Any Way You Slice It: Racial Segregation Statistics are Robust to Aggregation Bias*

Jacob R. Brown†
Department of Political Science
Boston University

Christopher T. Kenny‡
Data-Driven Social Science
Princeton University

Tyler Simko§
Department of Political Science
University of Michigan

August 27, 2025

Abstract

Residential segregation in the United States is widespread, persistent, and threatens economic opportunity and social cohesion. Most segregation metrics, however, rely on aggregate data with arbitrary spatial boundaries. The sensitivity of segregation measures to those boundaries is well theorized but rarely measured. Leveraging modern redistricting software, we simulate millions of alternative Census tract maps that satisfy Census guidelines, compute segregation for each, and recover the full probabilistic distribution of common indices. These simulations yield new estimates of racial segregation across U.S. cities and, crucially, measure the aggregation-induced variability hidden in conventional estimates. Encouragingly, values calculated with official tract definitions closely track the mean of the simulated distribution, exhibiting only a slight upward bias, and segregation metrics shift modestly across maps—though variability grows in smaller cities. While our findings confirm the practical robustness of Census tracts, our data and software provide a general framework for diagnosing and correcting spatial-aggregation error in other contexts.

Keywords segregation • measurement • simulation • Census data • measurement error

1 Introduction

Despite a diversifying population, racial segregation persists across American cities (Massey and Denton, 1993; Elbers, 2021), paralleled by increasing economic (Bischoff and Reardon, 2014; Mijs and Roe, 2021) and political (Rodden, 2019; Kaplan et al., 2022; Brown et al., 2024) segregation. These geographic divisions are associated with a broad set of negative societal outcomes: persistent economic inequality and racial wealth gaps (Ananat, 2011; Chetty et al., 2019), the propagation of inter-group prejudice by *reducing* inter-group contact and *heightening* group threat (Habyarimana et al., 2009; Enos, 2017), inequitable distributions of public goods (Alesina et al., 1999; Trounstine, 2016), and threats against the functioning of democratic governance (Putnam, 2007; Chen

*Thank you to Cory McCartan, Shiro Kuriwaki, Sergio Montero, Soichiro Yamauchi, and helpful commenters at the APSA 2024 annual meeting.

†Email: jbrown13@bu.edu. Website: https://www.jacobrbrown.com.
‡Email: ctkenny@princeton.edu. Website: https://christophertkenny.com/.
§Email: tsimko@umich.edu. Website: https://www.tylersimko.com.

and Rodden, 2013). Contemporary segregation is a function of historical legacies of de jure and de facto racial segregation (Rothstein, 2017; Logan and Parman, 2017) and economic constraints on housing and mobility (Reardon and Bischoff, 2011). Further, residential segregation is both exacerbated by and contributes to ongoing crises in housing supply and affordability in American metropolitan areas (Trounstine, 2018; Einstein et al., 2020).

Yet, researchers and policymakers face methodological challenges when measuring segregation. Foremost among these challenges is the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1983; White, 1983). Segregation can be understood as the uneven distribution of group populations across a geographic area (Massey and Denton, 1988). For segregation, the MAUP means that decisions about how to divide populations across that area can impact resulting measurements about the geographic distribution of those groups. The MAUP has two primary components—a "vertical" dimension where measuring spatial statistics at different scales (e.g. unit sizes) can lead to different results (Gehlke and Biehl, 1934; Jelinski and Wu, 1996), and a "horizontal" dimension where changing the precise definition of geographic units (e.g. changing border locations) changes results (Lee et al., 2025).

This vertical dimension can be conceptualized as a substantive choice: at what scale is segregation relevant for the research question (Arjona, 2019; Lee et al., 2025)? The horizontal component is a measurement error problem: conditional on choice of scale, how would our measures of underlying segregation differ if boundaries had been drawn differently? Most existing work diagnosing the MAUP for measuring segregation focuses on the vertical component (Fischer et al., 2004; Wong, 2004a; Dmowska and Stepinski, 2024), or tries to side-step aggregation challenges by using scale-invariant measures or individual-level location data (Lee et al., 2008; Reardon et al., 2008; Östh et al., 2014; Logan and Parman, 2017; Lee and Rogers, 2019; Brown and Enos, 2021). But the potential for large measurement differences from alternative boundaries matters in practice, because the vast majority of segregation research in the United States continues to rely on aggregate units—Census tracts—as neighborhood proxies (Sharkey and Faber, 2014). In fact, nearly all policy-relevant measures of segregation rely on Census tracts, and these statistics are used to allocate billions of dollars in public and private funding (Hotchkiss and Phelan, 2017; Kenny et al., 2021). This reliance persists despite long-standing concerns that MAUP-induced measurement error could be quite large (Wong, 2004b; Chen et al., 2022; Wildfang, 2025). Yet, the magnitude of this measurement error has yet to be quantified at scale.

In this article, we diagnose this measurement error by developing new methodological tools that measure segregation across millions of alternative possible Census tract maps. We evaluate how sensitive segregation measures are to aggregation using recent advances in algorithmic redistricting methods and software to redraw geographic tract boundaries (DeFord et al., 2021; Autry et al., 2019; McCartan and Imai, 2023; Kenny et al., 2024a). We sample alternative Census tract maps that follow Census guidelines: ensuring that maps are contiguous, respect natural boundaries, and meet certain population bounds. We then calculate segregation metrics for each map to construct probabilistic distributions of segregation indices. These distributions reflect the segregation estimates that we could observe across many reasonable Census tract definitions. We use these data to provide new estimates of racial segregation for every U.S. city with a population of over 100,000 people in 2020, averaging across the full distribution of how a Census tract map reasonably could have been drawn, rather than relying on any single snapshot. We further quantify the variability induced by aggregation measurement error, examining the width of our simulation distributions.

Our results demonstrate that official Census tract definitions are generally robust to aggregation measurement error. Compared to simulation estimates, estimates using official Census tract data slightly overstate the degree of racial segregation in the United States, but this difference is quite small. For example, we find that the most commonly used segregation measure—the Dissimilarity Index—with official Census tract data overstates Black-White segregation in our sample by 2.37% on average. We calculate quantities like these by comparing the observed segregation estimates to the averages from our simulated distributions. This means that most

alternative ways of grouping Census blocks together into Census tracts would lead to lower levels of measured segregation than the actual tract maps, but these differences are mostly quite small. We find similarly small differences between simulation and official estimates across other segregation indices including measures of multigroup segregation, racial exposure, and isolation. We replicate these results by conducting equivalent simulation analyses for 2000 and 2010 and find that the difference between official Census statistics and our estimates is consistent across recent decades.

While official Census tract definitions compare favorably to our simulation averages, wide variability across simulations indicates that aggregation measurement error creates noise in segregation estimates, rather than systematic biases (recurring upward or downward differences). We evaluate variability by taking the 95% credible intervals from our simulated distributions, and find that for the typical city the Black-White Dissimilarity index could reasonably be as much as 7.26% higher or lower than the simulation average. This measurement variability is most pronounced in small and medium sized cities. This variation is primarily driven by the number of geographic units: smaller cities have fewer Census tracts, so each redrawn map can more dramatically restate how segregated the city looks. By contrast, differences tends to be small and average out across simulations in larger cities with many Census tracts. Further, we conduct a bounding analysis using New Jersey to understand the practical magnitude of these differences. We find that, while large distortions in segregation are often possible through highly unusual spatial arrangements, in general possible alternative measures of segregation across neighborhoods are similar. Indeed, the median range of the simulations is only 21% as large as the corresponding range from the bounding analysis. This suggests that typical, alternative tracts are relatively constrained, but elucidates *why* the conventional wisdom expects segregation to be sensitive to aggregation choices.

We make four key contributions. First, we quantify measurement error in common segregation metrics resulting from aggregation choices. Previous research demonstrates the potential for aggregation challenges to distort geographic measurement (Gehlke and Biehl, 1934; Openshaw, 1983; White, 1983; Wong, 1997; Jelinski and Wu, 1996; Fotheringham and Wong, 1991), and some studies offer methods to circumvent the MAUP in certain contexts (Hennerdal and Nielsen, 2017). To the best of our knowledge, this is the first study to fully diagnose the extent of this "horizontal" measurement error for segregation. Some progress on aggregation issues to date comes from studies with access to individual-level data, which enables spatial measurements across many different contextual definitions (Östh et al., 2014; Logan and Parman, 2017; Brown and Enos, 2021). Such data, while increasingly available, are expensive and most segregation research must still rely on publicly-available demographics data. Many of the most cited recent studies of segregation's effects—in economics, political science, and sociology—use such aggregate data (for example, Ananat, 2011; Trounstine, 2016; Legewie and Schaeffer, 2016). Other studies make progress on different problems with common segregation metrics, such as aspatiality, lack of decomposibility, and sensitivity to baseline city demographics (Reardon and O'Sullivan, 2004; Lee et al., 2008; Mazza and Punzo, 2015; Roberto, 2018). While these other problems are related, none of their solutions solve problems of aggregation. These modern but less commonly used improvements thus still suffer from the challenges we diagnose in this article.

Second, despite concerns over the representativeness of Census tracts (Wong et al., 2012; Ansolabehere et al., 2025), we show that measures based on the U.S. Census Bureau's official Census tract definitions are robust to local area definitions. We do find some variability across simulations, particularly in small cities, but in general this variability is also small. This robustness is not because drawing dramatically different maps is impossible, but because when looking at the probabilistic distribution of simulated maps, the average map yields a similar segregation estimate as the official Census tract map and most maps offer only moderately different conclusions about levels of segregation. In our short-burst analysis we demonstrate that drawing maps with much higher or lower segregation estimates than official Tract estimates requires abandoning reasonable commitments to compactness, resulting in maps that are unrepresentative of the sampling distribution and thus underlying segregation. These results should reassure researchers and policymakers who regularly use Census tracts to

study segregation. Where measurement error does exist, our analysis provides recommendations for when, where, and with which metrics these issues are most apparent.

Our third contribution helps to resolve this measurement problem in future research by providing original data that researchers can use to measure segregation and its impacts. We offer nationwide simulated Census tract plans for every county and every town and city with a population over 100,000 in the United States for the years 2000, 2010, and 2020. These data are open-source, free to use, and can be readily adapted to simulate other geographic units and contextual variables.

Fourth, we provide a methodological approach with open-source software to diagnose the extent of horizontal measurement error in their own applications. Our sampling based approach estimates a distribution of alternatives, which allows researchers to be flexible in the comparisons they make. While enumerating every possible alternative boundary is computationally intractable for most applied problems, the algorithms underlying our sampling approach offer asymptotic convergence guarantees to an underlying target distribution. Based on our findings that measurements across small geographies are most sensitive, we particularly recommend researchers use our approach when their research question concerns segregation across small areas. If our pre-provided data does not suit their needs, researchers can use our provided template software to diagnose the sensitivity of their results to alternative ways that any aggregate geography—like Census tracts, voting precincts, school attendance zones, or neighborhoods—could have grouped the same population in different ways.

The rest of the paper proceeds as follows. Section 2 formalizes understandings of how aggregation choices leads to measurement error in segregation metrics. Section 3 explains how U.S. Census tracts are drawn. Section 4 describes our simulation analyses, including incorporation of previously discussed Census tract drawing rules, as well as performance statistics and validation. Section 5 presents results from these simulation analyses, comparing official U.S. Census statistics to simulations. Section 6 reports our simulation estimates over-time, examining segregation across 2000, 2010, and 2020. In Section 7, we extend our methodology to demonstrate what kind of (unrealistic and unrepresentative) maps have to be drawn in order to produce dramatically different segregation estimates than official numbers. Section 8 summarizes these analyses and concludes with broader takeaways and areas for future research.

2 Conceptualizing measurement error from aggregation

Measuring segregation requires the distillation of continuous information on the locations of groups in (at least) two-dimensional space into numerical summaries of how unevenly these groups are distributed across that space (Jahn et al., 1947; Gatrell, 1983). Despite some recent advances, in almost all applied work researchers use pre-aggregated data made by collapsing this continuous reality into administrative units like Census tracts to measure segregation. Understanding measurement error across specific realized administrative areas like Census tracts matters, therefore, as these units are used in a wide range of research and policy decisions.

The Modifiable Areal Unit Problem (MAUP) (Openshaw, 1983) is a key methodological issue for measuring segregation (Openshaw, 1983; White, 1983). The MAUP means that measuring statistics across aggregate spatial units can change when those aggregations are made at different scales (the "vertical" dimension) and different locations (the "horizontal" dimension). For segregation measures, this means that grouping the same population into larger/smaller units (vertical) or drawing boundaries in different locations (horizontal) can alter how segregated groups appear to be across these boundaries.

A broad literature across fields like geography, statistics, and the social sciences demonstrates the vertical dimension of MAUP can influence estimates (White, 1983; Jelinski and Wu, 1996; Bisbee and Zilinsky, 2023), and offers potential solutions (Sang-Il Lee and Griffith, 2019). Some work uses scale-invariant measure of segregation (Reardon et al., 2008; Lee and Rogers, 2019), while other researchers abandon aggregate units for individual-level data (Brown and Enos, 2021; Dmowska and Stepinski, 2024). While useful, these two approaches are infrequently

used because they can be prohibitively expensive for many applications, or the necessary data are simply not available. Generally, researchers choose a geographic scale across which they will measure segregation, and argue the specific level of aggregation is the appropriate substantive choice for their context (Arjona, 2019; Lee et al., 2025).

The horizontal component of MAUP is most often ignored by researchers. While researchers are often encouraged to run robustness checks on alternative definitions of their boundaries (Lee et al., 2025), in practice this is challenging because the number of potential alternative boundaries that could be drawn is unfathomably large for most problems of realistic size. We address this limitation in the next section.

2.1 Considering MAUP as a sampling problem

We propose to consider the horizontal component of MAUP as a sampling problem. For geographic areas of any reasonable size, there are near-infinite ways to segment an area into roughly equal population areas such as Census tracts. Each boundary could be drawn differently to place different areas in different units. For example, Fifield et al. (2020) estimate that even segmenting an 8x8 grid square into two connected components generates more than 1.2×10^{11} unique partitions. Each possible map may be a reasonable draw from the sampling distribution of all possible maps, but some maps are more representative of the typical draw than others.

In this framework, we cannot know from looking at any single Census tract map whether it is representative of all the other maps that could have been drawn under similar drawing rules, or whether it presents a distorted view of segregation. If we can observe the sampling distribution, however, we can know on average what segregation looks like by taking the average over the sampling distribution. A helpful analogy is that looking at a single redistricting plan alone cannot determine if it is an unfair "gerrymander" (Kenny et al., 2023). For example, unusual twists and turns may not be nefarious, but may instead reflect unusual residential patterns. Instead, comparing a single plan to a distribution of plans drawn under the same criteria can provide evidence that the relevant plan is 'unusual.' Further, we can measure the spread of the sampling distribution to get a sense of the variability of segregation estimates across different maps. Then, we can place any observed Census tract map within this distribution to see the difference in its estimate of segregation.

Figure 1 illustrates the logic of investigating sensitivity to MAUP by sampling. The figure shows the same geographic space with three different population groups (X's, triangles, and circles). The locations of these groups are fixed, but the top row shows two different ways to draw a "neighborhood plan" (which we use as shorthand to refer to the general problem of aggregating units) for these three groups. The measurement goal is to summarize the segregation of the space. The neighborhoods are drawn to be of equal population. In this stylized example, we use the two-group Dissimilarity Index, which measures the segregation (on a 0 = complete integregation to 1 = perfect segregation scale) between triangles and the two other groups (combined into a single "non-triangle" group).[1] The first neighborhood plan returns a triangle vs. non-triangle Dissimilarity Index of 0.25, while the second returns a Dissimilarity Index of 0.625.

The bottom row illustrates that these groupings are two of many potential ways this small space could be reorganized into three neighborhoods. The right subplot shows a histogram of the (triangle vs. non-triangle) Dissimilarity Index values from each draw. This distribution reflects all of the possible ways (in expectation) that the space could be divided up under the rule of equal population and three neighborhoods. From this distribution we see that the average Dissimilarity Index observed is much closer to 0.25 than 0.625. Still, the distribution shows substantial variance, with significant mass as much as ten percentage points from the center of the distribution. Therefore, we would conclude, in this example, that a segregation estimate based on the right

[1] A formal equation for the Dissimilarity Index and other details on how segregation is commonly measured is found in Supplementary Information Section A.3.

plan alone is very unusual compared to the simulated distribution of possible plan. The problem for applied researchers is clear—without access to such a reference distribution, it is difficult to know whether the single observed neighborhood plan offers an unusual estimate or segregation (right) or not (left).

