Computational and Bayesian Methods for Geographic Data in the Social Sciences

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Statistics

have examined a dissertation entitled

Computational and Bayesian Methods for
Geographic Data in the Social Sciences

presented by Cory McCartan

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Kosuke Imai

Signature _____

Typed name: Prof. Xiao-Li Meng

Signature _____

Typed name: Prof. Gary King

Date: April 5, 2023

Computational and Bayesian Methods for Geographic Data in the Social Sciences

A DISSERTATION PRESENTED
BY

CORY MCCARTAN

TO
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2023

Dissertation advisor: Professor Kosuke Imai Cory McCartan

Computational and Bayesian Methods for Geographic Data in the Social Sciences

Abstract

Research in the social sciences often involves the analysis of geographic data. These data may be explicitly spatial, with each observation tied to a specific point on or region of the globe, or they may include variables which are heavily influenced by geographic patterns of human activity and settlement. This dissertation advances computational and Bayesian modeling methods for analyzing geographic data in three important social scientific settings.

In Part I, we study legislative redistricting, where voters are geographically assigned to legislative districts. Analyzing redistricting plans is significantly complicated by the confounding role that geography plays: two plans in different states cannot be compared directly, because the political and social geography of each state may differ. We develop a sequential Monte Carlo sampling algorithm that allows researchers to generate counterfactual redistricting plans from a specific probability distribution. We then apply this tool as part of a larger probabilistic framework of individual and differential harm, which we advance to measure the impact of redistricting plans on different groups of voters.

In Part II, we develop a Bayesian model and accompanying online survey tool to record and study individuals' perceptions of their neighborhoods. While neighborhoods are recognized as important mediators and determinants of many social scientific outcomes, their inherent subjectivity, and the difficulties of modeling spatial regions, have posed a challenge for their quantitative analysis. The proposed Bayesian

model is both flexible and computationally efficient. In a sample of 2,527 voters in three metropolitan areas, we find that white voters, and party-identifying voters, are more likely to include areas in their neighborhood which have a higher share of co-racial and co-partisan residents, respectively.

In Part III, we tackle the problem of estimating racial disparities from individual-level data which contains no racial information. In this setting, researchers will often probabilistically impute race using individual surnames and addresses, a method known as Bayesian Improved Surname Geocoding (BISG). However, these probabilistic predictions are alone insufficient for unbiased estimation of racial disparities. We provide a highly plausible identifying assumption, as well as a class of Bayesian models, which we call Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that solve this problem. The models admit a highly efficient expectation-maximization algorithm for inference, and greatly reduce estimation error in a validation example from the North Carolina voter file.

Contents

This page intentionally left blank.

List of Figures

This page intentionally left blank.

Citations to Previous Work

Chapter 1 is based on McCartan and Imai (2023), which has been accepted for publication in the *Annals of Applied Statistics*:

> McCartan, C. and Imai, K. (2023). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *Annals of Applied Statistics*, Accepted

Replication code and data are available at `https://doi.org/10.7910/DVN/O38OCG`. Open-source software is available which implements the proposed methodology (Kenny et al., 2020).

Chapter 2 is based on McCartan and Kenny (2022), which has been submitted to the OSF SocArXiv:

> McCartan, C. and Kenny, C. T. (2022). Individual and differential harm in redistricting. *SocArXiv preprint nc2x7*

Replication code and data are available at `https://github.com/CoryMcCartan/harm-redistricting`. Appendix B.2 is excerpted and adapted from Kenny et al. (2023).

Chapter 3 is based on McCartan et al. (2021), which has been submitted for publication:

> McCartan, C., Brown, J. R., and Imai, K. (2021). Measuring and modeling neighborhoods. *arXiv preprint arXiv:2110.14014*

The experimental analysis was pre-registered with the Open Science Foundation; pre-registration materials can be found at `https://osf.io/xdumv/`. The open-source survey instrument for measuring subjective neighborhoods is available at `https://github.com/CoryMcCartan/neighborhood-survey`.

Chapter 4 is based on McCartan et al. (2023), which has been submited to the arXiv:

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

Replication code and data are available at `https://github.com/CoryMcCartan/birdie`. Open-source software is available which implements the proposed methodology (McCartan, 2023).

This page intentionally left blank.

Acknowledgments

What they say about raising children is true about completing Ph.D.s, too: it takes a village. I am deeply indebted to my advisors, mentors, teachers, colleagues, friends, and family for everything they have done to get me through these twenty-one years of formal education.

The work in this dissertation was partially supported by a grant from the Alfred P. Sloan Foundation, as well as the Sinnott-Wagner Geospatial Fund. I am grateful for this financial support.

I would like to thank my committee members: Gary King, Xiao-Li Meng, and chair Kosuke Imai. While Gary joined my committee quite recently, he has provided helpful and thought-provoking feedback on all of my work, often in the form of sharp questions during my presentations. Xiao-Li has been a familiar mentor throughout graduate school, first as an instructor in The Art and Practice of Teaching Statistics, and later as a research partner who always keeps me on my toes. I have greatly appreciated the opportunity to explore some of the foundations of statistics with him, especially as that has involved fascinating tangents and connections between very disparate areas of the field.

Not a week goes by where I don't feel so lucky to have landed, out of all the departments and statistics faculty in the country, with Kosuke as my advisor. From shortly after my admittance to the Ph.D. program, when he called to congratulate and welcome me, to late-night strategizing about the job market, Kosuke has dedicated an incredible amount of effort and commitment to my growth as a statistician, a researcher, and an academic. It has been an absolute pleasure to work with and learn from him. In particular, it has been the experience of a lifetime to go from developing a redistricting algorithm in the early months of the pandemic, to applying it to previously out-of-reach research projects with the ALARM team, to fighting gerrymandering through redistricting litigation in courts across the country. Through it all, Kosuke has provided invaluable guidance, feedback, insight, and advice.

The Ph.D. has been the final chapter of many in my educational journey, and I am extremely grateful to the many teachers and mentors that developed my skills and passion for mathematics, statistics, and social science, and provided advice and encouragement along the way. Ted Gooley and Shonda Kuiper, who gave me early guidance and inspiration to pursue statistics as a career, deserve particular thanks. I would also like to thank Dan Ho and Shiro Kuriwaki for their support and advice as I've navigated the job market this year, and for being fantastic research partners and collaborators.

My many wonderful colleagues and friends I have to thank for making the PhD much more than a purely academic adventure. I could not have made it through without them, and I will fondly look back

on all the experiences we shared together in these (too short) years. Ben Thorne and Chenchen Li were some of the first people I met the day I moved to Boston, and since then they have variously been pandemic roommates, outdoor partners, Pacific Northwest partisans, waffle connoisseurs, and great friends. Justin Bloesch and Beth Huang have been truly wonderful friends as well as examples of how to work towards justice in one's career and community.

Through Kosuke, I've been fortunate enough to meet some amazing people in the Government department. In particular I've had a blast working with Tyler Simko and Chris Kenny on innumerable redistricting projects. Chris has been a fantastic partner in software development (both useful and frivolous), paper writing, litigation analyses, not to mention a thoughtful and generous friend. The other members of the Imai Research Group have also been supportive and helpful in working on my job market presentations, and I've appreciated the chance to learn from them as they do great research. There are too many wonderful statistics Ph.D. students to mention individually, but I would be remiss if I did not thank them *en masse* for their sufferance of my dogmatic Bayesianism, which I made sure they never forgot. I'd also like to give special thanks to David Ham, Jonathan Che, James Bailie, Xiang Meng, and Biyonka Liang. Finally, a huge thank-you to the rest of my cohort, who in addition to being wonderful people also helped me survive the first year and the marathon of the qualifying exams.

A huge part of the last four years for me has been my involvement with the Harvard Graduate Students Union – U.A.W. Local 5118. In both bargaining committees, the organizing community, the interim benefits committee, various strike committees, and the larger membership, I have found an incredible set of friends and comrades. Their dedication, righteousness, organizing ability, empathy, and courage has been constantly inspiring and is an example to try to live up to. I'd especially like to thank Justin Bloesch, Rachel Sandalow-Ash, Lee Kennedy-Shaffer, Ege Yumuşak, Jenni Austiff, Cherrie Bucknor, Cole Meisenhelder, and Ellen Wallace, who welcomed me to the bargaining committee within a few weeks of my arrival on campus, and from whom I have learned so much. It's hard to put into words what the process of organizing, bargaining, and striking with all of these amazing people in order to build a better university has meant to me, but it's no exaggeration to say that it has been transformative. Solidarity forever.

Finally, I'd like to recognize my family for their constant care, encouragement, and support. My parents have given me so many opportunities, educational and otherwise, that have enabled me to be where I am today. And to my siblings, Anika and Julia, my grandparents, and my many aunts, uncles, and cousins—thank you for always believing in me, and for modeling such meaningful and thoughtful lives. For everything I've learned in school, I've learned so much more from you all about the most important thing of all: what it means to be a good person.

To my family.

0

Introduction

The word *statistics* entered the English language in the late 18th century via the German *Statistik*, itself derived from the Latin *statisticum* ("of the state").[1] In its original sense, *statistics* referred to a branch of political science, one which dealt with numerical and other facts about states (Yule, 1905). As late as 1881, the Statistical Society of London (yet to receive its royal charter) referred to the "two senses" of the term, one as "a method of scientific inquiry," and one as "a science dealing with the social life of man" (Hooper, 1881).

The social sciences, then, have long played a central role in the field of statistics. Compared to agricultural, biological, and other kinds of data that statisticians often analyze, social scientific data often bring distinct challenges. First, because of the complexity of human society, and the innumerable factors that affect any given social outcome, social scientific data are often noisy and far from sparse: any statistical

[1] Incidentally, the Latin form is itself derived from the Proto-Indo-European **steh₂-*, which also gave rise to the Persian suffix ستان (*-estān*, "country"), whence *Afghanistan*, *Turkmenistan*, etc.

signal is hidden beneath these competing factors and a large amount of individual heterogeneity. Second, individual observations are rarely independent from one another. Dependence can arise from social ties or other interactions, or from patterns in time and space. Geography in particular is an extremely important factor in structuring social scientific outcomes and data, with entire subfields of the major social sciences dedicated to its study.

The dual challenges of sample dependence and a complex, noisy generating process are a natural match for Bayesian methods. Bayesian models are often designed to optimally share statistical strength between observations, and can flexibly incorporate geographic and other dependence structures. The computational tools developed by Bayesian researchers to sample from posterior distributions can also prove useful in sampling complex geographic structures that are of interest in the social sciences.

The three parts of this dissertation provide new computational methods and Bayesian models for analyzing spatial and geographic data in three social scientific settings: redistricting, neighborhoods, and racial disparities. The rest of the introduction describes the problems and advances in each area in more detail. The value of fully probabilistic modeling, as well as the ways in which geography can both enable and complicate statistical analysis, are themes that recur throughout the work.

0.1 Redistricting

Part I studies legislative redistricting, which has gained increasing attention in recent decades as U.S. political parties have drawn increasingly sophisticated *gerrymanders*, which exploit political geography to amplify one party's voting power at the the other's expense. The analysis of redistricting plans is complicated by the lack of a baseline against which to measure plans. Plans in different states cannot be compared directly because physical and political geography, as well as laws and judicial rules, vary from place to

Figure 1: An adjacency graph of Census tracts in and around the city of Boston. Tracts are colored by the city or town to which they belong.

place. Laws themselves are too vague to arrive at specific quantitative measures, much less precise numerical benchmarks.

As Chapter 1 lays out, redistricting plans can be treated formally as graph partitions of an adjacency graph which captures spatial relationships between different voting units, usually precincts. An example of an adjacency graph, which is a core object of study in Chapters 1 and 3, is shown in Figure 1, with a possible graph partition corresponding to cities and towns indicated by coloring. Using this framework, random sampling of graph partitions under constraints has become a popular tool to address the challenge of incomparibility when evaluating legislative redistricting plans. Analysts detect partisan gerrymandering by comparing a proposed redistricting plan with an ensemble of sampled alternative plans. For successful application, sampling methods must scale to large maps with many districts, incorporate realistic legal

constraints, and accurately and efficiently sample from a selected target distribution. Unfortunately, most existing methods struggle in at least one of these areas.

We present a new Sequential Monte Carlo (SMC) algorithm that generates a sample of redistricting plans converging to a realistic target distribution. Because it draws many plans in parallel, the SMC algorithm can efficiently explore the relevant space of redistricting plans better than the existing Markov chain Monte Carlo (MCMC) algorithms that generate plans sequentially. Our algorithm can simultaneously incorporate several constraints commonly imposed in real-world redistricting problems, including equal population, compactness, and preservation of administrative boundaries. We validate the accuracy of the proposed algorithm by using a small map where all redistricting plans can be enumerated. We then apply the SMC algorithm to evaluate the partisan implications of several maps submitted by relevant parties in a recent high-profile redistricting case in the state of Pennsylvania. We find that the proposed algorithm converges to the target distribution faster and with fewer samples than a state-of-the-art MCMC algorithm.

The SMC algorithm is useful for more than identifying partisan gerrymanders. Because it generates samples from a known distribution, it can be used as an element in a larger probabilistic modeling framework. Chapter 2 develops one such framework, which allows researchers to separate various factors that affect the performance of redistricting plans, and to measure the impact of these plans on different groups of voters.

Social scientists have developed dozens of measures, probabilistic and otherwise, for assessing partisan bias in redistricting. But these measures are not easily adapted to other groups, including groups defined by race, class, or geography. Nor are they applicable to single- or no-party contexts, such as local redistricting. To overcome these limitations, we propose a unified framework of *harm* for evaluating the impacts of a districting plan on individual voters and the groups to which they belong. We consider a voter harmed if

their chosen candidate is not elected under the current plan, but would be under a different plan. Harm improves on existing measures by both focusing on the choices of individual voters and directly incorporating counterfactual plans. We discuss strategies for estimating harm, and demonstrate the utility of our framework through analyses of partisan gerrymandering in New Jersey, voting rights litigation in Alabama, and racial dynamics of Boston City Council elections.

0.2 Neighborhoods

Part II develops new methods for the study of individuals' neighborhoods. Similarly to redistricting plans, neighborhoods can be conceptualized formally as connected subgraphs of an adjacency graph that describes the spatial layout of city blocks. Rather than sampling certain types of subgraphs, however, the goal in this part is to perform inference on neighborhoods.

The availability of granular geographic data presents new opportunities to understand how neighborhoods are formed and how they influence attitudes and behavior. To facilitate such studies, we develop an online survey instrument for respondents to draw their neighborhoods on a map. We then propose a statistical model to analyze how the characteristics of respondents and geography, and their interactions, predict subjective neighborhoods. We illustrate the proposed methodology using a survey of 2,527 voters in Miami, New York City, and Phoenix. Holding other factors constant, White respondents tend to include census blocks with more White residents in their neighborhoods. Similarly, Democratic and Republican respondents are more likely to include co-partisan census blocks. Our model also provides more accurate out-of-sample predictions than standard distance-based neighborhood measures. Lastly, we use these methodological tools to test how demographic information shapes neighborhoods, but find limited effects from this experimental manipulation.

0.3 Racial Disparities

Part III deals with data that is not primarily spatial but which crucially includes a geographic element: individual addresses. Due to patterns of residential racial segregation, these addresses provide information on individual race which, with the appropriate methodology, may be used to estimate racial disparities.

The estimation of racial disparities in health care, financial services, voting, and other contexts is often hampered by the lack of individual-level racial information in administrative records. In many cases, the law prohibits the collection of such information to prevent direct racial discrimination. As a result, many analysts have adopted Bayesian Improved Surname Geocoding (BISG), which combines individual names and addresses with the Census data to predict race. Although BISG tends to produce well-calibrated racial predictions, its residuals are often correlated with the outcomes of interest, yielding biased estimates of racial disparities.

Part III proposes an alternative identification strategy that corrects this bias. The proposed strategy is applicable whenever one's surname is conditionally independent of the outcome given their (unobserved) race, residence location, and other observed characteristics. Leveraging this identification strategy, we introduce a new class of models, Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that estimate racial disparity by using surnames as a high-dimensional instrumental variable for race. Our estimation method is scalable, making it possible to analyze large-scale administrative data. We also show how to address potential violations of the key identification assumptions. A validation study based on the North Carolina voter file shows that BIRDiE reduces error by up to 84% in comparison to the standard approaches for estimating racial differences in party identification.

Part I

Redistricting

This page intentionally left blank.

1

Sequential Monte Carlo for Sampling Balanced and Compact Redistricting Plans

1.1 Introduction

In first-past-the-post electoral systems, legislative districts serve as the fundamental building block of democratic representation. In the United States, congressional redistricting, which redraws district boundaries in each state following the decennial Census, plays a central role in influencing who is elected and hence what policies are eventually enacted. Because the stakes are so high, redistricting has been subject to intense political battles. Parties often engage in *gerrymandering* by manipulating district boundaries in order to amplify the voting power of some groups while diluting that of others.

In recent years, the availability of granular data about individual voters has led to sophisticated partisan gerrymandering attempts that cannot be easily detected. At the same time, many scholars have focused their efforts on developing methods to uncover gerrymandering by comparing a proposed redistricting

plan with a large collection of alternative plans that satisfy the relevant legal requirements. A primary advantage of such an approach over the use of simple summary statistics is its ability to account for the characteristics of each state's physical and political geography and state-specific redistricting rules.

For its successful application, a sampling algorithm for drawing alternative plans must (1) be efficient enough to scale to maps with thousands of geographic units and a moderate or large number of districts, (2) simultaneously incorporate a variety of real-world legal constraints such as population balance (Section 1.3.1), geographical compactness (Section 1.3.3), and the preservation of administrative boundaries (Section 1.4.5), and (3) ensure these samples are representative of a specific target population, against which a redistricting plan of interest can be evaluated. Although some have been used in several recent court challenges to existing redistricting plans, all existing algorithms run into limitations of varying severity with regards to at least one of these three key requirements.

Optimization-based (e.g., Mehrotra et al., 1998; Macmillan, 2001; Bozkaya et al., 2003; Liu et al., 2016) and constructive Monte Carlo (e.g., Cirincione et al., 2000; Chen and Rodden, 2013; Magleby and Mosesson, 2018) methods can be made scalable and incorporate many constraints. But they are not designed to sample from any specific target distribution. As a consequence, the resulting plans tend to differ systematically, for example, from a uniform distribution under certain constraints (Cho and Liu, 2018; Fifield et al., 2020a,b). The absence of an explicit target distribution makes it difficult to interpret the ensembles generated by these methods and use them for statistical outlier analysis to detect gerrymandering.

MCMC algorithms (e.g., Mattingly and Vaughn, 2014; Wu et al., 2015; Chikina et al., 2017; DeFord et al., 2021; Carter et al., 2019; Fifield et al., 2020a; Cannon et al., 2022) can in theory sample from a specific target distribution, and incorporate constraints through the use of an energy function. In practice, however, existing algorithms struggle to mix and traverse through a highly complex sampling space, making scalability

difficult and accuracy hard to prove. Some of these algorithms make proposals by flipping precincts at the boundary of existing districts (e.g., Mattingly and Vaughn, 2014; Fifield et al., 2020a), rendering it difficult or even impossible to transition between points in the state space, especially as more constraints are imposed. More recent algorithms by DeFord et al. (2021) and Carter et al. (2019) use spanning trees to make their proposals, and this has allowed these algorithms to yield greater moves and substantially improve mixing. Yet recent theoretical results suggest that even these larger moves may not be enough to traverse the entire state space, and therefore may fail to converge to the correct distribution, if a realistic population balance constraint is imposed (Akitaya et al., 2022).

We contribute to the ongoing scholarly efforts to address the above three key challenges by developing a new Sequential Monte Carlo (SMC) algorithm, based on a similar but not identical spanning tree construction to DeFord et al. (2021) and Carter et al. (2019) (see Sections 1.3 and 1.4). Like MCMC algorithms, the SMC algorithm generates samples which approximate the target distribution arbitrarily well as the sample size increases. But like constructive Monte Carlo methods, the SMC algorithm draws many separate plans from scratch, rather than tweaking a single plan sequentially. This approach is better suited to the large discrete state space with a multimodal target distribution that characterizes redistricting problems. For example, in cases where existing MCMC proposals render the state space disconnected, the SMC algorithm can still converge to the target distribution. As we demonstrate in Sections 1.5 and 1.6 (see also Appendix A.2), this sampling approach translates to faster convergence and smaller standard errors for a given computational budget. For larger and more complex redistricting sampling problems, the SMC algorithm can be easily parallelized to facilitate efficient computation.

The proposed algorithm proceeds by splitting off one district at a time, building up the redistricting

plan piece by piece (see Figure 1.2 for an illustration). Each split is accomplished by drawing a spanning tree and removing one edge, which splits the spanning tree in two. We also extend the SMC algorithm so that it preserves administrative boundaries and certain geographical areas as much as possible, which is another common constraint considered in many real-world redistricting cases. An open-source software package, `redist`, is available for implementing the proposed algorithm (Kenny et al., 2020).

The SMC algorithm is not without limitations. Like existing MCMC approaches, the SMC algorithm only guarantees convergence to the target distribution as the sample size approaches infinity. Additionally, because the SMC algorithm involves repeated resampling with replacement, for a finite number of samples it can suffer from particle system collapse (Liu et al., 2001), where many of the sampled plans share a small number of common districts (which originate as common ancestor particles in the SMC scheme). This can significantly increasing sampling variability. Thus, it is important to understand the limitations of the proposed algorithm in finite samples, especially when dealing with large maps and many districts. In Section 1.4.4, we provide a set of diagnostics which can be used in practice to help identify if more samples are needed to reach convergence or if the constraints imposed by an analyst are too strong.

In Section 1.5, we validate the SMC algorithm using a small map for which all potential redistricting plans can be enumerated (Fifield et al., 2020b). We demonstrate that the proposed algorithm samples accurately from a realistic target distribution, and that our proposed diagnostics are a good proxy for total sampling error, which is generally unobservable. Section 1.6 applies the SMC algorithm to the 2011 Pennsylvania congressional redistricting, and also compares its performance on this problem with an MCMC algorithm with the same transition kernel as the proposed SMC algorithm. We find that the SMC algo-

rithm samples more efficiently than the MCMC approach. Section 1.7 concludes and discusses directions for future work.

1.2 The 2011 Pennsylvania Congressional Redistricting

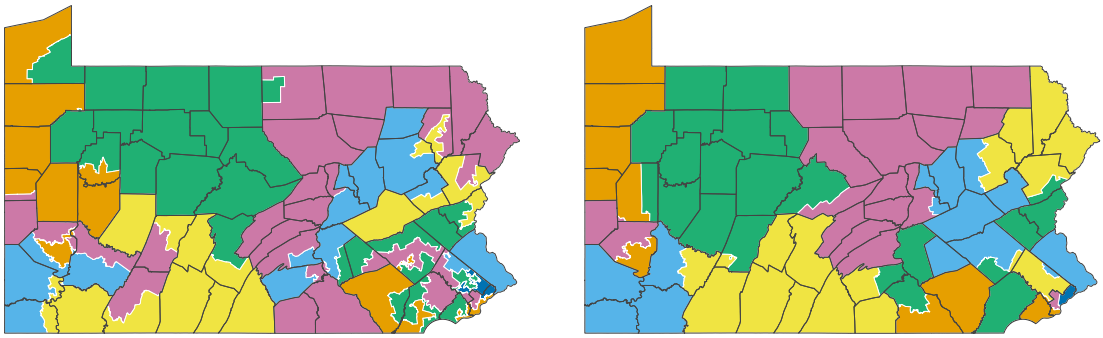
We study the 2011 Pennsylvania congressional redistricting because it illustrates the salient features of the redistricting problem. We begin by briefly summarizing the background of this case and then explain the role of sampling algorithms used in the expert witness reports.

1.2.1 Background

Pennsylvania lost a seat in Congress during the reapportionment of the 435 U.S. House seats following the 2010 Census. In Pennsylvania, the General Assembly, which is the state's legislative body, draws new congressional districts, subject to gubernatorial veto. At the time, the General Assembly was controlled by Republicans, and Tom Corbett, also a Republican, served as governor. In the 2012 election, which took place under the newly adopted 2011 districting map, Democrats won 5 seats while Republicans took the remaining 13. Under the previous plan, the split was 7–12.

In June 2017, the League of Women Voters of Pennsylvania filed a lawsuit alleging that the 2011 plan adopted by the Republican legislature violated the state constitution by diluting the political power of Democratic voters. The case worked its way through the state court system, and on January 22, 2018, the Pennsylvania Supreme Court issued its ruling, writing that the 2011 plan "clearly, plainly and palpably violates the Constitution of the Commonwealth of Pennsylvania, and, on that sole basis, we hereby strike it as unconstitutional." (League of Women Voters *v.* Commonwealth, 2018).

The court ordered that the General Assembly adopt a remedial plan and submit it to the governor,



(a) 2011 General Assembly map

(b) 2018 Pennsylvania Supreme Court map

Figure 1.1: Comparison of the 2011 map drawn by the General Assembly and the final map imposed by the Supreme court in 2018. County lines are shown in dark gray, and district boundaries that do not coincide with county boundaries are in white.

who would in turn submit it to the court, by February 15, 2018. In its ruling, the court laid out specific requirements that had to be satisfied by all proposed plans:

> composed of compact and contiguous territory; as nearly equal in population as practicable; and which do not divide any county, city, incorporated town, borough, township, or ward, except where necessary to ensure equality of population.

The leaders of the Republican Party in the General Assembly drew a new map, but the Democratic governor, Tom Wolf, refused to submit it to the court, claiming that it, too, was an unconstitutional gerrymander. Instead, the court received remedial plans from seven parties: the petitioners, the League of Women Voters; the respondents, the Republican leaders of the General Assembly; the governor, a Democrat; the lieutenant governor, also a Democrat; the Democratic Pennsylvania House minority leadership; the Democratic Pennsylvania Senate minority leadership; and the intervenors, which included Republican party candidates and officials. Ultimately, the Supreme Court drew its own plan and adopted it on February 19, 2018, arguing that it was "superior or comparable to all plans submitted by the parties." Figure 1.1 shows the remedial plan created by the Supreme Court as well as the 2011 map adopted by the General Assembly, which were found on the court's case page.

The constraints explicitly laid out by the court, as well as the numerous remedial plans submitted by the parties, make the 2011 Pennsylvania redistricting a useful case study that evaluates redistricting plans.

1.2.2 The Role of Sampling Algorithms

The original finding that the 2011 General Assembly plan was a partisan gerrymander was in part based on different outlier analyses performed by two academic researchers, Jowei Chen and Wesley Pegden, who served as the petitioner's expert witnesses. Chen randomly generated two sets of 500 redistricting plans according to a constructive Monte Carlo algorithm based on Chen and Rodden (2013). He considered population balance, contiguity, compactness, avoiding county and municipal splits, and, in the second set of 500, avoiding plans that placed more than two incumbents in the same district (at least one pair of incumbents in the same district was necessary, given that Pennsylvania lost a seat from 2000 to 2010). Pegden ran a reversible Markov chain similar to that used in the MCMC algorithm of Mattingly and Vaughn (2014) for one trillion steps, and computed upper bounds of p-values using the method of Chikina et al. (2017). This method was also used in a follow-up analysis by Moon Duchin, who served as an expert for Governor Wolf (Duchin, 2018). Both petitioner experts concluded that the 2011 plan was an extreme outlier according to compactness, county and municipal splits, and the number of Republican and Democratic seats implied by statewide election results.

The respondents also retained an expert academic witness, Wendy Tam Cho, who directly addressed the sampling-based analyses of Chen and Pegden. Cho criticized Chen's analysis for not sampling from a specified target distribution. She also criticized Pedgen's analysis by arguing that his Markov chain only made local explorations of the space of redistricting plans, and could not therefore have generated a representative sample of all valid plans, though the p-values computed using the Chikina et al. (2017) method

explicitly do not require mixing of the Markov chain (see also Cho and Rubinstein-Salzedo, 2019, and Chikina et al. (2019)). We do not directly examine the intellectual merits of the specific arguments put forth by the expert witnesses. However, these methodological debates are also relevant for other cases where simulation algorithms have been extensively used by expert witnesses (e.g., Rucho v. Common Cause (2019); Common Cause v. Lewis (2019); Covington v. North Carolina (2017); Harper v. Lewis (2020)), and highlight the difficulties in practically applying existing sampling algorithms to actual redistricting problems.

First, the distributions that some of these algorithms sample from are not made explicit, leaving open the possibility that the generated ensemble is systematically different from the true set of all valid plans. Second, even when the distribution is known, MCMC algorithms used to sample from it may be prohibitively slow to mix and cannot yield a representative sample. These challenges motivate us to design an algorithm that accurately samples from a specific target distribution and incorporates most common redistricting constraints, while being efficient and scalable.

1.3 Sampling Balanced and Compact Districts

1.3.1 The Setup

Redistricting plans are ultimately aggregations of geographic units such as counties, voting precincts, or Census blocks. The usual requirement that the districts in a plan be contiguous necessitates consideration of the spatial relationship between these units. The natural mathematical structure for this consideration is a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_m\}$ consists of m nodes representing the geographic units of redistricting and E contains edges connecting units which are legally adjacent.

A labeled redistricting plan on G consists of n districts, where each district is a collection of nodes. A labeled plan is described by a function $\xi : V \to \{1, 2, \ldots, n\}$, where $\xi(v) = i$ implies that node v is

in district i. In practice, we will be interested in *unlabeled* plans, since the assignment of sets of precincts to labels is arbitrary and has no impact on the real-world aspects of the district such as its population, demographic composition, or partisan lean. Define an equivalence relation by $\xi_1 \cong \xi_2$ if there exists a permutation σ such that $\xi_1(v) = \sigma(\xi_2(v))$ for all v. Then, an unlabeled redistricting plan can be viewed as an equivalence class under $\cong$, denoted $[\xi]$; nothing in what follows will depend on the particular labeled plan representative ξ.

We let $V_i(\xi)$ and $E_i(\xi)$ denote the nodes and edges contained in district i under a given redistricting plan ξ, so $G_i(\xi) = (V_i(\xi), E_i(\xi))$ represents the induced subgraph that corresponds to district i under the plan. We suppress the dependence on ξ when it is clear from context, writing $G_i = (V_i, E_i)$. Since each node belongs to only one district, we have $V = \bigcup_{i=1}^{n} V_i(\xi)$ and $V_i(\xi) \bigcap V_{i'}(\xi) = \varnothing$ for any redistricting plan ξ. In addition, we require that each district be contiguous, i.e., that G_i is a connected graph, for all i.

Beyond connectedness, redistricting plans are always required to have roughly equal population in every district. To formalize this requirement, let $\text{pop}(v)$ denote the population of node v. Then, the population of a district i may be written as $\text{pop}(V_i(\xi)) := \sum_{v \in V_i(\xi)} \text{pop}(v)$. We quantify the discrepancy between a given plan and the ideal of equal population in every district by the *maximum population deviation*,

$$\text{dev}(\xi) := \max_{1 \le i \le n} \left| \frac{\text{pop}(V_i)}{\text{pop}(V)/n} - 1 \right|,$$

where $\text{pop}(V)$ is the total population. Some courts and states have imposed hard maximums on this quantity, e.g., $\text{dev}(\xi) \le D = 0.05$ for state legislative redistricting (National Conference of State Legislatures, 2021).

The proposed algorithm samples plans by way of spanning trees on each district, i.e., subgraphs of $G_i(\xi)$ which contain all vertices, no cycles, and are connected. Let T_i represent a spanning tree for district i whose

vertices and edges are given by $V_i(\xi)$ and a subset of $E_i(\xi)$, respectively. The collection of spanning trees from all districts together form a spanning forest. Each node belongs to one spanning tree in the forest, and this assignment corresponds to a redistricting plan. However, a single redistricting plan may correspond to multiple spanning forests because each district may admit more than one spanning tree.

For a given redistricting plan, we can compute the exact number of spanning forests in polynomial time using the determinant of a submatrix of the graph Laplacian, according to the Matrix Tree Theorem of Kirchhoff (see Tutte (1984)). Thus, for a graph H, if we let $\tau(H)$ denote the number of spanning trees on the graph, we can represent the number of spanning forests that correspond to a redistricting plan ξ as $\tau(\xi) \coloneqq \prod_{i=1}^{n} \tau(G_i(\xi))$. This fact will play an important role in the definition of our sampling algorithm and its target distribution.

1.3.2 The Target Distribution

The algorithm is designed to sample an unlabeled plan $[\xi]$ with probability

$$\pi([\xi]) \propto \exp\{-J(\xi)\}\tau(\xi)^{\rho}\mathbf{1}_{\{\xi \text{ connected}\}}\mathbf{1}_{\{\text{dev}(\xi)\leq D\}}, \tag{1.1}$$

where the indicator functions ensure that the plans meet population balance and connectedness criteria, $\tau(\xi)$ measures the compactness of the districts in ξ (see Section 1.3.3), and J encodes additional constraints on the types of plans preferred. As done in Section 1.6, we often use a reasonably strict population constraint such as $D = 0.001$ and $D = 0.005$. The parameter $\rho \in \mathbb{R}_0^+$ is chosen to control the compactness of the generated plans.

