from the U.S. Census Bureau. Most researchers use data from the census tract level.

Previous studies using aggregated data have found that income inequality and social segregation have compound social effects on vulnerable populations. Although city populations consist of people with diverse backgrounds, they are becoming more and more segregated and homogeneous by income, education, occupation [9]. Bishop [3] refers to the phenomenon as "the big sort". Segregation has gradually built up through the process of "filtering" in the residential housing market, whereby well-off households tend to move to newly developed, high priced communities, leaving older communities to households with lower income levels [17]. Over time, low income racial minorities and high income racial majorities become geo-spatially segregated across the city [25]. As a result, low income residents not only suffer from lack of financial resources, but also from related neighborhood effects such as high rate of crime, pollution and chronic disease. These challenges inevitably interact with each other and have cumulative effects that prevent disadvantaged groups from changing their status [28].

Although the narratives on inequality and social segregation have been widely accepted in academia, the majority of these studies are built on highly aggregated data. We demonstrate in the following sections that variation in the unit of analysis will dramatically change the score of inequality (segregation) index.

3 PROBLEM FORMULATION

We have formulated the problem as two sub-problems corresponding to the "scale" problem and the "zone" problem of MAUP.

3.1 Assessment of the Scale Problem:

Input:

- Geo-located census records (X).
- Administratively-defined set of hierarchical spatial partitions (e.g., census tracts, census block groups) as shape-files ($\mathcal{P}$), where, $P \in \mathcal{P}$ is a space partitioning. Further, $\forall P_1, P_2 \in \mathcal{P}$, P_1 is hierarchically higher than P_2, if all the partitions $p_{1j} \in P_1$ are greater in size than $p_{2j} \in P_2$.
- Existing inequality measures (e.g., index of dissimilarity) denoted as M.

Output:

- Partitions' (say p_{ij}) rankings using metric M, where M is computed for all hierarchies lower than P_i.

Objective:

- Assess ranking sensitivity to the partition scale (e.g., tract to block group).

Constraints:

- Spatial partitions are rigid at each hierarchical level.
- The smallest spatial unit of the census data.

Figure 1 illustrates the input and output of the scale problem, where $X = 1, 2, ...20$, $\mathcal{P}$ has fine grain, coarse grain partitions, and gini index is the inequality measure M. The output is a set of rankings corresponding to each partition scale, as shown in the table at the bottom of the figure.

3.2 Assessment of the Zone Problem:

Input:

- Geolocated census records (X).
- Population limit of each zone (n).
- Zone generator (Z), where $Z(X, n) \rightarrow P_i$.
- Existing income inequality measures (e.g., Gini, IQSR) as M.

Output:

- Upper and lower bound of the measure (M) for a given partition scale governed by the parameter n.

Objective:

- Assess the sensitivity of inequality measurements to changes in partition zone (e.g., census tracts to county subdivisions).

Constraints:

- Random spatial partitions are subject to a population limit.
- The smallest spatial unit of the census data.

Figure 2 illustrates the the zone problem. As shown, census records are given at a fine grain scale which are inputs to a zone generator with population (or record) limit of 2. The figure shows two of the possible partitioning P_1, P_2, where $Z(C_1, 2) \rightarrow P_1, P_2$. The partitioning merge to form the coarse grain partitions each with different gini index (M) values (shown in the bottom left table of the figure). The potential output would be the lower-bound (LB) and upper-bound (UB) of the measure for all the possible partitioning at this scale.

4 MATHEMATICAL CHARACTERIZATION

In this section, we characterize results on two types of inequality measures and one type of segregation measure: an entropy based

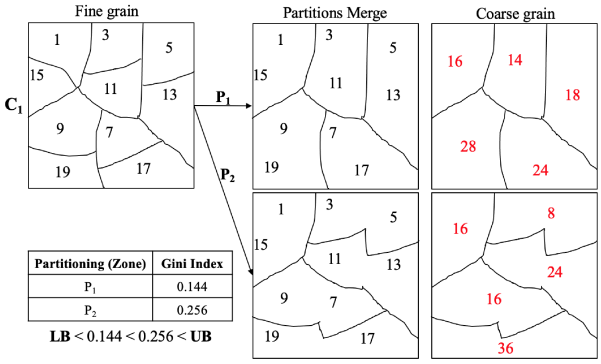


Figure 2: Illustration of the zone problem.

inequality measure (Gini index), a ratio-based inequality measure (IQSR), and a ratio-based segregation measure (Index of Dissimilarity). Three of our results are related to the assessment of the scale problem for Gini (Theorem 4.1), IQSR (Theorem 4.2), and Index of dissimilarity (Theorem 4.4). The remaining result is related to the assessment of the zone problem for IQSR (Theorem 4.3). For readability, the theorem proofs are moved to Appendix A.

Gini Index: The Gini coefficient is a measure of inequality developed by Corrado Gini in 1912 as an extension of work by Max Lorenz [11]. Assuming x_i is the income value for a population of size N. Gini can be calculated by taking the absolute sum of all the values normalized by a factor of the average value as shown below,

$$G_N = \frac{1}{2N^2\bar{x}} \sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|,$$

where $\bar{x}$ is the average income of the population. Theorem 4.1 gives the relation between the aggregated Gini Index ($\bar{G}_N$) and the non-aggregated Gini Index (G_N).

Aggregated Gini Index: $\bar{G}_N = \frac{1}{2N^2\bar{x}} \sum_{i=1}^{N} \sum_{j=1}^{N} |\bar{x}_i - \bar{x}_j|$, where $\bar{x}_i$ is the average income of the partition which contains the i^{th} person.

Theorem 4.1. *Gini produces a lower-bound estimate of overall inequality when calculated on group averages i.e.,* $\bar{G}_N <= G_N$.

Proof. Proof in Appendix A. □

Income Quintile Share Ratio (IQSR): IQSR is the ratio of the total income received by the highest 20% of income earners and income received by the lowest 20% of income earners.

Theorem 4.2 gives the relation between this metric calculated on aggregated units ($IQSR(A_P)$) to the metric calculated without aggregation (IQSR(X)). This result focuses on a special case where all partitions have the same size. Theorem 4.3 provides bounds on IQSR in the general case i.e., arbitrary space partitioning.