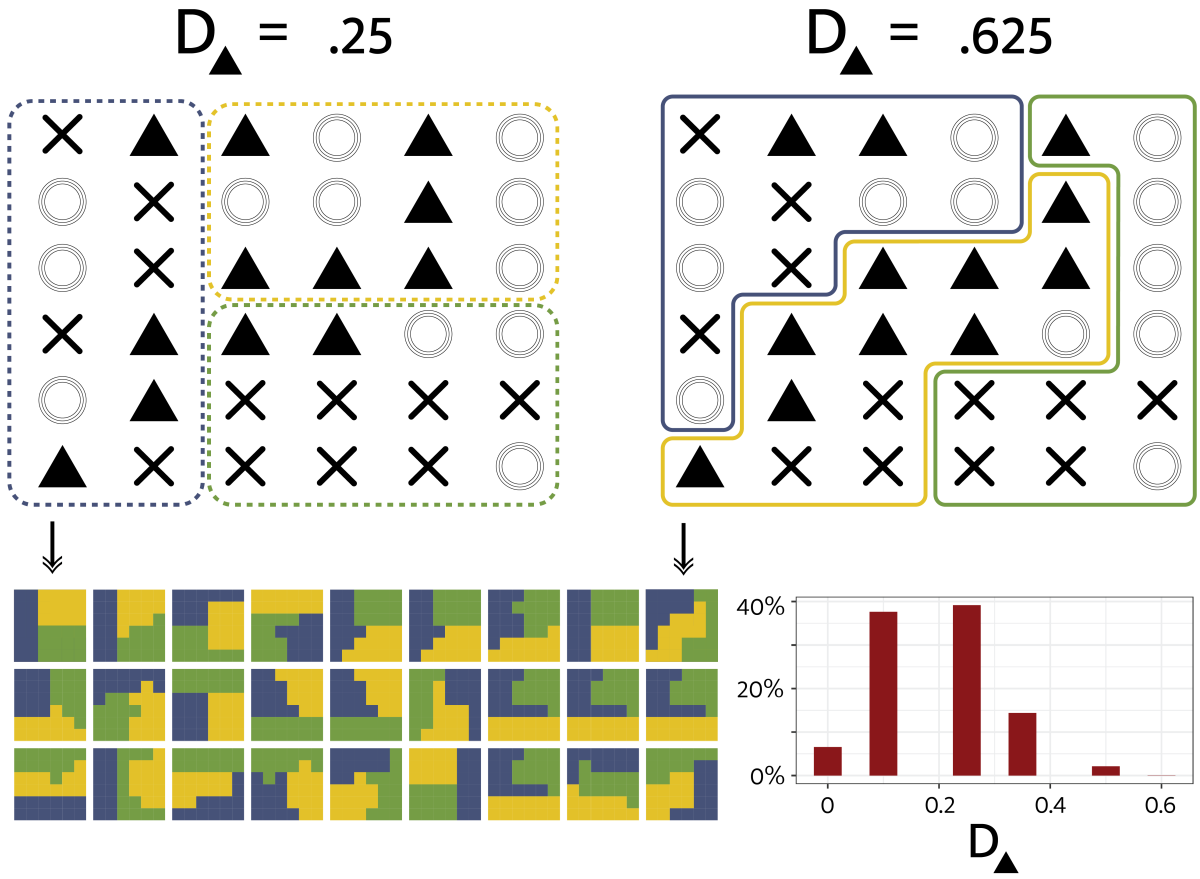

Figure 1: Different neighborhood maps, different segregation estimates. Top row shows two different neighborhood maps each with three equally-sized neighborhoods. The maps are labeled with their respective two-group (triangle versus non-triangles Dissimilarity Index. The bottom left plot illustrates the distribution of many different sets of 3 neighborhood maps. The right plot shows the histogram of the Dissimilarity Index across the distribution of feasible maps.

This figure also illustrates our methodological approach: we aim to simulate the Census tract drawing process to construct representations of the sampling distribution of segregation across many alternative maps. These maps are independently drawn and adhere to Census drawing rules and population targets. Below, we describe how the sampling algorithm we use ensures that the sample of maps drawn probabilistically reflects the true sampling distribution. Many plans will look quite similar, particularly those closer to the center of the distribution. We calculate the bias of any observed Census tract map as its deviation from the sampling distribution of plausible maps. For variability, under the assumption that each plan is independently drawn, we can use the quantiles of the sampling distribution to measure the spread of the distribution.

In the next section, we describe how Census tracts are created as aggregations from smaller administrative units in practice. Later, our methodological goal will be to approximate many features of this drawing process in our simulation framework to characterize measurement variability across alternative realistic definitions.

3 How Census tracts are drawn

Census tracts are "are small, relatively permanent geographic entities within counties (or the statistical equivalents of counties)" (U.S. Census Bureau, 2023).[2] Census tracts typically contain between 2,500 and 8,000 people, with a target population of about 4,000. Census tracts were first drawn as far back as the 1890 Decennial Census in a small number of cities, but were not administered at a larger scale until 1940, when the Census began publishing Census block data (the smallest available U.S. Census unit) for all cities above 50,000 people. This procedure expanded through the 1950-1980 Decennial Censuses. From the 1990 Decennial Census onward, the Census assigned Census blocks and tracts covering the the entirety of the U.S. Census tracts are widely used by researchers, policy-makers, and communities because they are free, standardized, exhaustive of US area, and available across several decades.

Census tracts are not completely redrawn with each Decennial Census. Rather, as Census tracts grow in population, they are eventually split into different Census tracts once the populations exceed the target thresholds. Thus, the Census tracts we observe today are from a path-dependence process that builds on the original definitions. At the time of original definition, Census tracts were drawn to hit the targeted population and to adhere to "visible" features of the geographic environment such as roads, waterways, railroads, and other major infrastructural barriers. Originally, tracts were also drawn to be as "homogeneous as possible with respect to population characteristics, economic status, and living conditions" (U.S. Census Bureau, 2023). This creates further opportunities for potential bias, if a more segregated variation of possible boundaries was intentionally chosen as an original tract definition. In practice, adherence to these infrastructural features is not always guaranteed as infrastructure and other geographic boundaries change. Census tracts are nested in the Census geographic "spine," meaning they never cross larger Census geographies like counties or states.

We follow and diagnose compliance with these population targets and infrastructural guidelines when designing our simulations. We design our simulations around the guidelines used for drawing Census tracts in practice because Census geographies are used for many consequential public policy decisions, like legislative redistricting and the allocation of over $600 billion in federal funds (Hotchkiss and Phelan, 2017; Kenny et al., 2024b). That is, we target a particular "sampling distribution" of alternative ways that Census tracts could be drawn under longstanding guidelines. We make this decision intentionally, as Census geographies like tracts are used for many other purposes beyond measuring segregation. Maintaining these guidelines is also important because they are unlikely to drastically change moving forward, due to the many other sources of federal, state, and local public policy for which Census geographies are used. In later analyses, we compare our results with these constraints to alternative analyses without them.

At the same time, our goal is not to follow the data generating process of the Census identically, but rather to adopt a set of rules inspired by the practical limitations the Census Bureau faces that are consistent and reasonable ways of dividing up geographic space. As such, we do not emulate the method of starting with original Census tract definitions and simulating the splitting process across decennial Censuses. Rather, for any given decennial Census we draw new Census tract maps under the guidelines described above. This removes any temporal dependence from our distributions and more directly reflects the quantities of interest described in previous sections.

4 Simulating millions of alternative Census tracts

We simulate alternative tracts from the smallest geographic unit published by the US Census Bureau: Census blocks.[3] Blocks are not defined by population, but typically contain about 40 people on average, although

[2] For a complete accounting of how Census tracts are drawn, see: https://www2.Census.gov/geo/pdfs/reference/GARM/Ch10GARM.pdf

[3] Census blocks are statistical areas defined by visible features such as roads, railroads, waterways, property lines, and municipal boundaries (U.S. Census Bureau, 201).

populations can range from 0 to over a thousand people. We collect Census block data from PL 94-171 Decennial Census Files that contain information on race (7 categories) and ethnicity (Hispanic/not). We supplement these data with information on Places (a Census definition for cities, towns, and villages) to use for later accuracy estimates.

We sample plans using "Merge-Split", an algorithm developed by Carter et al. (2019) (based on work by Deford et al. (2019)) for constructing partitions of units into groups. Merge-Split allows for algorithmic constraints that place limitations on how partitions are drawn like contiguity, population targets, compactness, and administrative boundary constraints. Within this constricted set of compliant plans, Merge-Split draws plans randomly. Merge-Split was originally created to assess gerrymandering by simulating redistricting plans, under the similar logic that any district plan can be compared to the sampling distribution of reasonable district plans (e.g, Kenny et al., 2023). We implement Merge-Split using the R software `redist` (Kenny et al., 2024a). This software integrates Merge-Split with a Markov Chain Monte Carlo simulation using the Sequential Monte Carlo sampler from McCartan and Imai (2023). This MCMC method edits two districts simultaneously and accepts them probabilistically, iteratively replacing them in the plan. We thin this, taking every k-th plan where k is large relative to the number of districts.

We ensure our simulations closely match Census boundaries in the following ways:

1. **Contiguity**: Census tracts, by definition, are contiguous shapes composed of Census blocks. We simulate contiguous tracts by creating adjacency lists for every Census block in the United States using (Kenny et al., 2024a). These adjacency lists indicate which Census blocks are contiguous pairs, and our simulations ensure that only contiguous pairs can be grouped into tracts. When blocks are completely discontiguous from any other blocks (e.g. islands, river segments), we connect them to their nearest neighbor block.
2. **Geographic Spine**: We ensure our simulated tracts never cross county lines by simulating each county separately. This is because tracts must fit into the broader "Geographic Spine," in which Census geographies nest into blocks, block groups, tracts, counties, and states.
3. **Population**: The Census generally draws tracts to have between 2,500 and 8,000 people, with an average of 4,000. We follow this guideline by setting population bounds on our simulated tracts. For each county, we set population bounds based on local populations: we use the population of the least populated Census tract in an area multiplied by 0.9 as the lower bound, and the most populated Census tract multiplied by 1.1 as an upper bound. This flexibility allows for slight variations in population targets that still respect Census guidelines, rather than strictly requiring the simulated tracts each have exactly the same populations as the existing tracts. We then also add a probabilistic constraint which encourages, but does not force, populations to sit within the original Census tract population goal.

Further, we monitor several diagnostics to ensure our simulations respect the following Census guidelines:

1. **Boundary Crossings**: We diagnose how often our simulated tracts cross interstate highways (e.g., I-90 or U.S. 1), all major railroads, and waterways defined as rivers, lakes, or channels using geographic Census definitions. The Census Bureau attempts to draw tract boundaries that respect visible infrastructural features like major highways and rivers. There is wide variation in how this guideline is adopted across the country due to geographic and infrastructural variation. The Census releases no public information on how this is operationalized in practice, though they say such boundaries are "generally" respected.
2. **Off-Spine Geographies**: We diagnose how often our simulated tracts split Census Place boundaries, which represent cities, towns, and villages. Although places bear no formal relationship to Census tracts by being off the main geographic spine, we calculate diagnostics to ensure places are split at similar rates in our simulations compared to enacted tracts.

3. **Compactness**: We use a probabilistic ("soft") constraint to nudge our simulated tracts into compact shapes. Census tracts generally have "compact," broadly rounded or rectangular shapes. There is no official documentation or criteria that we are aware of as to particular requirements on compactness. Instead, a preference for compact tracts is likely a consequence of other decisions, such as the path-dependency from earlier hand-drawn maps or decisions to follow major infrastructure. Specifically, we use the compactness constraint implemented in (Kenny et al., 2024a) set at the author's recommended value.

We run separate simulations by county[4] and decennial Census year (2000, 2010, 2020). For each county, we run between 25,000 and 4,000,000 simulations, increasing simulation iterations based on county size and convergence evaluation. In each iteration, Merge-Split creates a complete Census tract map for the county, creating the same number of Census tracts as the official Census tract plan with population constraints approximately equal to the population range of Census tracts in the official plan. We run four independent chains. Each chain runs for at least 50,000 steps and stops when converge diagnostics are met. We test both plan diversity (McCartan and Imai, 2023) and $\hat{R}$ statistics (Vehtari et al., 2021) to evaluate convergence and whether the simulation distribution reflects the theoretical sampling distribution of plans.[5] We perform a final thinning step to output 5,000 simulated plans for each county (1,250 from each chain). To illustrate this output, Figure 2 plots the official Census tract map (top left) for Elizabeth, NJ (a typical small city in our data) in 2020 compared to a randomly sampled simulated plan (top right).

We then calculate for each of these 5,000 stored plans the Dissimilarity Index for pairs of racial groups (Black-White, Hispanic-White), the H Index for Black, White, Hispanic, and an Other racial category, and the Isolation Index for each racial group.

4.1 Our simulations closely follow Census guidelines

Next, we show that our simulated tracts accurately reflect reasonable alternative Census tracts. In Table 1, we show that our simulated tracts are similar to real tracts across measures of compactness, population deviation, and place splitting. Table 1 reports the average levels of these metrics across all official plans and simulated Census tract plans for counties consisting in 2020 of more than one Census tract. Population deviation represents the maximum percent population deviation from a comparison plan where all tracts within a county have the same population. Compactness is based on a graph theoretic measure that calculates the fraction of possible edges in the adjacency graph (connections between units) are kept uncut. Place splits measure the number of Census Places which are "split" by having a Census tract contain area both inside and outside of the same Census place. Overall, our simulation procedure draws tracts with slightly smaller population ranges, similar compactness scores, and splits towns at a very similar rate.

Next, we show our simulations respect road, waterway, and railroad boundaries at similar rates to Census tracts. For each sampled plan and for the official Census tract plan, we calculate the proportion of the length of highways, railroads, and major waterways in that county that correspond with a Census tract border. [6] Table 1 reports these statistics, demonstrating that on average our simulations respect these boundaries at similar or higher rates to official Census tract maps.[7]

[4]For counties consisting of just 1 populated Census tract, of which there are 225, we do not simulate the Census tract map but rather return the official Census tract.

[5]We report convergence statistics for the simulations in the Supporting Information Section A.1.

[6]For this analysis, we limit analysis or roads to interstate highways and U.S routes. For railroads, we consider all rail lines listed in the Census nationwide shapefiles. For waterways, we limit our analysis to waterways classified as rivers, lakes or canals, excluding smaller waterways such as brooks or creeks.

[7]We calculate overlap statistics for a random sample 100 simulated plans from each county due to long computational times to measure these spatial overlays.

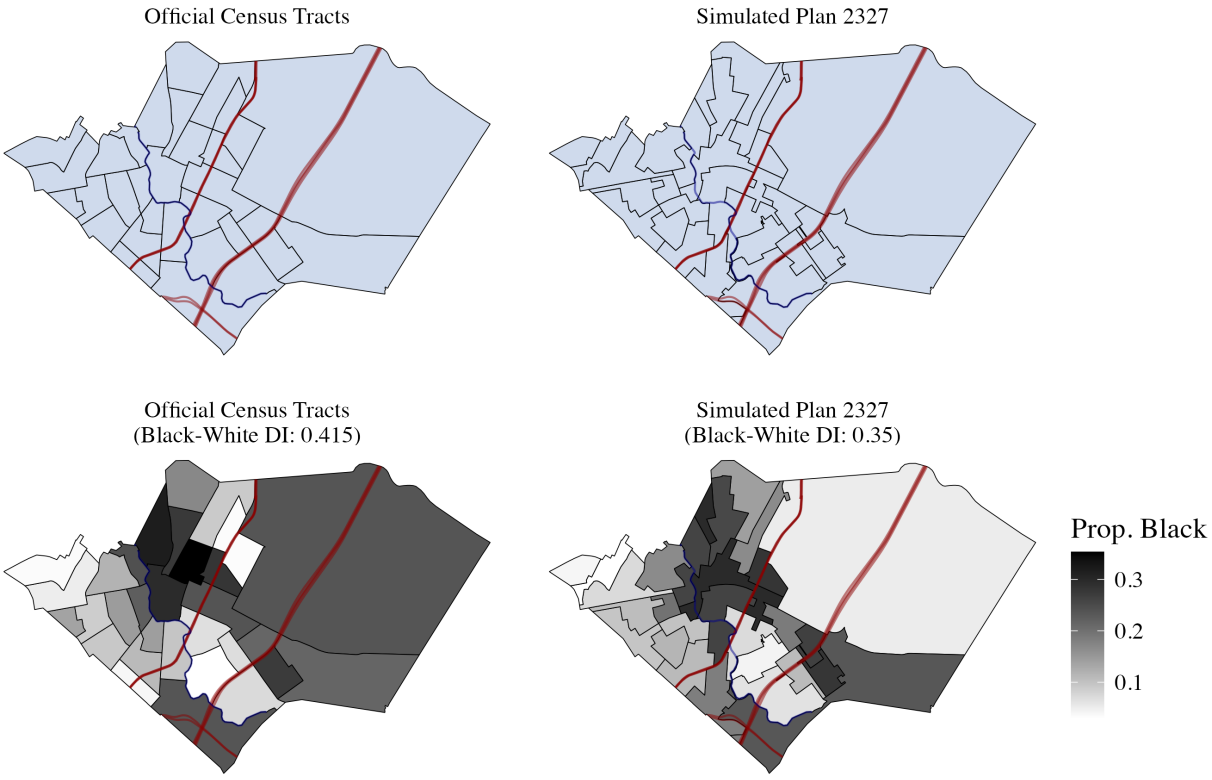

Figure 2: Elizabeth, NJ Official versus Simulated Tract Map. Left maps show the official tract map for Elizabeth, NJ and right maps show a simulated tract map. Top row illustrates the boundaries while the bottom row shades tracts by proportion Black population. Highways are plotted in in red and waterways in dark blue.

Table 1: Our simulated tracts have similar population deviation, compactness, and rates of place splitting as true Census tracts.

	Official (N=2,915)		Simulation (N=14,578,000)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Dev.	0.635	0.448	0.524	0.360	-0.111	0.008
Comp. Frac.	0.964	0.018	0.953	0.021	-0.011	0.000
Place Splits	4.495	7.907	4.646	8.768	0.151	0.146

Table reports the average and standard deviation of deviation, compactness, and place splits across official and simulated plans. Difference of means between official and simulated plans are reported in the fifth and sixth columns.

Table 2: Our simulated tracts respect boundaries like roads and waterways at similar rates to true Census tracts.

		Official		Simulation		Difference	
	Counties	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Highways	2,686	0.419	0.259	0.464	0.210	−0.045	0.177
Railroads	2,617	0.339	0.238	0.445	0.209	−0.106	0.110
Waterways	2,641	0.435	0.335	0.444	0.276	−0.009	0.175

Table reports the average and standard deviation of the proportion of highways, railroads, and waterways that overlap with official and simulated tract boundaries across counties. The number of counties with a major highway, railroad, or waterway in them is reported in the second column. The average difference and standard deviation of the difference is reported in the seventh and eighth columns. Statistics are estimated using a random sample of 100 plans from our 2020 simulations.