This target distribution π has both substantive and theoretical justifications. First, it incorporates two constraints which are almost always present in real-world redistricting: the support of π is restricted to

contiguous plans and those that meet a population deviation threshold. Second, it represents the unique maximum entropy distribution on the set of redistricting plans satisfying these two universal constraints and the moment conditions implied by the other constraints, i.e., $\mathbb{E}_\pi[\log \tau(\xi)] = \mu_\tau$ and $\mathbb{E}_\pi[J(\xi)] = \mu_J$ for some constants μ_τ and μ_J (see Cover and Thomas, 2006, Theorem. 12.1.1, originally of Boltzmann).

Thus, our target distribution ensures that all plans meet contiguity and population requirements, and *on average* satisfy a compactness standard as well as any other additional constraints (through the function J). It is no surprise, therefore, that this class of target distributions has been used by other work developing redistricting sampling algorithms (Herschlag et al., 2017; Fifield et al., 2020a).

The generality of the additional constraint function J is intentional, as its exact form and number imposed on the redistricting process varies by state and by the type of districts being drawn; any type of constraint may be incorporated by choosing a J which is small for preferred plans and large otherwise. For example, a preference for plans close to an existing plan ξ_{sq} may be encoded as

$$J_{\text{sq}}(\xi) = -\frac{\beta}{\log n}\text{VI}(\xi, \xi_{\text{sq}}) := -\frac{\beta}{2\log n}\sum_{i,j=1}^{n} \frac{P_{ij}}{\text{pop}(V)}\left(\log\left(\frac{P_{ij}}{\text{pop}(V_j(\xi_{\text{sq}}))}\right) + \log\left(\frac{P_{ij}}{\text{pop}(V_j(\xi))}\right)\right), \tag{1.2}$$

where $\beta \in \mathbb{R}^+$ controls the strength of the constraint and $P_{ij} = \text{pop}(V_i(\xi) \cap V_j(\xi_{\text{sq}}))$ is the population shared between district i of ξ and j of ξ_{sq}. The function $\text{VI}(\cdot, \cdot)$ represents the variation of information (also known as the shared information distance), which is the difference between the joint entropy and the mutual information of the distribution of population over the new districts ξ relative to the existing districts ξ_{sq} (Cover and Thomas, 2006). When ξ is any relabeling of ξ_{sq}, then $J_{\text{sq}}(\xi) = 0$. In contrast, when ξ evenly splits the nodes of each district of ξ_{sq} among the districts of ξ, then $J_{\text{sq}}(\xi) = \beta$. This distance measure will prove useful later in quantifying the diversity of a sample of redistricting plans (see also Guth et al., 2020).

There exist other formulations of constraints, and considerations in choosing a set of weights that balance constraints against each other (see e.g., Bangia et al., 2017; Herschlag et al., 2017; Fifield et al., 2020a). Here, we focus on sampling from the broad class of distributions characterized by Equation (1.1), which have been used in other work; we do not address the important but separate problem of picking a specific instance of this class for a given redistricting problem.

The flexibility of J can be deceptive, however. The algorithm operates efficiently only when the additional constraints imposed by J are not too severe. Even a small number of strong constraints incorporated into J can dramatically limit the number of valid plans and considerably complicate the process of sampling (Chatterjee and Diaconis, 2018). The Markov chain algorithms developed to date partially avoid this problem by moving toward maps with lower J over a number of steps, but in general including more constraints makes it even more difficult to transition between valid redistricting plans. Approaches such as simulated annealing (Bangia et al., 2017; Herschlag et al., 2017) and parallel tempering (Fifield et al., 2020a) have been proposed to handle multiple constraints, but these can be difficult to calibrate in practice and provide few, if any, theoretical guarantees.

In practice, we usually find that the most stringent constraints are those involving population deviation, compactness, and administrative boundary splits. As shown later, we address this issue by designing our algorithm to directly satisfy these constraints. Weak additional constraints do not generally have a substantial effect on the sampling efficiency, though there are exceptions. Monitoring the distribution of the weights and the overall sampling efficiency is crucial to obtaining a good sample, as we discuss later.

1.3.3 Spanning Forests and Compactness

One common redistricting requirement is that districts be geographically compact, though nearly every state leaves this term undefined. Dozens of numerical compactness measures have been proposed, with the Polsby–Popper score (Polsby and Popper, 1991) perhaps the most popular. Defined as the ratio of a district's area to that of a circle with the same perimeter as the district, the Polsby–Popper score is constrained to $[0, 1]$, with higher scores indicating more compactness.

Other scholars have proposed a graph-theoretic measure known as *edge-cut compactness* (Dube and Clark, 2016; DeFord et al., 2021). This measure counts the number of edges that must be removed from the original graph to partition it according to a given plan. Formally, it is defined as

$$\text{rem}(\xi) \;:=\; 1 - \frac{\sum_{i=1}^{n} |E_i(\xi)|}{|E(G)|},$$

where we have normalized to the total number of edges.

Plans that involve cutting many edges will necessarily have long internal boundaries, driving up their average district perimeter (and driving down their Polsby–Popper scores), while plans that cut as few edges as possible will have relatively short internal boundaries and much more compact districts. Additionally, given the high density of voting units in urban areas, plans which cut fewer edges will tend to avoid drawing district lines through the heart of these urban areas. This has the welcome side effect of avoiding splitting cities and towns, and in doing so helping to preserve "communities of interest," another common redistricting consideration.

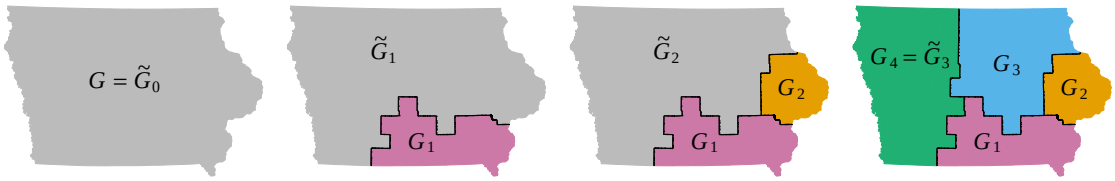
Empirically, this graph-based compactness measure is highly correlated with $\log \tau(G) - \log \tau(\xi)$ (we often observe a correlation in excess of 0.99). It is difficult to precisely characterize this relationship except in special cases because $\tau(\xi)$ is calculated as a matrix determinant (McKay, 1981). However, this quantity

is strongly controlled by the product of the degrees of each node in the graph, $\prod_{i=1}^{m} \deg(v_i)$ (Kostochka, 1995). Removing an edge from a graph decreases the degree of the vertices at either end by one, so we would expect $\log \tau(G)$ to change by approximately $2\{\log \bar{d} - \log(\bar{d} - 1)\}$ with this edge removal, where $\bar{d}$ is the average degree of the graph. This implies a linear relation $\log \tau(G) - \log \tau(\xi) \approx \text{rem}(\xi) \cdot 2\{\log \bar{d} - \log(\bar{d} - 1)\}$, and hence $\tau(\xi)^{\rho} \approx C_1 \exp(-C_2\, \rho\, \text{rem}(\xi))$, where C_1 and C_2 are some constants depending on the details of the map.

As a result, a greater value of ρ in the target distribution corresponds to a preference for fewer edge cuts and therefore a redistricting plan with more compact districts. This and the considerations given in the literature (Dube and Clark, 2016; DeFord et al., 2021) suggest that the target distribution in Equation (1.1) with $\rho = 1$ (or another positive value) is a good choice for sampling compact districts. The choice of $\rho = 1$ is computationally convenient, as it allows us to avoid calculating $\tau(\xi)$ as part of sampling (an asymptotic bottleneck), and yet usually produces satisfactorily compact districts. Of course, if another compactness metric is desired, one can simply set $\rho = 0$ and incorporate the alternative metric into J. This will preserve the algorithm's efficiency to the extent that the alternative metric correlates with the edge-removal measure of compactness. Setting $\rho = 0$ by itself, however, will make sampling intractable in most cases, just as choosing an extreme J will.

1.4 The Proposed Algorithm

The proposed algorithm samples redistricting plans by sequentially drawing districts over $n - 1$ iterations of a splitting procedure. This is fundamentally different from existing MCMC approaches, which change an existing plan according to some transition kernel. The iterations of the proposed algorithm are from district to district within a single plan whereas the iterations in an MCMC algorithm are from plan to plan.



(a) Initial map **(b)** Iteration 1 **(c)** Iteration 2 **(d)** Iteration 3

Figure 1.2: The sequential splitting procedure applied to the state of Iowa, where four congressional districts are created at the county level.

Our algorithm begins by partitioning the original graph $G = (V, E) = (\widetilde{V}_0, \widetilde{E}_0) = \widetilde{G}_0$ into two induced subgraphs: $G_1 = (V_1, E_1)$, which will constitute a district in the final map, and the remainder of the graph $\widetilde{G}_1 = (\widetilde{V}_1, \widetilde{E}_1)$, where $\widetilde{V}_1 = V \setminus V_1$ and $\widetilde{E}_1$ consists of all the edges between vertices in $\widetilde{V}_1$. Next, the algorithm takes $\widetilde{G}_1$ as an input graph and partitions it into two induced subgraphs, one which will become a district G_2 and the remaining graph $\widetilde{G}_2$. The algorithm repeats the same splitting procedure until the final $(n-1)$-th iteration whose two resulting partitions, G_{n-1} and $\widetilde{G}_{n-1} = G_n$, become the final two districts of the redistricting plan.

Figure 1.2 provides an illustration of this sequential procedure. To sample a large number of redistricting plans from the target distribution given in Equation (1.1), at each iteration, the algorithm samples many candidate partitions, discards those which fail to meet the population constraint, and then resamples a certain number of the remainder according to importance weights, using the resampled partitions at the next iteration. The rest of this section explains the details of the proposed algorithm.

1.4.1 The Splitting Procedure

We first describe the splitting procedure, which is similar to the merge-split Markov chain proposals of DeFord et al. (2021) and Carter et al. (2019). It proceeds by drawing a random spanning tree T, identifying the k_i most promising edges to cut within the tree, and selecting one such edge at random to create two

Algorithm 1 Splitting procedure to generate one district

Input: initial graph $\widetilde{G}_{i-1}$ and a parameter $k_i \in \mathbb{Z}^+$.

(a) Draw a single spanning tree T on $\widetilde{G}_{i-1}$ uniformly from the set of all such trees using Wilson's algorithm.

(b) Each edge $e \in E(T)$ divides T into two components, $T_e^{(1)}$ and $T_e^{(2)}$. For each edge, compute the following population deviation for the two districts that would be induced by cutting T at e,

$$d_e^{(1)} = \left| \frac{\sum_{v \in T_e^{(1)}} \text{pop}(v)}{\text{pop}(V)/n} - 1 \right| \quad \text{and} \quad d_e^{(2)} = \left| \frac{\sum_{v \in T_e^{(2)}} \text{pop}(v)}{\text{pop}(V)/n} - 1 \right|.$$

Let $d_e = \min\{d_e^{(1)}, d_e^{(2)}\}$, and index the edges in ascending order by this quantity, so that we have $d_{e_1} \leq d_{e_2} \leq \cdots \leq d_{e_{m_i-1}}$, where $m_i = |\tilde{V}_{i-1}|$.

(c) Select one edge e^* uniformly from the top k_i edges, $\{e_1, e_2, \ldots, e_{k_i}\}$, and remove it from T, creating a spanning forest $(T_{e^*}^{(1)}, T_{e^*}^{(2)})$ which induces a partition $(G_i^{(1)}, G_i^{(2)})$.

(d) If $d_{e^*}^{(1)} \leq d_{e^*}^{(2)}$, i.e., if $T_{e^*}^{(1)}$ induces a district that is closer to the optimal population than $T_{e^*}^{(2)}$ does, set $G_i = G_i^{(1)}$ and $\widetilde{G}_i = G_i^{(2)}$; otherwise, set $G_i = G_i^{(2)}$ and $\widetilde{G}_i = G_i^{(1)}$.

induced subgraphs. Spanning trees are an attractive way to split districts, as the removal of a single edge induces a partition with two connected components, and spanning trees can be sampled uniformly (Wilson, 1996).

As part of the full sampling procedure (Algorithm 2), after splitting, we check that the population of the new district G_i falls within the bounds $[P_i^-, P_i^+]$ where

$$P_i^- = \max\left\{ \frac{\text{pop}(V)}{n}(1-D),\ \text{pop}(\widetilde{V}_{i-1}) - \frac{n-i}{n}\text{pop}(V)(1+D) \right\} \quad \text{and}$$
$$P_i^+ = \min\left\{ \frac{\text{pop}(V)}{n}(1+D),\ \text{pop}(\widetilde{V}_{i-1}) - \frac{n-i}{n}\text{pop}(V)(1-D) \right\}.$$

These bounds also ensure that it will be possible for future iterations to generate valid districts out of $\widetilde{G}_i$.

If $\text{pop}(V_i) \notin [P_i^-, P_i^+]$, then the entire redistricting plan is rejected and the sampling process begins again. While the rate of rejection varies by map and by iteration, we generally encounter acceptance rates at each iteration between 5% and 30%, which are not so low as to make sampling from large maps intractable. Algorithm 1 details the steps of the splitting procedure, where at the first iteration we take $\widetilde{G}_0 = G$.

1.4.2 The Sampling Probability

The above sequential splitting procedure does not generate plans from the target distribution π. We denote the sampling measure by q, and write the sampling probability for a given connected plan ξ at iteration i as $q(G_i \mid \widetilde{G}_{i-1})$, since each new district G_i depends only on the leftover map area $\widetilde{G}_{i-1}$ from the previous iteration. This probability can be written as the probability that we cut an edge along the boundary of the new district, integrated over all spanning trees which could be cut to form the district, i.e.,

$$q(G_i \mid \widetilde{G}_{i-1}) = \sum_{T \in \mathcal{T}(\widetilde{G}_{i-1})} q(G_i \mid T)\, \tau(\widetilde{G}_{i-1})^{-1}, \tag{1.3}$$

where $\mathcal{T}(\cdot)$ represents the set of all spanning trees of a given graph, and we have relied on the fact that Wilson's algorithm draws spanning trees uniformly.

The key is that for certain choices of k_i (the number of edges considered to be cut at iteration i), the probability that an edge is cut is independent of the trees that are drawn. Let $ok(T)$ represent the number of edges on any spanning tree T that induce balanced partitions with population deviation below D, i.e.,

$$ok(T) \;\coloneqq\; |\{e \in E(T) : d_e \leq D\}|.$$

Then define $K_i \coloneqq \max_{T \in \mathcal{T}(\widetilde{G}_{i-1})} ok(T)$, the maximum number of such edges across all spanning trees. Furthermore, let $\mathcal{C}(G, H)$ represent the set of edges joining nodes in a subgraph G to nodes in a subgraph

H. We have the following result for the splitting probability for new districts whose populations lie inside the bounds defined above (Appendix A.1 for the proof).

Lemma 1.1. *The probability of splitting a* valid *new district G_i from an existing area $\widetilde{G}_{i-1}$ using Algorithm 1 with parameter $k_i \geq K_i$ is*

$$q(G_i \mid \widetilde{G}_{i-1}, \text{pop}(V_i) \in [P_i^-, P_i^+]) \;=\; \frac{\tau(G_i)\tau(\widetilde{G}_i)}{\tau(\widetilde{G}_{i-1})k_i}|\mathcal{C}(G_i, \widetilde{G}_i)|. \tag{1.4}$$

1.4.3 Sequential Monte Carlo

We follow a sequential Monte Carlo approach (Doucet et al., 2001; Liu et al., 2001) to generate draws from the target distribution, rather than simply performing $n-1$ iterations of Algorithm 1 and resampling or reweighting at the final stage. A sequential approach is also useful in operationalizing the rejection procedure to enforce the population constraint.

The proposed procedure is presented as Algorithm 2. The algorithm is governed by a parameter $\alpha \in (0, 1]$, which has no effect on the target distribution nor the asymptotic accuracy of the algorithm. Rather, α may be adjusted to maximize the efficiency of sampling. To generate S redistricting plans, at each iteration of the splitting procedure $i \in \{1, 2, \ldots, n-1\}$, we resample and split the existing plans one at a time, rejecting those which do not meet the population constraints, until we obtain S new plans for the next iteration. This rejection process can be viewed as a form of partial rejection control (Liu et al., 1998, 2001), or a version of the AliveSMC algorithm (LeGland and Oudjane, 2005; Peters et al., 2012).

The weights at each stage serve three purposes. The first is to account for the extraneous $|\mathcal{C}(G_i, \widetilde{G}_i)|$ term which appears in the splitting probability $q(G_i \mid \widetilde{G}_{i-1}, \text{pop}(V_i) \in [P_i^-, P_i^+])$ but not in the target distribution π. The second is to account for differences in the compactness parameter ρ; the splitting

procedure generates plans according to $\rho = 1$, but this may be different from the target value. The third purpose is to adjust for the imbalances between the labeled plans generated by the splitting procedure and the unlabeled plans which are the target of sampling. While there are $n!$ labeled plans corresponding to every unlabeled plan, not every labeled plan may be sampled according to the sequential procedure. In fact, the number of labeled plans which are in the support of the SMC proposal distribution varies from one unlabeled plan to another. It is this imbalance which necessitates an additional correction.

For a labeled plan ξ, let G/ξ denote the district-level quotient graph corresponding to the plan; i.e., the nodes are districts $\{1, \ldots, n\}$ and edges connect adjacent districts. Notice that the splitting procedure ensures that the leftover area after each split, $\widetilde{G}_i$, is a connected graph. This means that for every $1 \leq i \leq n-1$, the subgraph of G/ξ induced by the vertices $\{i+1, i+2, \ldots, n\}$ (the districts drawn after split i) is connected. We call any labeled plan ξ a *sequentially valid labeling* if this property holds, and denote by $\psi([\xi])$ the number of sequentially valid labelings corresponding to an unlabeled plan $[\xi]$. The SMC weights incorporate $\psi([\xi])$ as a correction term. Calculation of $\psi([\xi])$ is discussed below in Section 1.4.4.

From the output of Algorithm 2, one last resampling of S plans using the final weights can be performed to generate a final sample. Alternatively, the weights can be used directly to estimate the expectation of some statistics of interest under the target distribution, i.e., $H = \mathbb{E}_\pi(h(\xi))$, using the self-normalized importance sampling estimate $\widehat{H} = \sum_{j=1}^{S} h(\xi^{(j)}) w^{(j)} / \sum_{j=1}^{S} w^{(j)}$.

The sampled plans are not completely independent, because the weights in each step must be normalized before resampling, and because the resampling itself introduces some dependence. Precisely quantifying the amount of dependence is difficult. However, as we demonstrate in Section 1.5, the dependence is not large enough to cause measurable bias in summary statistics of interest.

It is difficult to precisely characterize the computational complexity of the entire SMC algorithm since

Algorithm 2 Sequential Monte Carlo (SMC) Algorithm

Input: graph G to be split into n districts, target distribution parameters $\rho \in \mathbb{R}_0^+$ and constraint function J, and sampling parameters $\alpha \in (0, 1]$ and $k_i \in \mathbb{Z}^+$, with $i \in \{1, 2, \ldots, n-1\}$.

(a) Generate an initial set of S plans $\{\widetilde{G}_0^{(1)}, \widetilde{G}_0^{(2)}, \ldots, \widetilde{G}_0^{(S)}\}$ and corresponding weights $\{w_0^{(1)}, w_0^{(2)}, \ldots, w_0^{(S)}\}$, where each $\widetilde{G}_0^{(j)} := G$ and $w_0^{(j)} = 1$.

(b) For each splitting iteration $i \in \{1, 2, \ldots, n-1\}$:

(1) Until there are S valid plans:

(i) Sample a partial plan $\widetilde{G}_{i-1}$ from $\{\widetilde{G}_{i-1}^{(1)}, \widetilde{G}_{i-1}^{(2)}, \ldots, \widetilde{G}_{i-1}^{(S)}\}$ according to weights $\left(\prod_{l=1}^{i-1} w_l^{(j)}\right)^{\alpha}$.

(ii) Split off a new district from $\widetilde{G}_{i-1}$ through one iteration of the splitting procedure (Algorithm 1), creating a new plan $(G_i, \widetilde{G}_i)$.

(iii) If the newly sampled plan $(G_i, \widetilde{G}_i)$ satisfies $\text{pop}(V_i) \in [P_i^-, P_i^+]$, save it; otherwise, reject it.

(2) Calculate weights for each of the new plans $w_i^{(j)} = \frac{\tau(G_i^{(j)})^{\rho-1}}{|\mathcal{C}(G_i^{(j)}, \widetilde{G}_i^{(j)})|}$.

(c) Calculate final weights

$$w^{(j)} = \exp\{-J(\xi^{(j)})\}\psi([\xi^{(j)}])^{-1}\left(\prod_{i=1}^{n-2} w_i^{(j)}\right)^{(1-\alpha)} w_{n-1}^{(j)}\left(\tau(G_n^{(j)})\right)^{\rho-1}. \tag{1.5}$$

(d) Output the S final plans $\{\xi^{(j)}\}_{j=1}^S$, where $\xi^{(j)} = (G_1^{(j)}, \ldots, G_{n-1}^{(j)}, G_n^{(j)})$, and the final weights $\{w^{(j)}\}_{j=1}^S$.

---

the rejection sampling introduces a random component, which depends on the difficulty of sampling a new district within the population bounds. This random complexity is also shared by existing MCMC approaches, which must redraw proposals if they are invalid. Appendix A.3 analyzes the complexity of

generating each sample without accounting for the rejection step. For both SMC and MCMC samplers, the total sampling time increases linearly in the sample size. In contrast with MCMC approaches, however, step (b)(1) of Algorithm 2 can be embarrassingly parallelized, since each resample-split-check requires interaction only with the previous set of samples.

The weights in the proposed algorithm are chosen to match existing general SMC algorithms with partial rejection control. These existing algorithms provide guarantees as to the convergence of the samples to the target distribution. One such result, which will suffice for our purposes, is the following central limit theorem.

Theorem 1.2. *Let* $\pi_S = \sum_{j=1}^{S} w^{(j)} \delta_{[\xi^{(j)}]}$ *be the weighted particle approximation generated by Algorithm 2. Then for all measurable* h *on unlabeled plans, as* $S \to \infty$*,*

$$\sqrt{S}(\mathbb{E}_{\pi_S}[h([\xi])] - \mathbb{E}_{\pi}[h([\xi])]) \xrightarrow{d} \mathcal{N}(0, V_{SMC}(h)),$$

for some asymptotic variance $V_{SMC}(h)$*.*

A proof is given in Appendix A.1. This central limit theorem implies consistency (in S) of any quantity derived from the weighted samples. However, since this convergence is in probability (w.r.t. the algorithm's sampling probability), the proposition does not establish that $\pi_S \xrightarrow{d} \pi$ almost surely. While the almost sure convergence result exists for a standard SMC algorithm (Del Moral et al., 2006), we do not know of an extension to the case of partial rejection control.

In some cases, the constraints incorporated into $J(\xi)$ admit a natural decomposition to the district level as $\prod_{i=1}^{n} J'(G_i)$—for example, a preference for districts which split as few counties as possible, or against districts which would pair off incumbents. In these cases, an extra term of $\exp\{-J'(G_i^{(j)})\}$ can be added

to the weights $w_i^{(j)}$ in each stage, and the same term can be dropped from the final weights $w^{(j)}$. This can be particularly useful for more stringent constraints; incorporating J' in each stage allows the importance resampling to "steer" the set of redistricting plans towards those which are preferred by the constraints. This same idea can be used to amortize the contribution of $\psi([\xi])$ to the final weights over the preceding SMC iterations; details can be found in Appendix A.3.

1.4.4 Practical Implementation and Use

The SMC algorithm described above allows for an accurate characterization of the target distribution given in Equation (1.1) as the sample size goes to infinity. In practice, of course, we must use a finite number of samples. Additionally, the algorithm relies on choosing k_i, which can be challenging since K_i is typically unknown, and on calculating $\psi([\xi])$, which has no closed-from expression.

This section discusses these practical implementation challenges, as well as selection of the parameter α. We also describe the use of diagnostics (Section 1.4.4) that can help identify when the algorithm is or is not performing well, and therefore if more samples or different constraints are required. Further details about implementation may be found in Appendix A.3.

Choosing k_i

The accuracy of the algorithm is theoretically guaranteed only when the number of edges considered for removal at each stage is at least the maximum number of edges across all graphs which induce districts G_i with $\text{dev}(G_i) \leq D$, i.e., $k_i \geq K_i$. Unfortunately, K_i is generally unknown in practice. We could conservatively set $k_i = |\tilde{V}_{i-1}| - 1$, the number of edges in each spanning tree but such a choice results in a prohibitively inefficient algorithm—the random edge selected for removal will with high probability induce an invalid partition, leading to a rejection of the entire map. In practice we estimate k_i before each

SMC stage by generating random spanning trees, as we detail in Appendix A.3. Theoretical support for this estimate is given by Proposition A.1, which bounds the inaccuracy of the approximation with a user-selectable parameter.

Calculating $\psi([\xi])$

The number of sequentially valid labelings $\psi([\xi])$ corresponding to an unlabeled plan $[\xi]$ can vary significantly across unlabeled plans, and has no closed-form expression. Unfortunately, $\psi([\xi])$ grows rapidly in n, but not nearly as fast as $n!$, the total number of labelings per unlabeled plan, which makes both direct and approximate calculation methods challenging. We adopt a hybrid approach in practice. When $n \leq 13$, we directly compute $\psi([\xi])$ with a recursive divide-and-conquer algorithm. When $n > 13$, we can estimate $\psi([\xi])$ to arbitrary precision using a particular importance sampling scheme. Both of these approaches, which are detailed in Appendix A.3, do not increase the runtime of the SMC algorithm appreciably.

Choosing α

As noted above, as long as $\alpha \in (0, 1]$ the SMC algorithm is asymptotically valid. Larger values are more aggressive in downweighting unlikely plans (those which are over-represented in q versus π), which may lead to less diversity in the final sample, while smaller values of α are less aggressive, which can result in more variable final weights and more wasted samples. Liu et al. (2001) recommend a default choice of $\alpha = 0.5$, which we have found appropriate, though in general we have not found the algorithm's performance to be very sensitive to α within the range $0.5 \leq \alpha \leq 0.9$.

Diagnostics

As with any complex sampler, application of the SMC algorithm in practice can be greatly facilitated by a set of diagnostic measures. Diagnostics can never prove that an algorithm is working correctly, but they can

identify situations when the algorithm fails and shed light on why. We discuss several useful diagnostics below, all of which are implemented in our open-source software.

Our primary recommendation is to perform multiple independent runs of the SMC algorithm, which enables both checking for nonconvergence to the target distribution as well as the calculation of standard errors for quantities of interest. For convergence, we adopt the Gelman-Rubin $\hat{R}$ statistic (Gelman and Rubin, 1992), which compares between-run variation to within-run variation. If the latter is an appreciable fraction of the former, then independent runs produce different results, and the algorithm has not converged. For our purposes, we will consider an algorithm not to have converged if $\hat{R}$ values for statistics of interest exceed 1.05, though higher or lower thresholds may be appropriate for given computational budgets and other applied considerations. If the algorithm has not converged for a particular S, analysts should run the algorithm again with a larger S, until $\hat{R}$ and other diagnostics discussed below indicate that the results are trustworthy. At this point, the samples from the multiple runs can be combined to increase the precision of the estimates (Gelman et al., 2013, Chapter 11).

We use the rank-normalized and folded $\hat{R}$ of Vehtari et al. (2019), which makes the statistic more robust to heavy tails and more sensitive to discrepancies in scale, not just location. For our purposes, $\hat{R}$ is advantageous, since it is computed on summary statistics and thus provides a measure of convergence tailored to practical quantities of interest. Additionally, it is unitless, and so can be compared across runs and indeed across algorithms, as we will do in Section 1.5.

Multiple runs can also be easily used to compute standard errors of expectations taken relative to the sampled plans. Recently, Lee and Whiteley (2018) and Olsson and Douc (2019), among others, have developed a method to estimate SMC variance by using information on the ancestry of each sampled particle (here, plan). While we have implemented this method in our open-source software and found it to be cor-

rect on average, the resulting estimates tend to be noisier than those obtained from a simple multiple-run standard error.

Other diagnostics are useful for assessing the overall quality of the sample and pinpointing where issues arise when the algorithm is not working well. One measure of quality is *sample diversity*, or how different the samples are from each other, which we measure using the variation of information metric shown in Equation (1.2). Similar plans will have a variation of information near zero, while plans which are extremely different will have a variation of information closer to 1. A non-diverse sample will have many copies of similar or identical plans, which tends to increase sampling error and reduces the interpretability of the generated samples.

When sampling problems arise, a closer look at each iteration of Algorithm 2 can help identify the cause. A low acceptance rate in step (b)(1)(3), or weights which are highly variable or have a heavy right tail in step (b)(2), can lower the sampling efficiency and diversity. Examining the variance of the log-weights at each iteration can be useful. Variances which increase consistently across iterations can be a sign to use a higher α, which will sample more aggressively earlier on to avoid a build-up of variable weights. We have also found that the number of unique plans from the previous iteration which survive to the subsequent iteration is a highly useful indicator of sampling problems, since it captures inefficiencies introduced by both the rejection and resampling steps. If the number of unique plans is substantially below what would be expected from a uniform sample with replacement from the same population, then some kind of bottleneck or other sampling problem is generally present.

All of the diagnostics presented here are sample statistics, in that they vary across runs of the random SMC algorithm. If the number of independent parallel runs were increased to infinity, these diagnostics would converge to a "population value" for the particular sampling problem and choice of algorithm pa-

rameters. But with a finite number of runs, there will be variation around this population value, which can lead to diagnostic values that are too optimistic or pessimistic. No diagnostic is foolproof. However, taken together, the set of diagnostics presented here is designed to catch insufficient sample sizes and too-strong constraints with high probability. Indeed, these diagnostics formalize and extend the so-called "multi-start heuristic" advocated by some MCMC redistricting practitioners (Cannon et al., 2022).

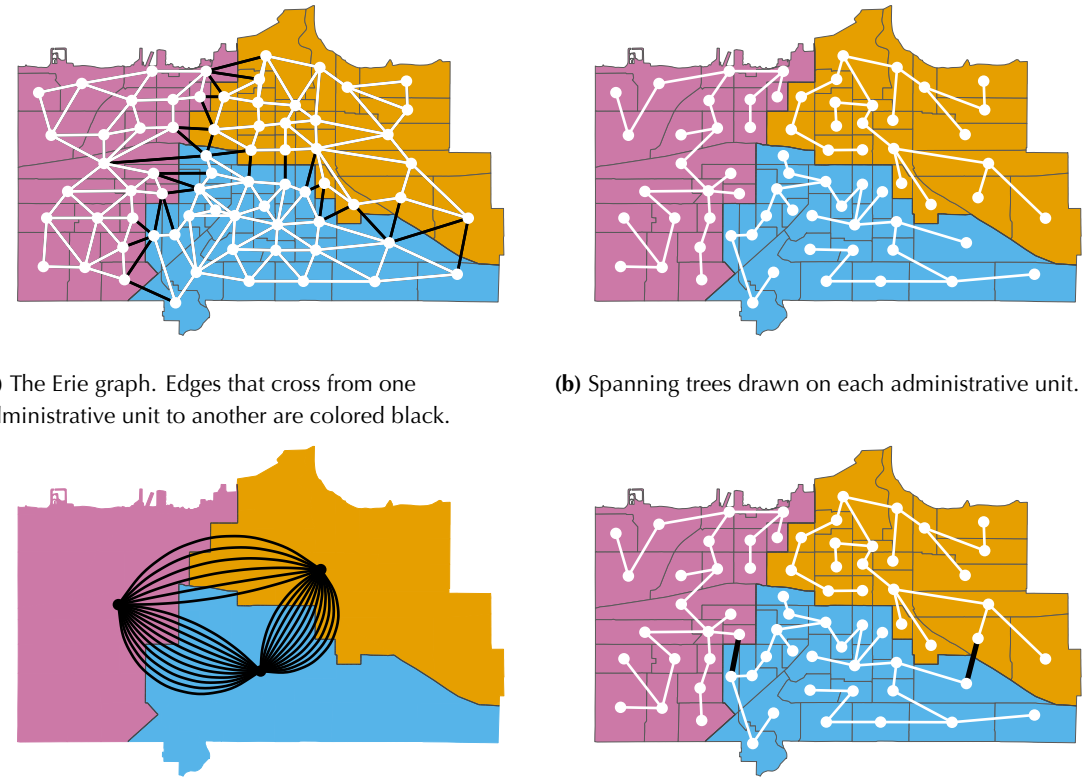
1.4.5 Incorporating Administrative Boundary Constraints

Another common requirement for redistricting plans is that districts "to the greatest extent possible" follow existing administrative boundaries such as county and municipality lines.[1] In theory, this constraint can be formulated using a J function which penalizes maps for every county line crossed by a district. In practice, however, we can more efficiently generate desired maps by directly incorporating this constraint into our sampling algorithm.