Theorem 4.2. *$IQSR(A_P) \leq IQSR(X)$, where X is a set of numbers with equi-cardinality partitioning, $P = \{p_1, p_2, ..., p_{C(P)}\}$, where p_i's are pairwise disjoint and their union yields X and $A_P = \{Avg(p_1), Avg(p_2), ..., Avg(p_{C(P)})\}$, where $Avg(p_i)$ is the arithmetic average of items in partition p_i, and $C(P)$ is the cardinality of P.*

Proof. Proof in Appendix A. □

Theorem 4.3. *IQSR(X) is lower bounded by the sum of $C(X)/5$ smallest values divided by the sum of $C(X)/5$ largest values and upper bounded by the sum of $C(X)/5$ largest values divided by the sum of $C(X)/5$ smallest values, where $C(X)$ is the cardinality of the given set X of incomes.*

Proof. Proof in Appendix A. □

Index of Dissimilarity (D): The index of dissimilarity [18] compares the distribution of a selected group of people (say X) with all others in that location (say Y). The more evenly distributed a group is compared to the rest of the population, the lower is the level of segregation. The index value range from 0 to 1, where 0 reflects no segregation and 1 reflects complete segregation. The dissimilarity index D can be expressed as follows,

$$D = \frac{1}{2} \sum_{i=1}^{N} \| \frac{x_i}{X} - \frac{y_i}{Y} \|,$$

where x_i is the number of individuals in the selected group in sub-partition i, X is the number of selected groups in the whole partition, y_i is the number of "others" in the sub-partition, and Y is the corresponding number in the whole partition. N is the total number of sub-partitions.

Assume that the sub-partitions are aggregated to $\bar{N}$ partitions and $\bar{D}$ is the index of dissimilarity for the new set of partitions as follows,

$$\bar{D} = \frac{1}{2} \sum_{i=1}^{\bar{N}} \| \frac{x_i}{X} - \frac{y_i}{Y} \|.$$

Then, the following theorem holds.

Theorem 4.4. *The index of dissimilarity diminishes as the scale of aggregation increases i.e., $\bar{D} \leq D$.*

Proof. Proof in Appendix A. □

The results show that the value of all the measures decrease monotonically with an increase in scale of the spatial unit. In addition, one of the results (Theorem 4.3) gives mathematical bounds as a function of the partition with the smallest and the largest sum of values. The results can help determine possible distortion for a set of partitions.

5 CASE STUDY: MULTI-SCALE RANKING BASED ON SEGREGATION INDEX

We used the index of dissimilarity (IOD) to conduct a small study on wealth segregation in the US at two different levels of analysis. We then compared our results with results from similar work published in the well-received book *Segregated City: The Geography of Economic Segregation in America's Metros* by Richard Florida and Charlotte Mellander [9].

Dataset: For this study we used 2010 American Community Survey (ACS) 5-Year Estimate data (2006-2010) [5] available in TIGER/Line shapefile format and metropolitan and micropolitan statistical areas (MMSA) shapefile [6]. The ACS publishes small area data using survey responses pooled over 5 years. The MMSA shapefile contains a simplified representation of 955 micropolitan and metropolitan regions.

We used the ACS table of household income data ($B19001E$), which reports the number of households divided into 16 income brackets for a given geographical unit [5]. Here, the smallest geographical unit was a census block group. Income data and the

boundary shapefiles were processed (Section 5.1) to calculate binned household income data across 366 metropolitan statistical areas at two scales: block group and census tract. Data for micropolitan areas was removed to be consistent with the methodology of the *Segregated City* study [9]. We took the *Segregated City* data for our study from several tables or "exhibits" in the book. The *Segregated City* results were based on census data aggregated at the census tract level.

Ranking: The binned income data was used to compute the total number of households in two categories, wealthy and non-wealthy, for each geographical unit. The two categories were used to rank the metropolitan areas based on their index of dissimilarity representing segregation of the wealthy. The ranking was limited to a single index of segregation simply to highlight the need for reporting rankings at multiple geographical scales.

Difference in ranking: The data analysis at two different scales (i.e., block group and census tracts) generated two sets of rankings. The rankings were then used to calculate the changes in ranking across the two scales.

5.1 Data processing

The ACS data files were divided by state. The initial processing involved reading the file for each state and dropping non-relevant tables. The filtered files were then merged to get the nationwide data. This data was spatially joined with MMSA shapefiles using the HAVE_THEIR_CENTER_IN spatial operation using ArcGIS Pro. The resulting file had the census data mapped to census blocks, tracts, counties, metropolitan areas and states.

Data aggregation: At the block group level, no aggregation was required to compute the index values. At the census tract level, household values were aggregated to compute the indexes.

Group definitions A wealthy household was defined as having an annual income greater than $200,000 which was calculated using column *B19001E17* or its aggregate at the census tract level. A non-wealthy household was the complement to a wealthy household and was calculated by subtracting it from the total number of households (Column *B19001E1*) in the corresponding spatial unit.

5.2 Results

Figure 3 shows US Metropolitan regions color coded based on their level of segregation. Dark blue indicates the highest levels of segregation, and yellow indicates the lowest levels. The levels were derived from their rankings based on the index of dissimilarity. The rankings were calculated at two different scales as described earlier. We find that there was a significant shift in ranking for the metropolitan regions when the analysis was done at the block level compared to the census tract level. The shift becomes more evident when the results are observed in a tabular form.