5 Census segregation statistics are robust to alternative tract maps

We find that both official and simulated Census tracts offer similar estimates of racial segregation across U.S. cities. Column 1 of Figure 3 shows a scatter plot of segregation estimates (by row: Black-White Dissimilarity Index, H Index, and the White Isolation Index, respectively) from our simulations (x-axis) and the official tracts (y-axis) for every city in the United States with a population over 100,000 people in 2020. Column 2 reports histograms of bias in official Census tract estimates. Bias is defined for each city as the segregation estimate from official Census tract definitions minus the average of our 5,000 simulated Census tract plans for that city. Starting with the top row, which reports results for Black-White Dissimilarity, we find minimal bias across cities, as official estimates tend to only slightly overstate Black-White Dissimilarity relative to simulation averages. On average, this bias is 0.81 percentage points. Proportionally, we find official statistics overstate Black-White Dissimilarity by 2.37% on average. We see similarly small estimations of bias for the H Index and for Black-White isolation. In the Supporting Information, we show these results are robust to examining segregation for counties rather than cities (SI Section B.3), considering other segregation indices like the Hispanic-White Dissimilarity index and isolation indices for Blacks and Hispanics (SI Figure B1), and alternative parameter values of our algorithmic constraints (SI Section B.5).

However, we do find modest measurement variability across simulated plans. To estimate measurement variability, we take the length of the 95% interval of our simulated Black-White Dissimilarity Index distributions (the 0.975 percentile minus the 0.025 percentiles). The longer this interval, the more uncertainty in our estimates of segregation due to aggregation measurement error. Figure 3 column 1 plots the distribution of 95% interval length. The average interval length is 4.37 percentage points. This means that our segregation estimates are on average consistent with segregation being approximately 2.33 percentage points higher or lower than the simulation averages. In percentage terms, this translates to segregation that is on average 7.26% higher or lower than the simulation average.

While we overall find only small bias and measurement variability in segregation estimates, we conduct further analyses to measure where these concerns may be most severe. Figure 4 plots the 335 cities with population over 100,000 as of 2020 in the continental United States. City points are colored blue to red based on the size and direction (darker blue indicates larger negative bias, darker red indicates larger positive bias) of the difference between the Black-White Dissimilarity Index using official Census definitions and the simulation average from our simulations. City point are sized by the amount of measurement variability across simulations for that city (larger points indicate larger measurement variability). The map also includes labels for a sampling of the cities with the largest and smallest exhibited bias and measurement variability. We see that large cities like Los

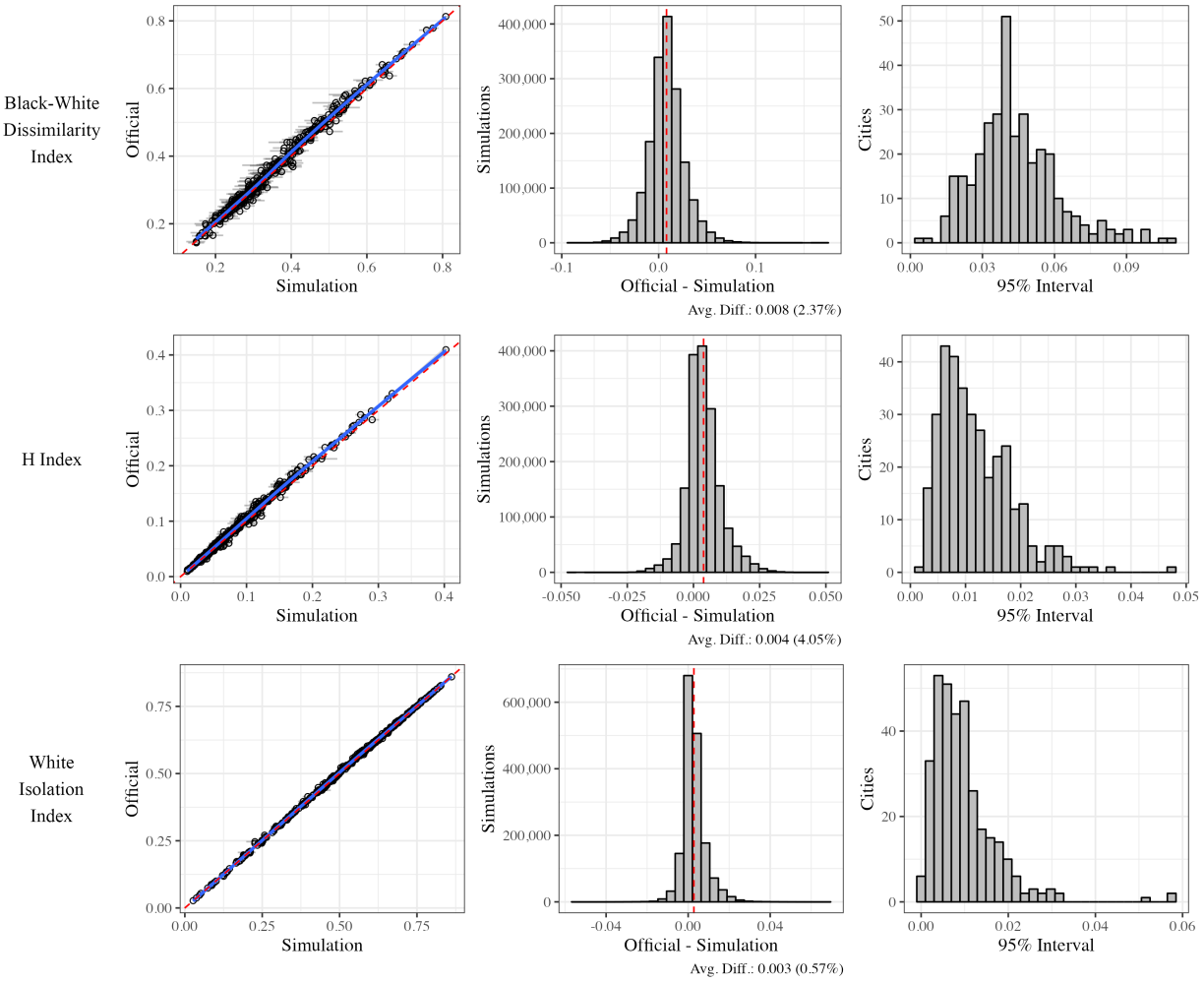

Figure 3: Measuring segregation using official Census Data versus simulations. Left plots show scatter plots of the relationship between official and simulation averages of segregation indices. Middle plots show histogram of official minus simulation averages. Right plots show histograms of 95% interval size across cities. Top row shows Black-White Dissimilarity Index plots, middle row shows H Index plots, and bottom row shows White Isolation Index.

Angeles, Chicago, or Philadelphia exhibit the smallest amounts of measurement variability, while smaller cities such as Sugar Land, TX, or High Point, NC have the largest amount of measurement variability.

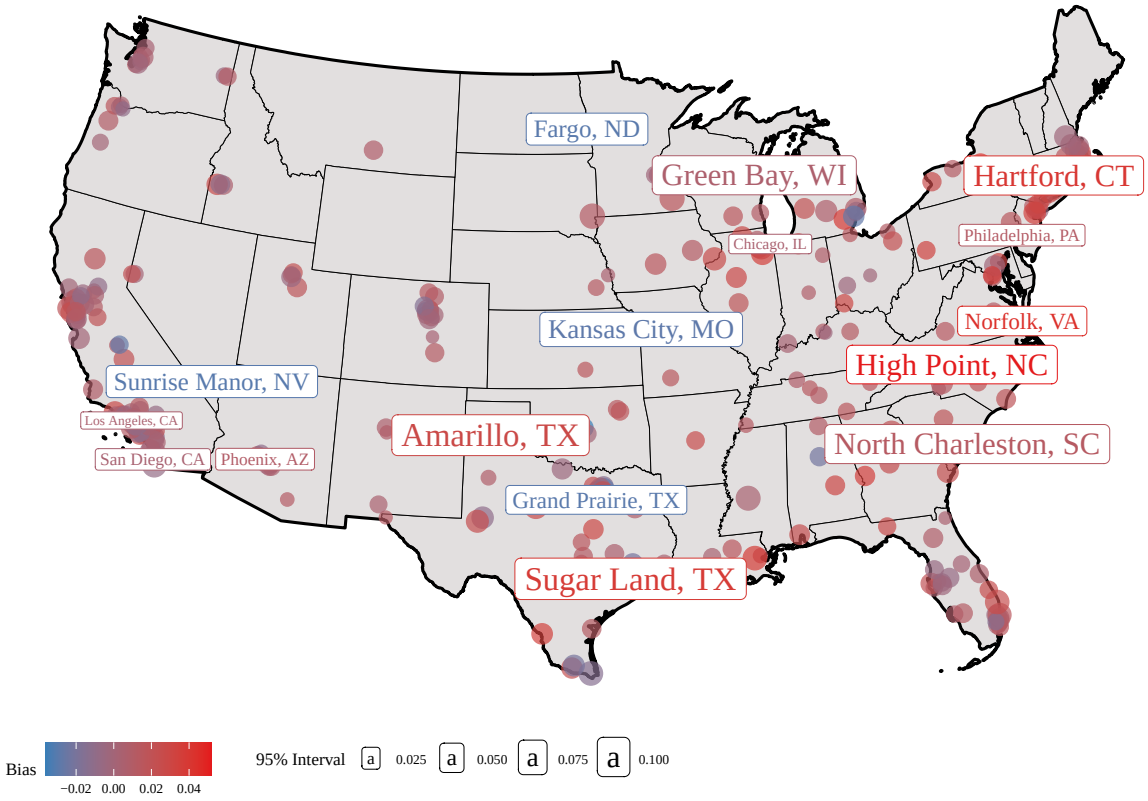

Figure 4: Map of Black-White Dissimilarity Index measurement error across cities. City dots and labels are colored blue-red based on size of the Black-White Dissimilarity Index bias, with darker blue values representing larger (in absolute magnitude) negative values and darker red values representing larger positive values. Dots and labels are sized proportional to the size of the city's 95% interval across simulated plans.

Figure 5 confirms this pattern across our sample by showing that measurement variability is negatively correlated with city population, while bias exhibits no correlation. Interval length is decreasing with logged city population, meaning that variability from aggregation-induced measurement error is predominantly a problem in small to medium-sized cities, and much less severe in the largest cities.

This relationship between uncertainty and city size is likely due to smaller cities having fewer Census tracts. For example, the contribution of each tract i to the Dissimilarity Index is $\left|\frac{a_i}{A_j} - \frac{b_i}{B_j}\right|$. In cities with fewer tracts, each tract contains a larger proportion of the citywide populations A_j and B_j. Therefore, changes in tract boundaries can more significantly influence the proportions $\frac{a_i}{A_j}$ and $\frac{b_i}{B_j}$ when n_j is small. In a city with many tracts, each tract represents a smaller proportion of the total population and any deviations lead to smaller changes in the proportions $\frac{a_i}{A_j}$ and $\frac{b_i}{B_j}$. Furthermore, the relative contribution of any given tract is smaller when n_j is large. By the Law of Large Numbers, as the Dissimilarity Index sums over n_j tracts, as n_j increases the effect of individual tracts tends to average out. Thus, the overall variability in D_{jk} decreases. Other proportion-based measures feature a similar relationship to the number of tracts in the target area.

In the Supporting Information Table 9, we present regression estimates of the relationship between city population and bias and measurement variability, including controls for city proportion White, Black, and

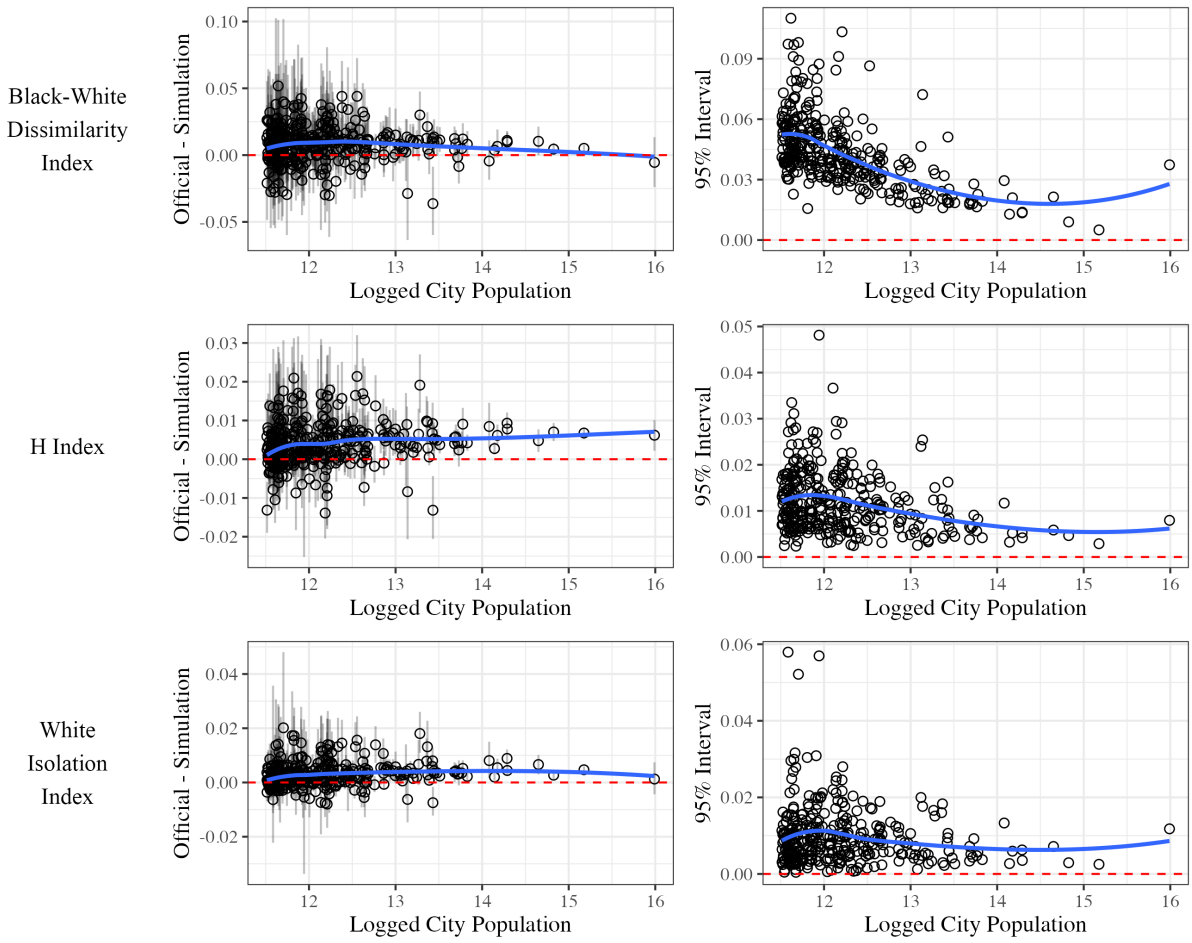

Figure 5: Measurement error by city population. Scatter plots show bias (left) and interval size (right) by logged city population. Blue lines plot the local regression line. Top row shos the Black-White Dissimilarity Index, middle row shows the H Index, and the bottom row shows the White Isolation Index.

Hispanic population and the the average segregation across simulated plans for each city (our best estimate of true segregation in the city). The regression coefficients support the decreasing relationship between measurement variability and city size. We do not find consistent patterns between bias or measurement variability and city demographics. We do see an increasing relationship between bias and measurement variability and how segregated a city is, although the inclusion of quadratic terms for simulation average segregation shows that this increasing relationship has diminishing returns, with the marginal effect decreasing as segregation increases.

6 Segregation trends over time are robust to alternative tract maps

While the previous section focused on the 2020 Decennial Census, we now investigate how our methodology alters understanding of how segregation has changed over time. To do so, we replicate our entire simulation procedure above to redraw Census tract plans for both the 2000 and 2010 Censuses. We then compare, for the 236 cities that had over 100,000 people in 2000, 2010, and 2020, changes in segregation over time in both simulated and official statistics. Figure 6 presents boxplots of simulation (all 5,000 draws for each city) and official (one for each city) segregation index estimates separately for the years 2000, 2010, and 2020.

In each year, the average of official statistics is slightly larger than the average of simulation statistics, particularly for the Black-White Dissimilarity Index. Thus, the differences observed in 2020 is relatively constant across the last three decennial Censuses. We also observe declining segregation from 2000 through 2020, both for official and simulation statistics and for both the Black-White Dissimilarity Index and the H index. Qualitatively, at least, the understanding that racial segregation is declining holds even when accounting for bias and uncertainty from aggregation measurement error. This consistency across decades is likely driven, in part, by the Census Bureau's decision to keep tracts relatively constant over time (even as the underlying population distribution shifts).

We use regression models to estimate exactly how much segregation has declined and to see how these estimates vary across official and simulation statistics over time. Specifically, we model a simulated or official plan's segregation estimate as a function of the year, whether the plan is official or simulated, the interaction of these two variables, and fixed effect terms for city. We fit this model to the set of cities with populations over 100,000 in all three decennial Censuses, stacking the data such that each city in each year has 5,001 associated rows (5,000 simulated plans and 1 official plan), treating each plan as an independent observation (as per the independence assumption in our simulation construction).