Fortunately, with a small modification to the proposed algorithm, we can sample redistricting plans proportional to a similar target distribution but with the additional constraint that the number of administrative splits not exceed $n - 1$. The hard constraint of $n - 1$ splits cannot be adjusted upwards or downwards, unfortunately, but a preference for even fewer administrative splits may be incorporated through the J function.

Let $\mathcal{A}$ be the set of administrative units, such as counties. We can relate these units to the nodes by way of a labeling function $\eta : V \to \mathcal{A}$ that assigns each node to its corresponding unit. This function induces an equivalence relation $\sim_\eta$ on nodes, where $v \sim_\eta u$ for nodes v and u iff $\eta(v) = \eta(u)$. If we quotient G by this relation, we obtain the administrative-level multigraph $G / \sim_\eta$, where each vertex is

[1] If a redistricting plan must always respect these boundaries, we can simply treat the administrative units as the nodes of the original graph to be partitioned.



(a) The Erie graph. Edges that cross from one administrative unit to another are colored black.

(b) Spanning trees drawn on each administrative unit.

(c) The quotient multigraph. Edges connecting boundary units in the original graphs become edges in the multigraph.

(d) The final spanning tree. A spanning tree on the quotient multigraph connects the trees on each administrative unit.

Figure 1.3: The two-step spanning tree sampling procedure applied to the city of Erie, Pennsylvania, with three arbitrary administrative units indicated by the colored sections of each map.

an administrative unit and every edge corresponds to an edge in G which connects two nodes in different administrative units. We can write the number of administrative splits as

$$\text{spl}(\xi) = \left(\sum_{a \in \mathcal{A}} \sum_{i=1}^{n} C(\eta^{-1}(a) \cap \xi^{-1}(i)) \right) - |\mathcal{A}|,$$

where $C(\cdot)$ counts the number of connected components in the subgraph $\eta^{-1}(a) \cap \xi^{-1}(i)$.

To implement this constraint, we draw the spanning trees in step (a) of Algorithm 2 in two substeps such that we sample from a specific subset of all spanning trees. First, we use Wilson's algorithm to draw

a spanning tree on each administrative unit $a \in \mathcal{A}$, and then we connect these spanning trees to each other by drawing a spanning tree on the quotient multigraph $\widetilde{G}_i / \sim_\eta$. Figure 1.3 illustrates this process. This approach is similar to the independently-developed multi-scale merge-split algorithm of Autry et al. (2020).

Drawing the spanning trees in two steps limits the trees used to those which, when restricted to the nodes $\eta^{-1}(a)$ in any administrative unit a, are still spanning trees. The importance of this restriction is that cutting any edge in such a tree will either split the map exactly along administrative boundaries (if the edge is on the quotient multigraph) or split one administrative unit in two and preserve administrative boundaries everywhere else. Since the algorithm has $n-1$ stages, this limits the support of the sampling distribution to maps with no more than $n-1$ administrative splits.

The two-step construction makes clear that the total number of such spanning trees is given by

$$\tau_\eta(\widetilde{G}_i) \;=\; \tau(\widetilde{G}_i / \sim_\eta) \prod_{a \in \mathcal{A}} \tau(\widetilde{G}_i \cap \eta^{-1}(a)), \tag{1.6}$$

where $\widetilde{G}_i \cap \eta^{-1}(a)$ denotes the subgraph of $\widetilde{G}_i$ which lies in unit a, and we take $\tau(\varnothing) = 1$. Replacing τ with τ_η in the expression for the weights $w_i^{(j)}$ and $w^{(j)}$ then gives the modified algorithm that asymptotically samples from

$$\pi_\eta([\xi]) \;\propto\; \exp\{-J(\xi)\}\tau_\eta(\xi)^\rho \mathbf{1}_{\{\xi \text{ connected}\}} \mathbf{1}_{\{\mathrm{dev}(\xi) \leq D\}} \mathbf{1}_{\{\mathrm{spl}(\xi) \leq n-1\}}. \tag{1.7}$$

This idea can in fact be extended to arbitrary levels of nested administrative hierarchy. We can, for example, limit not only the number of split counties but also the number of split cities and Census tracts to $n-1$ each, since tracts are nested within cities, which are nested within counties. To do so, we begin by drawing spanning trees using Wilson's algorithm on the smallest administrative units. We then con-

nect spanning trees into larger and larger trees by drawing spanning trees on the quotient graphs of each higher administrative level. (In this approach, cities which cross county boundaries must be split in two, and regions outside of cities must be treated as their own pseudo-city.) Even with multiple levels of administrative hierarchy, the calculation of the number of spanning trees is still straightforward, by analogy to Equation (1.6).

1.5 An Empirical Validation Study

Although the proposed algorithm has desirable theoretical properties, it is important to empirically assess its performance (Fifield et al., 2020b). We examine whether or not the proposed algorithm can produce a sample of redistricting maps that is actually representative of a target distribution, and how the diagnostics of Section 1.4.4 correlate with sampling accuracy.

Our validation setting is the 6-by-6 grid map shown in Figure 1.4. This map is small enough to obtain all possible redistricting plans with six contiguous districts using an efficient enumeration algorithm (see, e.g., Fifield et al., 2020b). Each precinct in the grid has an equal population. There are over 356 billion unlabeled plans with six districts, 451,206 of which have districts with exactly balanced populations.[2] We demonstrate that the proposed algorithm can efficiently approximate a realistic target distribution on this set of balanced plans. Appendix A.2 contains an additional validation study with four districts using a 50-precinct map taken from the state of Florida, where the population of each precinct varies.

We set the target distribution by choosing $\rho = 1$, which as discussed above will generate compact districts. Since this value of ρ makes the proposal and sampling distributions as close as possible, the SMC algorithm will operate at peak efficiency. The validation study in Appendix A.2 explores a wider range of

[2]The enumerated plans are publicly available at https://github.com/gerrymandr/trees/tree/main/enumerations.

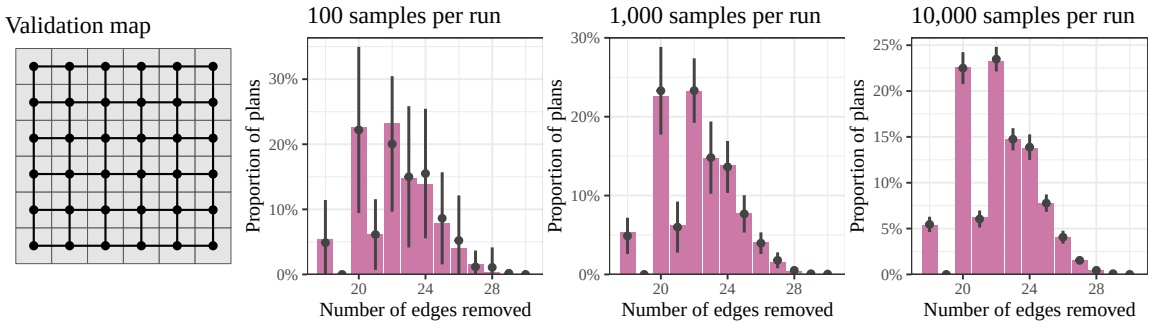


Figure 1.4: The 6-by-6 map used in validation (left). The remaining panels compare the distribution of the number of removed edges in the enumerated distribution (purple bars) and the sampled distribution (grey) over a number of sample sizes S. The dots are the mean histogram estimate across the 24 independent runs, and represent the algorithm's bias. The vertical lines are 90% confidence intervals for the histogram estimates, and represent the sampling variability. They are calculated from the standard deviation of the estimates across the runs.

values of ρ. As we have noted in Section 1.3, values of ρ other than 1 may not be computationally feasible for larger redistricting problems with many more precincts.

We run Algorithm 2 on a logarithmically-spaced range of sample sizes S ranging from 10 to 10,000. While $S = 10$ is far too small to use in any real analysis, we include it here for illustrative purposes to demonstrate the high variability and potential bias of the algorithm when S is too small. At each sample size, we perform 24 independent runs of the algorithm, each of size S, which allows for accurate calculation of standard errors and $\hat{R}$ for summary statistics.

To validate the algorithm, we compare the distribution of a summary statistics from each sample to the true distribution obtained by reweighting the enumerated plans according to the target measure, which is proportional to $\tau(\xi)$. The summary statistic used here is the number of edges that must be removed to form each set of districts, $\text{rem}(\xi) \cdot |E(G)|$. As discussed in Section 1.3.3, this statistic is highly correlated with the target $\tau(\xi)$, and so is a good test of the algorithm.

These comparisons are summarized in the right three panels of Figure 1.4, which show the true distribu-

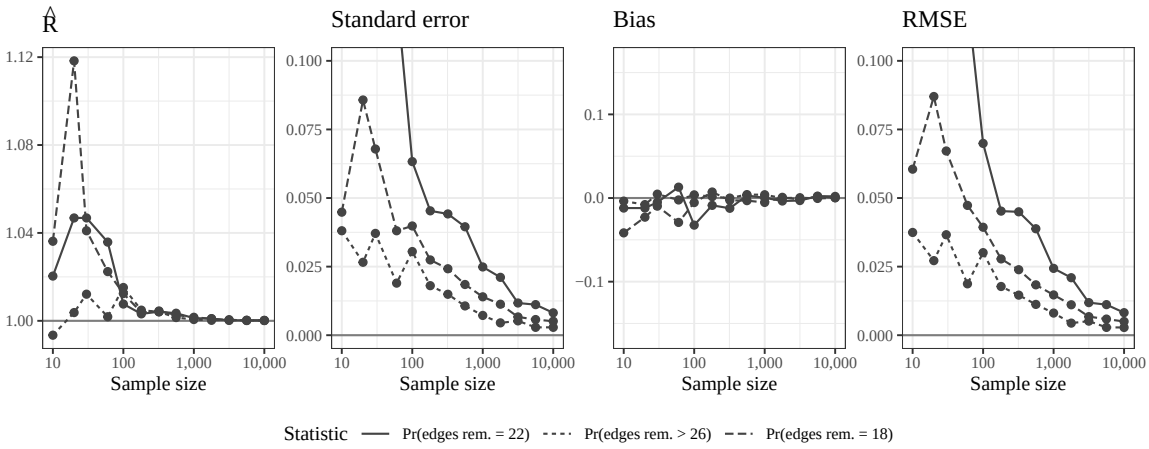


Figure 1.5: Convergence diagnostic $\hat{R}$ (on a log scale), standard errors, bias, and RMSE for the same three compactness summary statistics, calculated across 24 independent runs of the SMC algorithms. Values are plotted versus sample size S per run.

tion of the number of edges removed in purple, with the estimated histogram overlaid as solid points, for a range of sample sizes. Each histogram estimate produced by the SMC algorithm is accompanied by a 90% confidence interval produced using the standard errors calculated across the 24 independent runs (i.e., the standard deviations of the estimates for each run). With as few as $S = 100$ samples, the SMC estimates are already close to the true distribution. In fact, if we tested the null hypothesis that each estimate is centered at the true value, we would only reject for one of the twelve nonzero true values at the 10% level, within the range of what would be expected by chance given the sampling variability. As the sample size increases, the variance of the SMC estimates decreases, until by $S = 10,000$ samples it is mostly negligible. Overall, the agreement between the sampling and target distributions is excellent, indicating that the SMC algorithm is able to accurately sample from the target distribution in this setting.

We next study more quantitatively the quality of the SMC sample as the sample size increases, and in particular how well the diagnostics recommended in Section 1.4.4 can be used as a proxy for this quality. To

do so, we focus on three particular estimands that can be obtained from the distribution of edges removed: the probability of removing exactly 22 edges (the target median; true probability 0.2328), the probability of removing exactly 18 edges (true value 0.0542), and the probability of removing more than 26 edges (true value 0.0204). The latter two statistics are especially useful as tests of the algorithm's sampling accuracy, since they require estimating small tail probabilities which may correspond to very few redistricting plans. On the lower tail, there are just 2 possible unlabeled plans which remove exactly 18 edges, though together they comprise over 5% of the target distribution. The upper tail probability corresponds to the p-value a researcher would use for testing whether an enacted plan with 26 removed edges was significantly less compact than would be expected under the target distribution.

For each sample size from the SMC algorithm, we calculate the $\hat{R}$ diagnostic, and standard errors for these three summary statistics. Note that neither of these calculations requires access to the enumerated ground truth. We also compute the bias and the root mean-square error (RMSE) for the estimates, comparing to the true values from the enumerated target distribution. Ideally, the always-computable $\hat{R}$ and standard errors will well represent the generally uncomputable bias and RMSE.

Figure 1.5 summarizes the results of this experiment. As expected, as the SMC sample size S increases, $\hat{R}$, the standard errors, and the relative RMSE also decrease. The $\hat{R}$ values fall below our recommended cutoff of 1.05 by $S = 60$, indicating that for samples this large and greater, each independent run is producing similar estimates and the SMC algorithm has likely converged to the target distribution. Indeed, the estimates' bias is closely centered around zero in this range. However, standard errors are significantly larger for the smaller sample sizes. This tracks the pattern observed in the validation above: at small sample

sizes, the SMC estimates were on average unbiased for the target distribution, but had significant sampling variability.

Importantly, the calculated standard errors track the actual RMSE extremely closely, demonstrating their value as an observable proxy for this measure of sampling accuracy. In other words, with low $\hat{R}$ indicating likely algorithmic convergence (and therefore unbiasedness), the overall algorithmic error is driven almost entirely by the sampling variability, which is measurable with the standard errors.

1.6 Analysis of the 2011 Pennsylvania Redistricting

As discussed in Section 1.2, in the process of determining a remedial redistricting plan to replace the 2011 General Assembly map, the Pennsylvania Supreme Court received submissions from seven parties. In this section, we compare four of these maps to both the original 2011 plan and the remedial plan ultimately adopted by the court. We study the governor's plan and the House Democrats' plan; the petitioner's plan (specifically, their "Map A"), which was selected from an ensemble of 500 plans used as part of the litigation; and the respondent's plan, which was drawn by Republican officials.

1.6.1 The Setup

To evaluate these six plans, we drew reference maps from the target distribution given in Equation (1.7) by using the proposed algorithm along with the modifications presented in Section 1.4.5 to cap the number of split counties at 17 (out of a total of 67), in line with the court's mandate. We set $\rho = 1$ to put most of the sample's mass on compact districts, and enforced $\text{dev}(\xi) \leq 0.001$ to reflect the "one person, one vote" requirement.

This population constraint translates to a tolerance of around 700 people, in a state where the median precinct has 1,121 people. Like most research on redistricting, we use precincts because they represent the

smallest geographical units for which election results are available. To draw from a stricter population constraint we would need to use the 421,545 Census blocks in Pennsylvania rather than the 9,256 precincts, which would significantly increase the computational burden.

1.6.2 Comparison with an Analogous MCMC Algorithm

We first compare the computational efficiency of the SMC algorithm with a merge-split-type MCMC algorithm (Carter et al., 2019; DeFord et al., 2021). The merge-split algorithm here uses a spanning tree-based proposal identical to the splitting procedure described in Algorithm 1: it merges adjacent districts, draws a spanning tree on the merged district, and splits it to ensure the population constraint is met. A Metropolis-Hastings correction is used, analogous to the SMC reweighting step. In fact, the implementation (contained in our open-source `redist` package) shares much of the same code base, which allows for performance comparisons that better reflect the different algorithmic strategies, rather than different algorithmic implementations.[3]

The MCMC chains are each initialized with a random SMC sample, and run for 500 warm-up iterations before samples are recorded. We let the sample size vary from 100 to 4,000 for the SMC algorithm and from 100 to 8,000 for the MCMC algorithm. For each sample size, we perform four independent runs of each algorithm. The average acceptance rate for the MCMC algorithm across the four chains was 88.7%.

Figure 1.6(a) shows the $\hat{R}$ values of two summary statistics (Democratic seats and compactness) by sample size S per run. By this convergence heuristic, the SMC algorithm converges more quickly, with lower $\hat{R}$ values than the MCMC algorithm for a given sample size. According to the check that $\hat{R} \leq 1.05$, the SMC algorithm has approximately converged with $S = 1,600$, while the MCMC algorithm has still not

[3] A particularly performant implementation of a similar algorithm (Reversible ReCom, Cannon et al., 2022) may be found at https://github.com/pjrule/frcw.rs

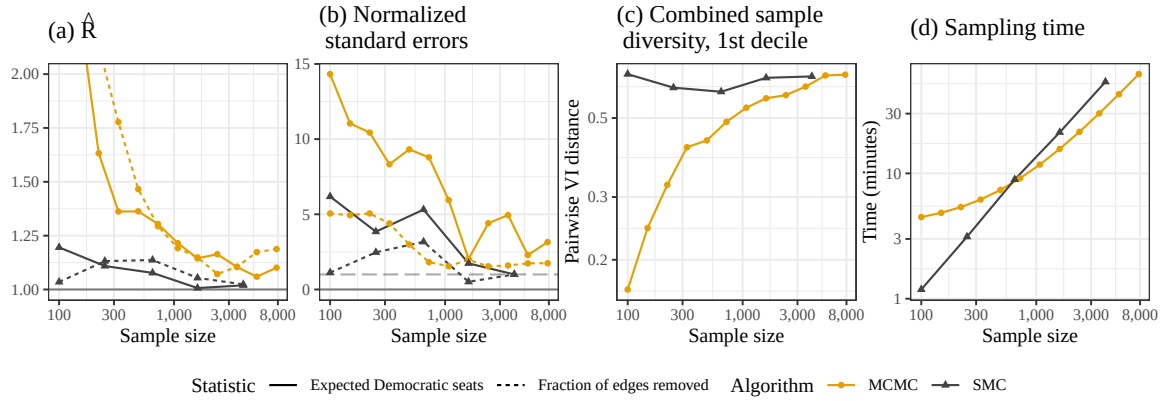


Figure 1.6: A comparison of sampling efficiency for SMC and MCMC algorithms using the same transition kernel. (a) Convergence diagnostic $\hat{R}$ for two summary statistics plotted against the sample size. (b) Standard errors for two statistics, normalized by the standard error for the SMC algorithm at 4,000 samples (to put the two statistics on the same scale), by sample size. (c) First decile of the pairwise variation of information distance (a measure of sample diversity) by sample size. Distance measures were computed with the combined sample of four independent runs. (d) Sampling time by sample size.

converged after with than 8,000 post-warmup samples (over 7,000 unique accepted proposals), at which point the MCMC algorithm has been running 2.4 times longer than the SMC algorithm.

We find a similar story in examining the standard errors of the two statistics, which are shown in Figure 1.6(b) after being normalized to be on comparable scales. With four independent runs per algorithm, standard error estimates themselves are rather noisy, but the overall trend mirrors the findings for $\hat{R}$. Even at large sample sizes, the MCMC standard errors are 1.5–5 times larger than their SMC counterparts, although the MCMC standard errors are not meaningful without the underlying chains having converged.

Figure 1.6(c) examines the sampling efficiency from the perspective of sample diversity. As discussed in Section 1.4.4, the pairwise variation of information (VI) distance provides a way to quantify how similar the plans in each sample are to one another. If the distribution of these pairwise VI distances has a long lower tail, then there are a significant number of plans which are very similar. Figure 1.6(c) shows the first decile of the pairwise VI distances in the combined sample (i.e., pooled across all four runs) for each

algorithm as the sample size increases. Because of the sequential nature of the MCMC algorithms, the sample diversity doesn't match that of the SMC algorithms until there are at least 3,000 samples. This is reflected in the lag-1 autocorrelation of the MCMC sample, which was 0.982 for Democratic seats and 0.984 for compactness. Autocorrelation is of course not a direct measure of sampling efficiency, but this comparison serves to underscore the differences in the types of samples generated by each algorithm for finite sample sizes.

Figure 1.6(d) shows how long each algorithm takes to run for a given number of samples. Because of the fixed warm-up cost of the MCMC algorithm, it is slower than the SMC algorithm for fewer than 600 samples. Asymptotically, both algorithms have runtime linear in the number of samples. But the per-sample cost of the MCMC algorithm is lower, since only two districts are redrawn, rather than the entire map.

1.6.3 Compactness and County Splits

While the results of the comparison above suggest that 1,600 SMC samples is enough for Pennsylvania, we perform the rest of our analysis with 4,000 samples from one of the SMC runs. Apart from the $\hat{R}$ diagnostic examined in detail above, the other SMC diagnostics of Section 1.4.4 all indicated no problems with the generated sample.

Figures 1.7(a) and (b) show distribution of the fraction of edges removed $\text{rem}(\xi)$ and the number of county splits $\text{spl}(\xi)$ across the reference maps generated by our algorithm. The figure also shows these values for each of the six plans. The 2011 General Assembly plan is a clear outlier for both statistics, being far less compact and splitting far more counties than any of the reference plans and all of the remedial plans. Among the remedial plans, the petitioner's is almost an outlier in being *too* compact, although this

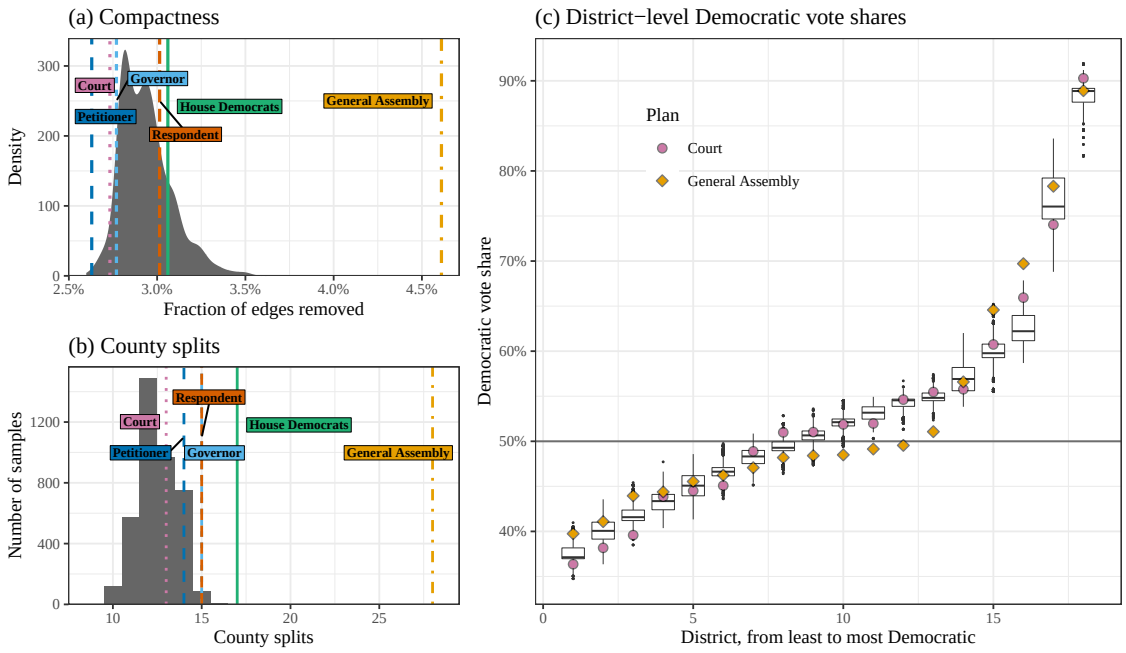


Figure 1.7: Summary statistics for the sampled and comparison plans: (a) compactness (smaller means more compact) and (b) county splits. Panel (c) shows the Democratic two-party vote share by district, where within each plan districts are ordered by Democratic vote share.

is perhaps not surprising—the map was generated by an algorithm explicitly designed to optimize over criteria such as population balance and compactness (Chen, 2017). The other plans' compactness falls within the range of the reference sample.

Figure 1.7(b) shows that all of the submitted remedial plans, as well as the reference sample, followed the court's instructions to limit the number of county splits. Yet around half of the reference maps split fewer than 13 counties, the minimum number of splits found in any of the remedial plans. This may be a result of the strict population constraint (all six plans were within 1 person of equal population across all districts), or a different prioritization between the various constraints imposed.

1.6.4 Partisan Analysis

While important, the outlier status of the General Assembly plan as regards compactness and county splits is not sufficient to show that it is a *partisan* gerrymander. To evaluate the partisan implications of the six plans, we take a precinct-level baseline voting pattern and aggregate it by district to explore hypothetical election outcomes under the six plans and the reference maps. The baseline pattern is calculated by averaging the vote totals for the three presidential elections and three gubernatorial elections that were held in Pennsylvania from 2000 to 2010.[4] These election data were also used during litigation. While being far from a perfect way to create counterfactual election outcomes, this simple averaging of statewide results is often used in academic research and courts. The values of $\hat{R}$ for the mean and various quantiles of the district-level vote shares were all within standard convergence ranges.

Then, within each plan, we number the districts by their baseline Democratic two-party vote share, so District 1 is the least Democratic and District 18 the most. Figure 1.7(c), analogous to Figure 7 in Herschlag et al. (2017), presents the distribution of the Democratic two-party vote share for each of the districts across the reference maps, and also shows the values for the General Assembly plan (orange triangles) and the court's adopted plan (purple circles). When compared to the reference maps and the court's plan, the General Assembly plan tends to yield smaller Democratic vote share in competitive districts (p-values of 0.0162, 0.0002, 0.0002, and 0.0002 for Districts 9–12 compared to the reference set) while giving larger Democratic vote share in non-competitive districts. The effect of this is to produce 6 Democratic seats, on average, under the General Assembly plan, compared to a range of 8–12 under the reference sample (p-value

[4] Data from Ansolabehere, S. and Rodden, J. (2011). Pennsylvania Data Files. Available at https://doi.org/10.7910/DVN/FJHHDS.

0.0002) and 11 under the Court plan. All of these findings provide evidence that the General Assembly plan was gerrymandered in favor of Republicans by packing Democratic voters into non-competitive districts.

1.7 Concluding Remarks

Redistricting sampling algorithms allow for the empirical evaluation of a redistricting plan by generating alternative plans under a certain set of constraints. Researchers and policymakers can compute various statistics from the redistricting plan of interest and compare them with the corresponding statistics based on these sampled plans. Unfortunately, existing approaches often struggle when applied to real-world problems, owing to the scale of the problems and the number of constraints involved.

The SMC algorithm presented here can provably asymptotically sample from a specific target distribution. Because of the constructive sampling approach, it does not face the problem of potential reducibility that MCMC algorithms have, and performs more efficiently in a realistic applied setting than a MCMC algorithm that uses an identical proposal distribution. The proposed method also incorporates, by design, the common redistricting constraints of population balance, geographic compactness, and minimizing administrative splits. We expect these advantages of the SMC algorithm to meaningfully improve the reliability of outlier analysis in real-world redistricting cases.

The proposed approach has some important limitations. Like other redistricting sampling algorithms, if the target distribution strays too far from the proposal distribution, the quality of the generated sample will suffer dramatically. Close examination of diagnostic measures like $\hat{R}$, standard errors, sample diversity, and per-iteration efficiency measures like weight variance and the number of surviving plans is required to provide confidence that the algorithm is sampling accurately and has converged for particular quantities

of interest. Our open-source software provides these diagnostics so that analysts can use the proposed algorithm successfully.

One important drawback particular to the SMC algorithm arises in situations with dozens or hundreds of separate districts. Like any other SMC or particle resampling scheme, as the number of iterations (districts) grows, eventually all of the samples will share a common ancestor. While this is not a problem in many SMC applications (such as Bayesian inference), for redistricting, this means that all of the sampled plans will share one or more districts that are completely identical. Any summary statistics which rely on these districts will have very large standard errors and may not converge even with a large number of samples. While this problem will be readily identified with the diagnostics we propose here, there is little the analyst can do to address the issue beyond increasing the sample size and possibly combining many independent runs of the algorithm. One interesting avenue for future research would be to examine whether several applications of an MCMC kernel at various points in the SMC algorithm could help refresh the sample and counteract the tendency to collapse to a single ancestor.

Future research should also explore the possibility of improving several design choices in the algorithm to further increase its efficiency. Wilson's algorithm, for instance, can be generalized to sample from edge-weighted graphs. Choosing weights appropriately could lead to trees which induce maps that are more balanced or more compact. And the procedure for choosing edges to cut, while allowing for the sampling probability to be calculated, introduces inefficiencies by leading to many rejected maps. Further improvements in either of these areas should allow us to better sample and investigate redistricting plans over large maps and with even more complex sets of constraints.

Finally, we believe that the application of these sampling algorithms to real-world redistricting problems is essential. We have applied the SMC algorithm to the 2020 Congressional redistricting in all 50 states

(McCartan et al., 2022), and this allowed us to evaluate the degree of partisan gerrymandering across states (Kenny et al., 2023). Such applications shed light on substantive political science theories, and highlight new methodological challenges.

This page intentionally left blank.

2

Individual and Differential Harm in Redistricting

2.1 Introduction

For as long as there have been district-based electoral systems, groups, individuals, and parties have manipulated them for political gain. Partisan gerrymandering, whether in the form of rotten boroughs or Elbridge Gerry's salamander-shaped district (Griffith, 1907, pp. 72–73), is perhaps the most consistently visible manifestation of this tendency. But gerrymandering takes other forms as well. In the United States, many states have historically used racial gerrymandering to maintain White political supremacy (Gilliland, 1969). At the local level, where most redistricting occurs, partisan dynamics are not always as salient, and the manipulation of the districting plan may be more subtle or may favor a narrower or less well-defined group (Cain and Hopkins, 2002; Halberstam, 2012). The diversity of types of gerrymandering is further apparent from a survey of major litigation over the past decades, which includes

allegations of gerrymandering involving party (League of Women Voters *v.* Commonwealth, 2018; Rucho *v.* Common Cause, 2019), race (Shaw *v.* Reno, 1993; Merrill *v.* Milligan, 2022), religion (United Jewish Organizations *v.* Carey, 1977; NAACP *v.* East Ramapo Central School District, 2020), incarceration status (Calvin *v.* Jefferson County, 2015; Davidson *v.* City of Cranston, 2016), and residency or citizenship (Burns *v.* Richardson, 1966; Evenwel *v.* Abbott, 2016).

The academic study of gerrymandering has unfortunately not reflected this diversity. The vast majority of research attention has been on partisan gerrymandering, and has particularly centered on how districting plans treat political parties, as captured in the *seats-votes curve* or notions of *partisan symmetry* (Tufte, 1973; Grofman, 1983; King and Browning, 1987; King, 1989; Gelman and King, 1994; Thomas et al., 2013; Katz et al., 2020). A large number of formulas for quantifying or identifying partisan gerrymanders have also been developed (Stephanopoulos and McGhee, 2015; McDonald and Best, 2015; Warrington, 2018). The study of racial gerrymandering has been largely shaped by the U.S. legal framework for challenging such gerrymanders: the Voting Rights Act of 1965 (VRA). This literature is characterized by a focus on identifying the voting preferences of racial groups through techniques such as ecological inference (O'Loughlin, 1982; Grofman and Handley, 1991; King, 1997; Greiner, 2006). Work on local redistricting is even more sparse, in part because of data limitations (but see Siegel-Hawley, 2013; Abott and Magazinnik, 2020; Vargas et al., 2021).

The dominance of the partisan framework for analyzing redistricting poses a problem even in cases of pure partisan gerrymandering. Existing tools, by focusing only on the effect of district lines on parties, are unable to examine the effect of manipulated lines on other groups. A redistricting plan that diminishes Democratic voting power will naturally have an effect on the relative voting power of White and Black vot-

ers, poor and wealthy voters, young and old voters, and urban or rural voters. These effects on individuals will be missed by any framework that only studies parties.

This paper addresses these limitations with a new framework of *individual and differential harm* for the unified study of all forms of gerrymandering. Fundamentally, gerrymandering involves the concentration or dilution of voting power to favor or disfavor a particular group. Our proposed framework focuses directly on measuring this concentration and dilution, and in doing so avoids the pitfalls of party-centric approaches.

The harm framework combines two key ideas, which are motivated at greater length in Section 2.2. First, it operates at the level of the individual voter and examines whether each voter is represented by the candidate of their choice. Doing so both unifies the approach to gerrymandering across contexts and also reflects a core goal of representative democracy: that elected officials represent the preferences of the voting public to the greatest extent possible (see also Brunell, 2010).