Table 1 shows three rankings of the ten most wealth segregated metro areas in the US with their IOD values. The first ranking was taken from Exhibit 2.2 (p. 18) in *Segregated City* [9] calculated at the 'census tract' level. The second and third rankings are the results of our case-study calculated at the census tract and block group level respectively. A comparison of the first and second rankings shows a high overlap of results, which indicates the use of similar methodology. However, there is only a 20% overlap between the

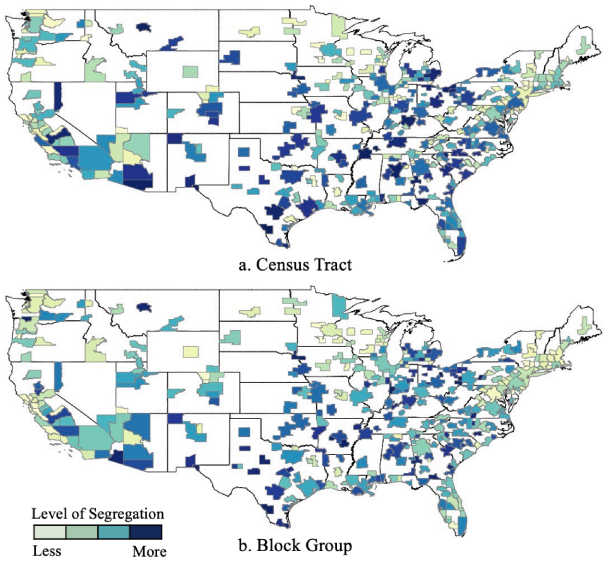


Figure 3: US Metropolitan regions color coded based on their lvel of segregation of the wealthy. Dark blue color indicates higher segregation whereas the pale Yellow color indicates low segregation. (Best in color)

second and third rankings (highlighted in bold). This indicates that the rankings are highly susceptible to the base spatial-unit of analysis. Also, IOD values at the census block level (3^{rd} ranking) are always higher than IOD values at the census tract level in accordance with our theoretical results (Theorem 4.4). Table 2 shows

Table 1: Three rankings of the ten most wealth segregated metro areas in the US with their spatial unit of analysis.

Rank	Census Tract	IOD	Census Tract	IOD	Census Block	IOD
1	**Laredo, TX**	0.646	**Laredo, TX**	0.646	Mansfield, OH	0.786
2	Jackson, TN	0.617	Jackson, TN	0.617	Wheeling, WV	0.752
3	**El Paso, TX**	0.611	**El Paso, TX**	0.612	**El Paso, TX**	0.738
4	Great Falls, MT	0.601	Great Falls, MT	0.601	Terre Haute, IN	0.732
5	Memphis, TN-MS-AR	0.582	Memphis, TN-MS-AR	0.582	**Laredo, TX**	0.730
6	Tucson, AZ	0.581	Tucson, AZ	0.581	Longview, TX	0.725
7	Columbus, GA-AL	0.578	Columbus, GA-AL	0.578	Sioux City, IA	0.712
8	Birmingham-Hoover, AL	0.576	Birmingham-Hoover, AL	0.576	Pine Bluff, AR	0.710
9	Louisville-Jefferson County, KY-IN	0.575	Louisville-Jefferson County, KY-IN	0.575	Steubenville-Weirton, WV-OH	0.701
10	San Antonio, TX	0.567	San Antonio, TX	0.567	Valdosta, GA	0.700

three rankings of large metropolitan areas (i.e., over one million people) having the highest level of wealth segregation by IOD score. Again, the first ranking is from *Segregated City* (Exhibit 2.1, p. 18) [9] and based on census tract level analysis. The second and third rankings are the results of our case-study calculated at the census tract and block group level respectively.

A high overlap of the first and second rankings indicates a similarity of methodology. More importantly, a comparison of the second and third rankings shows a significant empirical change in the rankings when the unit of analysis changes. For example, the Memphis region was ranked 5 based on the analysis at the census tract level but ranked 54 at the census block group level. Such changes can significantly impact the interpretation of rankings across the scale.

It is critical to note that even if there is a decline in the rankings at the block group level the corresponding IOD values are higher. Thus, if policymakers solely rely on rankings for interpretation

Table 2: Three rankings of ten large metros. The metros are the 10 most wealth segregated large metro regions taken from Exhibit 2.1 (p. 18) in *Segregated City* [9].

Large Metro	Rank	IOD	Rank	IOD	Rank	IOD
Memphis, TN-MS-AR	5	0.582	5	0.582	54	0.648
Birmingham-Hoover, AL	8	0.576	8	0.576	58	0.645
Louisville-Jefferson County, KY-IN	9	0.575	9	0.575	47	0.650
San Antonio-New Braunfels, TX	10	0.567	10	0.567	49	0.650
Cleveland-Elyria-Mentor, OH	13	0.560	12	0.561	97	0.624
Detroit-Warren-Livonia, MI	17	0.552	16	0.555	79	0.632
Nashville-Davidson-Murfreesboro-Franklin, TN	23	0.549	21	0.549	145	0.605
Columbus, OH	25	0.547	24	0.547	93	0.626
Charlotte-Gastonia-Rock Hill, NC-SC	29	0.541	29	0.541	151	0.603
Miami-Fort Lauderdale-Pompano Beach, FL	31	0.540	54	0.522	122	0.616

they may overlook the issue of segregation. Supporting IOD values may give a more realistic picture.

A particularly striking example concerns the city of Tallahassee, Florida (not shown in Table 2). Exhibit 3.2 (p. 23) in Richard Florida's book [9] ranked Tallahassee as the most segregated city in the U.S. on a measure of overall segregation. In our study it ranked 14 based on the segregation of the wealthy at the census tract level. And it ranked 92 based on the segregation of wealthy at the block group level. The IOD index in our study measures only one part of overall segregation. However, if a change in the scale of analysis changes the IOD value, then it must also change the ranking of overall segregation. As we will see next, this effect has many implications for policy-making.

5.3 Policy Implications

Our findings suggest that calculating a segregation index using different levels of spatial data may lead to significantly different conclusions about a city's segregation status and ranking. This can have broad social and political impacts on citizen satisfaction, turnover of public officials, and allocation of public resources. Richard Florida's report [9] on Tallahassee aroused intense discussion and debate among political leaders, city/county administration, and the public on whether the ranking of Tallahassee's economic segregation was fair [14]. The report investigated 350 metropolitan areas in the US and found that many small and medium sized cities suffered from high segregation because they were college towns and the university community was segregated from the service workers in the rest of the city.