Let $\mathbf{D}_{jtk}$ be the Black-White Dissimilarity Index for city j in year t from draw k (where k varies from 1 to 5,000 simulated draws and an "official" draw). Let $\sum_{t=2000}^{2020} I(T = t)$ be a series of 3 indicator variables taking values of 1 if a simulated plan is drawn in year t, 0 otherwise. The terms for the year 2000 are the omitted category in the regression estimation. Let Official_{jtk} be an indicator variable taking a value of 1 if plan k is the official Census tract estimate, 0 otherwise (and thus the plan is a simulation plan). γ_j is the city fixed effect, which constrains estimation to measure over-time and between official versus simulation statistics differences to plans in the same city. We cluster standard errors ϵ_j at the city level to account for correlated uncertainty within cities. We estimate the following linear model:

$$\mathbf{D}_{jtk} = \gamma_j + \sum_{T=2000}^{2020} \beta_T I(T = t) + \theta \text{Official}_{jtk} + \sum_{T=2000}^{2020} \tau_T I(T = t) \times \text{Official}_{jtk} + \epsilon_j \quad (1)$$

From this model, each β_T represents coefficients on whether segregation has increased among simulation estimates between 2000 and 2010, and 2000 and 2020. θ can be interpreted as the different between official and simulation estimates in 2000. The coefficients on the interaction of year and official versus simulation (τ_T) represent the extent to which official statistics differ from simulation statistics in their description of 2000-2010

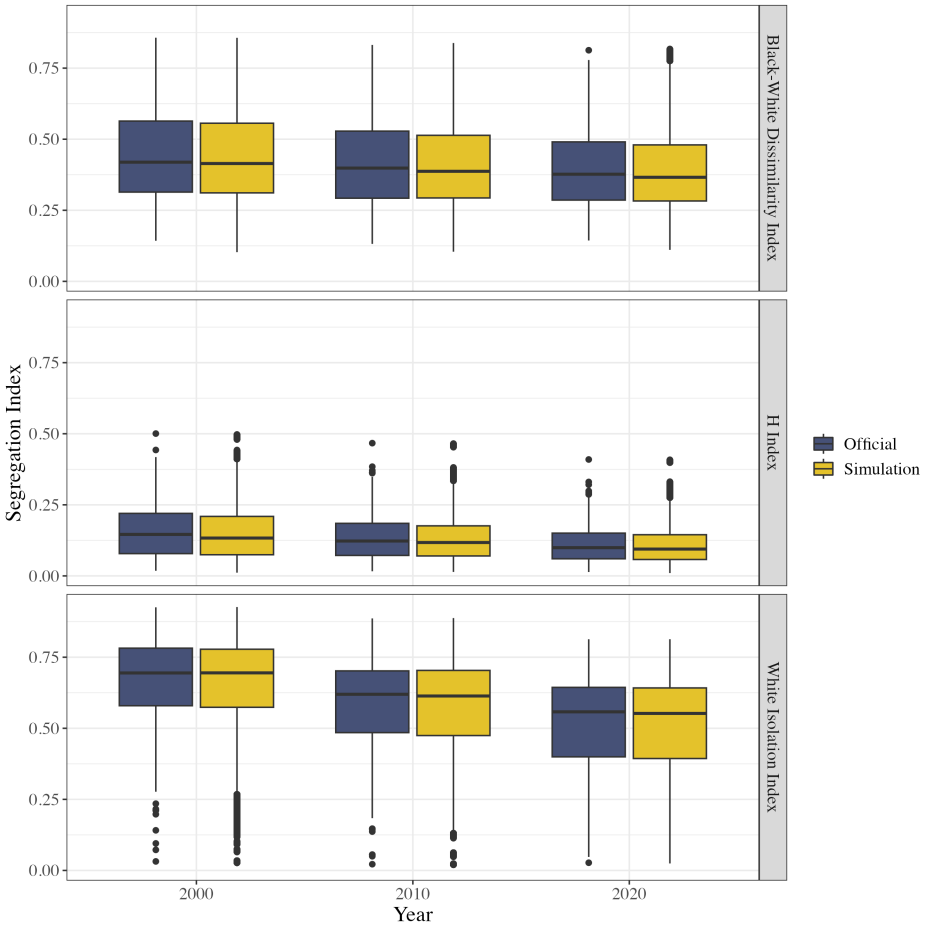

Figure 6: Official vs. simulation segregation, 2000-2020. Figure shows box and whisker plots reporting the distribution of each segregation index across blue (official) and simulated (yellow) plans for the years 2000, 2010, and 2020. Top row shows the Black-White Dissimilarity Index, middle row shows the H Index, and bottom row shows the White Isolation Index.

or 2000-2020 changes. We also estimate equivalent estimates for the H Index and the White Isolation Index, estimating similar equations as above but treating each segregation measure as a separate outcome.

We take a similar approach to assess how measurement variability has changed over time. Specifically, we calculate the deviation of each simulated plan for a given city and year from the average across all plans for that city in that year. Formally, we calculate $\text{Dev}_{jtk} = \sqrt{(\mathbf{D}_{jtk} - \bar{\mathbf{D}}_{jt})^2}$ for each plan. We regress this outcome on indicator variables for each year, including city fixed effects. The coefficients on the year indicator variables tell us whether measurement variability has increased within cities across time. Then, we estimate the following linear model for the same three segregation indices discussed above:

$$\text{Dev}_{jtk} = \gamma_j + \sum_{T=2000}^{2020} \beta_T I(T = t) + \epsilon_j \tag{2}$$

Table 3 reports the estimates from all models. We emphasize two major patterns: first, we find the broad pattern of declining segregation from 2000-2020 is found in both official and simulated estimates. In column (1) which reports coefficients from the model estimating the Black-White Dissimilarity Index, the coefficients for 2010 and 2020 are both negative and significant, showing that segregation has declined in simulation estimates since 2000. The coefficient for 2010 is -0.0273, while the coefficient for 2020 is -0.0451 percentage points. Second, we find the differences between official and simulation estimates of the Black-White Dissimilarity Index are consistently small from 2000 to 2020. The coefficient for the indicator variable measuring whether a plan is an official plan is 0.0077, meaning that official estimates in 2000 overestimate the Black-White Dissimilarity index by 0.77 percentage points. The interaction coefficients on 2010 × Official and 2020 × Official are not statistically distinguishable from zero at conventional significance threshholds ($\alpha < 0.05$).

We observe similar patterns in our estimation for the H Index and the White Isolation Index, with both of these indices decreasing from 2000 to 2020, and with a small upward bias for official statistics compare to simulation averages in 2000. The White Isolation Index model further returns no statistically significant estimates for the interaction coefficients for 2010 × Official and 2020 × Official. For the H Index, however, we do see negative and significant interaction coefficients, indicating that the bias in the H Index has decreased across time.

In columns 4-6 of Table 3, we report results from models estimating the change in measurement variability for each index across time. For the Black-White Dissimilarity Index, we observe no change across time, as the coefficients for 2010 and 2020 are not statistically distinguishable from zero. For both the H Index and the White Isolation Index, however, we find statistically significant negative estimates, showing that the spread of the simulated distributions is decreasing within cities across time. Therefore, measurement variability for these indices is smaller in 2010 and 2020 than it was in 2000. The 2020 coefficients are larger in absolute magnitude than the 2010 coefficients, indicating that measurement variability further declined from 2010 to 2020.

7 Extreme distortions in segregation measures are possible, but rare

Our previous analyses demonstrate that the bias and variability from MAUP in Census segregation statistics is relatively small. We arrive at this conclusion by measuring differences between enacted plans and alternative simulated plans drawn under Census drawing guidelines to respect targets for population, compactness, and both official and natural boundaries. However, it is difficult to assess the practical magnitude of these differences by comparing official and simulated tracts alone. This is because our simulated plans, which respect Census guidelines, are designed to approximate real tracts using realistic constraints.

We have shown that official tract maps do not produce systematically different segregation estimates than possible alternatives, but it remains unclear whether these differences are large or small relative to **possible**

Table 3: Modeling over-time changes in official versus simulation segregation.

	Segregation			Deviation		
	Black-White Dissimilarity Index	H Index	White Isolation Index	Black-White Dissimilarity Index	H Index	White Isolation Index
	(1)	(2)	(3)	(4)	(5)	(6)
2010	-0.0273***	-0.0186***	-0.0795***	0.0006	-0.0232***	-0.0356***
	(0.0030)	(0.0018)	(0.0030)	(0.0018)	(0.0014)	(0.0023)
2020	-0.0451***	-0.0432***	-0.1476***	0.0040	-0.0354***	-0.0671***
	(0.0042)	(0.0027)	(0.0044)	(0.0029)	(0.0022)	(0.0036)
Official	0.0077***	0.0069***	0.0035***			
	(0.0014)	(0.0007)	(0.0004)			
2010 × Official	0.0002	-0.0017***	-0.0001			
	(0.0012)	(0.0005)	(0.0003)			
2020 × Official	0.0017	-0.0021***	-0.0002			
	(0.0012)	(0.0006)	(0.0004)			
Observations	3,540,708	3,540,708	3,540,708	3,540,000	3,540,000	3,540,000
R^2	0.95342	0.95181	0.97624	0.85493	0.95375	0.96011
Within R^2	0.25782	0.48275	0.80849	0.00593	0.47834	0.56794
City FE	✓	✓	✓	✓	✓	✓

Column 1-3 report results from regressions modeling segregation estimates across time and official versus simulation maps. Columns 4-6 report results from regression modeling deviations from city-level averages for each simulated plan across time. Unit of analysis is the a given plan (official or one of 5,000 simulated plans for each city). Standard errors are clustered at the city-level. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

differences in segregation measures for a given geography. In this section, we assess the magnitude of these differences by comparing them to estimates of the maximum and minimum possible segregation to observe in a given area. Conventional wisdom suggests that maps could be drawn that dramatically overstate or understate segregation, especially if certain constraints are less strongly respected.

To understand how large these distortions could be, but also to demonstrate just how bizarre maps would have to be to yield majorly different segregation results, we conduct a case study of New Jersey where we optimize the minimum and maximum segregation across Census tracts in each county. We use a heuristic optimization approach which uses algorithmic trials called "short-bursts" to find increasingly more (or less) segregated maps (Cannon et al., 2023; Simko, 2024). Each short-burst is a short run of the Merge-Split algorithm described above. After each burst, the plan with the highest score (here the highest or lowest value of a Black-White dissimilarity index), is selected as the starting point for the next burst. This process is run 100,000 times, allowing us to capture plans which approximate the most and least segregated plans that are reasonably possible. Note that the algorithm is not guaranteed to locate the absolute global maximum or minimum, but works well in practice.

This algorithmic approach allows for all outputted tracts to still be contiguous and respect population constraints. The purpose of this exercise, unlike above, is not to draw realistically feasible maps. Instead, we use a smaller set of constraints to benchmark our more realistic estimates against possible, if less realistic, alternatives. As such, we do not include a compactness constraint, as intuitively the algorithm will have more flexibility to minimize or maximize racial separation between tracts under looser geographic constraints. Further, we do not encourage the algorithm to respect administrative boundaries for a similar reason. We run this algorithm separately for each county in New Jersey using the procedure described in Section 4.

Across nearly all counties, we find the range of possible segregation produced by short-bursts is considerably larger than our realistic alternative maps. This provides additional evidence that, in most cases, differences due to the kind of aggregation bias we examine here are relatively small. As Figure 7 demonstrates, the range of dissimilarity scores produced by the short-burst approach is very wide. In all counties, the range between the largest and smallest dissimilarity scores is at least 0.135 (on a possible 0-1 scale). The range is generally much larger for smaller counties, with Cumberland County having a range of 0.494. Across all counties, this range is considerably larger than the range of dissimilarity scores produced by the more realistic simulated plans used in our analyses above. To put this in perspective, the average ratio of the difference between the optimized dissimilarities to the corresponding difference between the two most extreme simulated plans is 4.49. The smallest ratio is 2.85 in Cape May County and the largest ratio is 6.39 in Ocean County. As we cannot guarantee that the short-burst approach finds the global maximum or minimum, these ratios should be interpreted as lower bounds on the potential range of dissimilarity scores.

Figure 7 also shows that the ability to reduce segregation is generally greater than the ability to increase it. This is shown by the larger distances from the observed segregation (grey) distribution to the lower bound than to the upper bound. This is likely due to several compounding factors. First, since Census blocks are for the most part not themselves racially homogeneous, it is easier to construct a completely balanced Census tract with respect to county White-Black population baselines than to create a entirely White or Black Census tract (which is impossible unless all the Census blocks are entirely White or Black themselves). Second, the Census guidelines may force homogeneous tracts to exist in official maps—and in our Merge-Split simulated distributions—so segregation may be closer to maximum potential values than to minimum potential value due to objective clustering of racial groups. Lastly, if sharp racial boundaries across official tracts exist, segregation can be reduced by crossing that sharp boundary with alternative tracts. As a result, there is a larger sample space wherein segregation can be reduced rather than increased on average. The impact of each of these factors would be further compounded in more populous counties, since more tracts must be drawn and it becomes harder to ensure that all of these tracts are sufficiently homogeneous to increase segregation to a great extent.

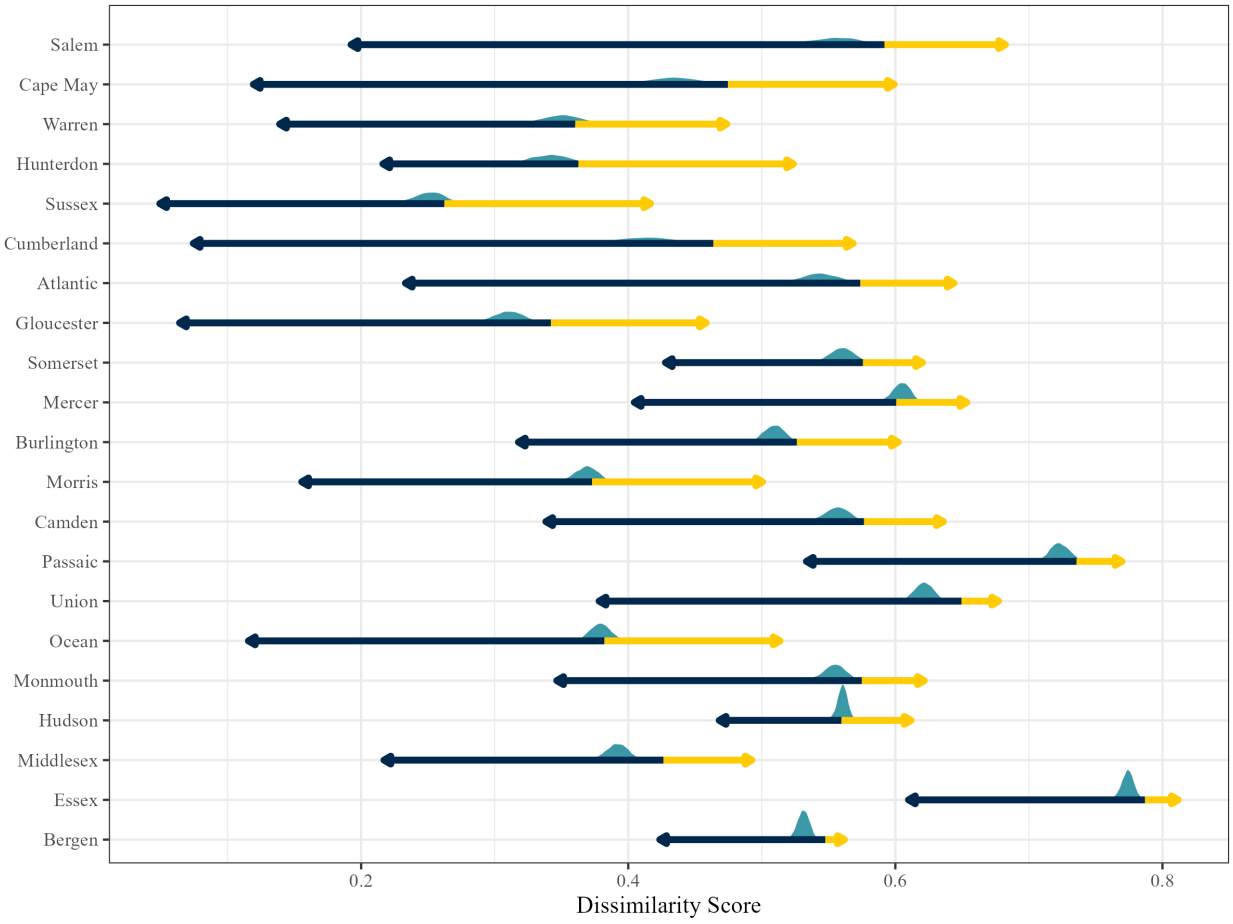

Figure 7: Optimized, simulated, and official segregation for New Jersey Counties. The teal arrow pointing to the left represents the range traversed by the minimizing short-burst, while the orange arrow pointing to the right represents the range traversed by the maximizing short-burst. The two arrows intersect at the dissimilarity score of the official tract map. The purple slabs show the distribution of the simulated plans. Counties are arranged by population, with largest population at the bottom.

The potential range of these dissimilarity scores aligns well with the conventional wisdom that the MAUP can cause large differences in estimates from alternative aggregations. However, the constraints we include above to make alternative tracts realistic considerably reduce the range of possible estimates. In our application, consider what it takes to make these scores maximize or minimize. First, compactness has to be allowed to drop significantly. Second, administrative boundaries have to be ignored. This allows for tracts to be drawn in ways which compromise communities or neighborhoods. In many cases, when maximizing segregation, the algorithm has to draw tracts that sit near the lower population bound. As Figure 8 shows, the optimized tract maps for Gloucester County, NJ, are quite extreme. To maximize or minimize the segregation, the algorithm draws noncompact tracts with long and jagged. In comparison, the official tract map boundaries are smooth, more rounded, and generally follow natural and administrative boundaries.