Second, the proposed framework moves beyond a single districting plan under study to also consider *counterfactual* districting plans. Many factors influence the performance and qualities of a districting plan. Without a well-defined comparison, it is impossible to differentiate whether certain aspects of the system, like its partisan performance, are intrinsic consequences of laws and political geography or result from improper manipulation of the system for political gain (Chen and Rodden, 2013). For example, in Massachusetts, it is currently impossible to draw a set of congressional districts that produces even one reliable Republican seat, despite Republican voters making up roughly 32% of the state's voters (Duchin et al., 2018). The 9-0 Democratic composition of the Massachusetts congressional delegation therefore reflects the homogeneity of the state's political geography and not any partisan gerrymandering. Once a set of districting rules has been agreed on, what matters is not how a plan measures up to a fixed ideal, but how

it performs in comparison to counterfactual plans. While many existing partisan gerrymandering metrics make use of an *implicit* counterfactual plan, such a plan may not be of substantive interest or even feasible at all. By making the comparison set explicit, we can better reflect legal and geographic realities, while making the interpretation of our measures clearer.

The combination of individual voter choices and counterfactual districting plans defines our notion of *individual harm*: informally, a voter is harmed if they are more likely to be represented by the candidate of their choice under a different districting plan. And by analyzing how harms fall differentially along partisan, racial, ideological, and other lines, we can connect individual harm to traditional notions of gerrymandering. Section 2.2 illustrates this with a simple example, and the general definitions are formalized in Section 2.3.

A key feature of the proposed framework is its clear separation of theoretical quantities of interest and empirical strategies for estimating them (Katz et al., 2021). This distinction is also missed by many critics of the efficiency gap (Veomett, 2018; DeFord et al., 2020). The challenge of estimating harm and differential harm from observed data is discussed in Section 2.4. Recent advances in redistricting simulation, which can generate a large number of random districting plans that respect relevant legal constraints, are critical to this effort (Fifield et al., 2020a; DeFord et al., 2021; Carter et al., 2019; McCartan and Imai, 2023, the latter being Chapter 1). These algorithmically-generated sets of districting plans can provide a realistic set of counterfactuals for our measures.

The harm framework has important similarities with the common-law approach to voting rights litigation. To proceed with a legal challenge to an alleged gerrymander, a voter must demonstrate that their individual right to vote has been harmed by the enacted district boundaries (Reynolds *v.* Sims, 1964; Baker *v.* Carr, 1962; Issacharoff and Karlan, 1998). This requires proof that *but for* the alleged gerrymander, the

voter would have been more likely to elect their chosen candidate (Vieth *v.* Jubelirer, 2004; League of Women Voters *v.* Commonwealth, 2018; Common Cause *v.* Lewis, 2019). Indeed, a primary factor in the Supreme Court's ruling that partisan gerrymandering is not justiciable was that existing redistricting measures are not clearly tied to individual injury (Gill *v.* Whitford, 2018). A more detailed discussion of the connections between the proposed framework and existing practices for analyzing racial and partisan gerrymandering may be found in Section 2.5.

Harm and differential harm are more than a useful conceptual unification—they are practically useful in a wide variety of redistricting analyses. A brief study of partisan gerrymandering of New Jersey congressional districts in Section 2.6 shows how existing gerrymandering metrics can make incorrect conclusions about the magnitude and even the direction of partisan gerrymanders. Section 2.7, studies voting rights litigation for Alabama congressional districts, and uses differential harm to separate the racial and partisan effects of the enacted districting plan. Section 2.8 demonstrates the application of our framework to local redistricting, and quantifies the benefits of race-aware city council district boundaries to Black voters in the city of Boston. Section 2.9 concludes.

2.2 Motivation

The core idea motivating individual and differential harm is a simple one. Consider *Minissouri*, a stylized representation of Missouri which is shown in Figure 2.1. The state's total population is 180, and it is divided into three districts with population between 54 and 66. Minissouri has two medium-sized cities and a substantial rural population with an overall partisan breakdown of 55.6% R to 44.4% D. Voters' preferences are assumed to be stable across elections, and turnout is 100%. In addition to partisan affiliation indicated by red or blue coloring, each voter also belongs to a racial group indicated by squares and circles.

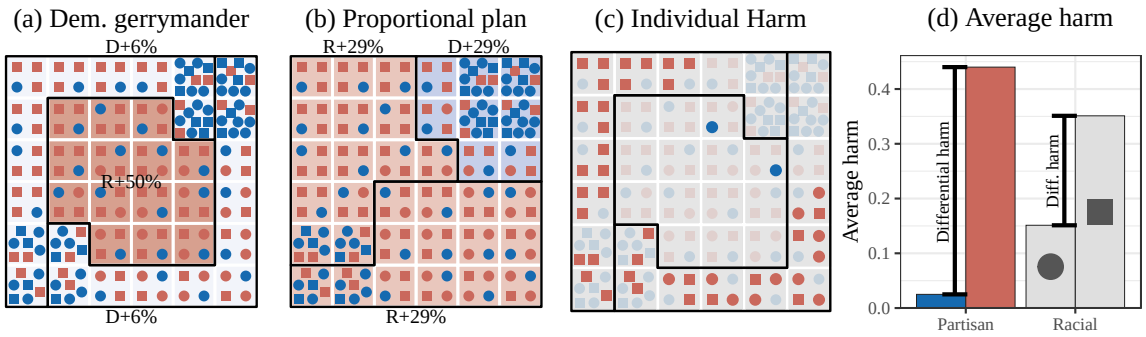


Figure 2.1: Two potential districting plans for Minissouri, a Democratic gerrymander (a), and a close-to-proportional plan (b). Panel (c) shows which voters are harmed by the Democratic gerrymander, using the proportional plan (b) as a counterfactual. Panel (d) summarizes this across precincts, showing the average harm done to voters of each partisan and racial group.

Now suppose the redistricting plan in Figure 2.1(a) has been enacted. Despite the statewide popular vote favoring Republicans, under this plan Democrats will always win two of the three seats. In contrast, a more proportional plan 2.1(b) elects two Republicans and one Democrat.

Intuitively, plan (a) unfairly packs Republicans into a rural district, depriving them of one seat. To turn this intuition into something more concrete, we can identify which voters are harmed by the enactment of plan (a) over the alternative (b). These voters are highlighted in Figure 2.1(c). For example, in the upper leftmost precinct, there are three Republican square voters who are placed into a Democratic district in plan (a) but would have been placed into a Republican district in plan (b), and thus are harmed. In contrast, Republican voters in the northeast city are not harmed, since they reside in a Democratic district in both the enacted and counterfactual plan.

After identifying which voters are harmed by plan (a) compared to counterfactual plan (b), it is natural to examine the overall composition of these harmed voters. Of the 46 harmed voters, 44 are Republicans and only 2 are Democrats. These represent 44% of all Republicans and 2.5% of all Democrats. We call the gap between these two rates the *differential harm*, and propose it as a general way to account for the

impact of a districting plan on particular groups. Here, then the partisan differential harm of plan (a) is -41.5%, meaning that the average Republican voter in Mininois is 41.5 percentage points less likely to be represented by a Republican than the average Democrat is to be represented by a Democrat, *compared to plan (b)*.

We need not limit our consideration to partisanship. We can also tally harm by *race*, which in Minissouri is correlated with both geography and partisanship. Plan (a) harms 13 circle voters and 33 square voters—a differential racial harm of -20%, where positive values indicate pro-square bias. In other words, square voters are 20 percentage points more likely to be harmed than circle voters. While this number cannot establish whether the drawers of plan (a) intended to harm square voters, or merely aimed to maximize Democratic representation, it does quantify the impact of plan (a) on these voters versus the alternative plan (b), and it does so on the same scale as differential partisan harm.

2.3 Individual and Differential Harm

We formalize and generalize the concepts in Section 2.2. Let $\mathcal{I}$ be the set of individuals, or *voters*, in the region and time period. Individuals vote for a candidate to represent districts, which are indexed by $\mathcal{D}$.

Individuals vote in elections, each of which produces a specific outcome v from the set $\mathcal{V}$ of all possible election outcomes. In an election, voters either vote for a candidate running in their district, or abstain from voting. We write $v(i)$ to mean the candidate that individual i votes for in election v. The set of all individual vote choices (which may include the choice to abstain from voting) fully determines the specific election outcome.

Every voter resides in some legislative district. This assignment defines a *districting plan* and is rep-

resented here by a function $q(i)$ indicating the district that individual i lives in. The set of all possible districting plans is $\mathcal{Q}$.

In an election v, the votes are tallied and the candidate with the most votes wins. Using potential outcomes notation, we denote the winner of voter i's district under districting plan q and election v as $\text{win}_i(v, q)$. This notation obscures the complex relationships and feedback effects between individuals, elections, and districting plans. For example, different districting plans may lead to different candidates running, depending on district partisanship and incumbent residency. The assumptions we make on these relationships will be made more explicit in Section 2.4, where we discuss how to estimate our measures in practice.

2.3.1 Harm

The key intuition from Section 2.2 is that to be harmed, a voter must have a better outcome in the counterfactual world where a different districting plan was in place. Letting $\tilde{q}$ denote a counterfactual districting plan, we define the *harm* to an individual i in election v held under districting plan q as

$$h_i(v, q, \tilde{q}) = \mathbf{1}\{v(i) \neq \text{win}_i(v, q)\}\mathbf{1}\{v(i) = \text{win}_i(v, \tilde{q})\}.$$

A voter is harmed only if they are not represented by the candidate of their choice under q ($\mathbf{1}\{v(i) \neq \text{win}_i(v, q)\}$) but they would have been represented by their candidate of choice under the counterfactual plan $\tilde{q}$ ($\mathbf{1}\{v(i) = \text{win}_i(v, \tilde{q})\}$).

Notice the fundamental asymmetry in this definition: $h_i(v, q, \tilde{q}) \neq h_i(v, \tilde{q}, q)$, as $\mathbf{1}\{v(i) \neq \text{win}_i(v, q)\} \neq \mathbf{1}\{v(i) \neq \text{win}_i(v, \tilde{q})\}$. Exchanging the role of the enacted plan and the counterfactual will not simply flip the direction of our conclusion.

In general, q harms some individuals versus $\tilde{q}$, and $\tilde{q}$ harms a distinct, though not necessarily disjoint, set of individuals, versus q. A symmetric version of $h_i(v, q, \tilde{q})$ would need to combine both of these groups: those harmed and those who gain by electing their chosen candidate under q but not under the counterfactual $\tilde{q}$. In practice, voters, courts, or politicians may differ in their views on how to balance preventing harm and maximizing gains.

Our definition of harm is limited in two regards. First, it considers only individuals who vote: people who abstain from voting have experienced zero harm. People who do not vote express no formal preference in electoral outcomes, even if they have informal preferences. The natural theoretical extension of this definition is to then separate individuals' candidate preferences from their vote choice; the latter defines election winners, but the former provides a more-broad harm measure. Naturally, implementing this extension in practice would require auxiliary data, such as a large-scale survey, since candidate preferences for non-voters cannot be directly inferred from election returns.

A second limitation is that $h_i(v, q, \tilde{q})$ ignores any preferences a voter may have between the candidates that they did not vote for. For example, a voter may consider candidates A and B nearly interchangeable, but may strongly dislike candidate C. If the voter ultimately voted for candidate A, this harm measure makes no distinction between candidate B or candidate C winning the voter's district. This reflects the simplistic nature of the cost function implicitly assumed in our proposed measure. The problem could be remedied by replacing the indicator functions in the measure with other cost functions that consider ordinal preferences. Ballot-level ranked-choice vote data, or candidate ideology scores, could be used to estimate such a cost function. Here, for simplicity and to avoid making further assumptions, we focus on the

indicator-based cost function. Section 2.8 further discusses the implications of this particular specification in the context of primary elections.

2.3.2 Multiple Elections or Districting Plans

As defined above, $h_i(v, q, \tilde{q})$ captures harm in a single election and relative to one counterfactual districting plan $\tilde{q}$. In general, we are interested in outcomes over multiple elections and a much larger set of counterfactual plans. Ideally, we would average individual harm over a set of future elections under the enacted districting plan and over a set of hypothetical districting plans that are consistent with districting criteria. We could also consider averaging over a set of individuals with some shared characteristics.

To allow for such averaging, we replace the fixed election v, counterfactual districting plan $\tilde{q}$, and individual voter i with random variables $V \in \mathcal{V}$, $\tilde{Q} \in \mathcal{Q}$, and $I \in \mathcal{I}$. The random variable V can be conceptualized as ranging over possible future elections—one realization could correspond to a heavily Democratic-favoring midterm year, while another could involve incumbents under-performing particularly badly. The random variable $\tilde{Q}$ captures the range of potential counterfactual districting plans. Rarely would an analyst have a single counterfactual plan in mind: different realizations of $\tilde{Q}$ correspond to different districting plans that are permissible under the relevant districting criteria. Finally, I is defined to be a voter taken uniformly at random from $\mathcal{I}$, with $I \perp\!\!\!\perp V, \tilde{Q}$.

The randomness in V and $\tilde{Q}$ allows us to compare the enacted districting plan to many plausible counterfactual alternatives, and across a range of electoral environments. At times we may wish to study on a particular election or a particular counterfactual plan. Treating V and $\tilde{Q}$ as random variables allows us to condition on particular values of each. A particular specification of the distribution of V and $\tilde{Q}$ is required only when *estimating* our harm measures in a particular application, a topic we return to in Section 2.4.

With these definitions, the harm for a (randomly selected) individual under plan q can be written as

$$H(q) \coloneqq h_I(V, q, \tilde{Q}).$$

$H(q)$ is itself a random variable, representing the harm to *some* individual in *some* election compared to *some* counterfactual plan. The random variable I here serves as a kind of "veil of ignorance" (Rawls, 2004): without picking a specific individual, a districting plan q which tends to produce a low $H(q)$ would be expected to be preferred over a plan which tended to produce a higher $H(q)$.

Despite being only a scalar quantity, $H(q)$ actually fully captures the harm created by the plan q. We can zoom in to specific dimensions of the plan by conditioning on values of V, $\tilde{Q}$, and I, or by taking expectations. Thus, for example, the harm for a single individual i becomes $H(q) \mid I = i$, and the electorate's average harm in a particular election v becomes

$$\mathbb{E}[H(q) \mid V = v] = \mathbb{P}[v(I) \neq \text{win}_I(v, q) \text{ and } v(I) = \text{win}_I(v, \tilde{Q})],$$

that is, the probability that a randomly selected voter in this election is misrepresented under the adopted plan but would not have been under a randomly selected counterfactual.

2.3.3 Differential Harm and Fairness

Single-member districts are often motivated by the value in representation to specific communities with shared interests. Preserving "communities of interest" is one of the most common statutory and constitutional requirements imposed by states on the plan-drawing process (National Conference of State Legislatures, 2021). It is therefore natural to consider not just the average harm to voters done by a plan, but also how this harm is distributed across certain groups of voters, defined by race, party, or other attributes.

To that end, let $\mathcal{G} = \{G_1, G_2, \dots\}$ be a set of mutually exclusive groups of voters, so that for each

voter i, we have $i \in G_k$ for some group G_k. In general we will not require the group identities to be fixed with respect to elections.[1] This is because some group identities, like partisan preference, may themselves change election-to-election.

Once we have defined a set of groups, we say that a plan q is *fair* for groups $\mathcal{G}$ in comparison to a counterfactual $\tilde{Q}$ if $\mathbb{E}[H(q) \mid I \in G_k]$ is constant across groups $G_k \in \mathcal{G}$. Fairness is analogous to the notion of *error rate balance* in the algorithmic fairness literature, which seeks to equalize false positive and false negative rates across protected classes (Zafar et al., 2017; Chouldechova, 2017). There, the implicit counterfactual is a perfect predictor. Compared to the perfect predictor, the actual decision process harms individuals whenever it makes a false positive or false negative error. Here, there can be no perfect districting plan, and the counterfactual instead consists of a collection of plausible alternatives.

Fairness is a strict condition: real-life plans are highly unlikely to be exactly fair. To quantify deviations from fairness, we introduce the concept of *differential harm*, defined as

$$\Delta_{G_k - G_l} H(q) \coloneqq \mathbb{E}[H(q) \mid I \in G_k] - \mathbb{E}[H(q) \mid I \in G_l],$$

for two groups G_k and G_l. So $\Delta_{G_k - G_l} H(q)$ will be positive if voters in group G_k are harmed at a higher rate than voters in group G_l, indicating that plan q favors G_l over G_k. Using our definition of fairness, $\Delta_{G_k - G_l} H(q) = 0$ for all $G_k, G_l \in \mathcal{G}$ if and only if q is fair.

Differential harm provides a unified framework for evaluating a districting plan across a variety of dimensions. It is derived from an individual-level measure of representation and it accounts for the space of counterfactual possibilities. Importantly, it separates *vote choice* from *group membership*, allowing one to

[1] Formally, we have random indicator variables $\mathbf{1}\{i \in G_k\}$ for all i and k, subject to the constraint that $\sum_k \mathbf{1}\{i \in G_k\} = 1$ for all i. These indicator variables may be correlated with V and/or $\tilde{Q}$, in general.

discuss how partisan election outcomes under a specific districting plan create harm for specific groups or their interactions, simply by conditioning $H(q)$ on different sets of voters.

Estimating these quantities from available data is another matter, of course, and requires additional assumptions, as we describe next.

2.4 Estimating the Harm Measures

Harm $H(q)$ and differential harm $\Delta H(q)$ as defined above require specification of a probability distribution on elections and districting plans. Depending on one's viewpoint, this is either a theoretical fiction which must be specified by a model, or an actual data-generating process we must estimate. Either way, past election data and an enacted districting plan alone are not enough to calculate these measures.

2.4.1 Modeling Districting Plans

It is difficult to construct a tractable probability distribution for districting plans, since the space of plans is discrete and lacks a natural metric structure. Thankfully, in recent years a number of algorithms have been developed which sample districting plans from specific distributions (Fifield et al., 2020a; DeFord et al., 2021; Carter et al., 2019; McCartan and Imai, 2023, see Chapter 1). These distributions can generally be characterized as maximum-entropy distributions on the space of all valid districting plans, in combination with several moment constraints.

For example, the algorithms of Carter et al. (2019) and McCartan and Imai (2023) (Chapter 1) sample from the maximum-entropy distribution of districting plans that are geographically contiguous, have district populations within a specified deviation from equality, and which have a certain average compactness value (according to a particular graph-theoretic compactness measure). These maximum-entropy distributions put as few conditions as possible on the districting sampling process, as long as a set of constraints

imposed by the researcher are met. If these constraints are precisely those which are required by law or traditional redistricting criteria, then samples from this distribution provide a natural set of counterfactual districting plans, uninfluenced by superfluous partisan or other considerations, to use in estimating $H(q)$ and $\Delta H(q)$. In practice, the translation from tradition and law into mathematical formulations of constraints is not always clear or even well-defined. There is certainly an element of subjectivity in sampling a set of districting plans. Nonetheless, sampling methods are widely used in practice (see e.g., Rucho *v.* Common Cause, 2019; League of Women Voters *v.* Commonwealth, 2018; Merrill *v.* Milligan, 2022; Becker et al., 2021; Chen and Rodden, 2013).

2.4.2 Modeling Election Outcomes

Election results present a different set of modeling challenges, given the large number of factors that influence election outcomes and the general unpredictability of future events. We limit our discussion here to the case of estimating harm and differential partisan harm in a two-party setting. Estimating differential racial harm in the two-party setting can be done by first generating ecological inferences or survey estimates of party preference by race, and then applying methods analogous to those outlined here. Multiparty or primary elections pose additional modeling challenges. In Section 2.8 we consider a spatial-racial model for primary elections with multiple candidates, to illustrate the application of our framework to these more complicated cases.

Motivated primarily by computational convenience and simplicity, we adopt the stochastic uniform partisan swing model of Gelman and King (1994), modeling the Democratic two-party vote share in district d and election v as

$$s_{dv} = \bar{s}_d + \delta_v + \varepsilon_{dv},$$

where $\bar{s}_d$ is the baseline Democratic vote share in the district, $\delta_v \sim \mathcal{N}\left(0, \sigma_\delta^2\right)$ is a election-specific uniform shift, and $\varepsilon_{dv} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is a district-election residual error term. The baseline $\bar{s}_d$ can be established by averaging the results of all available statewide elections over the past decade or other suitable period. This is done at the precinct level, since this is the most granular level that districting plans are usually built from and election returns are reported at.[2] Since $\bar{s}_d$ is calculated for the particular configuration of district d, this model takes into account the specific districting plan under consideration.

To estimate σ_δ^2 and σ_ε^2, we use historical U.S. congressional election returns from 1976-2020, provided by the MIT Election Data Science Lab (https://doi.org/10.7910/DVN/IG0UN2). We discard elections which were uncontested by either party, or which had more than one Democratic or Republican candidate, leaving 5,931 total elections. The data do not have complete coverage, and some states are missing or represented very sparsely. However, there are 38 states with at least 20 elections.

For each election, we compute the two-party Democratic vote share. We fit a linear mixed-effects model to these vote shares, with an overall intercept fixed effect, and random effects for districts and election years. To account for the fact that district boundaries change every decade, the district identifiers are built as a concatenation of the state, district number, and redistricting. For example, Washington's 7th congressional district receives one random effect from 2002-2010 and a different random effect from 2012-2020, since new district boundaries were enacted in the 2010-2012 period. We do not account for any mid-decade redistricting which may occur as a result of litigation.

The estimated standard deviations of the random effects are $\hat{\sigma}_\delta = 0.0266$ for the election-year shocks and $\hat{\sigma}_\varepsilon = 0.0653$ for the election-district-specific effects. We substitute these point estimates directly into

[2]When district boundaries split precincts, practitioners generally impute the returns in each component of the split precinct by population-weighted interpolation. Population information comes from Census blocks, which, when populated, are practically never split.

our election model. While this has the advantage of simplicity, it does not account for uncertainty in the estimates.

Since this model combines estimates from disparate sources, it is especially important to understand its calibration and predictive performance in practice. These elements were examined in more detail for a very similar model in Kenny et al. (2023); the relevant analysis has been included here as Appendix B.2.

More complicated election outcome models are certainly possible and are straightforward to use with the measures proposed here. For example, one could operate on the logit scale, model incumbency effects, or include additional data from local elections.

2.4.3 Calculating Harm and Differential Harm

With a model for election outcomes and districting plans in hand, calculation of $H(q)$ and $\Delta H(q)$ is straightforward. For each sampled counterfactual districting plan $\tilde{q}$, we use the election model to simulate vote shares in each district of q and $\tilde{q}$. We can then identify, for each precinct, how many Democratic or Republican voters were harmed.[3] We tally the total harm done to Democratic and Republican voters across the map and divide by the appropriate totals to get a single draw of $H(q) \mid I \in \textsc{dem}$ and $H(q) \mid I \in \textsc{rep}$. When other groups (such as racial groups) are used in $\Delta H(q)$, we must repeat these computations for each group. Finally, we average across all of the samples, and add or subtract across groups, to produce estimates of $E[H(q)]$ and $\Delta_{\textsc{dem}-\textsc{rep}} H(q)$.

[3]For this, we use the baseline count of Democratic or Republican votes. This is a slight simplification, since the vote share changes across random draws, and so the number of Democratic and Republican voters changes too. To deal with this issue completely, one must model elections at the precinct level, which requires more data and computational costs.

2.5 Connections to Current Practice for Analyzing Partisan and Racial Gerrymandering

A primary advantage of our proposed harm measures is that they are defined at the individual level, allowing for the study of redistricting for any defined group memberships. For political science, two of the most important group memberships are perhaps race and party. Even in these well-studied contexts, our measures provide an improvement over existing analytical approaches.

2.5.1 Racial Vote Dilution

The study of racial gerrymandering and vote dilution is considerably more complicated than that of partisan gerrymandering, due to the difficulty of understanding racial voting preferences when individual votes are secret. No simple formulas exist to measure the precise racial impacts of a districting plan. Indeed, much of the academic literature has focused on the statistical challenges of estimating vote preferences and turnout by race (King, 1997; Greiner, 2006; Imai and Khanna, 2016).

Modern analyses of vote dilution and racial gerrymandering are inseparable from the Voting Rights Act of 1965 (VRA) and the 14th and 15 Amendments. Section 2 of the VRA enables citizens to challenge districting plans in court for diluting the voting power of racial minorities. Under the current framework for vote dilution established by Thornburg *v.* Gingles (1986), plaintiffs must prove that (a) the minority group is large and geographically compact enough to form a majority in some hypothetical district, (b) the minority group is "politically cohesive," and (c) the majority group can and does vote to deny the minority group its choice of candidate (Hood III et al., 2018) (the so-called *Gingles* criteria). Criteria (a) is generally proven through a single example districting plan, while criteria (b) and (c) are shown using the statistical estimates of vote preferences and turnout by race.

While the specific legal requirements for a vote dilution claim will obviously remain front-and-center for court challenges of racial gerrymanders, our proposed harm measures unifies all three *Gingles* criteria. Criteria (b) and (c) capture the misrepresentation of minority voters by a candidate they do not support, while (a) requires the existence of a counterfactual where this would not be the case and (a) and (c) together establish that minority voters are harmed. But our proposed measures can also tackle trickier situations where the *Gingles* criteria may not be clearly met.

One such case arises when there are several distinct minority racial groups with similar preferences. In these cases, some courts have found that (a) may be weakened to consider "coalition" districts that still allow minority voters overall to elect candidates of their choice, but there is no clear rule (Stephanopoulos, 2014). Or when urban White voters have similar partisan preferences to minority voters, (b) and (c) may not be strictly satisfied, especially for two-party general elections. Some argue, however, that courts should still rely on the VRA in these cases to protect minority voters, especially in light of differing racial preferences in primary elections (Becker et al., 2021).

In both these cases, the proposed harm measures quantify the extent to which a challenged districting plan differentially harms minority voters. They account for geographic sorting, political cohesion, and the individual preferences of every voter. Sections 2.7 and 2.8 demonstrate this approach in practice.

2.5.2 Partisan Gerrymandering

The canonical framework for considering partisan fairness in redistricting is partisan symmetry (King and Browning, 1987). Partisan symmetry implies that the seats s won by a party with vote share v would be satisfied if $s(v) = 1 - s(1 - v)$. In practice, the partisan symmetry framework is applied via the partisan bias metric $\beta(v)$, which measures the seat deviation from symmetry (Tufte, 1973; Katz et al., 2020), i.e., the

number of seats that would need to be reallocated to reach partisan symmetry for a given statewide vote share v (usually $\frac{1}{2}$).

More recently, and particularly in the years since Vieth *v.* Jubelirer (2004), researchers have developed a veritable zoo of measures to identify partisan gerrymanders. The *mean-median* difference looks at the relationship between the vote share of the median seat in a delegation and the mean vote share of the delegation to make claims about how a map is skewed (McDonald and Best, 2015). *Declination* describes the relative rate at which districts become less competitive for each party (Warrington, 2018). Perhaps the new measure which has reached the most acclaim is the *efficiency gap* (Stephanopoulos and McGhee, 2015), which measures the difference in the number of the votes wasted by each party.[4]

The latter three metrics are grounded in intuition over how to reliably identify a partisan gerrymander, and not in the larger theoretical framework of partisan symmetry (Katz et al., 2020). As such, they make no distinction between the quantity they are trying to estimate and a particular estimator or formula used in calculations. This is in contrast to partisan symmetry as well as the proposed harm measures, which require practitioners to take the additional step of linking observed election data to the theoretical construct by way of an election model, as discussed in Section 2.4.[5] This is an important conceptual distinction that is often lost in day-to-day applications of various gerrymandering measures. Plugging in past election results without using an election model assumes too much about the stability of electoral dynamics and consequently can significantly understate the uncertainty in how a particular districting plan will perform into the future.

[4]Wasted votes are those cast for a losing candidate, as well as those cast for a winning candidate beyond the required 50% to win the election. They are exactly the votes which don't contribute to winning a seat.

[5]For partisan bias, both researchers and practitioners have largely relied on simple uniform-shift models similar to the one we use in this work.

Since none of the existing measures are centered on individual voters, they offer far less granular insights than the proposed harm framework, even in the case of studying partisan gerrymandering. For example, differential harm can be estimated at the precinct level and then plotted on a map to show not just *which party* is harmed by a gerrymander, but *where* this harm occurs.

All of these existing measures are constructed so that a value of zero indicates no bias, and a value of zero is at least implicitly held up as a normative standard for districting plans. In doing so, these measures allude to *implicit* counterfactuals. For example, with partisan bias, the implicit electoral counterfactual is an election in which the parties' vote shares have been reversed, and the implicit counterfactual districting plan is one which assigns equal vote shares to the parties under both the real-world and counterfactual electoral scenarios. Indeed, Katz et al. (2020) reference an implicit counterfactual in discussing the normative implications of partisan symmetry, writing that "[o]bviously, a process designed to be 'fair' that turns out to be substantially asymmetric [i.e., have nonzero partisan bias], if there is a fairer alternative, would not normally be regarded as a legitimate outcome." Similar implicit counterfactuals can be thought up for the other measures.

These implicit counterfactuals differ from the explicit counterfactuals we adopt in this work in that there is no guarantee they are possible. Katz et al. (2020) and others recognize that a nonzero value in a measure like partisan bias can only be considered problematic "if there is a fairer alternative." The inability of partisan bias and other existing measures to explicitly account for realistic counterfactual districting plans means that none of them may be interpreted in a vacuum. No particular value can be used as a threshold past which a plan becomes a gerrymander, because the measures confound the map-drawer's bias with the bias inherent to the electoral system's geography.

It is this confounding which explains many of the concerns that have been raised over the existing mea-

sures. The standard calculation of partisan bias, $\beta(\frac{1}{2})$, can give unintuitive results in more lopsided states, which, if not interpreted carefully, can lead practitioners to make incorrect conclusions about the presence and direction of a partisan gerrymander.[6] And none of the measures take into account any aspect of political geography, and therefore suffer in states like Massachusetts, where, for example, the efficiency gap takes a value generally indicating an extreme Democratic gerrymander.

Despite these differences, there are also empirical similarities in many cases. The efficiency gap is often highly correlated with estimates of differential partisan harm under a stochastic uniform partisan swing model (though the efficiency gap is much more discrete), as wasted votes and partisan harm are rooted in similar logic.[7] Similarly, in competitive states, where the baseline election has a statewide vote share near 50%, $\beta(\frac{1}{2})$ is highly correlated with the average number of seats won by each party and with estimates differential partisan harm as well. Correlations, of course, are not everything: it is the universal scale and centering of our harm measure which makes it more useful in evaluating partisan gerrymanders, as we demonstrate in Section 2.6.

Finally, it is worth comparing the proposed measures to the increasingly common application redistricting simulation techniques to cases of partisan gerrymandering. Standard practice in such litigation, for example, is to calculate a battery of gerrymandering metrics on the simulated plans, and show that the challenged plan is an outlier on all or most of the metrics. While this can help contextualize the challenged

[6]This has been called by some the "Utah paradox" (DeFord et al., 2020), though the paradoxical nature is contested as a mistake of using it for description, rather than inference (Katz et al., 2021). Even when interpreted carefully, however, partisan bias requires significant extrapolation from observed election outcomes in lopsided states—how realistic is it to evaluate the counterfactual where Democrats run even with Republicans statewide in Utah?

[7]The differences stem from the efficiency gap considering a vote to be wasted if it is in excess of the 50% threshold needed to win, while the harm framework does not penalize such votes, since every such voter is happily represented by their candidate of choice. This definition of wasted votes is also at the root of other criticisms of the efficiency gap, which note its tendency to penalize proportionality and reward 3-to-1 districts (Bernstein and Duchin, 2017; Tam Cho, 2017; DeFord et al., 2020).

plan, it is not foolproof—if the underlying metric is misleading, it will still be misleading when applied to thousands of simulated plans. In the worst cases, this can lead to directionally wrong conclusions, as we see in the next section.

2.6 Application to Partisan Gerrymandering in New Jersey

To illustrate the connections in the preceding section and to demonstrate the value of the proposed measures in a real-world setting, we provide a brief analysis of congressional districting in the state of New Jersey. Since 1995, New Jersey congressional redistricting has been handled by a 13-member commission, with 6 political appointees from each party and one independent chair with *de facto* tiebreaking authority. For the 2020 redistricting cycle, each party submitted a proposed plan to John Wallace Jr, the chair. Wallace ultimately selected the Democratic plan for enactment, "simply because in the last redistricting, the map was drawn by Republicans."[8] Here, in contrast, we evaluate the two plans on the basis of existing gerrymandering metrics as well as through the proposed harm framework.

Figure 2.2 shows the baseline Democratic vote share by precinct, along with the Democratic and Republican proposals.[9] Using an average of the 2016 presidential and 2018 senatorial elections as a baseline, the Democratic proposal yields 9 Democratic seats and 3 Republican seats, while the Republican proposal has 7 Democratic and 5 Republican seats.

We sample 10,000 districting plans across four independent runs of the SMC algorithm of McCartan and Imai (2023) (Chapter 1), as implemented in the R package `redist` (Kenny et al., 2020), setting con-

[8] See Matt Friedman, "Democrats prevail in New Jersey redistricting with map that could sacrifice Malinowski," *Politico,* December 22, 2021. https://www.politico.com/news/2021/12/22/new-jersey-redistricting-map-malinowski-525983 (accessed July 3, 2022).