Based on Florida and Mellander's calculation using census tract level data, Tallahassee, a mid-sized college town and the capital city of Florida, ranked as the most overall economically segregated metro in the US. City leaders refuted the report and argued for promoting the economic vitality of all communities. They called for a better way to measure income and economic segregation. The high ranking of economic segregation also made Tallahassee residents question the city's 30-year gentrification project for poor neighborhoods of the city. They also questioned whether the additional tax revenue that had been used to develop multi-use residential property in these neighborhoods should be shifted towards creating new job opportunities to reduce poverty and segregation. Our findings, however, show that when the finer block group level is used to calculate the segregation index, Tallahassee has relatively less wealth segregation than Richard Florida's study suggests. In fact, block group data may benefit larger cities even more due to an increased amount of aggregation at the census tract level.

6 CASE STUDY: MULTISCALE, MULTIZONE INCOME INEQUALITY ANALYSIS

In the second case study, we evaluated the effects of aggregation on income inequality statistics. First, we evaluated four different income inequality metrics at different scales. Then, we simulated different partitions constrained to a population limit. The purpose was to validate our mathematical characterization of inequality measures and to show the effect of MAUP on additional income inequality measures namely, Theil's index and Atkinson's index.

6.1 Data Sources

We used a highly granular synthetic dataset, generated from 2010 U.S. census income distributions at the block group level. The households were geo-located randomly within block groups while avoiding major bodies of water [34]. Location for each household was determined in two steps. First, each household was assigned to a block group based on a likelihood determined by 2010 block group median income. It was ensured that the block group lay within the household's Public Use Microdata Area (PUMA). Then, the assigned distribution was sampled to generate the final dataset [2].

In addition to geolocated income data, we used the standard hierarchy of spatial partitions from the 2010 U.S. Decennial Census. The dataset is in a shapefile format containing polygons representing census blocks. Each polygon contains the details of its block group, tract, county, and state ID, allowing aggregation to the necessary level [7]. This file contains over 11 million polygons representing the entire United States; however, we restricted our analysis to the state level and provide results for the state of Minnesota (approximately 260,000 census blocks). Household income data was also limited to Minnesota households, which numbered a little over 2 million in the dataset.

6.2 Hierarchical Aggregation

First, we calculated the Gini coefficient, IQSR, Theil's L, Theil's T, and Atkinson indexes with inequality aversion values of 0.25, 0.50, and 0.75. This was done for data within the state of Minnesota, averaged at different scales. At the smallest scale, ungrouped household median income was used as the income variable. These points were then spatially joined and assigned to census blocks, and income inequality statistics were calculated using the arithmetic mean of incomes within each block as the income variable. The process was repeated by taking the mean of incomes within block groups, tracts, and counties, resulting in 30 total inequality measures.

Computationally, this work was done in Python using the proprietary library "ArcPy" from the company Esri; NumPy and pandas were used for data processing. Special care was taken to avoid numerical overflow, which is probable when adding many five and six-figure integers, especially in the case of the Atkinson index. The income statistics were computationally efficient, all running in O(n) time except for Gini, which can be implemented with O(n log(n)) complexity [29].

6.3 Effects of Hierarchical Aggregation

Table 3 shows the values for various income inequality metrics for Minnesota at the various aggregation levels. As shown in the first two columns the income inequality reported by Gini and IQSR

Table 3: Minnesota income inequality indexes at various levels of aggregation.

Aggregation Level	Gini	IQSR	Theil's L	Theil's T	Atkinson ($\epsilon = 0.25$)	Atkinson ($\epsilon = 0.5$)	Atkinson ($\epsilon = 0.75$)
None	0.3146	4.7691	0.4352	0.3415	0.0836	0.1652	0.2505
Census Block	0.2381	3.3535	0.1520	0.1390	0.0343	0.0683	0.1031
Census Tract	0.1847	2.546	0.0550	0.0545	0.0136	0.0270	0.0403
County	0.0821	1.5104	0.0125	0.0135	0.0033	0.0065	0.0095

always decreases as the basic units of income merged with their neighbors and are replaced with group averages. This is in alignment with the theoretical results (Theorem 4.1, 4.2). We also find that the reduction in metrics is not equivalent across each aggregation step. For Gini, the largest reduction is at the county level whereas for IQSR the largest reduction is from no-aggregation to the census block level. However, in both cases not much inequality is lost by aggregating from blocks to tracts, suggesting little difference in income distribution between blocks.

6.4 Population Constrained Spatial Partitions

We considered the same inequality measurements as the first. Rather than strictly hierarchical spatial partitions, we generated a random sample of partitions at each scale in order to isolate the zone problem of the MAUP from the scale problem. Scale in this context refers to the approximate population size of a subgroup, and it acts as a constraint which limits the search space by which new partitions are selected.

Given an objective population size p, we used a genetic algorithm to construct a set of spatial partitions using census blocks as basic units. The algorithm works by first selecting a random block as a partition seed [8]. Partitions grow by agglomerating blocks with contiguous edges (a distance metric could also be used if growing from points) until the sum of block populations reaches p. This process repeats until every block is assigned to a partition, and the fitness of the zoning scheme is calculated using the sum of the squared population error:

$$Fitness_Z = \sum_{i=1}^{N} \left(\frac{p - p_i}{p} \right)^2$$

This process repeats 200 times, and all 200 solutions are ranked from lowest to highest fitness. The fittest 100 solutions are duplicated, and their duplicates have their partition seeds randomly crossed like chromosomes, with duplicate seeds removed and new seeds randomly selected. This results in 200 solutions at each generation, whereupon the growth process begins again. New seeds are randomly selected only if all the seeds inherited by the parent are claimed by grown partitions. The algorithm ends at the end of 20 generations, and the fittest solution is returned. We performed three simulations, each generating 100 sets of spatial partitions for population constraints of $p \in \{2000, 4000, 6000\}$. For each set of partitions, we calculated the arithmetic means of incomes within all zones, and used these as income variables for the four income inequality measures.