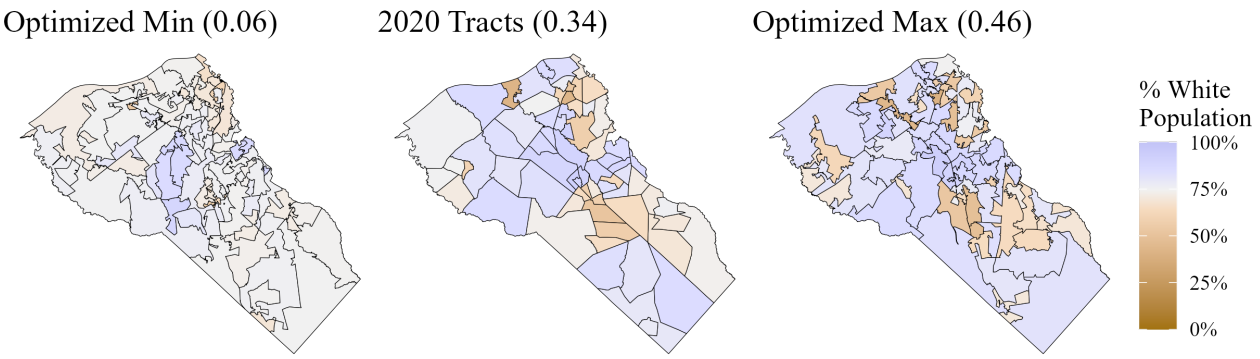

Figure 8: Example extreme and official tract maps for Gloucester County, NJ. The left and right maps show the optimized minimum and maximum dissimilarity maps. The middle map shows the official tract map for 2020. The fill for the maps shows the percentage white, with purple indicating that a tract is more white than average and orange indicating that a tract is less white than average. Lighter shades indicate that a precinct is closer to the average.

Altogether, this shows that aggregation biases from the MAUP are certainly possible, but only at great cost to feasibility constraints. Our original simulations that obey these constraints have a much narrower range than the range of all possible plans produced by short-burst. If tracts are intended to be used in analyses as some proxy for neighborhoods or other local communities, then the optimized plans show worst cases, but not reasonable alternatives of how tracts could have been redesigned. Instead, the plans drawn by Merge-Split, which respect compactness and administrative boundaries, likely serve as a better proxy for how the MAUP impacts measurements of segregation. This contextualizes our finding in Section 5 that biases and measurement variability is small in most cases, but larger in smaller cities.

8 Conclusion

Measurement is the foremost challenge of empirical social science research. Complex quantities of interest are difficult to estimate using available data, which can distort analyses of the drivers and consequences of social phenomena. Segregation is a prime example of these difficulties: researchers must organize continuous information in geographic space into aggregate buckets that distort true underlying spatial relationships. A long history of existing work has documented these challenges and offered alternative ways to measure segregation when facing them.

In this article, we investigate how a long-standing problem in geographic research—the Modifiable Areal Unit Problem—impacts the measurement of segregation using Census data. Using innovations in algorithmic redistricting software, we quantify the bias and uncertainty that aggregation choices produce in common segregation metrics that rely on official Census data. We simulate the data generating process for creating Census tract plans and estimate the sampling distribution of alternative ways that racial populations could be grouped into alternative tracts under Census guidelines. From these analyses, we derive bias-corrected estimates of segregation and measure the variability underlying these estimates.

This analysis reveals that official Census tract definitions yield estimates of segregation that well-represent racial residential segregation across U.S. cities. Looking at cities with populations over 100,000, we find that the Black-White Dissimilarity Index is biased upward in official Census estimates by a slight amount: 0.81 percentage points, or 2.37%. The Dissimilarity Index can be interpreted as the proportion of the minority group who would have to move across Census tracts to achieve complete integration. So based on our estimates, 0.81% fewer people need to move to achieve complete integration.

However, our analyses also suggest that segregation metrics are more sensitive to measurement error from the MAUP in smaller areas with fewer Census tracts. Across several analyses, we find that measurement variability is most pronounced in small and medium sized cities, like Richmond, CA or Elgin, IL. In contrast, the largest U.S. cities like Los Angeles, CA and Chicago, IL exhibit only small amounts of bias and uncertainty. The correlation between variability and city population is likely due to how consequential each Census tract decision is relative to the other Census tracts drawn in a given plan for a given city. In a large city, cracking or packing racial groups in one area is more likely to average out across many many neighborhoods. In cities with fewer Census tracts, each boundary decision represents a large proportion of the total boundary decisions made, and is thus more consequential. This dynamic has been observed in redistricting research as well, where at small scales there is potential for large distortions, but across large geographic scales biases tend to cancel out (Kenny et al., 2023).

We recommend researchers should be most concerned about aggregation measurement error in contexts where they have few geographic units across which to compare. This heterogeneity in measurement error by city population suggests an important relationship between the horizontal and vertical dimensions of the MAUP. When researchers choose geographic subunits to measure segregation, the choice of larger (and by definition fewer) units in a city will increase the potential for vastly different plans, and thus for greater aggregation measurement error to emerge. Conceptually, this is similar to statistical sampling uncertainty—with fewer sampled units, estimates are less certain. Conversely, choosing smaller geographic sub-unit means that more units must be redrawn to cause large changes, making vastly different plans less likely. This is potentially counterintuitive, since our bounding analysis in New Jersey shows that large changes are *possible* in small and large areas alike. However, while these vastly different measures are often possible, our sampling results show they are unlikely.

Some of our analyses point to contexts where official segregation metrics are more sensitive to measurement error. First, we find that measurement variability is most pronounced in small and medium sized cities, and the largest U.S. cities exhibit only small amounts of bias and uncertainty. Measurement error is more of a problem in places like Richmond, CA or Elgin, IL than in Los Angeles or Chicago. This relationship with city size is likely due to how consequential each Census tract decision is relative to the other Census tracts drawn in a given plan for a given city. In a large city, cracking or packing racial groups in one area is more likely to average out across many many neighborhoods. In cities with fewer Census tracts, each boundary decision represents a large proportion of the total boundary decisions made, and is thus more consequential. This dynamic has been observed in redistricting research as well, where at small scales there is potential for large distortions, but across large geographic scales biases tend to cancel out (Kenny et al., 2023).

This heterogeneity in measurement error by city population informs as to the relationship between the MAUP and the related problem of scale—i.e. the choice of which geographic unit to use a sub-geography—when

measuring segregation (White, 1983). The choice of larger (and by definition fewer) units in a city to measure evenness across will increase the potential for different plans to be drawn, and thus for greater aggregation measurement error to emerge. Conversely, a smaller unit means that more units must be redrawn, and aggregation problems are more likely to average. So researchers should be most concerned about aggregation measurement error in contexts where they have few geographic units across which to compare.

In total, our analyses contextualize commonly used segregation statistics by quantifying measurement error that is often not well understood and even ignored in applied research. Our findings to nod fundamentally alter current understanding of the state of racial segregation in the United States—we supplement existing work by finding that segregation is high and, despite decreasing across the past two decades, persistent.

Our findings are reassuring for researchers calculating segregation statistics, especially in areas with large population. When measurement bias bias and variability is correlated with other variables used in analyses for predicting segregation or measuring its effects, then severe problems may emerge (Knox et al., 2022; McCartan et al., 2023; Egami et al., 2023). Even classical measurement error—uncorrelated noise in variable measures, often the least harmful of measurement errors—is shown to downward bias effect estimates (Angrist and Pischke, 2008). If variability is correlated with variables of interest, then estimation of what drives segregation, or what it influences, will be even further biased. Our results provide comprehensive evidence that studies that rely on aggregate Census definitions to measure segregation are robust to such concerns.

References

Alesina, A., Baqir, R., and Easterly, W. (1999). Public goods and ethnic divisions. *The Quarterly Journal of Economics*, 114(4):1243–1284.

Ananat, E. O. (2011). The wrong side(s) of the tracks: The causal effects of racial segregation on urban poverty and inequality. *American Economic Journal: Applied Economics*, 3(2):34–66.

Angrist, J. and Pischke, J. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Ansolabehere, S., Brown, J. R., Enos, R. D., Shair, B., Simko, T., and Sutton, D. (2025). City-defined neighborhood boundaries in the united states. *Scientific Data*, 12(1):1031.

Arjona, A. (2019). Subnational units, the locus of choice, and concept formation. *Inside Countries: Subnational Research in Comparative Politics*, 214:214–242.

Autry, E., Carter, D., Herschlag, G., Hunter, Z., and Mattingly, J. C. (2019). Metropolized forest recombination for monte carlo sampling of graph partitions. *arXiv preprint arXiv:1911.01503*.

Bisbee, J. and Zilinsky, J. (2023). Geographic boundaries and local economic conditions matter for views of the economy. *Political Analysis*, 31(2):288–294.

Bischoff, K. and Reardon, S. F. (2014). Residential segregation by income, 1970–2009. *Diversity and disparities: America enters a new century*, 43.

Brown, J. R., Cantoni, E., Enos, R. D., Pons, V., and Sartre, E. (2024). A micro-level analysis of the increase and contributing factors of geographic partisan segregation. Working Paper, Harvard University.

Brown, J. R. and Enos, R. D. (2021). The measurement of partisan sorting for 180 million voters. *Nature Human Behaviour*.

Cannon, S., Goldbloom-Helzner, A., Gupta, V., Matthews, J., and Suwal, B. (2023). Voting rights, markov chains, and optimization by short bursts. *Methodology and Computing in Applied Probability*, 25(1):36.

Carter, D., Herschlag, G., Hunter, Z., and Mattingly, J. (2019). A merge-split proposal for reversible monte carlo markov chain sampling of redistricting plans. Technical report, Working Paper, Duke University.

Chen, J. and Rodden, J. (2013). Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*.

Chen, X., Ye, X., Widener, M. J., Delmelle, E., Kwan, M.-P., Shannon, J., Racine, E. F., Adams, A., Liang, L., and Jia, P. (2022). A systematic review of the modifiable areal unit problem (maup) in community food environmental research. *Urban Informatics*, 1(1):22.

Chetty, R., Hendren, N., Jones, M. R., and Porter, S. R. (2019). Race and Economic Opportunity in the United States: an Intergenerational Perspective*. *The Quarterly Journal of Economics*, 135(2):711–783.

Deford, D., Duchin, M., and Solomon, J. (2019). Recombination: A family of markov chains for redistricting. Technical report, Working Paper, Tufts University.

DeFord, D., Duchin, M., and Solomon, J. (2021). Recombination: A family of Markov chains for redistricting. *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/1ds8ptxu.

Dmowska, A. and Stepinski, T. F. (2024). Quantification and visualization of us racial geography using the national racial geography dataset 2020. *PLOS ONE*, 19(7):1–19.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217.

Egami, N., Jacobs-Harukawa, M., Stewart, B. M., and Wei, H. (2023). Using large language model annotations for valid downstream statistical inference in social science: Design-based semi-supervised learning.

Einstein, K., Glick, D., and Palmer, M. (2020). *Neighborhood Defenders: Participatory Politics and America's Housing Crisis*. Cambridge University Press.

Elbers, B. (2021). Trends in u.s. residential racial segregation, 1990 to 2020. *Socius*, 7:23780231211053982.

Enos, R. D. (2017). *The Space Between Us: Social Geography and Politics*. Cambridge University Press, New York.

Fifield, B., Imai, K., Kawahara, J., and Kenny, C. T. (2020). The essential role of empirical validation in legislative redistricting simulation. *Statistics and Public Policy*, 7(1):52–68.

Fischer, C. S., Stockmayer, G., Stiles, J., and Hout, M. (2004). Distinguishing the geographic levels and social dimensions of u.s. metropolitan segregation, 1960–2000. *Demography*, 41(1):37–59.

Fotheringham, A. S. and Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A: Economy and Space*, 23(7):1025–1044.

Gatrell, A. (1983). *Distance and Space: A Geographical Perspective*. Contemporary problems in geography. Clarendon Press.

Gehlke, C. E. and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29:169–170.

Habyarimana, J., Humphreys, M., Posner, D. N., and Weinstein, J. M. (2009). *Coethnicity: Diversity and the Dilemmas of Collective Action*. Russell Sage, New York.

Hennerdal, P. and Nielsen, M. M. (2017). A multiscalar approach for identifying clusters and segregation patterns that avoids the modifiable areal unit problem. *Annals of the American Association of Geographers*, 107(3):555–574.

Hotchkiss, M. and Phelan, J. (2017). Uses of census bureau data in federal funds distribution: A new design for the 21st century. Technical report, U.S. Census Bureau, Washington, D.C. Census Federal Funds Report.

Hwang, J. and McDaniel, T. W. (2022). Racialized reshuffling: Urban change and the persistence of segregation in the twenty-first century. *Annual Review of Sociology*, 48(Volume 48, 2022):397–419.

Jahn, J., Schmid, C. F., and Schrag, C. (1947). The measurement of ecological segregation. *American Sociological Review*, 12(3):293–303.

Jelinski, D. E. and Wu, J. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11(3):129–140.

Kaplan, E., Spenkuch, J. L., and Sullivan, R. (2022). Partisan spatial sorting in the united states: A theoretical and empirical overview. *Journal of Public Economics*, 211:104668.

Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T. R., Simko, T., and Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 u.s. census. *Science Advances*, 7(41):eabk3283.

Kenny, C. T., McCartan, C., Fifield, B., and Imai, K. (2024a). redist: Simulation methods for legislative redistricting. Available at The Comprehensive R Archive Network (CRAN).

Kenny, C. T., McCartan, C., Kuriwaki, S., Simko, T., and Imai, K. (2024b). Evaluating bias and noise induced by the us census bureau's privacy protection methods. *Science Advances*, 10(18):eadl2524.

Kenny, C. T., McCartan, C., Simko, T., Kuriwaki, S., and Imai, K. (2023). Widespread partisan gerrymandering mostly cancels nationally, but reduces electoral competition. *Proceedings of the National Academy of Sciences*, 120(25):e2217322120.

Knox, D., Lucas, C., and Cho, W. K. T. (2022). Testing causal theories with learned proxies. *Annual Review of Political Science*, 25(1):419–441.

Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., and O'Sullivan, D. (2008). Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73(5):766–791. PMID: 25324575.

Lee, D. W. and Rogers, M. (2019). Measuring geographic distribution for political research. *Political Analysis*, 27(3):263–280.

Lee, D. W., Rogers, M., and Soifer, H. D. (2025). The modifiable areal unit problem in political science. *Political Analysis*, pages 1–13.

Legewie, J. and Schaeffer, M. (2016). Contested boundaries: Explaining where ethnoracial diversity provokes neighborhood conflict. *American Journal of Sociology*, 122(1):125–161.

Logan, T. D. and Parman, J. M. (2017). The national rise in residential segregation. *The Journal of Economic History*, 77(1):127–170.

Massey, D. S. and Denton, N. A. (1988). The dimensions of residential segregation. *Social Forces*, 67(2):281–315.

Massey, D. S. and Denton, N. A. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Havard University Press, Cambridge, MA.

Mazza, A. and Punzo, A. (2015). On the upward bias of the dissimilarity index and its corrections. *Sociological Methods & Research*, 44(1):80–107.

McCartan, C., Goldin, J., Ho, D. E., and Imai, K. (2023). Estimating racial disparities when race is not observed. *arXiv preprint*, page 2303.02580.

McCartan, C. and Imai, K. (2023). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *The Annals of Applied Statistics*, 17(4):3300 – 3323.

Mijs, J. J. B. and Roe, E. L. (2021). Is america coming apart? socioeconomic segregation in neighborhoods, schools, workplaces, and social networks, 1970–2020. *Sociology Compass*, 15(6):e12884.

Openshaw, S. (1983). *Concepts and Techniques in Modern Geography*, chapter The Modifiable Areal Unit Problem. Geo Books, Norwich, UK.

Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century: The 2006 johan skytte prize lecture. *Scandinavian Political Studies*, 30(2):137–174.

Reardon, S. F. and Bischoff, K. (2011). Income inequality and income segregation. *American Journal of Sociology*, 116(4):1092–1153.

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

Reardon, S. F., Matthews, S. A., O'Sullivan, D., Lee, B. A., Firebaugh, G., Farrell, C. R., and Bischoff, K. (2008). The geographic scale of metropolitan racial segregation. *Demography*, 45(3):489–514.

Reardon, S. F. and O'Sullivan, D. (2004). Measures of spatial segregation. *Sociological Methodology*, 34(1):121–162.

Roberto, E. (2018). The spatial proximity and connectivity method for measuring and analyzing residential segregation. *Sociological Methodology*, 48(1):182–224.

Rodden, J. (2019). *Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide*. Basic Books.

Rothstein, R. (2017). *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing Corporation, a division of W.W. Norton & Company, New York ; London, first edition. edition. HOLLIS number: 990149136710203941.

Sang-Il Lee, Monghyeon Lee, Y. C. and Griffith, D. A. (2019). Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: a simulation study. *International Journal of Geographical Information Science*, 33(6):1135–1154.

Sharkey, P. and Faber, J. W. (2014). Where, when, why, and for whom do residential contexts matter? moving away from the dichotomous understanding of neighborhood effects. *Annual review of sociology*, 40(1):559–579.

Simko, T. (2024). School desegregation by redrawing district boundaries. *Scientific Reports*, 14(1):22097.

Theil, H. (1967). *Economics and Information Theory*. Studies in mathematical and managerial economics. North-Holland Publishing Company.

Trounstine, J. (2016). Segregation and inequality in public goods. *American Journal of Political Science*, 60(3):709–725.

Trounstine, J. (2018). *Segregation by Design: Local Politics and Inequality in American Cities*. Cambridge University Press.

U.S. Census Bureau (201). What are census blocks? Technical report.

U.S. Census Bureau (2023). Geographic areas reference manual. Technical report.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved r^ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2).

White, M. J. (1983). The measurement of spatial segregation. *American journal of sociology*, 88(5):1008–1018.

Wildfang, H. (2025). The maup effect: Spatial scale and the reliability of segregation indices. *Survey Methods: Insights from the Field*. Special issue 'Advancing Comparative Research: Exploring Errors and Quality Indicators in Social Research'.

Wong, C., Bowers, J., Williams, T., and Drake, K. (2012). Bringing the person back in: Boundaries, perceptions, and the measurement of racial context. *Journal of Politics*, 1(1):1–18.

Wong, D. W. (1997). Spatial dependency of segregation indices. *The Canadian Geographer*, 41(2):128–136.

Wong, D. W. (2004a). Comparing traditional and spatial segregation measures: a spatial scale perspective1. *Urban Geography*, 25(1):66–82.

Wong, D. W. S. (2004b). The modifiable areal unit problem (maup). In Janelle, D. G., Warf, B., and Hansen, K., editors, *WorldMinds: Geographical Perspectives on 100 Problems*, pages 571–575. Springer, Dordrecht.