[9] The enacted plan is sourced from All About Redistricting, and the GOP plan was provided by Harrison Neely, executive director for the GOP delegation to the New Jersey Congressional Redistricting Committee, following an Open Public Records Act request.

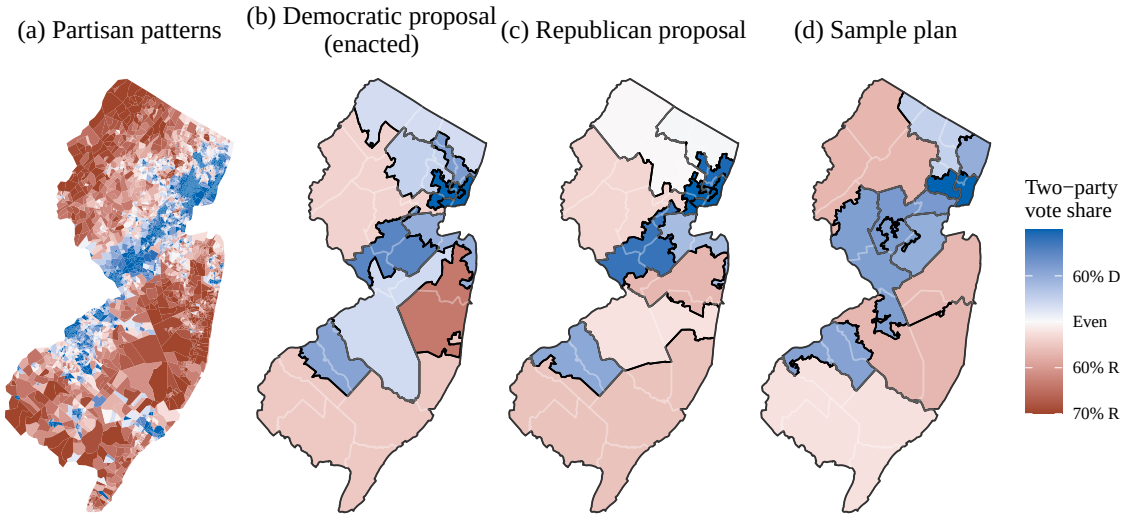


Figure 2.2: Baseline vote patterns (a), two hypothetical gerrymanders (b) and (c), and one randomly sampled plan (d). County lines are overlaid in white.

straints to favor compact plans with district boundaries that tend to follow county lines. One such plan is shown in Figure 2.2(d). For each of these sampled plans and the two gerrymanders, we estimate the average harm and differential partisan harm versus the simulated baseline,[10] and calculate the expected number of Democratic seats, the efficiency gap, partisan bias, the mean-median score, and partisan declination. The distribution of these measures is shown in Figure 2.3.

We first examine the values of each metric on the two proposals alone. The estimate of differential partisan harm aligns with the expected partisan skew of each plan, with a value of -3.6% for the Democratic proposal and 2.0% for the Republican proposal. These values are relatively directly interpretable. For example, the estimated differential harm of -3.6% means that an average Democratic voter is 3.6 percentage points less likely to be misrepresented than their average Republican counter-partisan. The efficiency gap, declination, and partisan bias measures agree with the differential harm estimates in this regard, though

[10] We choose to calculate $\Delta_{\text{DEM}-\text{REP}} H(q)$, so that positive values correspond to a plan that favors Republicans.

the former two are much more discrete. The efficiency gap and partisan bias estimate give roughly equal scores in absolute value to the Democratic and Republican plans, in comparison to differential harm and declination, which identify the Democratic proposal as slightly more extreme. While the mean-median measure identifies the Republican proposal as favoring Republicans, it also scores the Democratic proposal as favoring the Republicans, albeit by less.

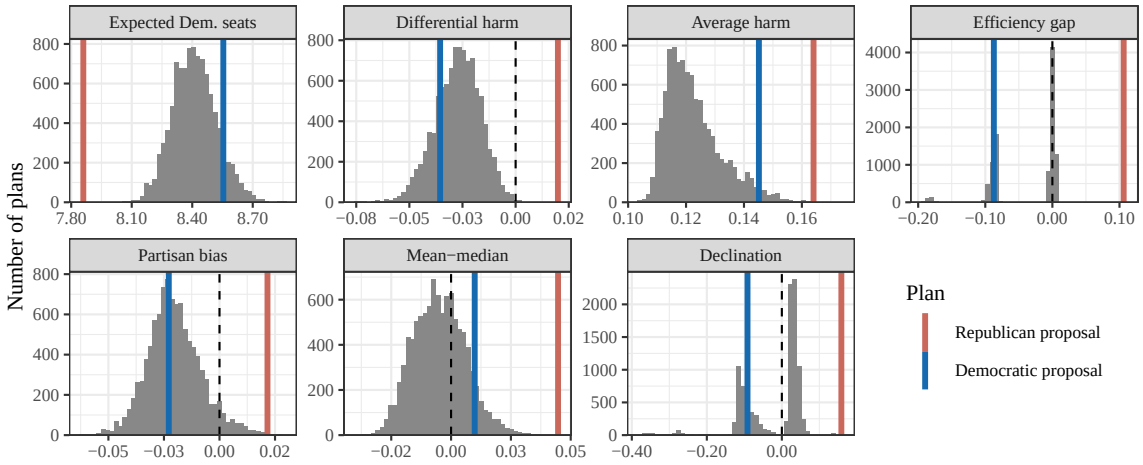


Figure 2.3: The distribution of the existing and proposed measures across a sample of 10,000 plans, with the values for two gerrymanders overlaid as vertical lines. The dashed line indicates a value of zero.

These values can be better contextualized in comparison to the distribution across the sampled plans. The Democratic proposal is expected to produce an average of 8.55 Democratic seats under our election model, more than 79% of all sample plans. The Republican proposal is an outlier, yielding fewer seats on average than every single sampled plan. This pattern is repeated for all of the four common gerrymandering measures, where the Republican plan is a clear outlier (all $p = 0.0002$, except $p = 0.002$ for partisan bias) while the Democratic plan is not ($p = 0.34$ for the efficiency gap, 0.50 for declination, 0.88 for partisan bias, and 0.17 for the mean-median score). However, both the partisan bias estimate and the

mean-median score again misidentify the direction of the bias of the Republican plan, classifying it as an extreme Democratic outlier.

The estimate of differential harm already uses the simulation output, and so comparing the proposals to the sampling distribution is not strictly required. However, it is notable that while the Democratic proposal has a more extreme estimate of differential partisan harm (1.6 percentage points higher), its value of -3.7% is not an outlier with respect to the distribution of differential harm for a randomly sampled plan ($p = 0.43$, using the same set of samples as a counterfactual), while the Republican plan is ($p = 0.0002$). In other words, a *typical* nonpartisan districting plan is not necessarily a *fair* one, even in comparison to other similarly generated districting plans. The sampled plan which minimizes estimated differential harm yields 8.18 expected Democratic seats, compared to a sample average of 8.41 seats, and the 6.8 seats that would be required for an outcome completely proportional to the parties' statewide vote share.

Both proposals are estimated to harm the average voter at a higher rate than most of the sampled plans, with the Republican plan again more extreme in this regard ($p = 0.0008$ versus $p = 0.061$ for the Democratic plan). The plans which minimize estimated total harm are also those which have near-zero estimated differential partisan harm. Thus, at least in New Jersey, overall harm reduction is consistent with a fair map on a partisan basis. Minimizing bias with other measures, however, could lead one astray: plans which score an efficiency gap of zero can have estimated average harm as high as 17.2% and estimated differential harm as extreme as -6.1%, nearly twice the value of the Democratic proposal. Plans with zero partisan bias or mean-median score likewise are estimated to have higher-than-average harm and to tend to differentially harm Republicans.

Figure B.1 in the appendix expands on Figure 2.3 to also show the correlations between these various measures. Expected Democratic seats, the differential harm estimates, the efficiency gap, and declination

are all moderately correlated with each other. Similarly, partisan bias and the mean-median score are moderately correlated, but very weakly correlated with the expected number of Democratic seats. Thus there exist many plans which give Democrats a higher-than-average share of the seats, but which have partisan bias and mean-median score favoring Republicans. New Jersey is just one state, and while some of the correlations between measures here are found in other states, some are not. Appendix B.1 records these correlations across simulated plans in eight states. The most consistent correlation is between estimated differential partisan harm and the expected number of seats won by each party, which averages 0.79 in the eight states we study. Seats are, of course, more discrete and cannot directly identify the local areas where gains and losses are derived.

2.7 Application to Voting Rights Litigation in Alabama

We next apply our proposed framework to Merrill *v.* Milligan (2022), ongoing VRA litigation in which civil rights groups have argued that the existing Black majority district packs Black voters, and so a second Black opportunity-to-elect district should be drawn.[11] Measuring differential harm by both race and party will allow us to cleanly separate the racial and partisan effects of the Alabama districting plan, something existing approaches are unable to do.

As discussed in Section 2.5.1, plaintiffs had to satisfy the first *Gingles* criterion by submitting one or more remedial maps demonstrating that a new Black opportunity-to-elect district could be drawn while meeting all other redistricting requirements. In *Merrill,* the plaintiff's expert, Moon Duchin, submitted four such remedial plans, one of which is shown in Figure 2.4(e). These plans form a natural counterfactual $\tilde{Q}$ to use in analyzing the challenged districting plan (Figure 2.4(d)), especially since the federal district court ruled in

[11] Relevant court filings are available at https://redistricting.lls.edu/case/milligan-v-merrill/.

favor of the plaintiffs and found that their remedial plans satisfied the first *Gingles* criterion. In Appendix B.4, we repeat the analysis performed here on a different counterfactual consisting of plans sampled with a redistricting simulation algorithm; the results are broadly similar.

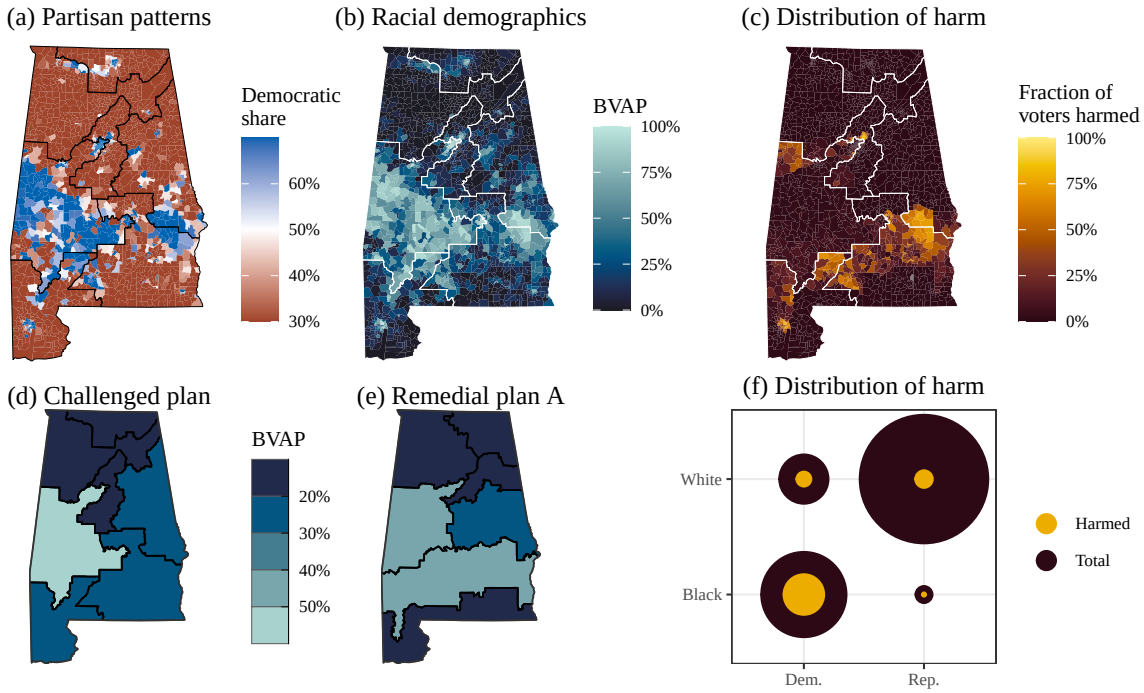


Figure 2.4: Top row: baseline vote patterns (a), fraction of the voting-age population which is Black (b), the expected fraction of voters in each precinct which are harmed by the 2020 enacted congressional districts versus the set of four remedial plans (c). 2020 enacted district boundaries are overlaid in black on the leftmost map and in white on the other maps. Bottom row: the challenged plan and a submitted remedial plan (d) and (e) colored by each district's Black share of the voting-age population, and the expected number of harmed voters in each race/partisan group, relative to the total number of such voters (f).

2.7.1 Election model used in Alabama

This analysis is slightly more complicated than in New Jersey, since in order to understand the racial distribution of harm we need to estimate voting preferences by race. As is custom in VRA analyses, we use an ecological inference model.

The model uses 2020 U.S. Census demographic data and election returns from seven recent statewide

elections: the 2016 races for president and U.S. senator, the 2018 races for governor, attorney general, and secretary of state, and the 2020 races for president and U.S. senator. Our ecological inference model used to produce the estimates is broadly similar to that of King (1997). However, it jointly estimates both turnout and Democratic support across all seven elections, rather than estimating each quantity separately for each election.

For precinct j in election v, let vap_j denote the voting-age population, n_{jv} denote the total number of votes cast for major-party candidates, s_{jv} denote the Democratic vote share, and p_{jr} denote the fraction of the voting-age population belonging to racial group r, where here $r \in \{\text{White}, \text{Black}, \text{Other}\}$. Our main parameters of interest are turn_{jr} and supp_{jr}, the baseline turnout and Democratic support rate by race in the precinct.

We model the observed turnout n_{jv} and vote counts $s_{jk}n_{jv}$ as

$$
\begin{aligned}
n_{jv} \mid \text{turn}_{jr}, \beta_v^{(t)} &\sim \text{Bin-logit}\left(\text{vap}_j, \text{logit}(\text{turn}_j^\top \mathbf{p}_j) + \beta_v^{(t)}\right) \\
\text{logit}(\text{turn}_j) \mid \mu^{(t)}, \Sigma^{(t)} &\sim \mathcal{N}\left(\mu^{(t)}, \Sigma^{(t)}\right) \\
s_{jv}n_{jv} \mid \text{supp}_{jr}, \text{turn}_{jr}, \beta_v^{(s)} &\sim \text{Bin-logit}\left(n_{jv}, \text{logit}\left(\text{supp}_j^\top \frac{\text{turn}_j \circ \mathbf{p}_j}{\text{turn}_j^\top \mathbf{p}_j}\right) + \beta_v^{(s)}\right) \\
\text{logit}(\text{supp}_j) \mid \mu^{(s)}, \Sigma^{(s)} &\sim \mathcal{N}\left(\mu^{(s)}, \Sigma^{(t)}\right),
\end{aligned}
$$

where the $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^{(s)}$ model election-specific logit shifts, Bin-logit indicates the Binomial distribution with the probability parameter specified on the logit scale, and $\circ$ indicates the element-wise (Hadamard) product. We give $\boldsymbol{\beta}^{(t)}$ an i.i.d. $\mathcal{N}(0, 0.6^2)$ prior,and $\boldsymbol{\beta}^{(s)}$ an i.i.d. $\mathcal{N}(0, 0.4^2)$ prior, and $\mu^{(t)}$ an i.i.d. $\mathcal{N}(0, 1.5^2)$ prior. We set priors independently on the variance and correlation components of $\Sigma^{(t)}$ and $\Sigma^{(s)}$. The former received $\text{Gamma}(1.5, 1.5/0.5)$ and $\text{Gamma}(1.5, 1.5/0.25)$ priors for turnout and sup-

port, while the latter received uniform priors (i.e., LKJ priors with shape parameter 1). Several reparametrizations are made for computational efficiency. The full Stan model is included in the replication code.

The mean support $\mu^{(s)}$ receives an informative prior based on the racially polarized voting estimates of Kuriwaki et al. (2021), with $\mu^{(s)} \sim \mathcal{N}(\text{logit}((0.18, 0.90, 0.35)), (0.03^2, 0.03^2, 0.2^2))$. These estimates were derived through multilevel regression and poststratification from survey data, calibrated to actual election results.

The model was fitted with a mean-field variational approximation for computational efficiency. Posteriors of turn_{jr} and supp_{jr} are summarized in Appendix B.3

2.7.2 Results

The model estimates, as expected, reveal that voting is highly polarized by race outside of the cities of Birmingham and Montgomery. With these estimates in hand, we can calculate the average number of voters harmed in each precinct and racial group. Figure 2.4(c) shows the estimated geographic distribution of harm. Even without the ecological estimates, it is immediately clear, with reference to the partisan and racial patterns shown in Figures 2.4(a) and (b), that harm is concentrated among Black and Democratic voters in Birmingham, Montgomery, and the southern end of the Black Belt region. These areas are excluded from the Black-majority district under the challenged plan, but are included in a Black-majority district in the average remedial plan.

This intuition is confirmed by a closer analysis, summarized in Figure 2.4(f). Across the state, the estimated average harm is 7.9%, which translates to around 159,000 voters. Of these harmed voters, around 123,000 are Black, while 36,000 are White. Similarly, 137,000 are Democratic, while only 22,000 are Republican. Of course, the total number of each of these groups of voters is very different. Normalizing by

group sizes, the expected harm for a Black Democrat is 22%, for a Black Republican is 3.7%, for a White Democrat is 8.3%, and for a White Republican is 1.7%. Thus while the challenged plan differentially harms both Black voters and Democratic voters, it *especially* harms Black Democrats, who have an estimated differential harm of 13.8% versus White Democrats and of 18.3% versus Black Republicans.

These calculations are made by aggregating harm within racial groups at the precinct level, and so they automatically take into account individual voter preferences. Unlike in a traditional ecological/*Gingles* analysis, there is no need to identify a particular party as the uniform statewide "choice" of the minority racial group (though vote choice by race and geography must still be estimated), nor is there any cause for concern because of lower racial voting polarization in urban areas. The harm measure treats voters of every race and location equally, recording just their inability to elect their preferred candidate. Only once we have aggregated these individual harms does it become apparent that the challenged plan concentrates the harm among certain groups of voters.

2.8 Application to Primary Elections in Boston

The proposed harm measures are particularly easy to calculate and interpret in two-party elections, since the preferences of each kind of partisan are directly observable as partisan vote shares. However, the same framework can be fruitfully applied to more complex settings, providing insights that go beyond those available through traditional methods.

A major question in voting rights law and research is the extent to which voting must be racially polarized before race should enter into the redistricting equation and statutory protections like the VRA should kick in (Elmendorf et al., 2016). This question is particularly acute in cities, where general elections are characterized by Democratic dominance and low levels of racial polarized voting. When the locus of political

conflict shifts to primary elections, racial patterns are harder to observe and generalize. Thus there remains uncertainty about the extent that race is a factor in primary elections, and consequently the role it should play in designing urban legislative maps.

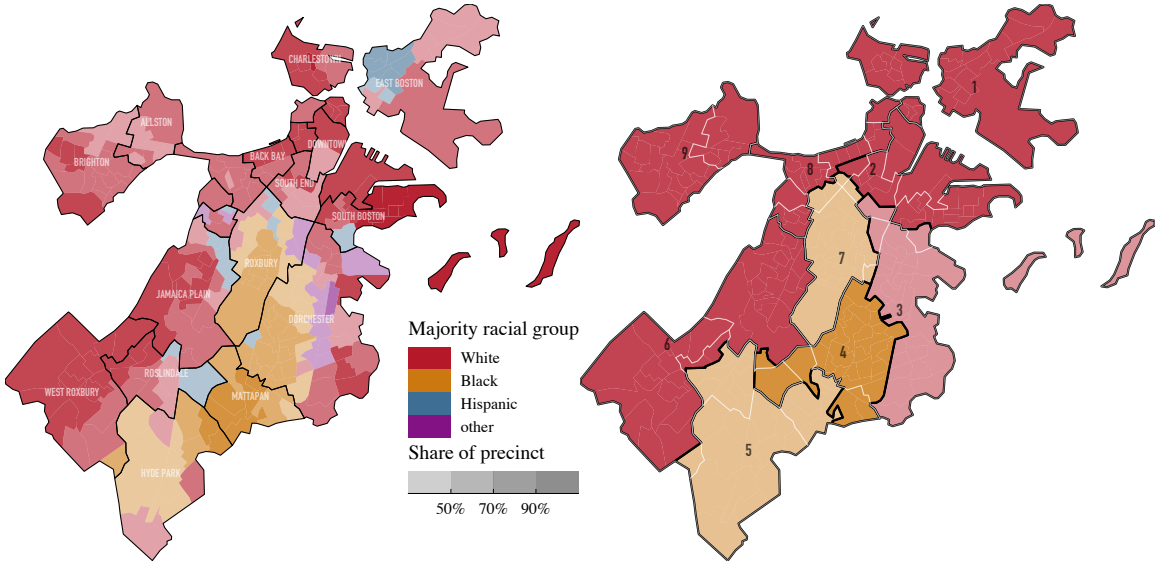


Figure 2.5: Racial composition of the city of Boston (left) and 2010 city council districts (right). Neighborhood boundaries are overlaid in black on the left map and in white on the right map.

The city of Boston is governed by a mayor and a nonpartisan 13-member city council. Nine of the councilors are elected from contiguous city council districts, and the remaining four councilors are elected at-large, with all councilors serving two-year terms. Council district boundaries are redrawn every ten years following the decennial census, and must be equipopulous, be compact, and preserve neighborhood and community boundaries.[12] This section will study the extent to which the current district boundaries differentially harm voters of different racial groups, in comparison to a completely race-neutral districting

[12]City of Boston charter. Available at https://www.boston.gov/sites/default/files/embed/b/boston-city-charter-2007.pdf.

process. Figure 2.5 shows the racial composition of the city as well as the boundaries of the 2010 city council districts and the city's official neighborhoods.

2.8.1 Approach

To apply our proposed framework to study the council district boundaries, we must be able to predict individual voting preferences under counterfactual districting plans. This is more difficult than in both New Jersey and Alabama, for two reasons. First, the city council is officially nonpartisan, and city politics are dominated by the Democratic Party, so voting choices are more complex than a simple partisan preference. To estimate election outcomes, we must map observed support for particular candidates in one election onto counterfactual support for a hypothetical slate in another election. Second, as in Alabama, we aim to examine the racial impact of the district system, and so must estimate these voting preferences by race. Even if we could, for instance, divide candidates into a more "liberal" and a more "conservative" group, the vote counts for each group by precinct will not alone show how different racial groups voted.

To solve both of these problems, we develop a *latent factor ecological inference model*, which simultaneously estimates (a) turnout by precinct and race, (b) candidate positions in a latent two-dimensional political space,[13] and (c) voter preferences in this space by precinct and race, using precinct-level demographic data and vote totals from the March 2020 presidential primary and September 2021 mayoral preliminary elections. While the turnout component of the model is identical to that used for Alabama, the voter preference model expands on that used in Alabama: it estimates voter preferences in a two-dimensional latent space and also estimates candidates' positions in this space.

Continuing the notation from Section 2.7 above, for precinct j in election v, let vap_j denote the voting-

[13] Exploratory analyses found that three latent dimensions did not provide a substantially better fit to the data.

age population, n_{jv} denote the total number of votes cast for major-party candidates, s_{jk} denote the vote share for candidate k (who implicitly belongs to a particular election v_k), and p_{jr} denote the fraction of the voting-age population belonging to racial group r, where here $r \in \{\text{White, Black, Hispanic, Other}\}$ ("Other" consisting primarily of Asian voters in Boston). Our main parameters of interest are turn_{jr}, the baseline turnout by race and precinct, F_{jqr}, the preferences in latent dimension q by race and precinct, and L_{kq}, the loading of each candidate on latent dimension q. Here $q \in \{1, 2\}$, since preliminary analyses found no improved fit when expanding to three dimensions.

We model the observed turnout n_{jv} identically as in Alabama, above:

$$\begin{aligned} n_{jv} \mid \text{turn}_{jr}, \beta_v^{(t)} &\sim \text{Bin-logit}\left(\text{vap}_j, \text{logit}(\text{turn}_j^\top \mathbf{p}_j) + \beta_v^{(t)}\right) \\ \text{logit}(\text{turn}_j) \mid \mu^{(t)}, \Sigma^{(t)} &\sim \mathcal{N}\left(\mu^{(t)}, \Sigma^{(t)}\right). \end{aligned}$$

Because of the latent factor model, we do not require that vote shares by race and precinct, when tallied by race, necessarily equal the observed vote share in the precinct. As a result we choose to model the vote shares directly, rather than the vote counts, which is more computationally efficient. Because each precinct has dozens or hundreds of voters, a Normal approximation to the Binomial likelihood is appropriate, and so we model the vote shares s_{jk} as

$$\text{logit}(s_{jk}) \mid \pi_{jk}, \varepsilon \sim \mathcal{N}\left(\pi_{jk}, \frac{\varepsilon}{1 + n_{jv_k}}\right),$$

where π_{jk} is the preference of precinct j for candidate k on the logit scale, as detailed below. The Normal approximation variance is inversely proportional to the total number of votes in the precinct (the additional 1 is to avoid divide-by-zero errors). The multiplicative factor ε captures the additional variance that isn't accounted for by the latent factors.

The logit candidate preference itself is modeled as

$$\pi_{jk} = \beta_k^{(s)} + \alpha_k F_j^\top (\text{turn}_j \circ \mathbf{p}_j)^\top \text{diag}(\boldsymbol{\rho}) \mathbf{L}_k$$
$$F_{jq} \mid \Sigma^{(p)} \sim \mathcal{N}\left(0, \Sigma^{(p)}\right).$$

Note that F_j is a $q \times r$ matrix of preferences by race and dimension within precinct j, and $\text{turn}_j \circ \mathbf{p}_j$ is a vector of the percentages of the total VAP in the precinct that belongs to voters of each race, so $F_j^\top (\text{turn}_j \circ \mathbf{p}_j)$ is a vector with q elements containing the preferences by dimension of the precinct's voters, after marginalizing out race. The vector $\mathbf{L}_k$ contains the loading of candidate k onto the latent space: When $\mathbf{L}_k$ and F_j are nearby in latent space, then their product will be large, and the precinct's support for the candidate, π_{jk}, will be large. The matrix $\text{diag}(\boldsymbol{\rho})$ captures the relative scale or importance of each dimension to the overall outcomes. The $\boldsymbol{\alpha}$ parameters are "fudge factors" that adjust the scale of the deviations from each candidate's average support.

We give $\boldsymbol{\beta}^{(t)}$ an i.i.d. $\mathcal{N}(0, 0.5^2)$ prior,and $\boldsymbol{\beta}^{(s)}$ an i.i.d. $\mathcal{N}(0, 4.0^2)$ prior, and $\text{logit}(\mu^{(t)})$ an i.i.d. $\text{Beta}(10, 30)$ prior. We set priors independently on the variance and correlation components of $\Sigma^{(t)}$ and $\Sigma^{(p)}$. The former has $\text{Gamma}(1.5, 1.5/0.25)$ and $\text{Gamma}(1.5, 1.5/1.0)$ priors for turnout and support, while the latter have LKJ priors with shape parameters 4.0 and 20.0. A larger LKJ shape parameter places more weight on correlation matrices closer to diagonal, i.e., it regularizes away from high correlations. The scale parameters $\boldsymbol{\rho}$ have i.i.d. $\text{Gamma}(1.5, 1.5/4.0)$ priors, the fudge factors $\boldsymbol{\alpha}$ have i.i.d. $\text{Gamma}(5.0, 5.0)$ priors, and ε has an $\text{Expo}(1/4.0)$ prior. Overall, these priors weakly identify the location and scale of the parameters. Several reparametrizations were made for computational efficiency. The full Stan code for the model is available as part of the replication code.

Like all probabilistic factor models, ours is unidentifiable due to the rotation, location, and scale invari-

ance of the factors in the likelihood. To identify the location and scale, we place $\mathcal{N}(0, 1)$ priors on most components of $\mathbf{L}$. To identify the rotation, we place more informative priors on the loadings of certain candidates. This was accomplished through somewhat of an interactive process, where a preliminary principal components decomposition of the precinct vote shares helped suggest the latent factors that might be found by the model. We chose to align the first dimension with the Sanders-Biden axis, and chose to make the second dimension roughly orthogonal along the Janey-Buttigieg axis. This translated to a prior of $\mathcal{N}((-1, 0), (0.1^2, 0.25^2))$ for Sanders, $\mathcal{N}((1, 0), (1.0^2, 0.5^2))$ for Biden, and $\mathcal{N}((-0.25, -1), (0.5^2, 2^2))$ for Janey. We additionally set the prior correlation between Sanders' and Biden's loadings to be $(-0.5, 0.8)$.

Like all ecological inference models, this one makes several assumptions, verifiable and unverifiable. We assume no spatial correlation in residuals, when some is almost certainly present. We place distributional assumptions on how voter preferences by races vary across precincts (importantly, that these preferences are centered around a common location). And we of course impose a latent two-dimensional factor structure, which must always be a major simplification of reality.

To test the suitability of the model for representing the outcomes of hypothetical future elections, we use the model fitted to the primary elections to predict the outcome of the mayoral general election in November 2021. The results of this analysis are summarized in Appendix B.5.

Figure 2.6 shows the key outputs from the latent factor ecological inference model. The estimated loading matrix $\mathbf{L}$ is visualized on the left, and F is visualized for Black and White voters on the right. As expected one dimension (the first) primarily aligns with a traditional liberal-conservative spectrum, with progressives like Bernie Sanders, Elizabeth Warren, and current Boston mayor Michelle Wu on the left, and centrists like Joseph R. Biden and Wu's opponent, Annissa Essaibi-George, on the right. The second

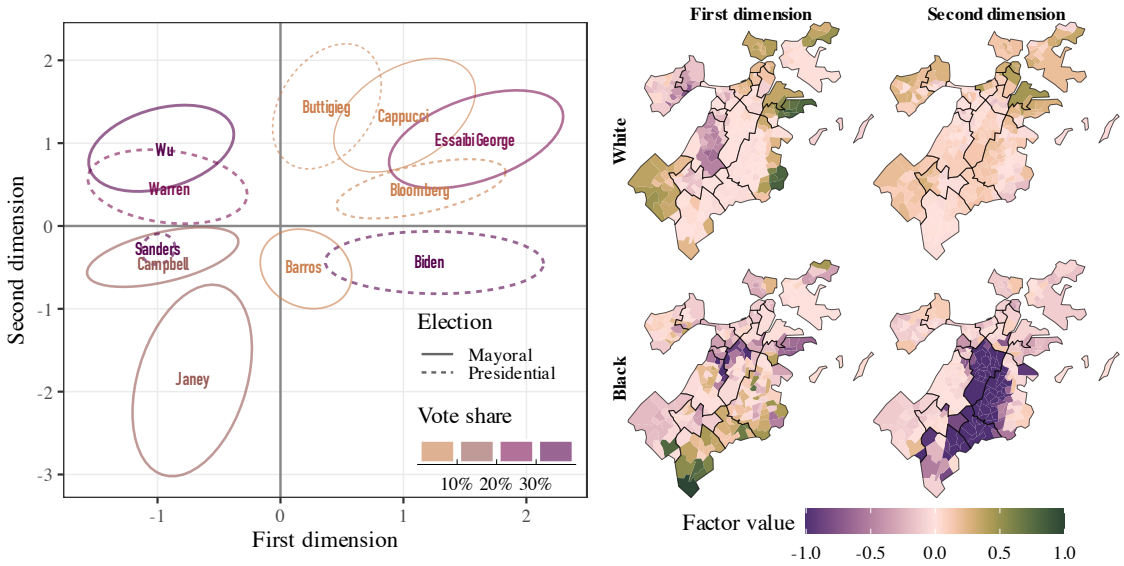


Figure 2.6: Candidate factor loadings, with approximate 80% credible intervals shown as ellipses (left), and geographic variation in factor values by race (right). Neighborhood boundaries overlaid in black. Factor values in the Black second-dimension map are truncated so that variation in the other maps is visible.

dimension appears to capture racial differences in support—as the factor value maps in Figure 2.6 show, minority-majority and White areas of the city take opposite values on the second dimension.

Like all ecological inference models, ours relies on several unverifiable assumptions. The latent factor component imposes additional assumptions about the structure of voting in Boston elections. As Appendix B.5 discusses, these assumptions are almost certainly not *strictly* met. However, the model is a useful simplification of reality that captures the major electoral dynamics relevant to our question of interest. For example, using the estimated latent positions and support levels for Michelle Wu and Annissa Essaibi-George, the two mayoral candidates who advanced to the general election, we predict the precinct-level general election turnout and vote shares with considerable accuracy (Spearman correlation $\approx$ 0.85, see Figure B.7 in Appendix B.5). The accuracy of these validation predictions tracks with previous work using factor models to predict runoff outcomes after a first-stage election (Kamakura et al., 2006). We stress that

the latent factor ecological inference model is not necessary for this kind of analysis; it is one out of several approaches which could provide estimated vote choices for individuals. Local survey data with ideological preferences and racial information would be a sufficient alternative.