Computationally, the genetic algorithm was implemented using the "Build Balanced Zones" tool of the ArcPy python library. Additional data processing was done with pandas and NumPy. Each simulation was fairly expensive, with complexity $O(n_i n_g n_c n_p)$ where n_i is the number of iterations, or number of zonations sampled, n_g is the number of generations, n_c is the number of candidate zonations per generation, and n_p is the average number of partitions generated, approximately equal to the ratio of the Total Population and Objective Population .

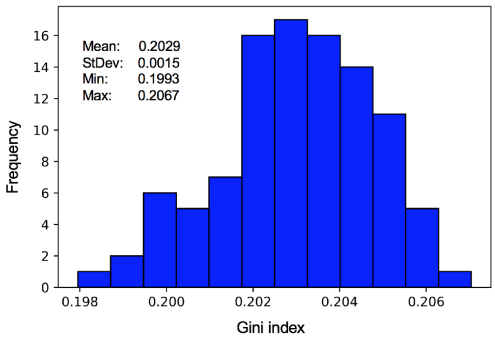


Figure 4: Histogram of 100 randomly-sampled Gini indices subject to a population constraint of 4,000.

6.5 Effects of Constrained Spatial Partitions

We first focus on the distribution of the Gini coefficient under a simulated population constraint of 4, 000 individuals, which is the optimal population suggested by the Census Bureau for census tracts. Actual average tract population according to the dataset is 3, 854. We first checked the effects of the optimization process on the samples. We assumed at least one globally optimal organization of partitions. If the algorithm finds this solution every time, then the distribution of Gini will have 0 variance. In fact, we found that no two Gini measurements out of 100 iterations were the same, meaning no solution was ever repeated. This suggests that the search space of the algorithm is sufficiently large for our purposes.

A histogram of Gini samples is shown in Figure 4, with sample statistics inlaid. The mean of the sample distribution is 0.2029 , lower than the observed Gini on census tracts for Hennepin County (0.2187). In fact, assuming that Gini is asymptotically normal, we can set up a hypothesis test. Let H_0 be that the Gini observed on Hennepin County census tracts is drawn from the simulated distribution of partitions with similar population constraints. We use a p-value of 0.001 , requiring very strong evidence against H in order to reject it. Then $z = \frac{0.2029 - 0.2187}{0.0015}$ is drawn from $Z \sim N(0, 1)$. $2 * Pr(Z < z) = 7.3 \times 10^{-27}$. This is much smaller than our acceptance criteria, so we reject H_0 and are left with the conclusion that census tracts are not drawn by the same process as the population-optimizing algorithm.

The shape of the histogram is relatively normal, with some distortion due to bin size, and a somewhat surprisingly small standard deviation. The presence of any variation suggests that aggregation effects are at play, but the correlation between Gini and average partition population across each simulation is -0.07 , making the scale

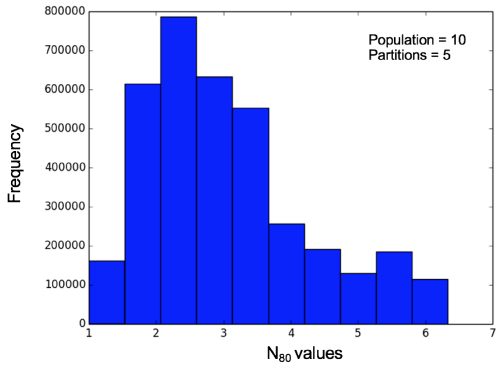


Figure 5: Histogram based on IQSR values for all the possible equi-cardinality partitions subject to a population constraint of 10 and partition size of 2.

problem of the MAUP an unlikely culprit. A more likely conclusion is that Gini is sensitive to the zone problem of the MAUP.

Nevertheless, it seems clear that the sensitivity of Gini to the zone problem is secondary to its sensitivity to the scale problem. This is further illustrated by Table 4, which shows the distribution means for all income inequality measures as the population constraint increases. The results also show that the variation in sample means due to scaling population ±50% is 4-5x greater than the standard deviation at each aggregation level.

Figure 5 shows the histogram on IQSR values for all possible equi-cardinality partitions subject to a population constraint of 10 and partition size of 2. As shown, the value of IQSR is always within the theoretical bounds of 1 and 6.33 (Theorem 4.3). We also observe that a majority of the partitions favor lower IQSR values and the resultant inequality value is usually an underestimate.

7 DISCUSSION

When using AI to inform policy-making, transparency refers to providing enough information on how different measurements or indicators are calculated, their performance and limitations, so that policymakers can better select and use these measures. There's no single best solution. However, transparency is an important value in policy-making to allow stakeholders understand the strengths and weakness of each option so that they can make a decision after weighing the trade-offs of different options.

There could be different ways to address the impact of partitions on different measures and resultant rankings. The first is to acknowledge the issue. This can be done in different ways, for example, by showing the change in rankings with change in partitions if it is clear there is a significant change in rankings, policymakers would need to go back to the base data which is accessible at census enclaves.

Another option is to use different measures and potentially select the measures with lower variability across different choices of partitioning. However, this may raise other macroeconomic issues regarding how effective the measures are. The overall effectiveness of different measures of inequality is a much debated topic in the field of public policy and any choice will involve trade-offs.

Policymaking involves various stages, one of which is studying rankings based on socio-economic measures. Here, we argue that comparisons (e.g., rankings) based on socio-economic measures for a given set of partitions may not be accurate. Policymakers need to be made aware of this and researchers (supplying the rankings) need to report their spatial unit of analysis towards mitigating the effects of inaccurate rankings.

It may be intuitive that well-behaved measures (e.g., Gini function) are smooth as granularity becomes smoother for hierarchically partitioned datasets such as the census. However, the impact of this smoothing on ranking is not intuitive. For example, Tallahassee ranked 14 on segregation of the wealthy at the census tract level, but ranked 92 at the block group level. In addition, there are ratio-based measures (e.g., disparity ratios) which are not monotonic with the change in spatial scale of the space partitions [32].