Östh, J., Malmberg, B., and Andersson, E. K. (2014). *Seven: Analysing segregation using individualised neighbourhoods*, pages 135 – 162. Policy Press, Bristol, UK.

Table 4: $\hat{R}$ quantiles - 2020

	$\hat{R}$					
		Quantile				
Stat	Avg.	0.05	0.25	0.50	0.75	0.95
Compactness	1.020	1.000	1.000	1.001	1.009	1.117
Deviation	1.002	1.000	1.000	1.000	1.001	1.005
Black-White Dissim.	1.004	1.000	1.000	1.001	1.003	1.019
Place Splits	1.002	1.000	1.000	1.000	1.001	1.007
Black Pop.	1.005	1.000	1.000	1.001	1.004	1.023
Hispanic Pop.	1.004	1.000	1.000	1.001	1.003	1.019
White Pop.	1.003	1.000	1.000	1.001	1.002	1.012

Table reports the mean and quantiles of $\hat{R}$ for simulation convergence across counties.

A Simulation statistics

A.1 Convergence

Here, we present $\hat{R}$ metrics evaluating convergence for our simulations. To do so, we examine district and city-level statistics and calculate the ratio of between chain and within chain variance in these average outcomes across iterations for each chain (Vehtari et al., 2021). The $\hat{R}$ measures whether chains are sampling from the same areas of the latent distribution (not drawing totally different plans). Values closer to 1 indicate better convergence and generally values at or below 1.05 are commonly accepted benchmark.

We calculate these statistics for compactness, deviation, Black-White dissimilarity, place splits, tract Black population, tract Hispanic population, and tract White population. For tract-specific statistics we compare average statistics across the first Tract drawn in each simulation.

Table 4 reports the average $\hat{R}$ for each statistics and the quantiles (5th, 25th, 50th, 75th, and 95th) of the $\hat{R}$ distribution across 2020 county simulations. Average and median values are all at or below and 1.05 and even the 95th percentiles are all belwo 1.05, excpet for compactness (1.117, although the average value is 1.001). Similarly levels of convergence are observed for 2010 (Table 5) and 2000 (Table 6) simulations.

A.2 Comparison of compactness, population deviation and place splitting

Table 7 and Table 8 present equivalent comparisons of deviation, compactness, and place splits across official and simulation tract maps as Table 1 in the manuscript, but for the 2010 and 2000 simulations, respectively.

A.3 How segregation is commonly measured

In many applications, aggregation issues like the MAUP are ignored because researchers lack proper tools to diagnose how sensitive measurements are to aggregation challenges. In the United States, most studies of segregation use racial demographic data from the U.S. decennial census or American Community Survey. Local segregation is then calculated by metrics that compare how different racial group populations are distributed across sub-geographies (usually Census tracts). These metrics often focus on what Massey and Denton (1988) refer to as evenness, the dimension of segregation encompassing how equally (or unequally) racial groups are

Table 5: $\hat{R}$ quantiles - 2010

	$\hat{R}$					
		Quantile				
Stat	Avg.	0.05	0.25	0.50	0.75	0.95
Compactness	1.018	1.000	1.000	1.001	1.003	1.121
Deviation	1.005	1.000	1.000	1.000	1.001	1.005
Black-White Dissim.	1.004	1.000	1.000	1.001	1.001	1.015
Place Splits	1.001	1.000	1.000	1.000	1.001	1.003
Black Pop.	1.004	1.000	1.000	1.001	1.002	1.014
Hispanic Pop.	1.004	1.000	1.000	1.001	1.002	1.013
White Pop.	1.003	1.000	1.000	1.001	1.001	1.009

Table reports the mean and quantiles of $\hat{R}$ for simulation convergence across counties.

Table 6: $\hat{R}$ quantiles - 2000

	$\hat{R}$					
		Quantile				
Stat	Avg.	0.05	0.25	0.50	0.75	0.95
Compactness	1.015	1.000	1.000	1.001	1.002	1.104
Deviation	1.005	1.000	1.000	1.000	1.001	1.006
Black-White Dissim.	1.005	1.000	1.000	1.001	1.001	1.017
Place Splits	1.001	1.000	1.000	1.000	1.001	1.004
Black Pop.	1.004	1.000	1.000	1.001	1.002	1.015
Hispanic Pop.	1.004	1.000	1.000	1.001	1.002	1.013
White Pop.	1.004	1.000	1.000	1.001	1.001	1.009

Table reports the mean and quantiles of $\hat{R}$ for simulation convergence across counties.

Table 7: Comparison of compactness, population deviation and place splitting - 2010

	Official (N=2,903)		Simulation (N=14,511,000)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Dev.	0.662	0.535	0.492	0.408	-0.170	0.010
Comp. Frac.	0.972	0.016	0.963	0.019	-0.009	0.000
Place Splits	3.907	7.465	4.208	8.128	0.301	0.139

Table reports the average and standard deviation of deviation, compactness, and place splits across official and simulated plans. Difference of means between official and simulated plans are reported in the fifth and sixth columns.

Table 8: Comparison of compactness, population deviation and place splitting - 2000

	Official (N=2,932)		Simulation (N=14,660,000)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Dev.	0.671	0.512	0.479	0.392	-0.192	0.009
Comp. Frac.	0.971	0.017	0.960	0.020	-0.011	0.000
Place Splits	4.416	8.303	4.550	8.597	0.135	0.153

Table reports the average and standard deviation of deviation, compactness, and place splits across official and simulated plans. Difference of means between official and simulated plans are reported in the fifth and sixth columns.

distributed across geographic units. Other dimensions of segregation include isolation[8] (extent to which group members live mostly around other in-group members), clustering (are homogeneous neighborhoods located close to one another), concentration (how much land area do different groups inhabit), and centralization (how close to the geographic center do minority groups live) (Massey and Denton, 1988). For the most part, when researchers measure segregation they focus on either isolation or evenness metrics (Hwang and McDaniel, 2022). The most common measure of evenness is the Dissimilarity Index, which measures two-group segregation and can be interpreted as what proportion of the minority group in the city would need to move to achieve complete integration between two groups (proportions of group in each unit that match the proportions of those groups in the city as a whole) (Duncan and Duncan, 1955).

Let $\mathbf{D}_j$ be the Dissimilarity Index for a given city j with n Census tract. Let A_j and B_j be the populations of Group A and B, respectively in the city, and a_i and b_i be the population of Groups A and B in each Census tract such that $\sum_{i=1}^{n} a_i = A_j$ and $\sum_{i=1}^{n} b_i = B_j$. The Dissimilarity Index is calculated by:

$$\mathbf{D}_j = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{a_i}{A_j} - \frac{b_i}{B_j} \right| \tag{3}$$

The Dissimilarity-Index only calculates segregation between two groups. Given the racial history of the U.S., historically, this index has been commonly applied to measure Black-White segregation. As the U.S. diversifies, particularly with large increases in Hispanic populations, many segregation studies increasingly focus on multi-group segregation (Reardon and Firebaugh, 2002). These indices measure diversity across multiple groups and compare diversity across (typically) Census tracts. The most commonly used of these is the entropy-based H Index, developed by Theil (1967). This metric measures the entropy score (measuring multi-group diversity) of each Census tract and then compares the weighted average of Census tract entropy to the entropy of the city as a whole.

Let E_i be the entropy for neighborhood i, and E_j be the entropy for city j. For each racial group r, let p_{ri} be the population of racial group r in neighborhood i, and let P_{rj} be the population of racial group r in the city. Let p_i be the population of neighborhood i and let P_j be the population of the city, such that $\sum_{r=1}^{R} p_{ri} = p_i$ and $\sum_{i=1}^{n} \sum_{r=1}^{R} p_{ri} = \sum_{i=1}^{n} p_i = P_j$. Neighborhood entropy, city entropy, and the H Index are calculated through the following three equations:

$$\mathbf{E}_i = \sum_{r=1}^{R} \frac{p_{ri}}{p_i} \times \mathbf{ln}\left(\frac{p_i}{p_{ri}}\right) \tag{4}$$

[8]The inverse of isolation is exposure, which is defined as on average how many people of group B live in the same Census tract as the typical person in Group A.

$$\mathbf{E}_j = \sum_{r=1}^{R} \frac{P_{rj}}{P_j} \times \mathbf{ln}\left(\frac{P_j}{P_{rj}}\right) \tag{5}$$

$$\mathbf{H}_j = \sum_{i=1}^{n} \frac{p_i}{P_j}\left(\frac{E_j - E_i}{E_j}\right) \tag{6}$$

We also estimate measures of racial isolation, defined as the extent to which members of different racial groups live in Census tracts comprised of members of their own racial group. To calculate isolation, let I_{aj} be the racial isolation of racial group a in city j. Let t_i be the total population of neighborhood i. We define racial isolation of as follows:

$$\mathbf{I}_{aj} = \sum_{i=1}^{n} \frac{a_i}{A_j}\frac{a_i}{t_i} \tag{7}$$

Racial isolation measures how exposed in the average member of a given group is to in-group members, using the Census tract as the definition of neighborhood exposure. Thus, it is a function of overall diversity (in a homogeneous city, the majority group will have high isolation and the minority group will have low isolation regardless of spatial distribution of these groups) and the spatial distribution of groups across the city (a segregated city will increase isolation). We calculate racial isolation for Whites, Blacks, and Hispanics. The results across these groups are similar and we focus on White isolation in our main analyses but present results for other groups in the Supporting Information Section B.2.

In the main text, we focus on the Black-White Dissimilarity index, the H Index measuring multi-group segregation for Black, Hispanic, White, and an Other racial category, and the White Isolation Index. While commonly used, these indices face other limitations beyond the MAUP. These includes problems of aspatiality, lack of decomposibility, and sensitivity to baseline city demographics (Reardon and O'Sullivan, 2004; Lee et al., 2008; Mazza and Punzo, 2015; Roberto, 2018). Despite such critiques, these indices remain widely used in both scholarly and popular investigations of segregation. As such, we focus on measurement error in the most commonly used measures of segregation.

B Additional results

B.1 Estimating differences in official versus simulation estimates

Here, we present city-level regressions of modeling bias and variability of segregation indices as a function of city logged population, proportion White population, proportion Black population, proportion Hispanic population, average simulated segregation, and average simulated segregation squared. We estimate a linear model across cities separately for bias and variability as the outcome. The results for the 2020 simulations are shown in Table 9. Columns 1 and 2 report results for the Black-White Dissimilarity Index, columns 3 and 4 report results for the H Index, and columns 5 and 6 report results for the White Isolation Index. We see consistent evidence of declining variability (95% interval length) as city population increases for each index. The effect of racial demographics on bias or variability is less consistent. We find a non-linear relationship between average simulation segregation, with an initially increasing relationship between bias and variability and the overall segregation of a city, but the marginal effect of higher segregation diminishes at higher segregation values (as evidenced by the negative quadratic term). Table 10 report results for other segregation indices (Hispanic-White Dissimilarity Index, Black Isolation Index, Hispanic Isolation Index), with similarly consistent results. We further report similar estimates for the years 2010 (Table 11 and Table 12) and 2000 (Table 13 and Table 14).

Table 9: Modeling bias and variability as function of city demographics – 2020

	Black-White Dissimilarity Index		H Index		White Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0218	0.2069***	-0.0032	0.0685***	-0.0064	0.0498***
	(0.0171)	(0.0185)	(0.0062)	(0.0061)	(0.0050)	(0.0066)
Logged Pop.	-0.0024*	-0.0150***	-2.25×10^{-5}	-0.0056***	0.0003	-0.0039***
	(0.0013)	(0.0014)	(0.0005)	(0.0005)	(0.0004)	(0.0005)
Prop. White	-0.0023	-0.0013	0.0036	0.0006	-0.0085	-0.0367***
	(0.0078)	(0.0084)	(0.0028)	(0.0028)	(0.0062)	(0.0083)
Prop. Black	0.0072	0.0049	0.0069**	0.0066**	0.0083***	0.0191***
	(0.0087)	(0.0094)	(0.0032)	(0.0031)	(0.0028)	(0.0036)
Prop. Hispanic	-0.0077	0.0043	-0.0017	4.21×10^{-5}	-0.0008	0.0027
	(0.0076)	(0.0082)	(0.0028)	(0.0027)	(0.0024)	(0.0032)
Avg. Sim. Est.	0.0760**	0.0879***	0.0796***	0.1466***	0.0253***	0.0558***
	(0.0304)	(0.0328)	(0.0137)	(0.0133)	(0.0066)	(0.0087)
Avg. Sim. Est. Sq.	-0.0684*	-0.0947**	-0.1898***	-0.3251***	-0.0165***	-0.0301***
	(0.0348)	(0.0376)	(0.0426)	(0.0414)	(0.0063)	(0.0083)
Observations	337	337	337	337	337	337
R^2	0.08435	0.33559	0.24632	0.49503	0.21899	0.51877
Adjusted R^2	0.06771	0.32351	0.23261	0.48585	0.20479	0.51002

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

Table 10: Modeling bias and variability as function of city demographics – other indices, 2020

	Hispanic-White Dissimilarity Index		Black Isolation Index		Hispanic Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-0.0068	0.1773***	0.0047	0.0572***	-0.0070	0.0658***
	(0.0132)	(0.0166)	(0.0066)	(0.0090)	(0.0052)	(0.0092)
Logged Pop.	0.0001	-0.0133***	-0.0007	-0.0052***	0.0003	-0.0051***
	(0.0009)	(0.0012)	(0.0005)	(0.0007)	(0.0004)	(0.0007)
Prop. White	0.0020	-0.0116	0.0053*	0.0096**	0.0055**	0.0025
	(0.0058)	(0.0073)	(0.0032)	(0.0044)	(0.0025)	(0.0044)
Prop. Black	0.0057	0.0009	-0.0150**	-0.0642***	0.0052**	0.0048
	(0.0058)	(0.0073)	(0.0075)	(0.0102)	(0.0025)	(0.0043)
Prop. Hispanic	-0.0067	-0.0094	-0.0003	0.0068	-0.0123*	-0.0843***
	(0.0058)	(0.0073)	(0.0031)	(0.0043)	(0.0067)	(0.0118)
Avg. Sim. Est.	0.0709***	0.1370***	0.0469***	0.1434***	0.0230***	0.1004***
	(0.0248)	(0.0313)	(0.0062)	(0.0085)	(0.0054)	(0.0096)
Avg. Sim. Est. Sq.	-0.0859**	-0.1565***	-0.0307***	-0.0957***	-0.0079	-0.0254***
	(0.0370)	(0.0468)	(0.0070)	(0.0096)	(0.0053)	(0.0094)
Observations	337	337	337	337	337	337
R^2	0.07557	0.33583	0.29691	0.54792	0.08385	0.30697
Adjusted R^2	0.05876	0.32376	0.28412	0.53970	0.06719	0.29437

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

Table 11: Modeling bias and variability as function of city demographics – 2010

	Black-White Dissimilarity Index		H Index		White Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-0.0032	0.2115***	0.0036	0.0720***	-0.0088	0.0506***
	(0.0211)	(0.0229)	(0.0093)	(0.0091)	(0.0067)	(0.0080)
Logged Pop.	-0.0007	-0.0167***	-0.0007	-0.0064***	0.0006	-0.0041***
	(0.0015)	(0.0016)	(0.0007)	(0.0007)	(0.0005)	(0.0006)
Prop. White	-0.0009	0.0086	0.0038	0.0092**	-0.0006	-0.0323***
	(0.0100)	(0.0109)	(0.0044)	(0.0043)	(0.0073)	(0.0088)
Prop. Black	0.0077	0.0175	0.0103**	0.0186***	0.0109***	0.0191***
	(0.0111)	(0.0120)	(0.0050)	(0.0049)	(0.0037)	(0.0045)
Prop. Hispanic	-0.0163	0.0235**	-0.0028	0.0085*	-0.0020	0.0061
	(0.0103)	(0.0112)	(0.0046)	(0.0045)	(0.0035)	(0.0042)
Avg. Sim. Est.	0.0949***	0.1159***	0.0895***	0.1462***	0.0176**	0.0596***
	(0.0334)	(0.0364)	(0.0168)	(0.0166)	(0.0081)	(0.0097)
Avg. Sim. Est. Sq.	-0.0893**	-0.1157***	-0.1856***	-0.2840***	-0.0168**	-0.0340***
	(0.0369)	(0.0401)	(0.0430)	(0.0423)	(0.0076)	(0.0091)
Observations	282	282	282	282	282	282
R^2	0.13356	0.34305	0.25498	0.46899	0.20659	0.50844
Adjusted R^2	0.11466	0.32872	0.23872	0.45741	0.18928	0.49771

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

Table 12: Modeling bias and variability as function of city demographics – other indices, 2010

	Hispanic-White Dissimilarity Index		Black Isolation Index		Hispanic Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0028	0.1620***	0.0133	0.0707***	-0.0047	0.0549***
	(0.0177)	(0.0211)	(0.0092)	(0.0143)	(0.0087)	(0.0125)
Logged Pop.	-3.47×10^{-5}	-0.0133***	-0.0014**	-0.0068***	-0.0003	-0.0050***
	(0.0012)	(0.0015)	(0.0006)	(0.0010)	(0.0006)	(0.0009)
Prop. White	-0.0086	0.0068	0.0049	0.0140*	0.0098**	0.0152**
	(0.0081)	(0.0096)	(0.0046)	(0.0072)	(0.0043)	(0.0062)
Prop. Black	-0.0015	0.0159*	-0.0215**	-0.0857***	0.0102**	0.0145**
	(0.0080)	(0.0095)	(0.0091)	(0.0143)	(0.0042)	(0.0061)
Prop. Hispanic	-0.0124	0.0078	-0.0010	0.0127*	-0.0044	-0.0643***
	(0.0084)	(0.0100)	(0.0047)	(0.0073)	(0.0104)	(0.0150)
Avg. Sim. Est.	0.0641*	0.1522***	0.0623***	0.1790***	0.0213***	0.0967***
	(0.0328)	(0.0391)	(0.0074)	(0.0115)	(0.0077)	(0.0110)
Avg. Sim. Est. Sq.	-0.0795*	-0.1744***	-0.0442***	-0.1113***	-0.0084	-0.0323***
	(0.0460)	(0.0547)	(0.0086)	(0.0134)	(0.0085)	(0.0123)
Observations	282	282	282	282	282	282
R^2	0.03419	0.29578	0.31724	0.53742	0.04704	0.26828
Adjusted R^2	0.01312	0.28041	0.30234	0.52732	0.02625	0.25232