2.8.2 Results

As Figure 2.6 shows, primary voting preferences in Boston are polarized by race, especially for the second latent dimension. But there is significant variation within racial groups as well, such as the split between more conservative White voters in South Boston and more liberal White voters in Jamaica Plain (the first latent dimension appears to capture a progressive-moderate divide). It is certainly possible, then, that the particular way districts are drawn will have an effect on the ability of different racial groups to elect their candidate of choice.

We consider two different scenarios: one in which all of the council races are contested along the first latent dimension, and one in which the races are contested along the second. Specifically, we draw voter choices between two candidates from the posterior predictive distribution of our model, setting the candidates' loadings to -1 and 1 on the chosen dimension and 0 on the other dimension. We draw the overall level of support from a Beta$(15, 15)$ distribution, which means that we will study election scenarios where the overall vote share for each bloc of candidates ranges from roughly 30 to 70 percent. The posterior predictive draws yield counts of voters for each candidate in each precinct by race, from which we can calculate winners at the district level, under any configuration of districts.

Our hypothetical elections involve only two candidates, to simplify the challenges of measuring harm with multiple candidates, over which voters may have different ranked preferences. However, our use of latent factors provides one way to extend our harm framework to the multiple-candidate setting. One

could weight voter harm by the distance in latent space between each voter's preferred candidate and the elected candidate, thus dropping the implicit assumption in the definition of harm that all candidates the voter did not vote for are interchangeable.

We sample 2,000 districting plans across two independent runs of the SMC algorithm of McCartan and Imai (2023) (Chapter 1), setting constraints to favor compact plans that respect neighborhood boundaries. Since we do not provide any racial information to the algorithm, this set of plans forms a good approximation of a race-neutral districting process.

For the first scenario, where council races are contested along the ideological axis, we estimate that the enacted plan is quite fair, differentially harming White voters at a rate of only 0.79% compared to the race-neutral baseline. For the second scenario, where the contests are oriented along the racial dimension, we estimate that the enacted plan differentially harms White voters at a rate of 4.1%, compared to the race-neutral baseline.

These two harm estimates reflect the harm of the current boundaries compared to a race-neutral baseline. The more policy-relevant direction might be in evaluating a change from the status quo. What is the harm of switching to race-neutral districting from the current race-aware boundaries? We can flip the above setup and compute harm by averaging over the simulated plans with the enacted plan as the counterfactual. For this reverse comparison, along the first dimension, we estimate that a race-neutral baseline would differentially harm Black voters at a rate of 3.1% compared to the enacted plan Along the second dimension, the differential harm to Black voters would be 1.8%.

Taken together, these findings suggest that despite the dominance of the Democratic party in city politics, and the shared preferences between certain groups of Black and White voters, race-aware districting still serves an important role in empowering Black voters across the city to elect their chosen candidates.

A switch to race-neutral boundaries would have a material impact on Black voting power. Further study could focus on the racial impact of the four at-large members in comparison to a system with 13 districts and no at-large members.

2.9 Discussion

This paper offers a new framework for studying redistricting that brings together a focus on individual voter preferences and the causal-inference logic of counterfactuals. The combination of these ideas imbues the analysis of districting plans with historical understandings of legal standing and harm. This focus naturally leads to a new normative standard of plan fairness, which is achieved when the necessary harms of a districting plan are distributed evenly across voters. In other words, your personal characteristics should not impact your ability to be represented.

Compared to existing alternatives, differential harm more accurately identifies partisan gerrymanders—both in sign and in existence. The exact same approach applies beyond partisanship and allows for the calculation of differential harms by race and other protected groups. By centering the framework on individual voters' choices, we sidestep the need to blindly associate voters' preferences with their group memberships. Analysts can model vote choice separately, estimate the harm to each individual, and then aggregate by any group of interest. As Section 2.6–2.8 showed, this opens up space for more nuanced analyses that can extend to the local level as well.

The proposed framework is far from the end of the road. Its flexibility comes at a cost—most notably, the need to specify a set of counterfactual districting plans, the choice of which has a significant impact on conclusions. Fundamentally, harm and differential harm are unable to measure plans against an absolute benchmark, the way that other measures such as partisan symmetry are able to do. And while modern sim-

ulation methods are a valuable starting point for defining counterfactuals, these approaches involve analyst discretion and interpretation of often-murky legal standards. When subjective choices are not represented consistently, different analyses may arrive at different conclusions about which voters are harmed and at what rate.

Modeling future elections poses another set of challenges. Simple models are easy to build from commonly-available data and are computationally tractable, but are gross simplifications of reality. More realistic models require more analyst thought and can add a significant burden to the calculation of our harm measures. And as discussed in Section 2.3.1, multi-candidate elections where voters have strong ordinal preferences provide further challenges that the proposed framework does not fully address. Each of these limitations present opportunities for future study.

At its core, redistricting is a multifaceted and high-dimensional problem that must balance political, legal, and administrative constraints to allow citizens to exercise their rights. A districting plan can never be boiled down to a single number. But we hope our approach provides a modular and flexible set of measures that can help decompose how various components of the districting plan combine to produce observed patterns and election outcomes. Researchers can and should tinker with our assumptions and push our framework further to better apply it to multi-candidate elections, more realistic election models, and beyond district-based electoral systems.

Part II

Neighborhoods

This page intentionally left blank.

3

Measuring and Modeling Neighborhoods

3.1 Introduction

How do people perceive the local communities in which they reside? Which area do they include or exclude as part of their neighborhood, and why? These questions represent a long-standing focus of social science research that has implications for the structure of interpersonal networks (Fischer, 1982; Huckfeldt and Sprague, 1987; Onnela et al., 2011), social cohesion and inter-group conflict (Putnam, 2007; Christ et al., 2014; Enos, 2017), public goods provision (Habyarimana et al., 2007; Wong, 2010; Hopkins, 2011; Trounstine, 2015), urban planning (Storper, 2013), and the conceptions of social and political identity (Paddison, 1983; Sampson, 1988).

In recent years, the availability of granular geographical data, together with increasing computing power, have provided researchers with new opportunities to gain insights on how people's neighborhoods affect their attitudes and behavior (Enos, 2015; Chetty and Hendren, 2018a,b; Brown et al., 2021; Nuamah and

Ogorzalek, 2021; de Kadt and Sands, 2021). One key methodological challenge, however, is the question of how to *measure* one's neighborhood. Many studies use either administrative units such as census place or measures based on distance and population density (Hopkins, 2010; Legewie and Schaeffer, 2016; Velez and Wong, 2017; Brown and Enos, 2021).

A major problem of these measurements is that the definition of neighborhood is inherently subjective — two individuals who live at the same address may identify different local communities as their neighborhoods (Paddison, 1983; Chaskin, 1997; Sampson et al., 2002). Foundational work conceptualizes neighborhoods as natural sub-units of larger geographies (such as cities or towns) that arise from population grouping, infrastructure, land use, and economic forces (Park et al., 1925; Suttles, 1972). The inherent ambiguity of neighborhood definition, therefore, allows for variation not just across place, but also across people, as different people may experience and perceive the same local geography differently (Paddison, 1983; Chaskin, 1997; Sampson et al., 2002).

Measuring this variation is key to understanding how people connect to the places in which they live, and construct their own definitions of community and belonging. Furthermore, perceptions of local geography play an important role in how neighborhoods influence attitudes, behavior, and other socio-political outcomes (Sampson, 1988; Fowler and Christakis, 2010; Sharkey and Faber, 2014).

In a pioneering study, Wong et al. (2012) address this problem by asking survey respondents to draw their own neighborhoods on a map (see also Wong et al., 2020). We follow their innovative measurement strategy and develop an easy-to-use, open-source survey instrument to measure subjective neighborhoods. Our instrument is highly customizable and can be easily incorporated into standard online survey platforms such as Qualtrics, facilitating its use by other researchers.

The main contribution of this paper is the development of a new statistical model that takes full advan-

tage of this survey measurement. Existing studies, including those that measure subjective neighborhoods, do not directly model how respondent and geographic characteristics, and their interactions, relate to one's neighborhood. Instead, they almost exclusively rely upon descriptive statistics of observed neighborhoods such as racial and economic demographics, neighborhood size, and agreement with administrative boundaries to describe subjective neighborhood definitions (Coulton et al., 2001; Hipp et al., 2012; Wong et al., 2012; Coulton et al., 2013; Wong et al., 2020; Hjorth et al., 2021). The absence of a formal statistical model means, for example, that researchers cannot predict one's neighborhood given their characteristics and residence location.

We propose a Bayesian hierarchical model based on the probability that survey respondents include each small local area (e.g., census block) in their subjective neighborhoods. The proposed model quantifies the degree to which different characteristics of respondents, those of relevant local areas, and their interactions shape their subjective neighborhoods. The respondent characteristics can include demographic attributes and any attitudinal or behavioral measures, whereas the area characteristics may include census statistics, and the location of landmarks such as churches, parks, schools, and highways.

This new statistical model offers several concrete improvements over simpler methods such as regressing neighborhood summary statistics on a set of predictors.

- *The model uses more information.* We model the probability that each constituent Census block is included in a neighborhood, leveraging granular block-level characteristics. A summary-statistic-based regression typically average these characteristics, losing valuable information contained in the respondent's decision to include or exclude each Census block, especially those near the boundaries of subjective neighborhoods.

- *Estimation uncertainty is quantified.* Our proposed Bayesian approach naturally accounts for estimation uncertainty in the model parameters, which is propagated through posterior predictions and other post-analysis summaries.

- *All characteristics of the neighborhood can be modeled simultaneously.* Because the model is at the level of the actual Census blocks, any higher-level neighborhood summary statistic can be calculated for the model predictions and fitted values, allowing formal statistical quantification of differences in these statistics.

- *The model can be used to make individual-level predictions for neighborhoods or portions thereof.* We can sample new neighborhoods from the posterior of the model, including for out-of-sample respondents or for counterfactual covariate values. For example, the model allows one to predict the neighborhood that would be drawn by a survey respondent with a certain set of characteristics if they lived at a different address.

The proposed model allows researchers to answer a host questions about how people understand the places they live, and how they psychologically organize their local area. For example, how does infrastructure and other physical aspects of the built environment shape where people think their neighborhood ends and another begins? How do local institutions shape this consideration, such as churches, community centers, or libraries? What about zoning or administrative boundaries? Individual characteristics of neighbors may also be influential: and people may define their local geography differently based on the race, religion, class, or even partisanship of the people they live around. This model allows for quantification of how each of these characteristics (or any other variable that can be measured geographically) influences whether someone includes an area of certain characteristics in their neighborhood.

We use the proposed statistical model to analyze an original survey of 2,527 registered voters across three major metropolitan areas: Miami, New York City, and Phoenix. From these data, we demonstrate the model's capacity to quantify how individual and contextual characteristics are related to drawn neighborhoods. These characteristics include respondent race, party, age, gender, education, homeownership, and neighborhood infrastructure (i.e. location of parks, churches, schools, and major roadways), administrative boundaries, and racial, partisan, and economic demographics.

Our analysis shows that race and partisanship are significant predictors of subjective neighborhoods. In particular, voters tend to draw neighborhoods that are same-race and co-partisan. For example, net of other factors, White respondents are 6.1 to 16.9 percentage points more likely to include in their neighborhoods a census block composed entirely of White residents compared to one with no White residents. Democratic and Republican respondents are 8.6 to 19.2 percentage points more likely to include an entirely co-partisan census block compared to one consisting entirely of out-partisans. These predictive effects are found even after accounting for other socio-economic demographics, local infrastructure, and administrative boundaries and survey respondent characteristics.

Our open-source survey software allows for easy customization, including embedding experimental components. To demonstrate the wide applicability of our proposed methodology, we also conduct a survey experiment by randomly assigning respondents to draw their neighborhoods on maps with or without information about the racial and partisan makeup of surrounding areas. We find that while the aforementioned patterns of racial and partisan homophily are present across experimental conditions, such information does not fundamentally change how voters draw their neighborhoods. In total, our analysis

demonstrates the potential to gain new understandings of how various characteristics of people and their residential areas shape their sense of shared space and common communities.

The rest of the paper proceeds as follows. We begin by briefly describing our survey instrument that can be used to measure subjective neighborhoods and introducing our survey of registered voters in the aforementioned cities. We then propose our statistical model and lay out the empirical analysis strategies based on this model. Using our survey data, we show that the model yields more accurate out-of-sample prediction of neighborhoods than the standard distance-based measures. Finally, we conduct an empirical analysis of the survey data and examine the racial and partisan factors that shape subjective neighborhoods.

3.2 Measuring Subjective Neighborhoods

In this section, we describe the setup of the survey we conducted and explain how our mapping tool allows ones to measure subjective neighborhoods.

3.2.1 Survey setup

Data for this study comes from an original survey of 2,527 respondents across three U.S. cities: Miami ($n = 474$), New York City ($n = 453$), and Phoenix ($n = 1,600$). These cities were chosen to provide a variety of political and regional contexts, with the aim to collect large enough samples for each city to conduct within-city analysis.

Survey respondents were recruited via e-mail using a list of email addresses attached to registered voter records. The list was provided to the researchers by the vendor Target Smart. We randomly sampled a total of 1,514,612 potential respondents from this list within each city (533,333 in Miami, 476,605 in New York City, and 504,674 in Phoenix). We sent an email invitation to the sampled registered voters on non-holiday weekdays between December 21, 2020 and February 19, 2021. Of these e-mails, 38.5% failed to be delivered

to the potential respondent's inbox, due to the email address either being invalid or the receiving email server rejecting the email. In total, 930,839 voters received survey invitations (329,624 in Miami, 275,449 in New York, and 325,766 in Phoenix). Those who did not respond to the initial invitation were sent up to 3 weekly reminder e-mails. The invitation informed the potential respondent of the purpose of the survey, that they would be asked to draw their neighborhood on a map, and provided information on the researcher's affiliations and contact information. No compensation was offered or provided to respondents and no deception was employed in this study.

Among voters who received a survey invitation, response rate to the survey invitations was 0.8% (0.5% among total voters sampled), which is slightly lower than previous survey recruitments using another email list from the same vendor (see e.g., Brown and Enos, 2021). The Phoenix sample exhibited a higher response rate than the Miami and New York City sample, with a 0.49% response rate in Miami and New York and a 1.3% response rate in Phoenix. In total, we collected 7,691 responses, split evenly between 5 experimental conditions (see Section 3.6 for the details of the experimental conditions), but limit analysis to the 2,527 that drew their neighborhood on our mapping tool. We further limit our analysis sample to the 96.3% of those respondents who verified that the voter record to which their email had been attached corresponded to their correct identity. We use the data from the control group to introduce and illustrate the statistical model while we discuss the details of the experiment and present experimental results in Section 3.6.

Respondents who accepted the invitation to take the survey were presented with a consent form informing them they were taking part in a research study. No deception was used in this study and respondents were not compensated for participation. The consent form was followed by demographic questions including partisanship, race, age, income, employment, homeowner status, whether they had children, marital

status, and how long they had lived at their current residence. Tables C.1 and C.2 in the Supplementary Materials (SM) displays the summary statistics of the survey sample. Across cities, the sample is approximately evenly self-identified Democrats or Republican but with a majority reporting voting for President Biden in the 2020 general election. The sample is also on the whole disproportionately White, wealthy, and residentially stable compared to the broader U.S. electorate.

3.2.2 Mapping tool

Next, respondents were presented with an embedded mapping application where they enter their residential address, at which point the maps zooms to a centered view of their address. Then, the underlying census block grid was shown on the map over the road base map, and respondents used the brush tool to select the census blocks that they considered a part of their "local community." We use this terminology, mirroring previous surveys that ask respondents to draw their own neighborhoods (Wong et al., 2020). These authors have shown that the phrase "local community" is a tangible concept in people's minds, and further demonstrate the consistency of drawn neighborhoods when re-contacting survey respondents.

Our mapping application is comparable to these previous surveys in functionality. One difference is that in this case, rather than having respondents use a drawing tool to draw a circle around their residence that constitutes their neighborhood, the application offers a brush tool to shade in the census blocks around the residence that are included in the neighborhood. Respondents could zoom in or out on the map, and were able to make edits to their neighborhood after the initial shading. The only constraint was that neighborhoods had to be contiguous. Figure 3.1 shows a screen shot of the map drawing tool. We make our map drawing tool publicly available so that other researchers can use it for their own surveys (`https:`

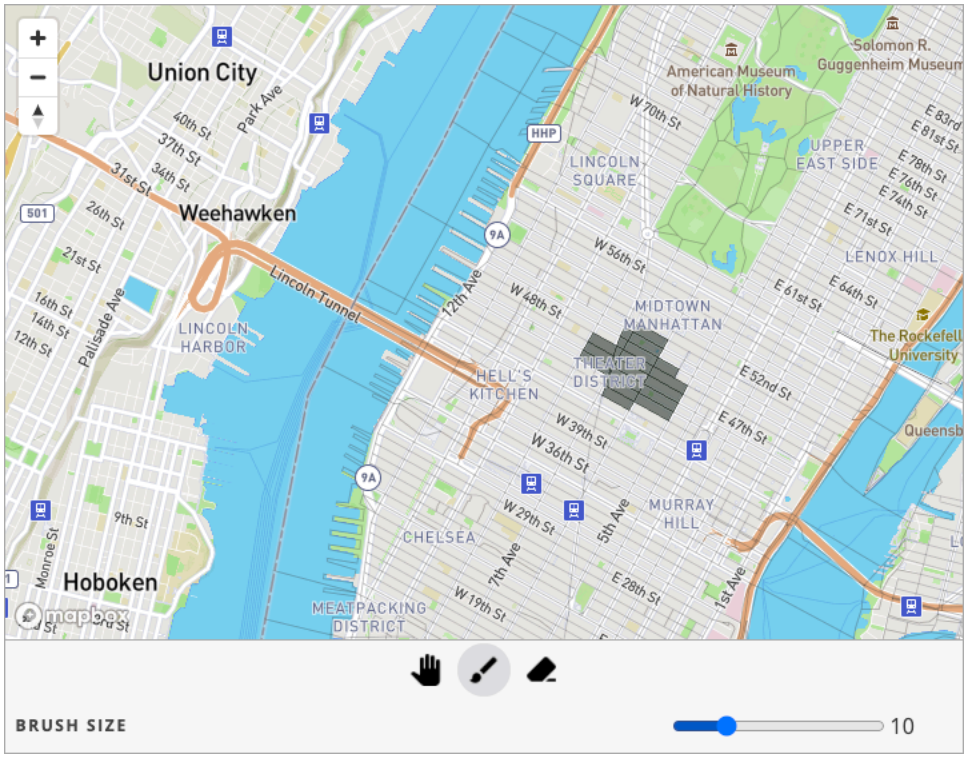


Figure 3.1: Map with brush tool used to draw neighborhoods.

//github.com/CoryMcCartan/neighborhood-survey). In particular, the tool can be embedded into a popular survey platform such as Qualtrics as done in our survey.

3.2.3 Descriptive statistics of drawn neighborhoods

Figure 3.2 shows the central tendencies and range of the population, area, proportion Democrat, proportion Republican, proportion White, and proportion Black of all drawn neighborhoods, broken out by city as well as pooled. Partisanship is measured using Target Smart voter records of every registered voter in the three cities, using information on the residential location of each voter to create aggregate counts by census block. The population, area, and racial demographics are measured using the 2010 census.

We find a wide range of neighborhood sizes, both in terms of population and land area. The median

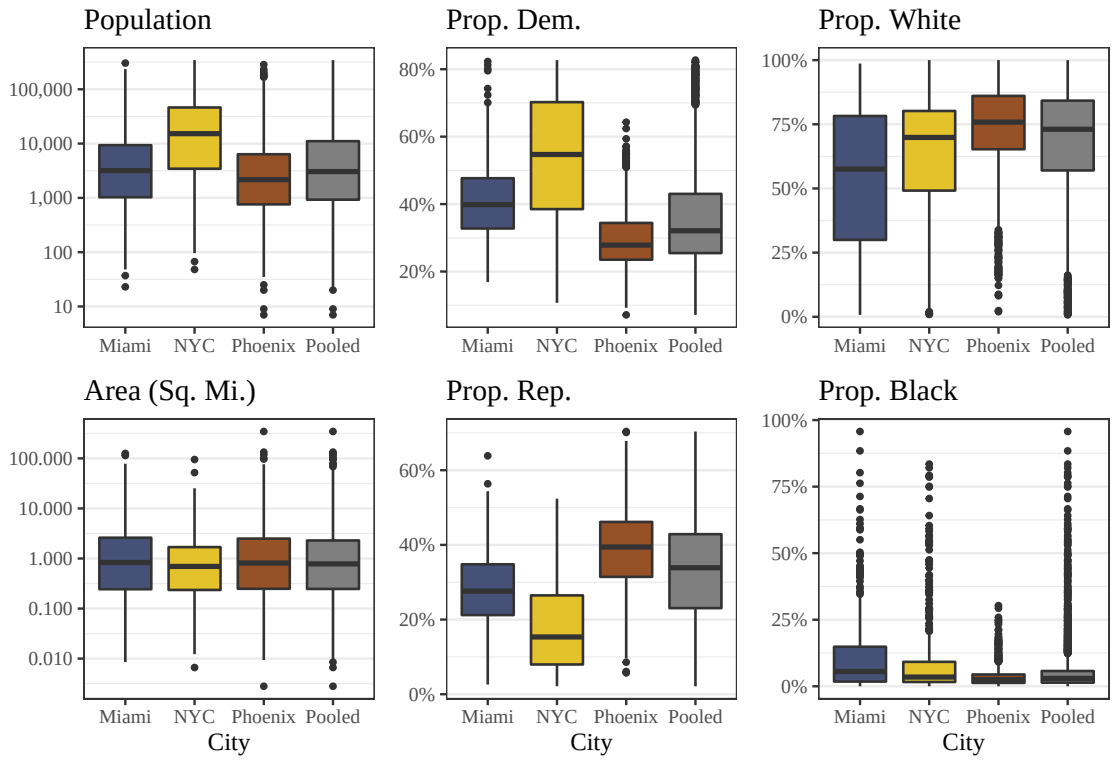


Figure 3.2: Descriptive statistics for respondent neighborhoods.

number of residents contained in a drawn neighborhood is 3,020 residents, while its range extends from single digits to over 340,000 residents. Similarly, the median area is 0.78 square miles, but the entire distribution ranges from 0.01 to 344.74 square miles. This variation in size of neighborhoods is indicative of the variation in how different individuals experience their local geography and define their neighborhood. Accounting for this heterogeneity is critical in our statistical model we introduce in the next section.

We also find substantial variation in the demographics of the drawn neighborhoods. Figure 3.3 shows the distribution of proportion White and proportion Black in drawn neighborhoods separately for White and Non-White survey respondents. Drawn neighborhoods from White respondents are on average 21.1 percentage points Whiter than those from Non-White respondents. Figure 3.4 contains the distribution

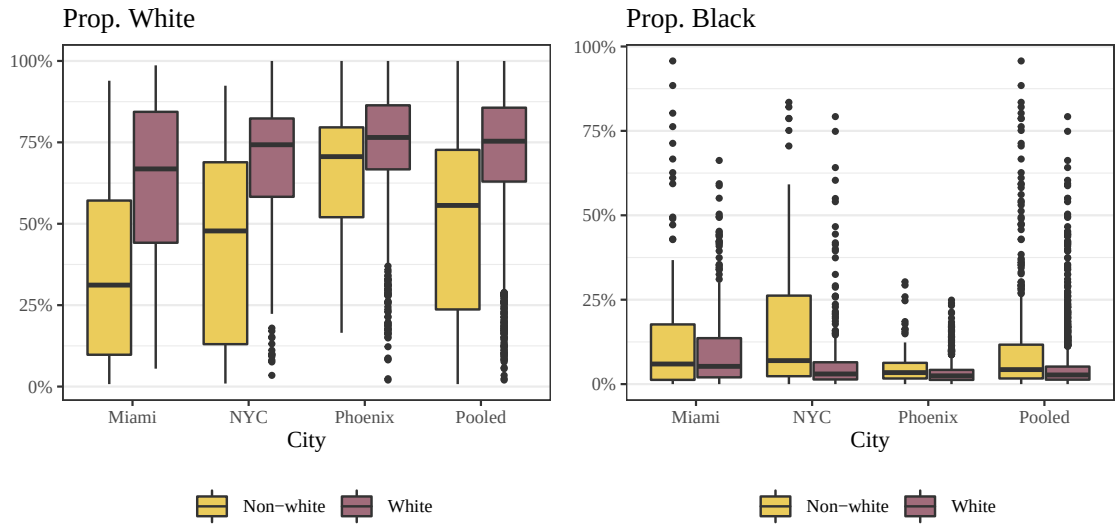


Figure 3.3: Racial demographics for respondent neighborhoods by respondent race.

of proportion Democrat and Republican separately by respondent party registration, with Democrats drawing consistently more Democratic neighborhoods and Republicans drawing more Republican neighborhoods.

These differences likely reflect objective differences in racial and partisan exposure across race and party, but may also be influenced by conscious or subconscious motivations for respondents to construct their subjective neighborhoods to include more members of their own racial or partisan in-group. Our statistical model can quantify the extent to which, net of other variables that may determine whether someone includes an area in their neighborhood, the added predictive effect of racial or partisan demographics on neighborhood inclusion. We now turn to our proposed statistical model of subjective neighborhoods.

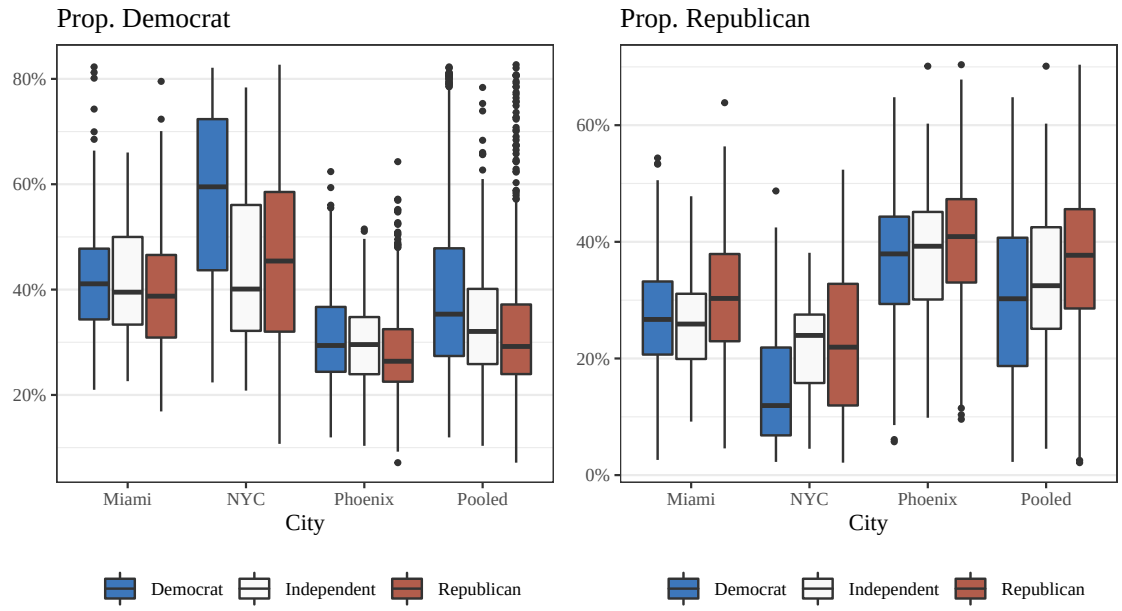


Figure 3.4: Party demographics for respondent neighborhoods by respondent party.

3.3 Modeling Subjective Neighborhoods

To analyze the collected survey data, we propose a generative model for neighborhood drawing that incorporates respondent characteristics, geographic factors, and their interaction. The model predicts the likelihood of including a given census block in a voter's neighborhood. In addition, the coefficients of the model represent the direction and magnitude of predictive effects different variables have on this inclusion probability. Using this model, one can measure the degree to which the characteristics of respondents and geographic factors together predict subjective neighborhoods of different types.

Though the model is developed from explicitly spatial principles, as we show in Section 3.3.2, ultimately it reduces to a generalized linear mixed-effects model (GLMM) with a particular link function, where every observation is a census block. Spatial information enters the model through the use of distance as a covariate, and implicitly in deciding which census blocks are included in model fitting and which are in-

cluded. The simplification to a GLMM means that even without specialized software, practitioners can implement the model using existing statistical packages. We provide an open-source implementation that is computationally efficient and tailored to the particular use case here.

We limit our first analysis to the 471 respondents in the control group who drew a map that consisted of more than one census block. Because the census block that contained the residential address is highlighted by default, we cannot distinguish between respondents who selected single block neighborhoods and those who entered a residential address but decided not to draw a neighborhood. In Table C.6 in the SM, we report the relative balance of covariates across respondents who did and did not draw usable neighborhoods, and further demonstrate that the experimental conditions did not produce substantively different rates of drawing usable maps.

3.3.1 Notation and setup

For ease of notation, we begin by describing the model for a single neighborhood drawn by one respondent. We start with an undirected graph $G = (B, E)$ representing the layout of the city or town, with each vertex B_i corresponding to a census block and the edges corresponding to block adjacency. We write $B_i \sim B_j$ if the edge $(i, j) \in E$, i.e., if B_i and B_j are adjacent. We use $K \coloneqq |B|$ to denote the total number of blocks. In Figure 3.5, for example, the block with respondent's residence is adjacent to four blocks labeled as "1a", "1b", "1c", and "1d". We do not consider two blocks that are touching one another with a single point as adjacent blocks (i.e., no point contiguity). Thus, the block with respondent's residence is not adjacent to blocks "2a", "2b", and "2c".

Without loss of generality, we number the blocks in order of their distance from the block where the survey respondent resides, according to some distance function $d : B \to \mathbb{R}$, so that B_0 is this block,

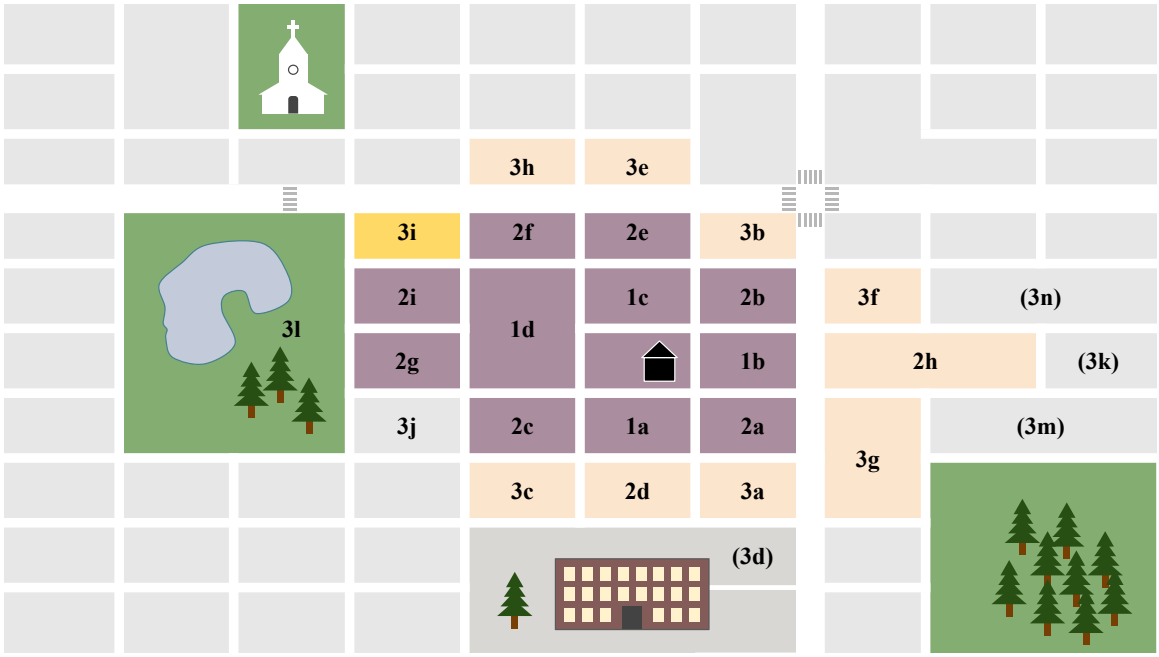


Figure 3.5: Model schematic. The respondent's location is indicated by the black house in the center. Blocks are labeled in the order in which they are considered for inclusion in the neighborhood, with the number indicating the graph-theoretical distance and the letter the spatial distance tiebreaker. Blocks shaded purple have been included in the neighborhood, while blocks shaded light orange have been excluded. Parentheses around a block label indicate that it will not be considered for inclusion in the neighborhood because none of its neighboring blocks which are closer to the respondent belong to the neighborhood.