Other Related Work: Some previous work has analyzed the sensitivity of income inequality measures in particular to space partitioning. Portnov and Felsenstein [21] addressed the sensitivity of various income inequality measurements, including both unweighted and population-weighted Gini, to changes in aggregate zone construction objectives. They conducted a controlled experiment to compare Gini, Theil, Atkinson, and other indices' reaction to the zone problem, but did not address the scale issue or verify on a realistic dataset. Briant et al. [4] showed that the unweighted Gini coefficient is sensitive to scale and zoning problems in one experiment, but did not provide a theoretical explanation and invited future work to confirm their findings on data outside of the French zoning system. Our work takes up this call. Finally, Rey [26] confirmed the issue of MAUP while calculating regional inequality and later with Smith [27] provided a spatial decomposition of the Gini index to better capture the spatial variation across partitions. In contrast, our work proposes sensitivity analysis on the Gini Index and IQSR.

8 CONCLUSION AND FUTURE WORK

Spatial data brings an important dimension to algorithmic transparency. Many inequality and segregation based analyses use aggregated census data but do not report the sensitivity of results to choice of spatial partitioning (e.g., census block group, census tract). We show that values of many measures (e.g., Gini index, dissimilarity index) diminish monotonically with increasing unit size in a hierarchical space partitioning (e.g., block, block-group, tract), however the ranking remains sensitive to the scale of spatial partitions (e.g., block, block group). Our findings highlight the importance of collecting and using fine-scale data to inform policies that address various social equity issues.

In the future, we plan to analyze the sensitivity of other measures of income inequality and segregation to choice of spatial partitioning. Further, we plan to investigate computational methods to address the spatial dimensions of algorithmic transparency. In addition, we plan to consider other aggregation problems such as boundary effects. Computationally, random partitioning techniques can be improved to reduce complexity and produce more optimal zones. Other techniques such as Monte Carlo simulations can be used to estimate the effects of space partitioning for different zones.

Table 4: Sample averages for all six statistics under increasing population constraints.

Target Population	Gini	Theil's L	Theil's T	Atkinson ($\epsilon = 0.25$)	Atkinson ($\epsilon = 0.5$)	Atkinson ($\epsilon = 0.75$)
2000	0.2125	0.0766	0.0706	0.0179	0.0361	0.0548
4000	0.2029	0.0696	0.0643	0.0163	0.0329	0.0499
6000	0.1962	0.0648	0.0601	0.0152	0.0307	0.0466

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1737633. We thank spatial computing research group for their helpful comments and refinements. Finally, we thank the anonymous reviewers for their insightful comments to further refine the paper.

REFERENCES

[1] General Assembly. 2015. Sustainable development goals. *SDGs Transform Our World* 2030 (2015).
[2] Richard J Beckman, Keith A Baggerly, and Michael D McKay. 1996. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice* 30, 6 (1996), 415–429.
[3] Bill Bishop. 2009. *The Big Sort: Why the Clustering of like-Minded America Is Tearing Us Apart.* Houghton Mifflin Harcourt.
[4] Anthony Briant, P-P Combes, and Miren Lafourcade. 2010. Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations? *Journal of Urban Economics* 67, 3 (2010), 287–302.
[5] U.S. Census Buereau. 2011. The 2006-2010 ACS 5-Year Summary File Technical Documentation. https://www2.census.gov/acs2010_5yr/summaryfile/ACS_2006-2010_SF_Tech_Doc.pdf
[6] U.S. Census Buereau. 2011. TIGER/Line with Selected Demographic and Economic Data. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-data.2010.html
[7] US Census Bureau. 2010. TIGER/Line shapefiles.
[8] ArcGIS Pro Documentation. 2019. How Build Balanced Zones works. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/learnmore-buildbalancedzones.htm
[9] Richard Florida and Charlotta Mellander. 2015. *Segregated City: The Geography of Economic Segregation in America's Metros.* Martin Prosperity Institute.
[10] A Stewart Fotheringham and David WS Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and planning A* 23, 7 (1991), 1025–1044.
[11] Corrado Gini. 1912. Variabilità e mutabilità. Reprinted in E. Pizetti & T. Salvemini. *Memorie di metodologica statistica* (1912).
[12] Rick Grannis. 2005. T-Communities: pedestrian street networks and residential segregation in Chicago, Los Angeles, and New York. *City & Community* 4, 3 (2005), 295–321.
[13] Robin Jacob. 2016. Using Aggregate Administrative Data in Social Policy Research. *OPRE Report. Retrieved from https://www. acf. hhs. gov/sites/default/files/opre/opre_brief_draft_dec2016_finaldraftjacob_clean_508. pdf* (2016).
[14] Robert Kenon. 2015. Tallahassee: A Tale of Two Cities. (2015). http://capitaloutlook.com/site/tallahassee-a-tale-of-two-cities/
[15] Simon Kuznets. 1955. Economic Growth and Income Inequality. *The American Economic Review* 45, 1 (1955), 1–28. http://www.jstor.org/stable/1811581
[16] Mag Christian Lagona. 2017. Geographical Equity of the EU's Agricultural Subsidies in Belgium. (2017).
[17] Ira S. Lowry. 1960. Filtering and Housing Standards: A Conceptual Analysis. *Land Economics* 36, 4 (1960), 362–370.
[18] Douglas S Massey and Nancy A Denton. 1988. The dimensions of residential segregation. *Social forces* 67, 2 (1988), 281–315.
[19] Tara O'Neill. 2008. Subsidized housing, private developers and place: A spatial analysis of the clustering of Low Income Housing Tax Credit properties in the 25 largest US cities. (2008).
[20] Stan Openshaw. 1976. An empirical study of some spatial interaction models. *Environment and Planning A* 8, 1 (1976), 23–41.
[21] Boris A Portnov and Daniel Felsenstein. 2010. On the suitability of income inequality measures for regional analysis: Some evidence from simulation analysis and bootstrapping tests. *Socio-Economic Planning Sciences* 44, 4 (2010), 212–219.
[22] Charles Chapman Pugh and CC Pugh. 2002. *Real mathematical analysis.* Vol. 2011. Springer.
[23] Cynthia Putnam and Lorna Chong. 2008. Software and technologies designed for people with autism: what do users want?. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility.* 3–10.
[24] Martin Ravallion. 2014. Income inequality in the developing world. *Science* 344, 6186 (2014), 851–855.
[25] Sean F. Reardon and Kendra Bischoff. 2011. Income Inequality and Income Segregation. *Amer. J. Sociology* 116, 4 (Jan. 2011), 1092–1153. https://doi.org/10.1086/657114
[26] Sergio J Rey. 2004. Spatial analysis of regional income inequality. *Spatially integrated social science* 1 (2004), 280–299.
[27] Sergio J Rey and Richard J Smith. 2013. A spatial decomposition of the Gini coefficient. *Letters in Spatial and Resource Sciences* 6, 2 (2013), 55–70.
[28] Patrick Sharkey and Bryan Graham. 2013. Mobility and the Metropolis: How Communities Factor into Economic Mobility. *Pew Charitable Trusts* (2013).
[29] Buck Shelegeris. 2016. A dynamic programming algorithm for the Gini coefficient. http://shlegeris.com/2016/12/29/gini
[30] Yan Song, Louis Merlin, and Daniel Rodriguez. 2013. Comparing measures of urban land use mix. *Computers, Environment and Urban Systems* 42 (2013), 1–13.
[31] Cheri Speier and MF Price. 1998. Using aggregated data under time pressure: a mechanism for coping with information overload. In *Proceedings of the thirty-first Hawaii International Conference on System Sciences*, Vol. 2. IEEE, 4–13.
[32] Kangkang Tong, Anu Ramaswami, Corey Kewei Xu, Richard Feiock, Patrick Schmitz, and Michael Ohlsen. 2021. Measuring social equity in urban energy use and interventions using fine-scale data. *Proceedings of the National Academy of Sciences* 118, 24 (2021).
[33] Danny Wende. 2019. Spatial risk adjustment between health insurances: using GWR in risk adjustment models to conserve incentives for service optimisation and reduce MAUP. *The European Journal of Health Economics* 20, 7 (2019), 1079–1091.
[34] William D Wheaton, James C Cajka, Bernadette M Chasteen, Diane K Wagener, Philip C Cooley, Laxminarayana Ganapathi, Douglas J Roberts, and Justine L Allpress. 2009. Synthesized population databases: A US geospatial database for agent-based models. *Methods report (RTI Press)* 2009, 10 (2009), 905.