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

Table 13: Modeling bias and variability as function of city demographics – 2000

	Black-White Dissimilarity Index		H Index		White Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0218	0.2069***	-0.0032	0.0685***	-0.0064	0.0498***
	(0.0171)	(0.0185)	(0.0062)	(0.0061)	(0.0050)	(0.0066)
Logged Pop.	-0.0024*	-0.0150***	-2.25×10^{-5}	-0.0056***	0.0003	-0.0039***
	(0.0013)	(0.0014)	(0.0005)	(0.0005)	(0.0004)	(0.0005)
Prop. White	-0.0023	-0.0013	0.0036	0.0006	-0.0085	-0.0367***
	(0.0078)	(0.0084)	(0.0028)	(0.0028)	(0.0062)	(0.0083)
Prop. Black	0.0072	0.0049	0.0069**	0.0066**	0.0083***	0.0191***
	(0.0087)	(0.0094)	(0.0032)	(0.0031)	(0.0028)	(0.0036)
Prop. Hispanic	-0.0077	0.0043	-0.0017	4.21×10^{-5}	-0.0008	0.0027
	(0.0076)	(0.0082)	(0.0028)	(0.0027)	(0.0024)	(0.0032)
Avg. Sim. Est.	0.0760**	0.0879***	0.0796***	0.1466***	0.0253***	0.0558***
	(0.0304)	(0.0328)	(0.0137)	(0.0133)	(0.0066)	(0.0087)
Avg. Sim. Est. Sq.	-0.0684*	-0.0947**	-0.1898***	-0.3251***	-0.0165***	-0.0301***
	(0.0348)	(0.0376)	(0.0426)	(0.0414)	(0.0063)	(0.0083)
Observations	337	337	337	337	337	337
R^2	0.08435	0.33559	0.24632	0.49503	0.21899	0.51877
Adjusted R^2	0.06771	0.32351	0.23261	0.48585	0.20479	0.51002

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

Table 14: Modeling bias and variability as function of city demographics – other indicies, 2000

	Hispanic-White Dissimilarity Index		Black Isolation Index		Hispanic Isolation Index	
	Bias	95% Interval	Bias	95% Interval	Bias	95% Interval
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.0113	0.1827***	0.0082	0.0859***	0.0009	0.0537***
	(0.0244)	(0.0287)	(0.0185)	(0.0275)	(0.0102)	(0.0178)
Logged Pop.	-0.0015	-0.0155***	-0.0013	-0.0097***	-0.0008	-0.0055***
	(0.0016)	(0.0019)	(0.0013)	(0.0019)	(0.0007)	(0.0012)
Prop. White	-0.0006	0.0145	0.0104	0.0366**	0.0082	0.0215**
	(0.0118)	(0.0138)	(0.0096)	(0.0143)	(0.0053)	(0.0092)
Prop. Black	0.0052	0.0390***	-0.0131	-0.0668***	0.0126**	0.0194**
	(0.0114)	(0.0133)	(0.0169)	(0.0251)	(0.0052)	(0.0092)
Prop. Hispanic	0.0024	0.0136	-7.98×10^{-5}	0.0300**	-0.0064	-0.0706***
	(0.0125)	(0.0146)	(0.0101)	(0.0150)	(0.0133)	(0.0232)
Avg. Sim. Est.	0.0523	0.1405**	0.0736***	0.2370***	0.0347***	0.1335***
	(0.0472)	(0.0554)	(0.0125)	(0.0186)	(0.0081)	(0.0143)
Avg. Sim. Est. Sq.	-0.0470	-0.1314*	-0.0606***	-0.1712***	-0.0215**	-0.0617***
	(0.0623)	(0.0731)	(0.0163)	(0.0242)	(0.0105)	(0.0185)
Observations	244	244	244	244	244	244
R^2	0.02402	0.28190	0.20869	0.48792	0.09840	0.29740
Adjusted R^2	-0.00069	0.26372	0.18866	0.47496	0.07557	0.27961

Columns report coefficient estimates from models estimating the average bias (odd number columns) and variability (even number columns) by city against city-level demographics. Unit of analysis is a city. Coefficients are marked with asterisks to denote statistical significance: $***$ < 0.001, $**$ < 0.01, $*$ < 0.05.

B.2 Hispanic-White Dissimilarity, Hispanic Isolation Index, Black Isolation Index

Figure B1 shows the same plots as in Figure 3 in the manuscript, but for the Hispanic-White Dissimilarity, Hispanic Isolation Index, Black Isolation Index. For each of these additional indices, we see similar patterns as with the indices in the manuscript. Namely, we observe some variability across cities but limited bias, with simulation averages well-adhering to official estimates.

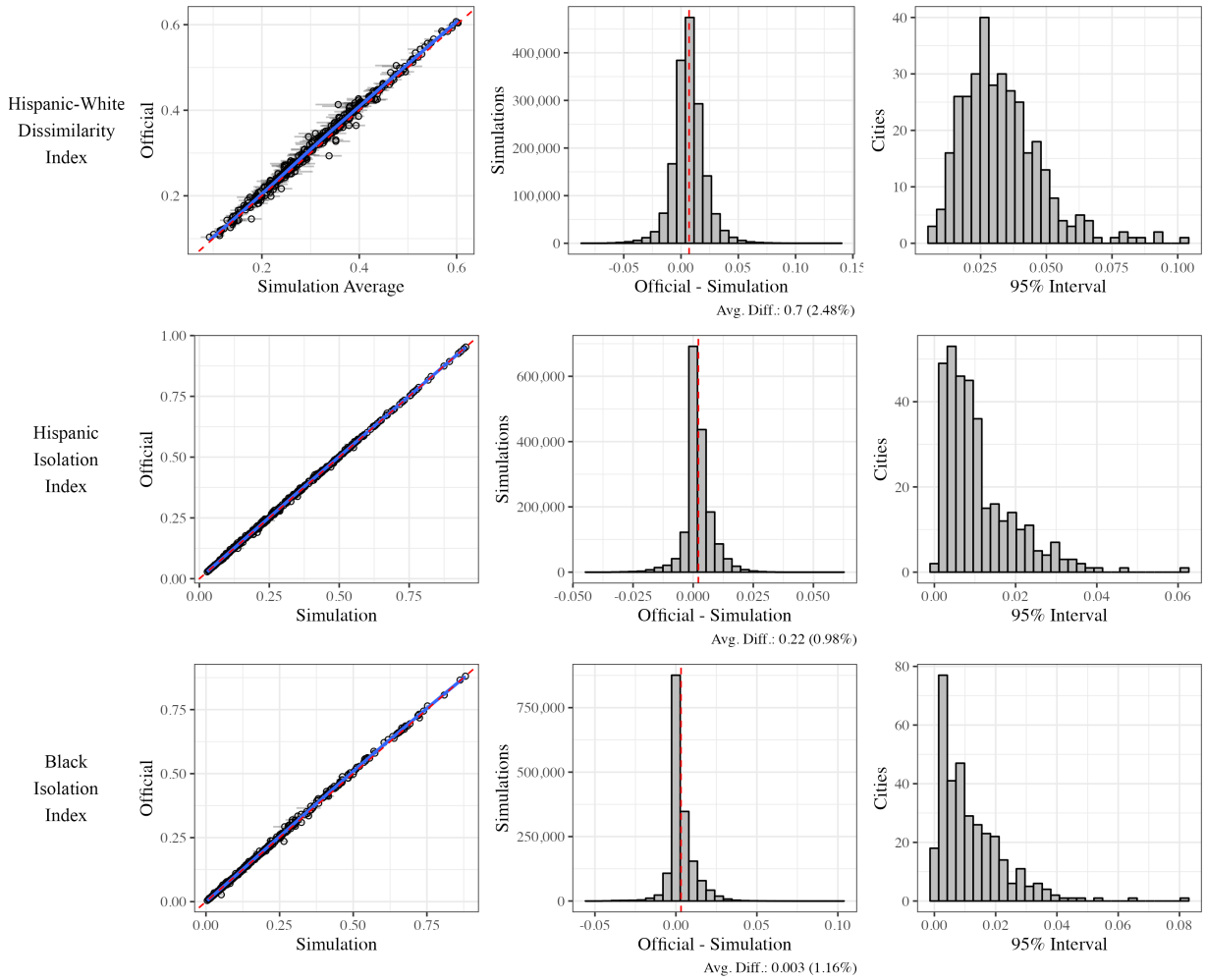

Figure B1: Measuring segregation using official Census Data versus simulations – other indices. Left plots show scatter plots of the relationship between official and simulation averages of segregation indices. Middle plots show histogram of official minus simulation averages. Right plots show histograms of 95% interval size across cities. Top row shows Hispanic-White Dissimilarity Index plots, middle row shows Hispanic Isolation Index plots, and bottom row shows Black Isolation Index.

Figure B2 shows the same plots as in Figure 5 in the manuscript, but for these alternative indices. These indices also see declining variability with city population.

B.3 County results

Our main results report analysis of aggregation measurement bias for city-level segregation. Here, we present equivalent analyses across counties. As with cities, we limit our analysis to counties with populations over

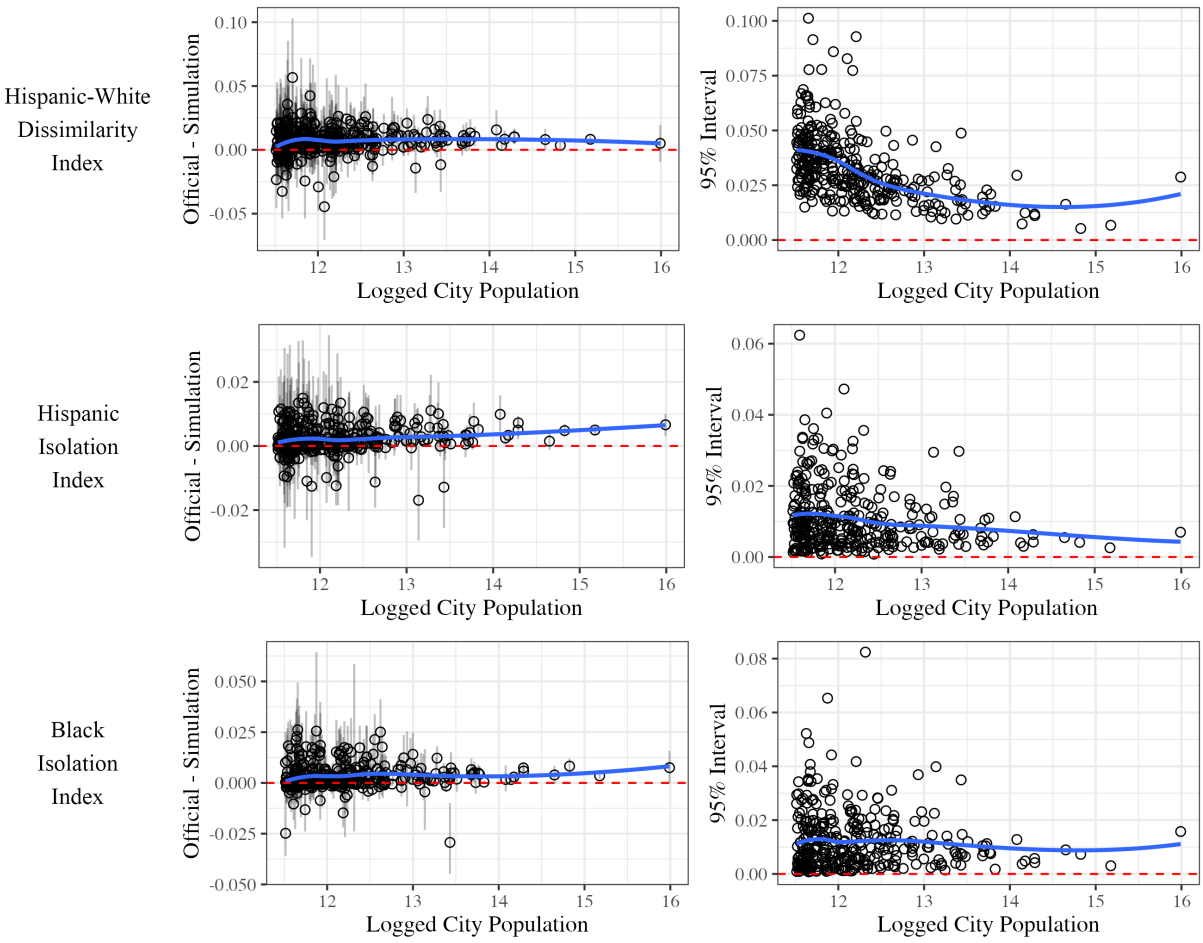

Figure B2: Measurement error by city population. Scatter plots show bias (left) and interval size (right) by logged city population. Blue lines plot the local regression line. Top row shos the Hispanic-White Dissimilarity Index, middle row shows the Hispanic Isolation Index, and the bottom row shows the Black Isolation Index.

100,000 in 2020. Figure B3 recreates Figure 3 for county comparisons, showing similar takeaways county analyses.

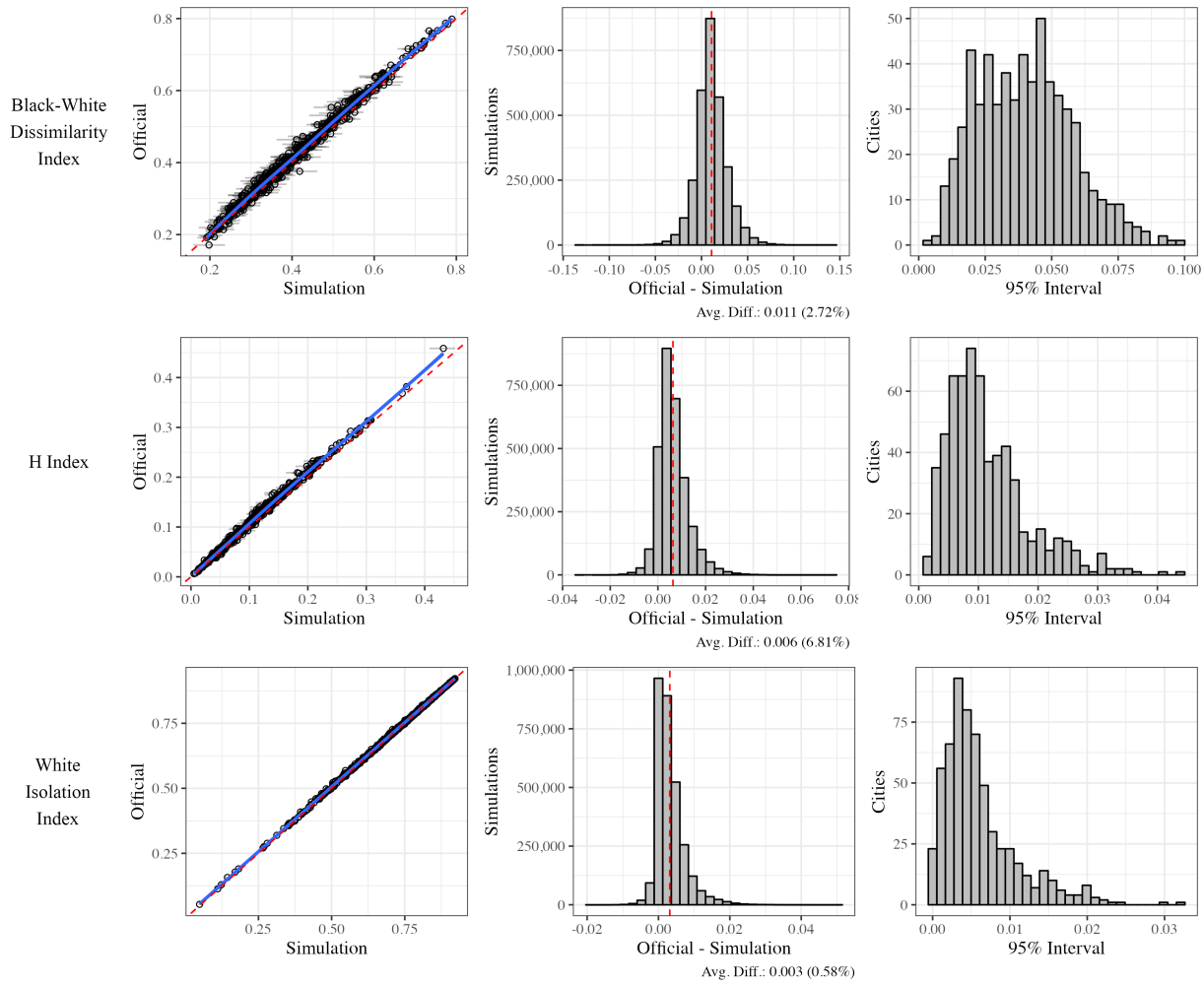

Figure B3: Measuring county segregation using official Census Data versus simulations. Left plots show scatter plots of the relationship between official and simulation averages of segregation indices. Middle plots show histogram of official minus simulation averages. Right plots show histograms of 95% interval size across counties Top row shows Black-White Dissimilarity Index plots, middle row shows H Index plots, and bottom row shows White Isolation Index.

Likewise, Figure B4 shows a similarly declining relationship between variability and city size as in Figure 5.

We recreate our over-time analyses in Figure B5 and Table 15. As with previous analyses, our results are consisten across city and county analyses.