B_1 its closest neighbor, and so on. Thus, if $i < j$ then $d(B_i) < d(B_j)$. In this application, we take d to be the graph-theoretical distance (i.e., the minimum number of edges between two nodes), with ties broken by spatial distance (i.e., the distance between centroids of census blocks). Figure 3.5 illustrates this ordering scheme. For example, block "1d" in the figure is numbered "1" because it is one step away from the respondent's block, and "d" because among all the one-step-away blocks, it is the fourth closest to the respondent spatially.

Let Y_i be an indicator variable for the inclusion of B_i in the respondent's neighborhood, so that the neighborhood itself may be defined as the set of blocks with $Y_i = 1$, i.e., $Y = \{B_j : Y_j = 1\}$. Since we require any neighborhood to contain the block where respondent's residence is located, we always have

$Y_0 = 1$. We define a following connectivity indicator C_i to be 1 if block B_i is connected to the respondent's neighborhood by way of any closer blocks. Formally,

$$C_i = \mathbf{1}\{\text{there exists a } j < i : B_i \sim B_j \text{ and } Y_j = 1\}.$$

This indicator checks whether, among the blocks which are closer to B_0 than B_i is, if any are in the respondent's neighborhood. In Figure 3.5, blocks with parenthetic labels have $C_i = 0$. For instance, block "3d" would be expected to be considered before "3e", "3f", etc. based on its location, but since none of "2d", "3a", and "3c" are in the neighborhood, "3d" is never considered.

3.3.2 The model

Under the proposed model, a neighborhood is generated sequentially, starting with B_0 and adding blocks in order of increasing distance from B_0 according to a probability. This probability is heavily influenced by the graph-theoretic distance between the block under consideration B_i and B_0, and its connectivity. In particular, we assume that the neighborhood is connected and our survey tool does not allow respondents to draw disconnected neighborhoods.

The core of the model is

$$Y_i \mid Y_0, \ldots Y_{i-1} \sim \text{Bernoulli}(\pi_i \cdot C_i),$$

where π_i is the inclusion probability of block B_i into one's neighborhood provided that it is connected. As long as $\pi_i \to 0$ as $d(B_i) \to \infty$, the generated neighborhoods will be bounded around B_0 almost surely. Figure 3.5 illustrates the state of the neighborhood partway through the generation process, when block "3i" (shaded gold) is under consideration. The process concludes once the neighborhood is surrounded by light

orange blocks, since then there are no blocks left which could be added while keeping the neighborhood contiguous.

The specification of π_i determines the type of neighborhoods that are generated. Let $\mathbf{X}$ be a $m \times K$ matrix of predictors, not including an intercept, with $\mathbf{x}_i$ the column vector of m predictors for block $i = 1, 2, \ldots, K$. These may include the characteristics of the respondent, those of the graph or map (e.g., the demographics of blocks, locations of landmarks and roads), and their interactions. The inclusion probability can also depend on the inclusion of blocks whose distance to B_0 is less than that of the block under consideration B_i. This means that the predictors can include the information about the partially drawn neighborhood. The factorization formulation above, however, precludes the possibility that $\mathbf{x}_i$ depends on $\{Y_j : j \geq i\}$, i.e., the inclusion of farther-out blocks.

We model the inclusion probability using a kernel function that smoothly decays as the distance between B_i and B_0 grows:

$$\pi_i \;=\; \pi(\mathbf{x}_i, \boldsymbol{\beta}, \alpha, L, \sigma) \;=\; \exp\left(-\left|\frac{d_{\text{sp}}(B_i)}{L}\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon\right)\right|^{\alpha}\right) \quad \text{and} \quad \varepsilon \;\sim\; \mathcal{N}\left(0, \sigma^2\right),$$

where ε is the respondent random effect, $d_{\text{sp}}(B_i)$ represents spatial distance between B_0 and B_i, L controls the scale of decay, and α controls its rate. In particular, α represents the sharpness of the neighborhood boundary. Along with L, the random effect ε plays an important role in addressing the heterogeneity of neighborhood size across individual respondents. Figure 3.6 visualizes the kernel function we use where we choose an arbitrary length scale (horizontal axis) for illustration. As the value of α increases, the inclusion probability decays faster as a function of distance.

Conveniently, the model reduces to a Bernoulli generalized linear mixed-effects model (GLMM) with

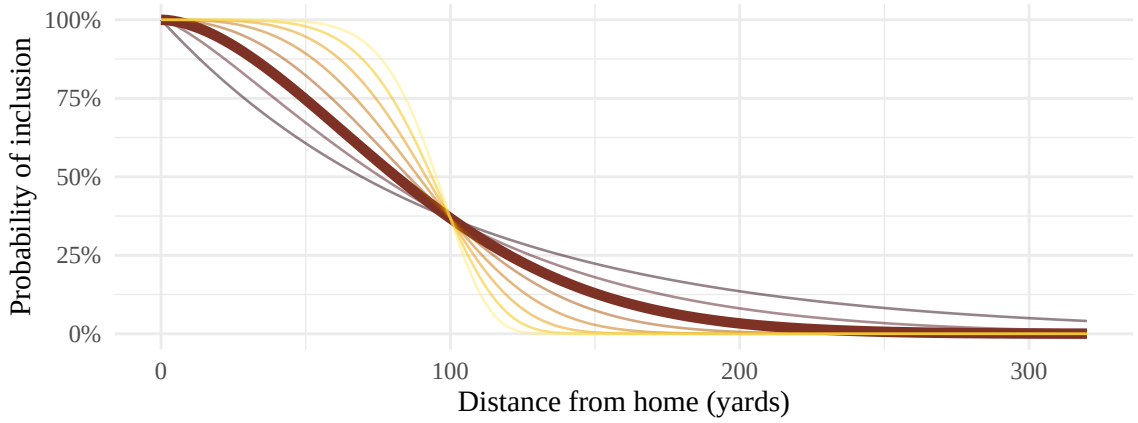


Figure 3.6: Illustration of kernel function across a range of values of the α parameter, indicated by different colors. The length scale shown here is arbitrary; in the model, it is estimated as the L parameter.

complementary log-log (cloglog) link function for the exclusion probability,

$$1 - \pi_i \;=\; 1 - \exp\left\{ - \exp\Big(\alpha \log d_{\text{sp}}(B_i) - \alpha \log L + \alpha \mathbf{x}_i^\top \boldsymbol{\beta} + \alpha\varepsilon \Big) \right\}. \tag{3.1}$$

That is, we can fit the model by regressing the non-inclusion indicators on the log distance, any covariates, and an individual random effect, using the cloglog link function. As discussed above, we do not include every census block in the model—just those that are included in the neighborhood, and those which are not but border the neighborhood (i.e., those for which $C_i = 1$). This reflects the sequential generation process that underlies the model.

Since the link function is nonlinear, the marginal effect of each covariate varies by its underlying value, and is not simply equal to the value of the coefficient. We choose to focus our interpretation on the effect of each covariate on the "margin" of the neighborhood, where the probability of a block's inclusion is 50%. Specifically, we interpret the coefficient estimates by calculating

$$\exp\big(-\exp\big(\mu_{0.5} + \beta_j\big)\big) - \exp(-\exp(\mu_{0.5}))$$

for each coefficient β_j, where $\mu_{0.5} \coloneqq \log\log 2$ is the value of the linear predictor which corresponds to an inclusion probability of 0.5. This quantity reflects the percentage point change in the probability of inclusion for a one-unit change in the covariate $\mathbf{X}_j$, at the margin.

The model is completed with the following prior distributions,

$$\alpha \log L \sim \mathrm{t}_3(0, 2.5) \quad \text{and} \quad \sigma \sim \mathrm{t}_3(0, 2.5).$$

Prior distributions for the coefficients are formed indirectly by taking a QR decomposition of the centered covariate matrix (including the log distance variable). The coefficients on the QR-decomposed (centered) covariates are given a $\mathrm{t}_2(0, 2.5)$ prior. This setup implies priors for the actual coefficients of interest which are weakly informative and adapted to the scale and correlation of the covariates. We have used importance sampling to fit the model under alternative prior specifications and found no measurable changes in the posterior distribution of the parameters.

Because of the sequential generation and the indicator function, the posterior distribution simplifies to

$$\begin{aligned} p(\theta \mid G, Y, \mathbf{X}) &\propto p(\theta) \prod_{i=1}^{K} p(Y_i \mid Y_0, \ldots, Y_{i-1}, \mathbf{x}_i, \theta) \\ &= p(\theta) \prod_{i:C_i=1} p(Y_i \mid Y_0, \ldots, Y_{i-1}, \mathbf{x}_i, \theta) \\ &= p(\theta) \prod_{i:C_i=1} \pi(\mathbf{x}_i, \theta)^{Y_i} \{1 - \pi(\mathbf{x}_i, \theta)\}^{1-Y_i}, \end{aligned}$$

where $\theta = (\boldsymbol{\beta}, \alpha, L, \sigma)$ is shorthand for the parameters, and $p(\theta)$ is its prior distribution. This formulation only requires the computation of the likelihood for the blocks in the drawn neighborhood and all of their adjacent blocks.

We assume individual responses are exchangeable, allowing us to simply multiply their likelihoods to

create a joint model for all responses. The β, L, σ and α are shared across responses, but each respondent's response would has its own error term ε common to all of its blocks. As appropriate, β may also contain hierarchical terms that vary by demographic categories, or metropolitan area or subdivision.

3.3.3 Computational details

As shown by Equation (3.1), the model can be expressed as a Bernoulli GLMM with complementary log-log link function. In this setup, every block that is included the neighborhood, as well as those blocks at the boundary of the neighborhood which could have been included ($C_i = 1$) but are not, becomes a separate GLMM observation. As a result, even moderate respondent sample sizes lead to many more block-level observations, although these block-level observations are of course dependent. In our data, for example, 309 respondents in the control group from Phoenix translate to 57,242 block-level observations.

The large number of block-level observations means that estimates will generally be precise for coefficients which vary at the block level and are shared across all respondents. At the same time, it can create computational efficiency problems for traditional Bayesian posterior sampling methods. Consequently, for our analysis we use a Normal approximation corrected by importance resampling, as implemented by the Stan modeling package (Stan Development Team, 2020). This approximation centers a Normal distribution at the posterior mode with covariance matrix the inverse of the curvature of the log posterior density at the mode. It subsequently samples 1,000 draws from this Gaussian approximation and performs importance resampling so that the draws better approximate the posterior.

3.4 Empirical Analysis

To illustrate the proposed model, we apply it to our survey data using the control group alone. We begin by fitting the model by including individual and location characteristics as well as their interactions. We then show how to assess the quality of model fit.

3.4.1 Model specification

We fit a full model, which includes demographic information, as well as a baseline model, which includes only geographic information. Comparing the predictions of these two models will allow us to quantify the extent to which demographics contribute to the prediction of subjective neighborhoods. In the full model, we include as predictors individual characteristics consisting of voter race, political party, homeowner status, educational attainment, income, age, retirement status, and length of residence in current home. Individuals that differ along these characteristics may view their local area differently, and we quantify the predictive power of these factors about drawn neighborhoods.

We also include geographic characteristics of census blocks including race, party, and education demographics, whether the block contains a school, park, or church, and the distance to the closest of each of these features, whether the block is in the same block group as the voter's residence, the same census tract, whether the block is bounded by the same major roads and railroads as the respondent's residence, block population, and block land area. These aggregate characteristics account for features of place that may influence whether respondents include census blocks in their neighborhood. In particular, indicators for *same block group*, *same census tract*, and *same road/rail regions* should help disentangle demographic effects from the effects of physical boundaries, which can often align with sharp transitions in demographic composition. Block groups and tracts generally group blocks together following natural boundaries like

existing neighborhood designations, highways, or bodies of water. The custom indicator for road/rail regions is designed to have the same effect.

In our analysis, we limited these infrastructure variables to those which could be computed from national data such as Census TIGER shapefiles, but researcher could also incorporate more specific geographic data, often available from municipalities, such as by using road speed or width data, to achieve even greater precision.

The model specification also includes interactions between respondent race and racial demographics, respondent party and party demographics, and respondent educational attainment and education demographics. Table C.3 in the SM describes the model specifications, all variables we use, and any of their transformations and interactions.

Our main coefficients of interest correspond to the three variables that measure the fraction of people in each block who belong to the same racial, partisan, and educational category as the respondent, respectively. We allow these coefficients to vary by the categories of each variable as well, to understand differences between groups. For example, the coefficient for the *same race category* variable can differ between white and minority respondents.

Given the geographical and demographic heterogeneity, we expect model parameters to vary significantly by city. Since there are only three cities, we choose to fit a separate model to each city's respondents. This decision is in part based on the fact that there are a moderate number of respondents for each city, leading to relatively precise parameter estimates. Linking all three cities through a single hierarchical model is also possible, but fitting such a model to the entire data would substantially increase computational cost without clear benefits.

3.4.2 Findings

To interpret the fitted models, we use the method described in Section 3.3.2. Specifically, we compute the posterior estimate of the percentage point change in the predicted probability that a respondent will include a census block in their neighborhood when increasing the value of the corresponding covariate by one unit, over a baseline probability of 50%, while holding other variables in the model constant. Figure 3.7 presents these posterior means and credible intervals (95% and 50%) for selected coefficients from the full model (see Section C.4 of the SM for the posterior summaries of all coefficients from the full and baseline models). Posterior summaries are plotted separately by city.

The results for racial similarity demonstrate that White respondents are consistently (across city samples) more likely to to include census blocks in their neighborhoods that contain a larger proportion of White residents. Holding other variables in the model constant, a White respondent is 6.1 to 16.9 percentage points more likely to include a census block composed entirely of White residents compared to one with no White residents. We cannot be confident that this preference for racial homogeneity occurs for neighborhoods drawn by minority (non-White) respondents, as the credible intervals for each city sample overlap with zero.

Partisan similarity exerts analogous predictive power for subjective neighborhoods, as both Democrats and Republicans are more likely in each of the city samples to include census blocks with larger proportions of co-partisan residents. In other words, Democrats are more likely to include Democratic neighbors in their neighborhoods and Republicans are more likely to include Republican neighbors, holding other variables in the model constant. These findings are particularly interesting since they are net of the influence of race or class homogeneity. They also suggest that one's neighborhoods are in part based on factors

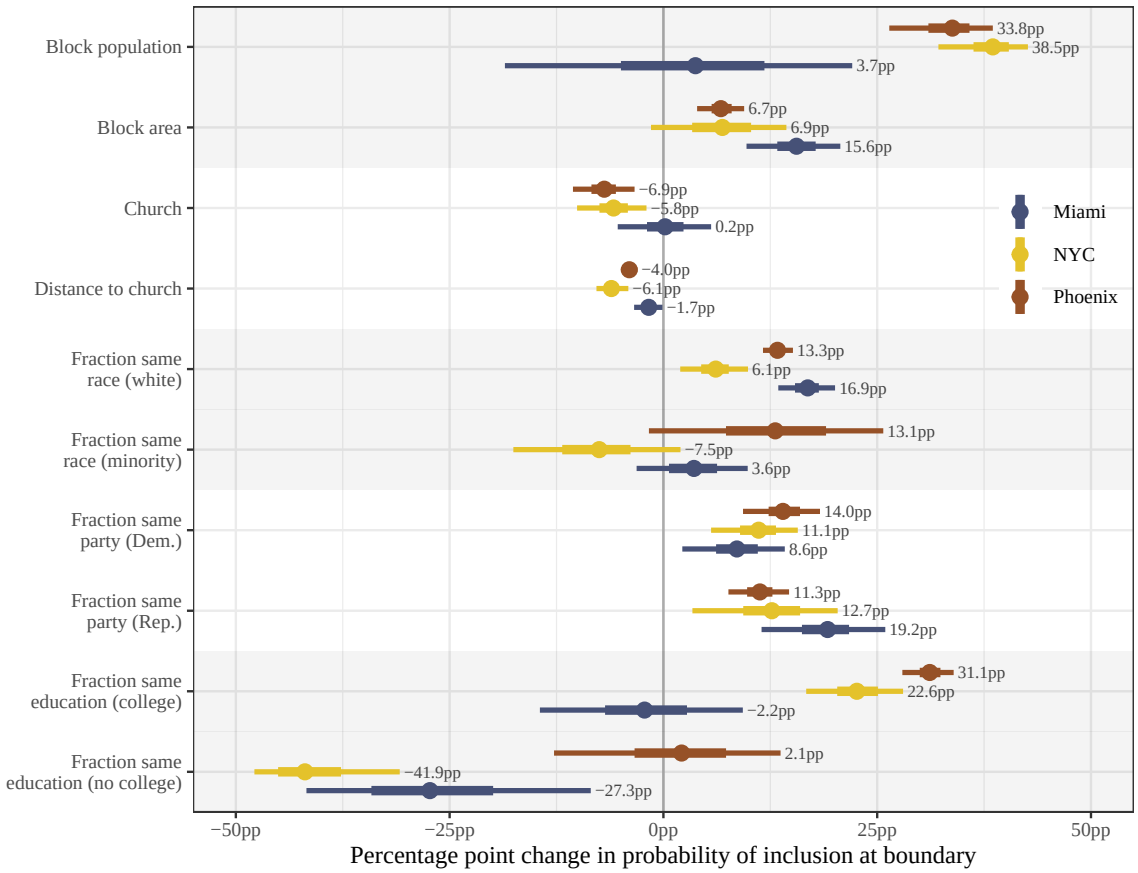


Figure 3.7: Selected full model coefficient posteriors, scaled to show the percentage point change in probability of a block's inclusion for a baseline probability of 50%. Plotted are 95% and 50% credible intervals, with posterior medians displayed to the right of each interval. Section C.4 of the SM contains the full results table.

not as immediately visible as race. This is consistent with recent research demonstrating that voters tend to give accurate answers about the partisanship of their neighbors and the relationship between actual and perceived partisanship in local areas is getting stronger (Brown, 2021).

We do not observe consistent results for educational similarity, with both college educated and non-college educated respondents seemingly preferring to include areas with more college educated residents, although the credible intervals for some of the city samples overlap with zero and the medians vary considerably across cities. Differential results across cities could be due to contextual differences between cities

but could also be due to sampling noise. Additionally, it is difficult to determine with the data what specific characateristics of cities might produce differential results.

Other neighborhood features influence inclusion probability as well. The population of a census block has a large predictive influence on census block inclusion in the New York and Phoenix samples; in these samples, respondents are much more likely to include populous census blocks in their neighborhoods. This interval is less precise and smaller in magnitude in the Miami sample. The size of census block is consistently positively associated with neighborhood inclusion, as in all three city samples larger census block are more likely to be included in neighborhoods.

We also present results on the predictive influence of whether a church is present in a census block, and how far the census block is from the nearest church. Local institutions, such as churches, schools, or community centers, may shape how people conceive of their neighborhood, either serving as alternative centers that expand the size or shape of one's neighborhood or as barriers the define the edges of neighborhoods. The inclusion of church presence in a census block and distance from a church demonstrates how our model can flexibly incorporate spatial information on local institutions and infrastructure, to quantify how these features of the built environment shape conceptions of local geography.

In our results, the presence of a church in a census block is negatively associated with inclusion of the census block in a neighborhood in the New York and Phoenix samples, with no predictive effect in the Miami sample. Distance to a church is also negatively associated with inclusion in each of the samples, meaning that census blocks that are closer to churches are more likely to be included. These results may speak to respondents opting to include residential census blocks over ones with churches in them, but their neighborhoods still being shaped by proximity to churches.

3.4.3 Quality of model fit

Before we describe how to use the fitted model for predicting neighborhoods, we examine the quality of model fit. There is significant heterogeneity in respondents' neighborhoods, as reflected in the wide range of neighborhood areas and demographics shown in Figure C.1. Neighborhoods range in size from less than 0.01 to over 100 square miles, and much of this variation in size is not captured by demographic variables. Any model will consequently struggle to make accurate predictions, especially for respondents not included in the data to which the model was fitted.

Despite these challenges, both the full and baseline models are more effective in predicting respondents' neighborhoods than a naive approach based on circles centered around their residence locations. We measure predictive accuracy by first generating 100 posterior predictions for each respondent's neighborhood. This is accomplished by taking a random sample of parameter values from the posterior, and then sequentially sampling census block inclusions according to the model's data generating process.

For each neighborhood prediction, we compute the precision and recall for the constituent census blocks, and then take their median values over predictions. Precision measures the fraction of the predicted neighborhood that is in the original neighborhood, while recall measures the fraction of the original neighborhood that is in the prediction. Both the baseline and full models have an in-sample median precision of 0.52 and recall of 0.62. Out of sample, precision falls to 0.39 and 0.37 for the baseline and full models, respectively. The out-of-sample recall is 0.53 and 0.65. So while the out-of-sample precision is about the same for the baseline and full models, the full model has superior recall.

However, due to the contiguity requirement and the sequential nature of the neighborhood model, precision and recall in this context are driven largely by the size of the predicted neighborhood. As the

neighborhood grows larger, the recall will increase at the cost of precision. In addition, by shifting the intercept of our model, we can grow or shrink the predicted neighborhood while maintaining the same discrimination with regards to predictive covariates.

Thus, to better contextualize this performance, we compare each posterior prediction (before averaging) to a circular neighborhood of the same radius as the prediction. We find this radius by taking the smallest circle centered on the respondent's home which covers their drawn neighborhood. Using the same fixed-radius circle for comparison with all model predictions would not be appropriate given the wide variation in neighborhood sizes, but allowing the circular neighborhood to exactly match the modeled radius gives this naive approach a significant leg up—it can leverage all the information learned in the model about the neighborhood's radius. We might therefore expect the baseline and full models to only minimally improve upon the circular neighborhoods, especially for out-of-sample predictions, where individual random effects have not been fit.

For the 100 predictions for each respondent, we take the median of the difference in the F1 score, which is the harmonic mean of precision and recall, between the prediction and the matching circle. Figure 3.8 shows the results of this comparison, broken out by city. Both models outperform the circular-neighborhood approach by around 0.15 in-sample, and 0.03 out-of-sample on average. Importantly, only for a few respondents do the naive approaches meet or outperform the modeled approaches, as indicated by the bulk of each boxplot lying above the x-axis. And even in these cases, as shown above, the model is able to provide uncertainty quantification and coefficient estimates, which a naive approach cannot.

The substantial heterogeneity in neighborhood sizes makes accurate neighborhood predictions difficult in general. However, the model's use of local and individual covariate information allows it to improve on

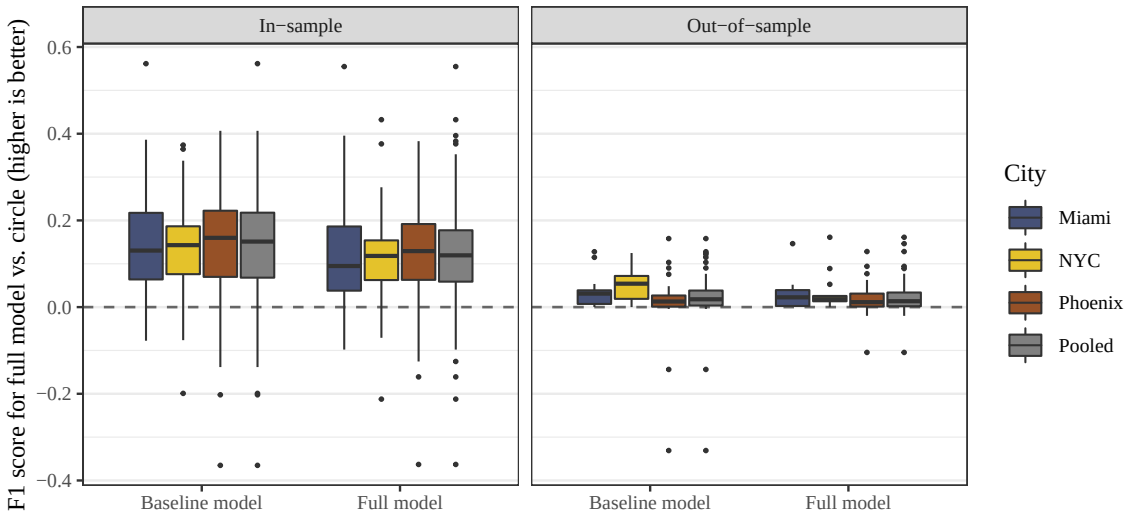


Figure 3.8: Posterior median of the difference in F1 scores between a neighborhood predicted by the model and a circular neighborhood of the same radius. The boxplot shows the variation in this median difference accros the respondents included in the model fitting (left plot) and excluded from the model fitting (right plot). Positive values indicate the model outperforming the ciruclar baseline, on average, for a particular respondent. The baseline model includes geographic information only while the full model also includes demographic information.

purely distance-based measures, even when these are well calibrated by matching the radius of a circular neighborhood to that of the model-based prediction.

3.5 Predicting Subjective Neighborhoods

The fitted model can also be used for posterior prediction of neighborhoods both at the respondent and aggregate levels, as we illustrate in this section. We show how to predict subjective neighborhoods for individual respondents and compute aggregate-level predictions.

3.5.1 Respondent-level prediction

First, we demonstrate the predictive influence of race on census block inclusion probability using a single respondent in Miami. This voter is white, female, and is not registered to a major political party. Figure 3.9 maps the racial demographics surrounding the respondent's residential address in the left panel (each

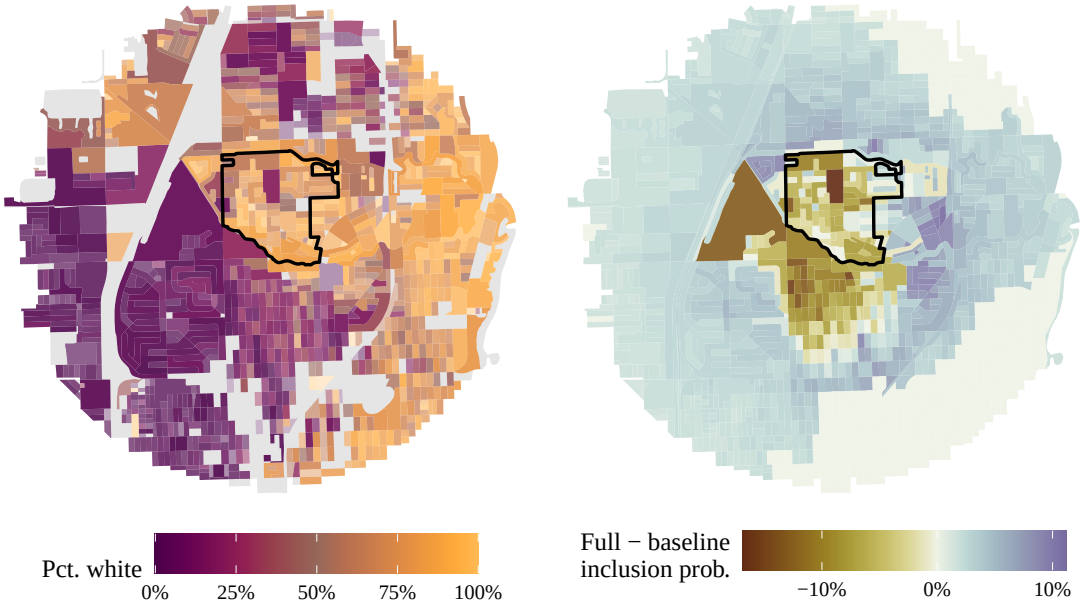


Figure 3.9: The left plot shows the racial demographics of area surrounding the example respondent. The subjective neighborhood drawn by this resopndent is indicated by the solid black line and each census block is shaded based on the percent White of its population. The right plot shows the difference in the posterior probability of a block being included in the respondent's neighborhood between the full and baseline models. The baseline model includes geographic information only while the full model also includes demographic information. Blue areas are relatively more likely to be included under the full model, while orange areas are relatively less likely to be included.

census block shaded based on the percent White of its population), and the change in the posterior probability of inclusion for each census block comparing the full model to the baseline model in the right plot. The respondent lives in a mixed but majority White area (indicated by light orange color in the left plot) that is just to the north of areas comprised largely of minority residents (dark purple color). Her drawn neighborhood (represented by a black solid line) adheres sharply to this stark southern boundary. The posterior probability map shows how these majority non-White areas are less likely to be included when demographics are accounted for in the model (indicated by dark brown color in the right plot).

Depending on one's substantive questions of interest, other quantities may be of interest and can be directly computed from the fitted model though it may require additional causal and other assumptions.

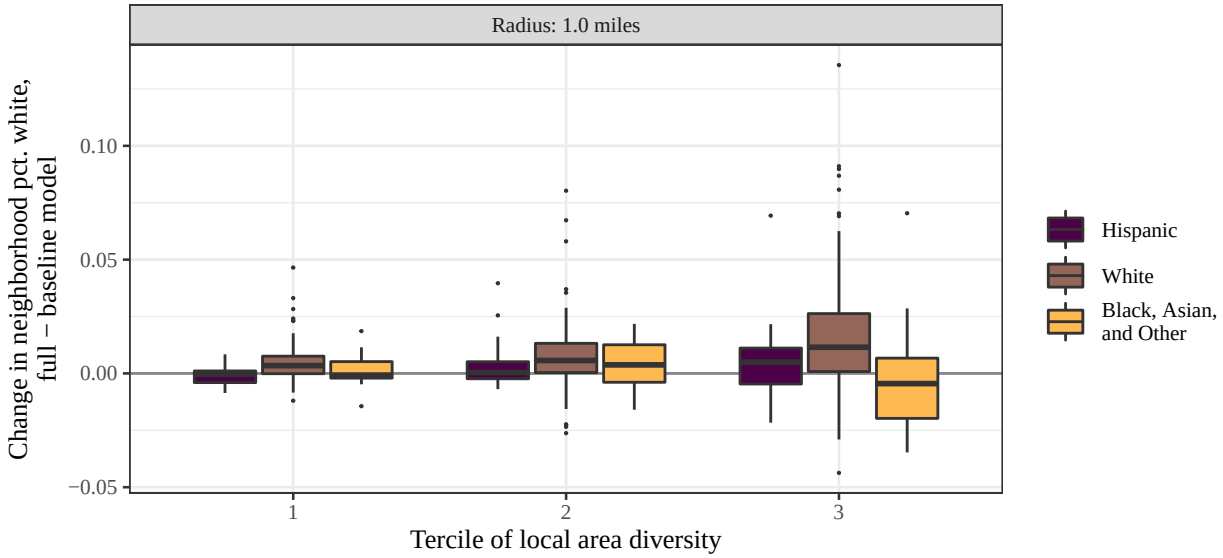


Figure 3.10: For each respondent, we compare the posterior mean of the fraction white in their predicted neighborhoods between the full and baseline models. Positive values indicate that using demographic information (full model) leads to a predicted neighborhood that is more white, on average. These differences in neighborhood fraction white are broken out by tercile of local area diversity and by the race of the respondent.

Examples include the probability of including one block, given that another block is or is not included; the change in an individual respondent's posterior predictive neighborhood if their demographics were different; or how a change in the demographics of one block (say, by a new housing development) could influence the shape and size of a respondent's neighborhood.

3.5.2 Aggregate-level prediction

Next, we demonstrate how, in aggregate, the modeled relationship between co-racial and co-partisan demographics and census block inclusion produces predicted neighborhoods with different racial and partisan makeups across voters of different races and parties. Figure 3.10 presents boxplots showing the median and interquartile ranges of the change in proportion White in predicted neighborhoods, comparing the baseline model to the full model. The full model considers demographic information, while the baseline

model does not, so the difference between the two is evidence of how much more homogeneous subjective neighborhoods become when demographics are considered.

This comparison is plotted separately by tercile of local racial diversity, to illustrate that, when voters live in areas where it is plausible to include or exclude out-group neighbors, they tend to do so. We measure local racial diversity as the standard deviation in the White percentage of each block within a fixed radius of a voter's residence. Since respondents must include or exclude whole blocks, measuring diversity according to block-level statistics is appropriate. Higher standard deviations indicate more block-level variation in racial composition and thus more opportunity to exclude out-group neighbors. Figure C.2 in the SM shows the variation in local diversity across respondents.

Three quarters of White respondents' predicted neighborhoods contain greater proportions of White residents under the full model compared to the baseline model, with the disparity increasing as neighborhood diversity increases. We further see evidence, in the most mixed neighborhoods, of predicted neighborhoods for Black and Asian respondents with lower numbers of White residents, further evidence of the impact of demographics on subjective racial neighborhoods. Figure 3.10 shows the results for a one-mile radius, but the results are robust to different specifications; Figure C.3 in the SM repeats the same plot across three different radii used in the calculation of local racial diversity.

In Figure 3.11, we plot the same comparison for partisan demographics, showing the boxplots of the difference in proportion Democrat between the full and baseline models across terciles of partisan diversity. Boxplots are shown separately by the self-reported partisan identification of the voter, ranging from strong Republican to strong Democrat. Figure C.4 in the SM repeats the same plot across three different radii used in the calculation of local racial diversity.