A MATHEMATICAL PROOFS

Theorem 4.1. Gini produces a lower-bound estimate of overall inequality when calculated on group averages i.e., $\bar{G}_N <= G_N$.
Proof. As the normalization term (i.e., $\frac{1}{2N^2\bar{x}}$) is same for both G_N and $\bar{G}_N$, it is sufficient to prove that,

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\bar{x}_i - \bar{x}_j\right| = \sum_{p,q\in P}\sum_{x_i\in p}\sum_{x_j\in q}\left|\bar{x}_i - \bar{x}_j\right| \leq \sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right|$$

where, P is the set of partitions (or groups).
Case 0. When x_i, x_j are within the same partition they have the same mean, then, $\left|\bar{x}_i - \bar{x}_j\right| \leq \left|x_i - x_j\right|$.
Case 1. When x_i, x_j are in different equi-cardinal partitions (say p, q), which is having the same population (size = n). Then, the aggregated Gini can be written as,

$$\sum_{p,q\in P}\sum_{x_i\in p}\sum_{x_j\in q}\left|\bar{x}_i - \bar{x}_j\right| = \sum_{p,q\in P}\sum_{x_i\in p}\sum_{x_j\in q}\left|\frac{\sum_{x_i\in p} x_i}{n} - \frac{\sum_{x_j\in q} x_j}{n}\right|.$$

As, there are n^2 pairwise subtractions in the inner two double summation, the aggregated gini can be further simplified as,

$$\sum_{p,q\in P} n^2\left|\frac{\sum_{i\in p} x_i - \sum_{j\in q} x_j}{n}\right| = \sum_{p,q\in P} n\left|\sum_{i\in p} x_i - \sum_{j\in q} x_j\right|.$$

By triangle inequality of absolute numbers [22], we know that $|a+b| \leq |a| + |b|$, where $a, b \in \mathbb{R}$. Now, if we denote any pair of

$(x_i - x_j)$ as y_{ij}, then, the aggregated gini can be written as,

$$\sum_{p,q \in P} n \left| \sum_{i \in p, j \in q} y_{ij} \right| \leq \sum_{i=1}^{N} \sum_{j=1}^{N} |y_{ij}| = G_N.$$

$\implies \bar{G}_N \leq G_N$

Case 2. When x_i, x_j are in different non-equi-cardinal partitions (say p, q), with the cardinality of $p = m$ and $q = n$. The aggregated gini can be simplified as follows,

$$\begin{aligned}
\sum_{p,q \in P} \sum_{i \in p} \sum_{j \in q} |\bar{x}_i - \bar{x}_j| &= \sum_{p,q \in P} \sum_{i \in p} \sum_{j \in q} \left| \frac{\sum_{i \in p} x_i}{m} - \frac{\sum_{j \in q} x_j}{n} \right| \\
&= \sum_{p,q \in P} \sum_{i \in p} \sum_{j \in q} \left| \frac{n \sum_{i \in p} x_i - m \sum_{j \in q} x_j}{mn} \right| \\
&= \sum_{p,q \in P} mn \left| \frac{n \sum_{i \in p} x_i - m \sum_{j \in q} x_j}{mn} \right| \\
&= \sum_{p,q \in P} \left| n \sum_{i \in p} x_i - m \sum_{j \in q} x_j \right|
\end{aligned}$$

Again, aggregated gini can be written in terms of y_{ij} as follows,

$$\sum_{p,q \in P} \left| \sum_{\substack{i \in p \\ j \in q}} y_{ij} \right| \leq \sum_{i=1}^{N} \sum_{j=1}^{N} |y_{ij}| = G_N.$$

$\implies \bar{G}_N \leq G_N$ □

Theorem 4.2. IQSR(A_P) ≤ IQSR(X), where X is a set of numbers with equi-cardinality partitioning, $P = \{p_1, p_2, ..., p_{C(P)}\}$, where p_i's are pairwise disjoint and their union yields X and $A_P = \{Avg(p_1), Avg(p_2), ..., Avg(p_{C(P)})\}$, where $Avg(p_i)$ is the arithmetic average of items in partition p_i, and $C(P)$ is the cardinality of P.