B.4 Relationship between bias and uncertainty

In Figure B6, we present scatter plots of the relationship between city-level bias and variability across years and segregation indices. We see a generally increasing relationship between bias and variability, although there are non-linearities to these relationships.

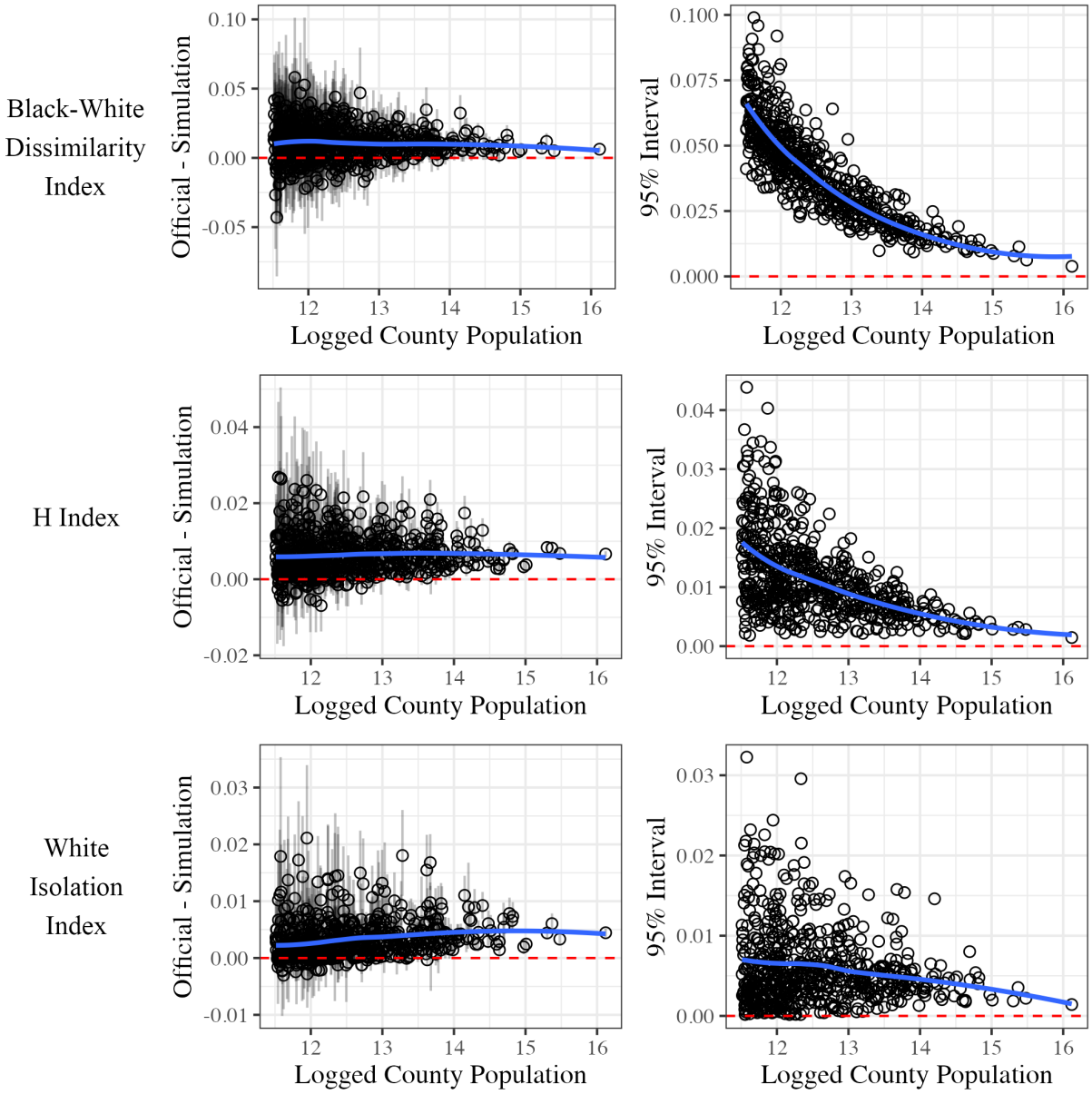

Figure B4: Measurement error by county population. Scatter plots show bias (left) and interval size (right) by logged county population. Blue lines plot the local regression line. Top row shos the Black-White Dissimilarity Index, middle row shows the H Index, and the bottom row shows the White Isolation Index.

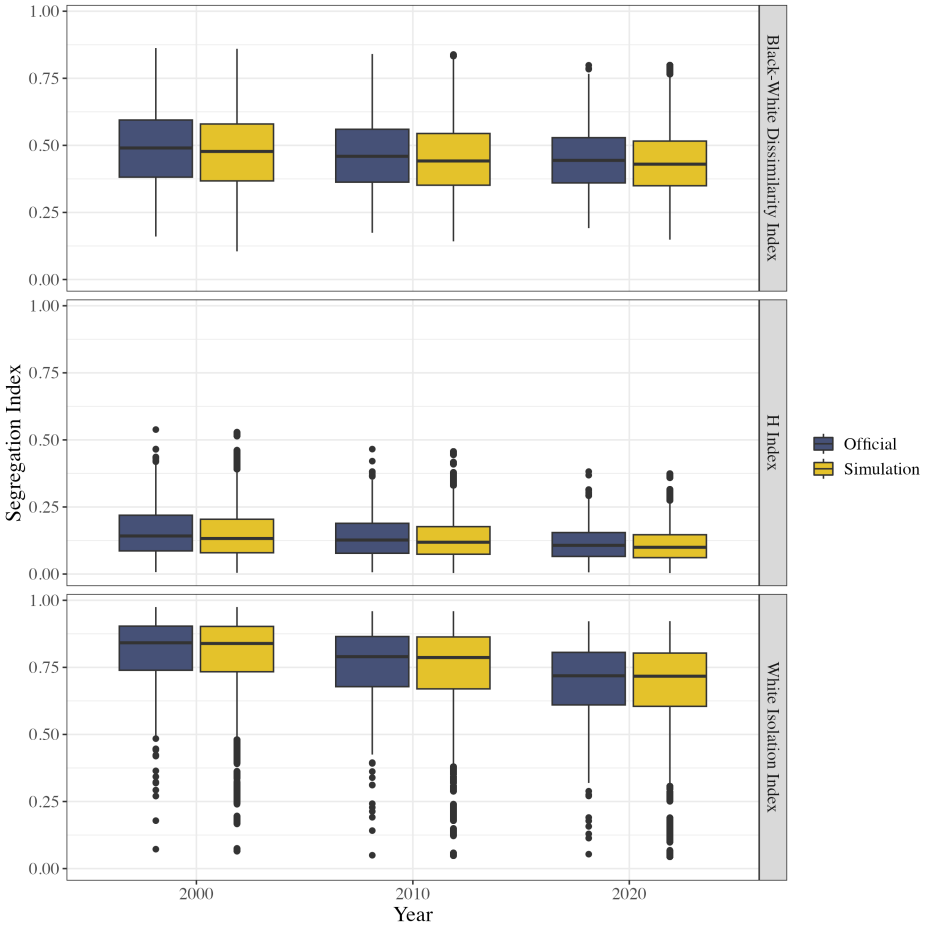

Figure B5: Official vs. simulation county segregation, 2000-2020. Figure shows box and whisker plots reporting the distribution of each segregation index across blue (official) and simulated (yellow) plans for the years 2000, 2010, and 2020. Top row shows the Black-White Dissimilarity Index, middle row shows the H Index, and bottom row shows the White Isolation Index.

Table 15: Modeling over-time changes in official versus simulation county segregation

	Segregation			Deviation		
	Black-White Dissimilarity Index	H Index	White Isolation Index	Black-White Dissimilarity Index	H Index	White Isolation Index
	(1)	(2)	(3)	(4)	(5)	(6)
2010	-0.0292***	-0.0169***	-0.0507***	0.0007	-0.0154***	-0.0176***
	(0.0021)	(0.0011)	(0.0013)	(0.0011)	(0.0008)	(0.0013)
2020	-0.0422***	-0.0397***	-0.1129***	0.0011	-0.0252***	-0.0464***
	(0.0028)	(0.0018)	(0.0019)	(0.0016)	(0.0013)	(0.0020)
Official	0.0129***	0.0103***	0.0036***			
	(0.0006)	(0.0004)	(0.0002)			
2010 × Official	-0.0018***	-0.0024***	-4.69×10^{-5}			
	(0.0006)	(0.0003)	(0.0001)			
2020 × Official	-0.0014**	-0.0036***	-0.0002			
	(0.0006)	(0.0003)	(0.0001)			
Observations	7,816,563	7,816,563	7,816,563	7,815,000	7,815,000	7,815,000
R^2	0.94439	0.95229	0.98461	0.83402	0.96164	0.97411
Within R^2	0.23889	0.45651	0.85783	0.00048	0.38022	0.49171
State-County fixed effects	✓	✓	✓	✓	✓	✓

Column 1-3 report results from regressions modeling segregation estimates across time and official versus simulation maps. Columns 4-6 report results from regression modeling deviations from county-level averages for each simulated plan across time. Unit of analysis is the a given plan (official or one of 5,000 simulated plans for each county). Standard errors are clustered at the county-level. Coefficients are marked with asterisks to denote statistical significance: * * * < 0.001, ** < 0.01, * < 0.05.

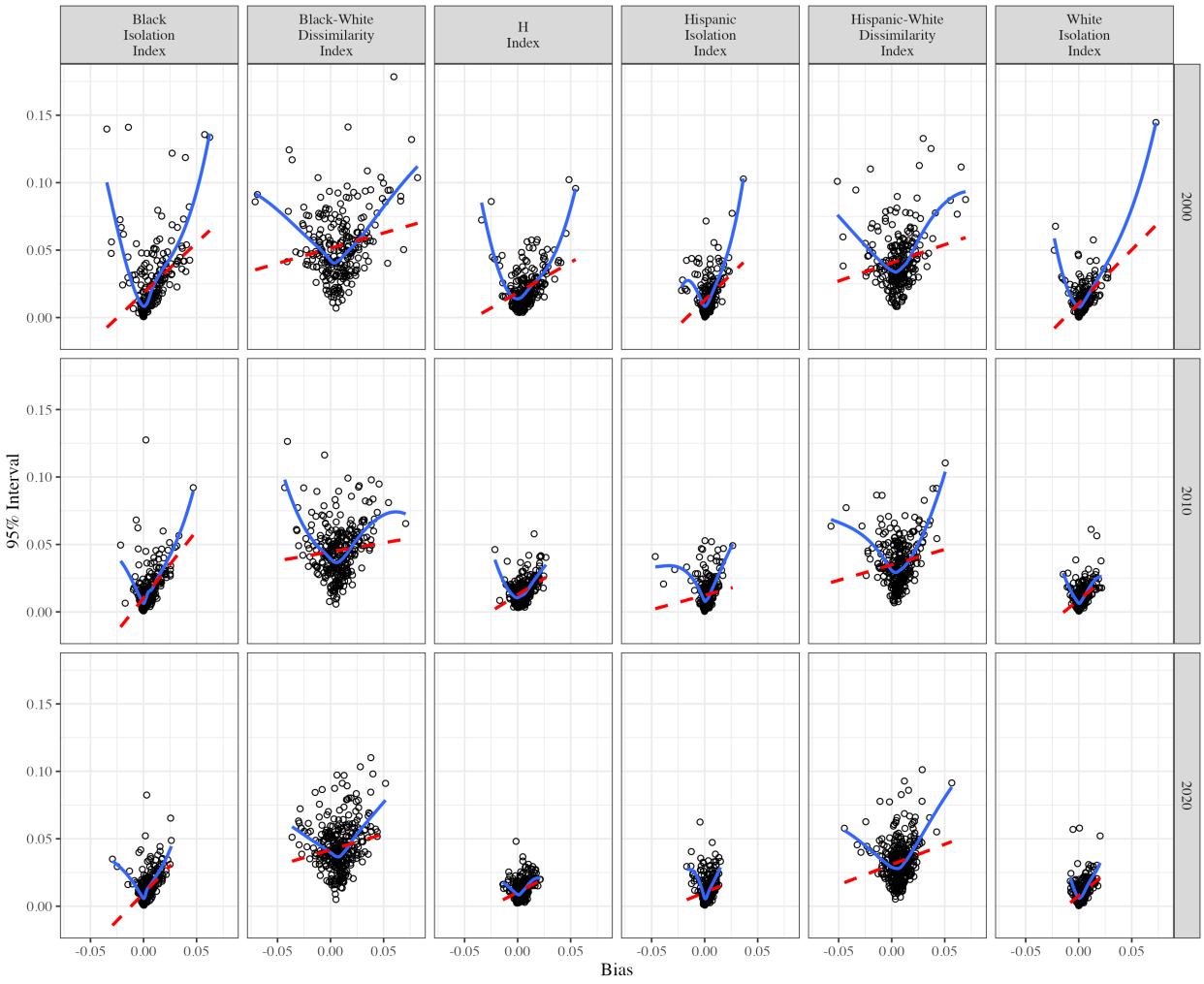

Figure B6: Relationship between bias and variability, 2000-2020. Figure shows scatter plots of the relationship between bias and variability across cities for all segregation indices for the years 2000, 2010, 2020. Red lines plot linear lines of best fit and blue lines plot local regression lines.

B.5 Results from simulation with higher compactness constraint

In our main results, we incorporate a compactness constraint set to the default for the simulation. To test the sensitivity of our simulation to a higher compactness constraint, we conducted an equivalent simulation analysis on 2020 cities with a 30% higher compactness constraint. We replicate the main results for this alternative set of simulated tracts across cities and present those results in Figure B7. These results are quite similar to those in the main paper, with an average bias for the Black-White Dissimilarity Index of 0.0032, and an average 95% interval size of 0.0387 (compared to 0.0081 bias and 0.0437 interval size in the main results.)

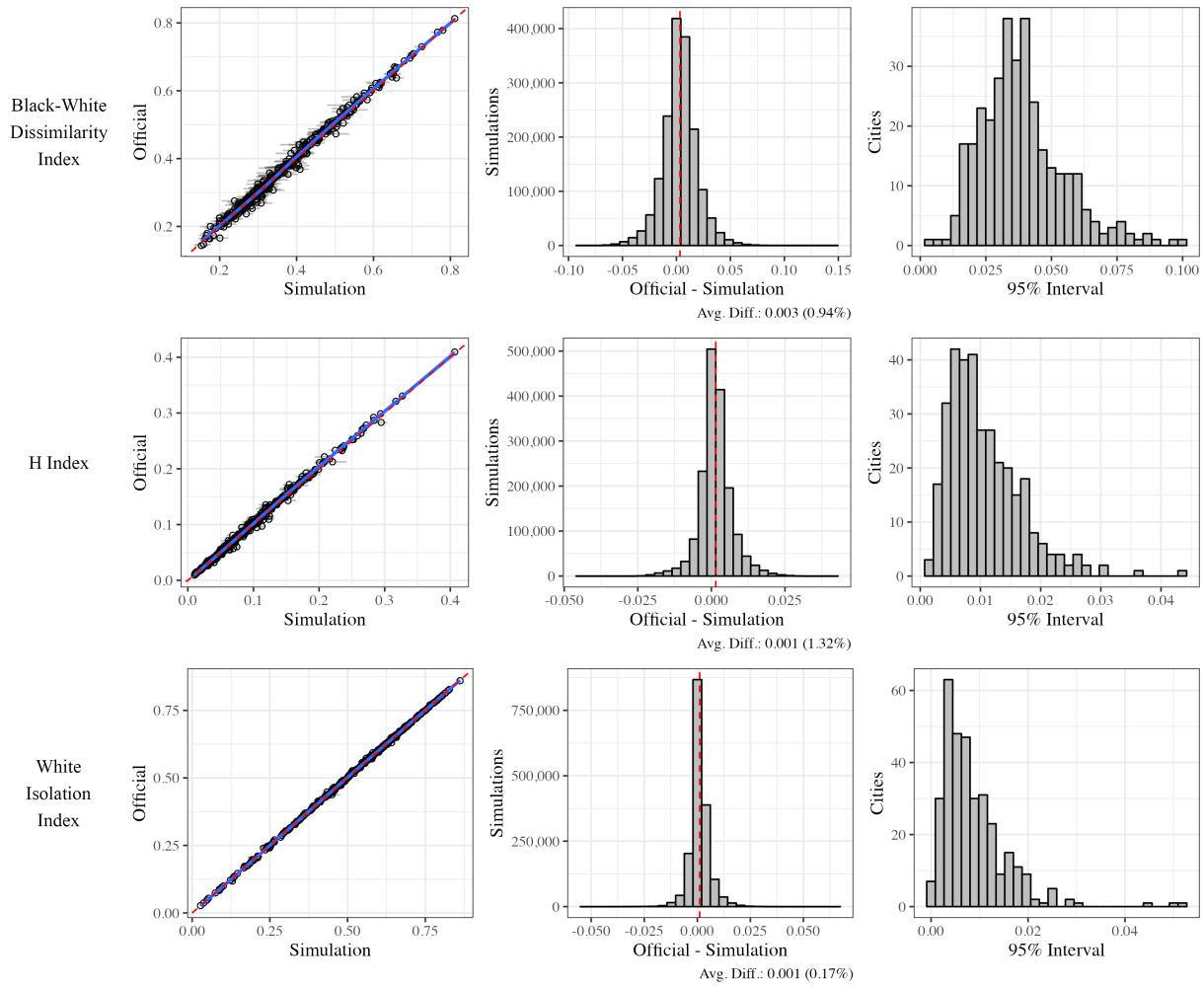

Figure B7: Measuring segregation using official Census Data versus simulations - higher compactness simulation. Left plots show scatter plots of the relationship between official and simulation averages of segregation indices. Middle plots show histogram of official minus simulation averages. Right plots show histograms of 95% interval size across cities. Top row shows Black-White Dissimilarity Index plots, middle row shows H Index plots, and bottom row shows White Isolation Index.