Proof. Let,

C(X) be the cardinality of X.

$S_L(X)$ be the sum of $\frac{C(X)}{5}$ lowest values in X.

$S_H(X)$ be the sum of $\frac{C(X)}{5}$ highest values in X.

By definition,

IQSR(X) = $\frac{S_H(X)}{S_L(X)}$.

$C(A_P)$ = Cardinality of A_P = $C(P)$(i.e., Number of partitions).

$C(p_i)$ = Cardinality(p_1) = ... = Cardinality(p_K) = $\frac{C(X)}{C(A_P)}$.

$S_L(A_P)$ = Sum of $\frac{C(A_P)}{5}$ lowest values in A_P.

$S_H(A_P)$ = Sum of $\frac{C(A_P)}{5}$ highest values in A_P.

For readability, if $C(A_P)$ is used as a subscript, we denote it by K.

The proof has three parts:

Lemma 1. For equi-cardinality spatial partitioning P of set X, numerator $IQSR(A_P) * C(p_i)$ is upper bounded by numerator IQSR(X), i.e., $S_H(A_P) * C(p_i) \leq S_H(X)$.

Proof sketch. Without loss of generality assume that the partition set P is sorted by $Avg(p_i)$, i.e., $Avg(p_1) \geq Avg(p_2) \geq ... \geq Avg(p_K)$. For simplicity, assume $C(A_P)$ is divisible by 5. In this case, $S_H(A_P) * C(p_i) = (Avg(p_1) + Avg(p_2) + ... + Avg(p_{\frac{K}{5}})) * C(p_i) = (Sum(p_1) + Sum(p_2) + ... + Sum(p_{\frac{K}{5}})) \leq S_H(X)$.

Lemma 2. For equi-cardinality spatial partitioning P of set X, denominator $IQSR(A_P) * C(p_i)$ is lower bounded by denominator IQSR(X), i.e., $S_L(A_P) * C(p_i) \geq S_L(X)$.

Proof sketch. Without loss of generality assume that the partition set P is sorted by $Avg(p_i)$, i.e., $Avg(p_1) \geq Avg(p_2) \geq ... \geq Avg(p_K)$. For simplicity, assume $C(A_P)$ is divisible by 5. In this case, $S_L(A_P) * C(p_i) = (Avg(p_K) + Avg(p_{K-1}) + ... + Avg(P_{\frac{4K}{5}+1})) * C(p_i) = (Sum(P_K) + Sum(P_{K-1}) + ... + Sum(P_{\frac{4K}{5}+1})) \geq S_L(X)$.

Lemmas 1 and 2 imply that equi-cardinality partitioning diminishes IQSR and can never increase it because $IQSR(A_P) = \frac{S_H(A_P)}{S_L(A_P)} = \frac{S_H(A_P) * C(p_i)}{S_L(A_P) * C(p_i)} \leq \frac{S_H(X)}{S_L(X)} = IQSR(X)$. □

Theorem 4.3. IQSR(X) is lower bounded by the sum of $C(X)/5$ smallest values divided by the sum of $C(X)/5$ largest values and upper bounded by the sum of $C(X)/5$ largest values divided by the sum of $C(X)/5$ smallest values, where $C(X)$ is the cardinality of the given set X of incomes.

Proof. Let,

C(X) be the cardinality of X.

$S_L(X)$ be the sum of $\frac{C(X)}{5}$ lowest values in X.

$S_H(X)$ be the sum of $\frac{C(X)}{5}$ highest values in X.

The proof has two parts:

Lemma 1. IQSR(X) is upper bounded by the sum of $C(X)/5$ largest values divided by the sum of $C(X)/5$ smallest values.

Proof sketch. By definition,

IQSR(X) = $\frac{S_H(X)}{S_L(X)}$.

Then, Numerator (IQSR(X)) ≤ $S_H(X)$ and $S_L(X)$ ≤ Denominator (IQSR(S)). Therefore, $\frac{S_H(X)}{S_L(X)} \geq IQSR(X)$.

Lemma 2. IQSR(X) is lower bounded by the sum of $C(X)/5$ smallest values divided by the sum of $C(X)/5$ largest values.

Proof sketch. As, $S_L(X)$ ≤ Numerator (IQSR(X)) and Denominator (IQSR(X)) ≤ $S_H(X)$. Therefore, $\frac{S_L(X)}{S_H(X)} \leq IQSR(X)$. □

Theorem 4.4. The index of dissimilarity diminishes as the scale of aggregation increases i.e., $\bar{D} \leq D$.

Proof. Let there be two partitions p and q where $\{x_p, x_q\}$, $\{y_p, y_q\}$ are the distributions of the selected group and "others" respectively. Then,

$$D = \left| \frac{x_p}{X} - \frac{y_p}{Y} \right| + \left| \frac{x_q}{X} - \frac{y_q}{Y} \right|$$

If the partitions are merged, the new dissimilarity index $\bar{D}$ is

$$\bar{D} = \left| \frac{x_p + x_q}{X} - \frac{y_p + y_p}{Y} \right| = \left| \frac{x_p}{X} - \frac{y_p}{Y} + \frac{x_q}{X} - \frac{y_q}{Y} \right|$$

By triangle inequality of absolute numbers [22], we know that $|a + b| \leq |a| + |b|$, where $a, b \in \mathbb{R}$. Therefore,

$$\left| \frac{x_p}{X} - \frac{y_p}{Y} + \frac{x_q}{X} - \frac{y_q}{Y} \right| \leq \left| \frac{x_p}{X} - \frac{y_p}{Y} \right| + \left| \frac{x_q}{X} - \frac{y_q}{Y} \right| \implies \bar{D} \leq D$$

The proof holds true for any set of partitions P aggregated to a set of partitions $\bar{P}$, where $p_i \in \bar{P}$ are pairwise disjoint and their union yields P. □