On average, predicted neighborhoods for Democrats are slightly more Democratic in the full model

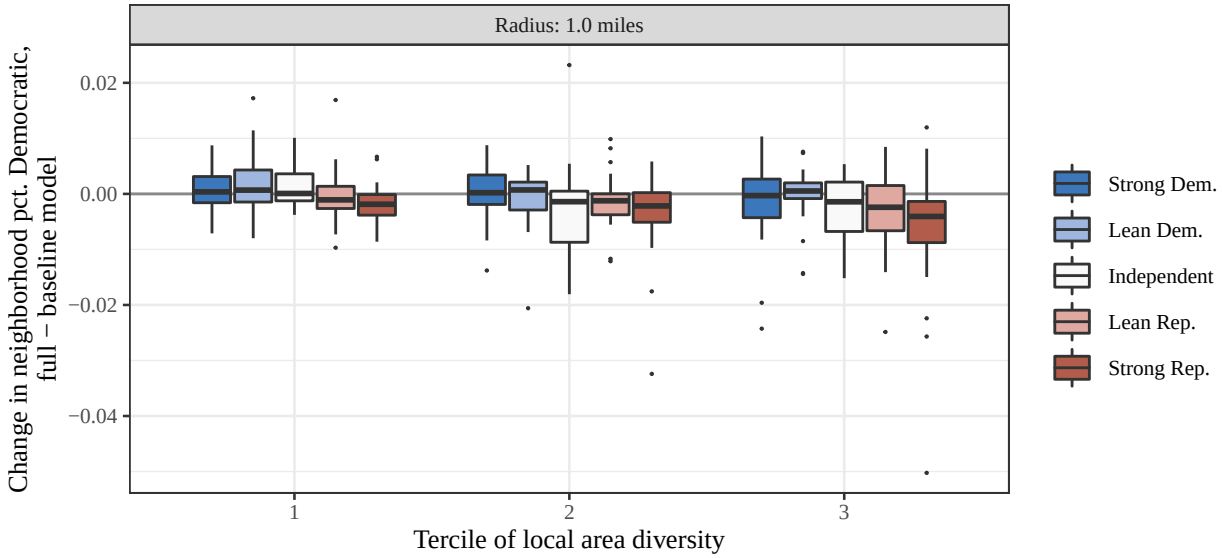


Figure 3.11: For each respondent, we compare the posterior mean of the fraction of Democrats in their predicted neighborhoods between the full and baseline models. Positive values indicate that using demographic information (full model) leads to a predicted neighborhood that is more Democratic, on average. These differences in neighborhood fraction Democratic are broken out by tercile of local area diversity and by the party identification of the respondent.

compared to the baseline model, although the interquartile ranges overlap with zero across diversity terciles. Predicted neighborhoods for Republicans, on the other hand, are noticeably less Democratic in the full model compared to the baseline model, and the largest disparity is seen for strong Republicans, where the gap reaches 0.37 percentage points in the most politically diverse areas. Similar to the racial comparison, the degree of difference is increasing with partisan diversity, and thus the potential to draw neighborhoods more differentiated by partisanship.

For both race and partisanship, the changes in neighborhood composition as a result of factoring in demographics are small in magnitude. This reflects, we believe, the overwhelming influence of residential segregation and sorting. Voters' preferences for homogeneous neighborhoods are already reflected in their choice of residence; all that we measure here is the marginal predictive effect of this preference on their subjective definition of neighborhood, given that residence.

3.6 Experimental Application

To further explore how race and partisanship influence the drawing of subjective neighborhoods, we design and implement an experiment where we randomly vary whether survey respondents are provided with information as to the racial or partisan demographics of the potential areas to include in their neighborhood. The goal with the experiment is to see whether providing concrete information as to which census blocks are more racially or politically diverse might change how overtly exclusive survey respondents are when defining their neighborhood. When voters get information on where out-group members live, they may deliberately draw them out of their neighborhoods. It is also possible that voters are not purposely exclusive but the influence of racial and partisan demographics is subconsciously incorporated into how voters' neighborhood definitions. This survey experiment, therefore, tests these two possibilities by examining the extent to which voters are deliberately exclusive when defining their neighborhoods.

3.6.1 Experimental design

The treatment for the experiment was administered by overlaying a different type of information over the map while the respondent was drawing their neighborhood. There are five experimental conditions, which were randomly assigned to survey respondents:

Control (C)	No information
Party Placebo (PH)	Partisan information, but not identified as such
Party (P)	Partisan information, identified as such
Race Placebo (RH)	Racial information, but not identified as such
Race (R)	Racial information, identified as such

The control condition C contains no overlaid shading. For the Party Placebo and Party conditions, blocks were shaded on a gradient scale based on the proportion of major party registrants who are Democrats

(Democrats/(Democrats + Republicans)). Respondents in the Party Placebo condition were just shown the colored map, with no explanation for the coloring. Respondents in the Party condition were shown the colored map and an explanation that the purple areas are more Republican, while the orange areas are more Democratic. For the Race and Race Placebo conditions, the blocks were colored according to the majority racial group (White, Black, or other) and shaded by the proportion of the census block population that is that group. As with the partisan conditions, the difference between the RH and R conditions is whether respondents were told that the colors represent the racial composition of different map areas. Figure 3.12 shows the overlaid color schemes for each of the five experimental conditions, and the accompanying text that was shown to respondents in the Party and Race groups.

The reason for splitting the partisan and racial conditions in two, creating the intermediate color-but-no-information condition, is that a map coloring in and of itself may change how respondents draw their neighborhood: it would be natural for one's neighborhood boundaries to closely match the boundaries created by the artificial map coloring, similar to how people rely on anchors and heuristics in the absence of more concrete information when making decisions or estimating or estimating numerical quantities (Tversky and Kahneman, 1974). Creating two different conditions allows us to separate this effect from our main effect.

After completing the mapping module, respondents next were asked whether they would support or oppose a ban on the construction of new housing in their neighborhood. This outcome is meant to measure a policy preference that tracks to general questions of exclusivity and NIMBYism in one's residential environment. Lastly, the respondents were asked to answer three questions comprising a trust battery: measuring the level of trust they express towards their neighbors—which we explicitly define for them as the people

living in the neighborhood they just drew. The questions were adapted from well-validated survey items designed to capture general trust (Reeskens and Hooghe, 2008; Dinesen and Sønderskov, 2015). These questions test whether defining one's neighborhood along explicit racial or partisan dimensions cause respondents to express greater trust in their neighborhood (relative to outside their neighborhood) and express greater desire to prevent more housing (and thus new residents) into their defined neighborhood. We hypothesize that the treatment conditions that provide racial or partisan information will reduce willingness and that defining one's neighborhood along partisan or racial dimensions (i.e. being assigned to the racial or partisan information condition) will cause people to express greater trust in their neighbors in their drawn neighborhood.

3.6.2 Findings

To measure the effect of racial and partisan information on exclusivity, we refit the model on the full sample ($n = 2,527$ across the three cities) and interact treatment group indicators with the co-partisan, co-ethnic, and co-class interaction variables.[1] From this, we calculate the following posterior quantities of interest:

- The difference in the respondent-block co-partisan interaction coefficient between the Party and Party Placebo groups
- The difference in the respondent-block co-ethnic interaction coefficient between the Race and Race Placebo groups

[1]Due to numerical stability issues in fitting this model to the full Phoenix sample, we used a different inference routine to fit the GLMM (the `bam` function in the `mgcv` R package). While this routine fits the same model likelihood, the implied priors are different, and due to the details of the inference function we are unable to perform corrective importance sampling. Given the large number of block-level observations in the sample, we do not believe that final inferences were meaningfully affected. Indeed, using this alternative fitting routine on the Miami and New York samples yielded no noticeable differences other than a narrower credible interval on the intercept parameter.

These quantities of interest are shown in Figure 3.13, along with the difference in the respondent-block co-ethnic and co-partisan interaction coefficient between the Race, Party, and Control groups. Appendix C.6 in the SM contains the full results table. In general, we find little overall difference in the influence of co-ethnicity or co-partisanship. Respondents in Miami and New York who were assigned to view the map shaded to show racial demographics, do not exhibit a stronger preference for co-ethnicity than respondents assigned to the placebo condition where they are shown the same shaded map but not told what the shading signifies. In Phoenix, however, we do find a statistically significant effect, with the racial information increasing the coefficient on racial homophily by 2.6 percentage points. For minority voters, there are conflicting results across cities, with little treatment effect in Miami, a positive (more exclusivity) effect in New York, but a negative effect in Phoenix. We see little effect of partisan information for Democrats in each of the cities, but conflicting results for Republicans. Republicans in New York become much more exclusionary along partisan lines in response to partisan information, while Republicans in Miami and Phoenix become much less exclusionary.

These null or conflicting results do not support the conclusion that voters are explicitly motivated to draw more homogeneous neighborhoods. When given the information to more starkly define their neighborhood along racial or partisan definitions, voters do not generally do so. But across treatment groups, the model demonstrates that—net of other variables in the model—racial and partisan homophily are important determinants of how neighborhoods are defined in voters' minds. This is consistent with the influence of local demographics being already subconsciously incorporated into how voters form attachments to their local area. In Figure C.6 of the SM, we present the treatment effects for these outcomes, which are unchanged across treatment groups.

3.7 Concluding Remarks

The study of political and social geography is often impeded by persistent measurement challenges. Progress on substantive questions necessitates methodological advancements in the construction and analysis of geographic data. We provide an open-source software package that allows researchers to measure subjective social and political geography. Our survey module can be easily modified to directly measure various geographies of interest using different prompts, designs, and instructions. For example, researchers could use different sub-geographic units (such as housing parcels, rather than census blocks) as the building blocks for subjective geographies. Another possibility is to add or remove certain information about geographical units, buildings, and landmarks.

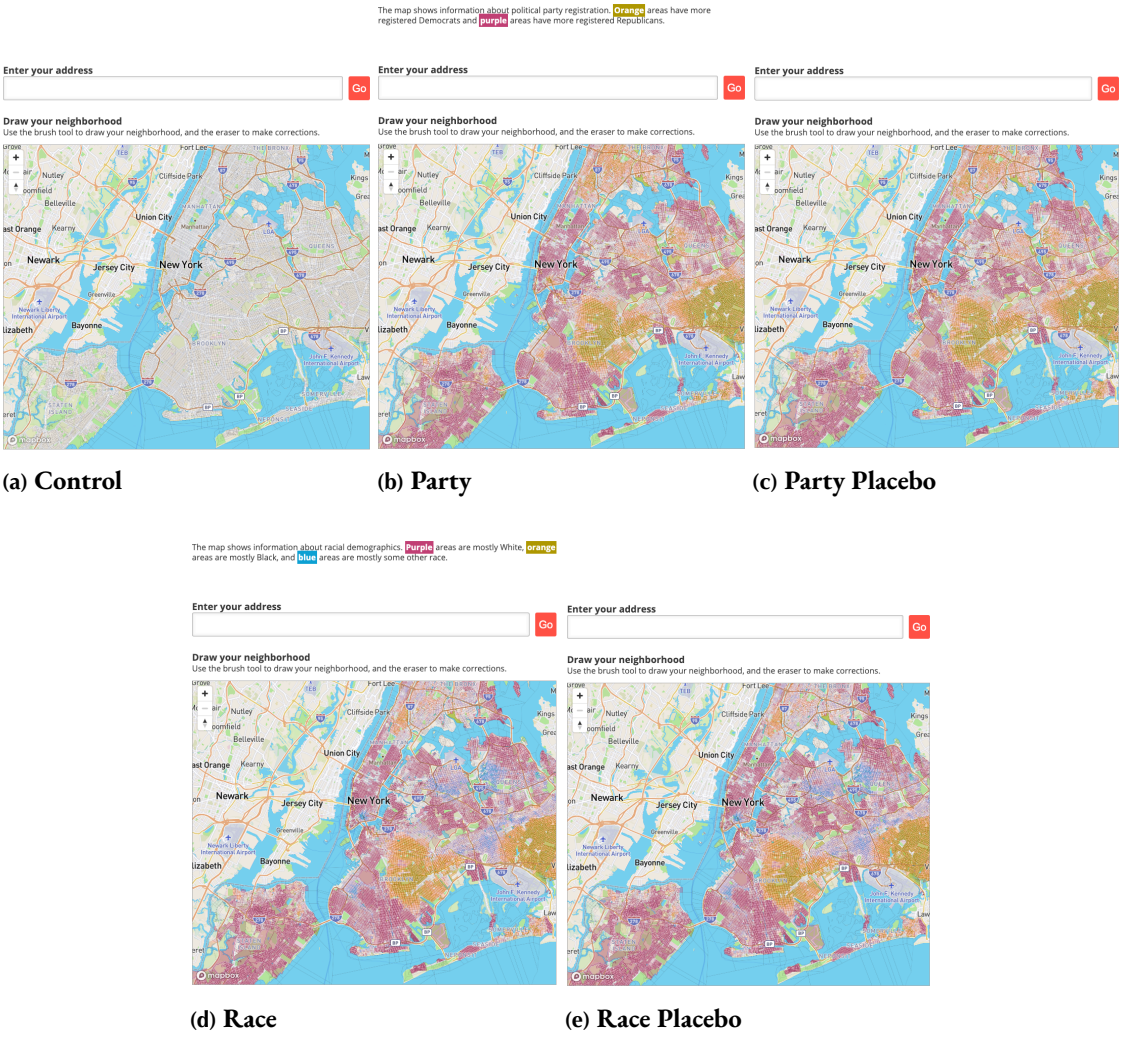
We also propose a statistical model that can be used to analyze the data obtained from our survey module. The model helps us better understand how people perceive their local geography, and this perception in turn informs the investigation of how perceived geography may influence social, political, and economic behaviors. In terms of explanatory variables of the model, any information that is spatially measured can be incorporated into the analysis, and the model can produce estimates of its influence on neighborhood inclusion. The capability to predict subjective neighborhoods out-of-sample offers further opportunities to test at scale the relationship between subjective neighborhood and socio-political outcomes.

The proposed methodology allows for quantification of the influence of any geographic variable (infrastructure, demographic, institutions, etc.) on how people define geographic areas. In addition, researchers can quantify the influence of these geographic variables in interaction with individual attributes, providing insight into how different types of people respond to different types of geographic features. Furthermore, the applications of these methodological tools are not limited to defining "neighborhoods", researchers

can ask respondents about any type of geography, and then measure how respondents define that geography differently. For example, redistricting in many states tries to keep "communities of interest" together, but policymakers and citizens differ on how to define this term. The methods illustrated in this article allow for evaluation of these subjective definitions.

Our substantive application illustrates the potential uses of this methodology, and demonstrates a striking relationship between racial and partisan demographics and subjective neighborhoods. Even after accounting for individual characteristics, aggregate socio-economic variables, and infrastructural characteristics, voters are more likely to include census blocks that consist of greater numbers of same-race or same-party residents. Variation in racial or partisan homophily produces sizable changes in inclusion probability, which in turn produce substantive differences in the kinds of subjective neighborhoods for respondents of different parties and races. These patterns spur further questions about the role of inter-ethnic and inter-party relations in shaping social geography. They are also consistent with a large body of research demonstrating how racial demographics shape psychology and political behavior (Gay, 2006; Putnam, 2007; Wong, 2010; Enos, 2017), and more recent work on the influence of partisan neighborhoods (Perez-Truglia, 2017; Brown, 2021; McCartney et al., 2021).

In this contemporary political moment of heightened partisan polarization, racial divisions in American politics, and entrenched residential segregation, understanding how inter-group conflict and geographic divisions exacerbate each other is of paramount importance for researchers, policy-makers, and engaged citizens. Future research can further investigate the extent of this influence, and the methodology demonstrated in this article provides new opportunities for this scientific pursuit.



(a) Control **(b) Party** **(c) Party Placebo**

(d) Race **(e) Race Placebo**

Figure 3.12: Map color schemes and identifying information for the five experimental conditions.

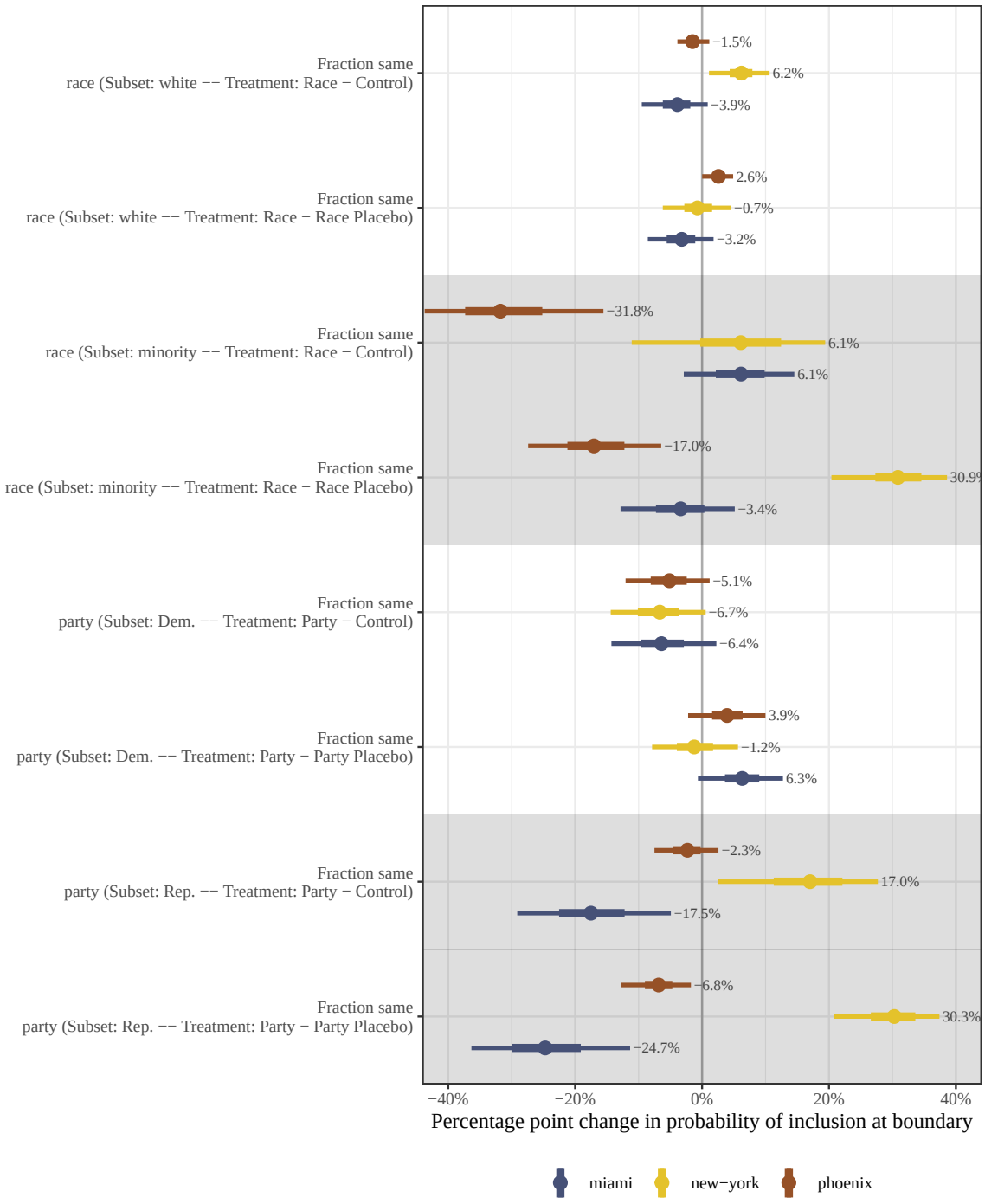


Figure 3.13: Posterior difference in racial and partisan homophily coefficients between treatment groups, scaled to show the percentage point change in probability of a block's inclusion for a baseline probability of 50%. Plotted are 95% and 50% credible intervals, with posterior medians displayed to the right of each interval. Section C.6 in the SM contains the full results table.

This page intentionally left blank.

Part III

Racial Disparities

This page intentionally left blank.

4

Estimating Racial Disparities when Race is Not Observed

4.1 Introduction

The identification and estimation of racial disparities is of paramount importance to researchers, policymakers and organizations in a variety of areas including public health (Van Ryn and Fu, 2003; Williams and Jackson, 2005), employment (Conway and Roberts, 1983; Greene, 1984), voting (Gay, 2001; Hajnal and Trounstine, 2005; Barreto, 2007), criminal justice (Berk et al., 2021; Chouldechova, 2017; Dressel and Farid, 2018), economic policy (Brown, 2022), taxation (Black et al., 2022; Elzayn et al., 2023), housing (Kermani and Wong, 2021), lending (Chen, 2018), and technology and fairness (Alao et al., 2021). Within the U.S. government, efforts to identify and remedy racial disparities have taken on greater urgency with the recent issuance of Executive Order 13985, which in part directs agencies to conduct equity assessments by developing appropriate methodology.

In many of these areas, however, racial information is not available at the individual level. The unavailability of individual racial information makes it impossible for analysts to simply tabulate variables of interest against race to identify disparities among different racial groups. In fact, in some areas, the law explicitly prohibits the collection of racial information even as it demands fair treatment on the basis of race (see, e.g., the U.S. Equal Credit Opportunity Act). This creates a dilemma for organizations who wish to measure possible disparities in order to monitor the fairness of their decision-making or service provision.

To estimate racial disparities without individual-level racial data, some researchers have turned to ecological inference methods (Goodman, 1953; King, 1997; King et al., 2004; Wakefield, 2004; Imai et al., 2008; Greiner and Quinn, 2009). These methods, however, require strong assumptions, which can be difficult to verify and may provide misleading results (Cho and Manski, 2008). Additionally, they all rely on accurate marginal information about race, which may not always be available.

Where the analysis of racial disparities involves large-scale administrative data, many analysts have adopted Bayesian Improved Surname Geocoding (BISG), which generates individual probabilities of belonging to different racial groups using Bayes' rule applied to last names and geolocation (Fiscella and Fremont, 2006; Elliott et al., 2008; Imai and Khanna, 2016; Imai et al., 2022). BISG leverages residential racial segregation and the hereditary association between self-reported race and surname to produce generally accurate and calibrated predictions of self-reported individual race (Kenny et al., 2021; DeLuca and Curiel, 2022).

Unfortunately, calibrated BISG racial prediction alone is not sufficient for unbiased estimation of disparities. Indeed, the most common techniques for estimating racial disparities using BISG probabilities are known to be biased when race is correlated with the outcome even after controlling on name and location (Chen et al., 2019; Argyle and Barber, 2022). These approaches include the use of BISG probabilities as weights and thresholding the BISG probabilities to produce point estimates of individual race. As formally

discussed in Section 4.2, these standard methods of racial disparity estimation based on BISG predictions also require individuals' race to be conditionally independent of the outcome given their residence location, surnames, and other observable attributes. This key identification assumption, however, is unlikely to hold because race affects many aspects of society even after accounting for residence location, surnames, and other observable attributes. Other researchers have noted the implausibility of this assumption and have proposed methods to address it, but these alternative approaches are more general than the racial disparity setting and require additional data such as a partially labeled subset (Fong and Tyler, 2021).

To address this challenge, in Section 4.3, we propose an alternative identification strategy. Specifically, we assume that the outcome is conditionally independent of surname given (unobserved) individual's race, residence location, and other observed attributes. This assumption is a type of exclusion restriction where surname serves as an instrumental variable for unobserved race. It implies that for two individuals who live in the same area, belong to the same racial group, and share the observable attributes, their surnames have no predictive power of the outcome. Somewhat counterintuitively, the high-dimensionality of surnames aids rather than hinders identification. We argue that this new identification assumption is more credible unless surname is directly used to determine the outcome of interest (i.e., name-based discrimination). Leveraging this identification strategy, we introduce a new class of models, Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that accurately estimates racial disparities. This approach includes built-in flexibility for researchers to make problem-specific modeling choices. We propose an EM algorithm for fitting BIRDiE models that can scale to hundreds of thousands or millions of observations. BIRDiE also produces updated BISG probabilities that incorporate the outcome variable, and are more accurate than the BISG probabilities based only on surnames and geolocation. Finally, we show how to

address potential violations of the key identification assumptions, such as occurs with name-based discrimination, by exploiting auxiliary information about the relations between names and more specific ethnic groups. All of the proposed methodology is implemented in an open-source software package, `birdie`, that is made available with the paper.[1]

In Section 4.4, we validate the proposed methodology using real-world data taken from the voter file in North Carolina, where self-reported individual-race is observed and can be used to construct the ground-truth of racial disparities. BIRDiE substantially outperforms existing estimators across different error measures, two different outcome variables, and multiple levels of geolocation specificity. For example, the most popular existing BISG disparity estimator pegs the gap at Democratic party registration between White and Black voters at 20.8 percentage points, while the actual gap is 54.6 percentage points—more than double. Our preferred BIRDiE model yields an estimate of 49.2 percentage points. This represents about a 84% reduction in bias.

Section 4.5 gives concluding remarks.

4.2 Bias of the Standard Methodology

In this section, we review the assumptions of the standard BISG-based methodology for estimating racial disparities when individual race is not observed. We show that the racial disparity estimates based on the standard methodology are biased unless the outcome variable is independent of race given surname, residence location, and other observed covariates. We argue that this assumption is likely to be violated given the significant role race plays in many aspects of our society.

[1] The software and accompanying documentation are available at https://corymccartan.com/birdie/.

4.2.1 Setup and BISG Procedure

Suppose that we have an i.i.d. sample of N individuals from a population of infinite size. For each individual $i = 1, 2, \ldots, N$, we define a tuple $(Y_i, R_i, G_i, X_i, S_i)$, where $Y_i \in \mathcal{Y}$ is the outcome of interest for individual i, $R_i \in \mathcal{R}$ is the (unobserved) race of the individual, $G_i \in \mathcal{G}$ is the (geo)location of the individual's residence, $X_i \in \mathcal{X}$ are other observed characteristics, and $S_i \in \mathcal{S}$ is the individual's surname. When we are not referring to a particular individual, we will drop the subscripts for notational simplicity. Note that individual race is unobservable but all other variables are assumed to be observed. The availability of particular (or any) X is not required for either the standard or proposed methodology.

We assume throughout that these variables are discrete, i.e., $|\mathcal{Y}|$, $|\mathcal{R}|$, $|\mathcal{G}|$, $|\mathcal{X}|$, and $|\mathcal{S}|$ are all finite. Note that typically S is high-dimensional as there exist a large number of unique surnames. In practice, residence location G is also discrete, since joint information about location, race, and other variables is generally only available down to the Census block level. For the sake of simplicity, we assume that the outcome variable Y is also discrete, though it is possible to extend the standard and proposed methodologies to continuous outcome variables.

Typically, BISG relies on data from the decennial Census or the American Community Survey (ACS), which provide information on the joint distribution of R and G (and any other covariates X, such as gender or age). It then combines this information with data from the Census Bureau's surname tables (U.S. Census Bureau, 2014), which provide information on the joint distribution of R and S. We summarize this set of information from the Census by two conditional probabilities, $\mathbf{q}_{GX|R}$ and $\mathbf{q}_{S|R}$, and one marginal probability, $\mathbf{q}_R$.

The BISG estimator of the probability that individual i belongs to race $r \in \mathcal{R}$ can then be written as

(Fiscella and Fremont, 2006; Elliott et al., 2008)

$$\hat{P}_{ir} := \frac{q_{G_i X_i|r}\, q_{S_i|r}\, q_r}{\sum_{r' \in \mathcal{R}} q_{G_i X_i|r'}\, q_{S_i|r'}\, q_{r'}}, \tag{4.1}$$

where, for example, $q_{G_i X_i|r}$ indicates the estimated conditional probability of residence location G_i and covariates X_i given race r, taken from the Census table $\mathbf{q}_{GX|R}$.

The BISG estimator relies on two key assumptions. The first is the following conditional independence relation between an individual's surname and residence location (as well as other characteristics) given their unobserved race.

Assumption A1 (Conditional independence of name and other proxy variables)**.** *For all i,*

$$S_i \perp\!\!\!\perp \{G_i, X_i\} \mid R_i.$$

Assumption A1 implies, for example, that once we know an individual is White, knowing their surname is Smith tells us nothing about their residence location and other observed characteristics. Although this assumption appears to be reasonable, the lack of granularity in the coding of race may lead to its violation. For example, people with Chinese, Indian, Filipino, Vietnamese, Korean, or Japanese are all coded as one racial group "Asian" in the census. These groups, however, have varying sets of surnames and have different demographic and geographic distributions. For instance, unlike the Smith example, knowing that an Asian individual's surname is Gupta makes it more likely that they have a higher income and live in the Eastern U.S (Budiman et al., 2019).

The second assumption required by BISG is that the Census tables reflect the true population distributions of R, S, G, and X.

Assumption A2 (Data accuracy)**.** *For all i,*

$$\mathbb{P}(R_i = r) = q_r$$

$$\mathbb{P}(S_i = s \mid R_i = r) = q_{s|r}$$

$$\mathbb{P}(G_i = g, X_i = x \mid R_i = r) = q_{gx|r}$$

Like Assumption A1, Assumption A2 may never hold exactly in practice. Despite the best efforts of the Census Bureau, the decennial census has intrinsic error, including undercounting minority racial groups (U.S. Census Bureau, 2022; Anderson and Fienberg, 1999; Strmic-Pawl et al., 2018), as well as error introduced by privacy-preserving mechanisms (Abowd et al., 2022). And because of births, deaths, and moves, census data are often out-of-date from the moment of publication. These errors have led further extensions of the BISG estimator to account for some of this measurement error (Imai et al., 2022), which can help with accuracy for smaller racial groups. The plausibility of Assumption A2 is stretched even further when the study population is a subset of the whole U.S. population, and so is not covered by national census data. In these cases, analysts should set $\mathbf{q}_R$ to the known or estimated marginal racial distribution in the study population, rather than the national racial distribution. It may be more plausible then to assume that the conditional distributions $\mathbb{P}(S \mid R)$ and $\mathbb{P}(G, X \mid R)$ match the census distributions, even if $\mathbb{P}(R)$ does not.

Even though Assumptions A1 and A2 may not hold exactly, researchers find that BISG produces accurate and generally well-calibrated estimates in practice (Imai and Khanna, 2016; Kenny et al., 2021; DeLuca and Curiel, 2022). We observe this pattern as well in the validation study in Section 4.4. New methods con-

tinue to be developed that improve the calibration of BISG probabilities, including some machine learning methods based on labeled data (Imai et al., 2022; Argyle and Barber, 2022; Decter-Frain, 2022).

Under Assumption A1, by Bayes' Rule,

$$\mathbb{P}(R_i = r \mid G_i, X_i, S_i) \propto \mathbb{P}(G_i, X_i \mid R_i = r)\, \mathbb{P}(S_i \mid R_i = r)\, \mathbb{P}(R_i = r).$$

This justifies the estimator given in Equation (4.1), and provides us with the following immediate result.

Proposition 4.1 (Accuracy of BISG)**.** *Under Assumptions A1 and A2, the BISG estimator is accurate. That is, we have* $\hat{P}_{ir} = \mathbb{P}(R_i = r \mid G_i, X_i, S_i)$.

4.2.2 Bias of Racial Disparity Estimates based on BISG Probabilities

Unfortunately, accurate and calibrated estimates of individual race predictions alone are not sufficient for unbiased estimation of racial disparities using standard methodologies. The reason is that the prediction error of race probabilities may be correlated with the outcome variable of interest. Specifically, consider the following weighting estimator commonly used by researchers, which attempts to capture the uncertainty inherent in race prediction:[2]

$$\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) = \frac{\sum_{i=1}^{N} \mathbf{1}\{Y_i = y\} \hat{P}_{ir}}{\sum_{i=1}^{N} \hat{P}_{ir}}.$$

Chen et al. (2019) show that the asymptotic bias of this weighting estimator is controlled by the residual correlation of Y and R after adjusting for G, X, and S. We reproduce this result here.

[2]An alternative is a *thresholding* or *classification* estimator, which deterministically assigns individuals to a predicted racial category based on the BISG estimates $\hat{\mathbf{P}}_i$ (either the largest $\hat{P}_{ir}$ or the one which exceeds a predetermined threshold). Unlike the weighting estimator, this classification estimator does not take into account prediction uncertainty, and its bias is harder to reason about than that of the weighting estimator (Chen et al., 2019).

Theorem 4.2 (Theorem 3.1 of Chen et al. 2019)**.** *If race is binary (so* $\mathcal{R} = \{0, 1\}$*), then as* $N \to \infty$*,*

$$\hat{\mu}^{(wtd)}_{Y|R}(y \mid r) - \mathbb{P}(Y = y \mid R = r) \xrightarrow{a.s.} -\frac{\mathbb{E}[\mathrm{Cov}(\mathbf{1}\{Y = y\}, \mathbf{1}\{R = r\} \mid G, X, S)]}{\mathbb{P}(R = r)}.$$

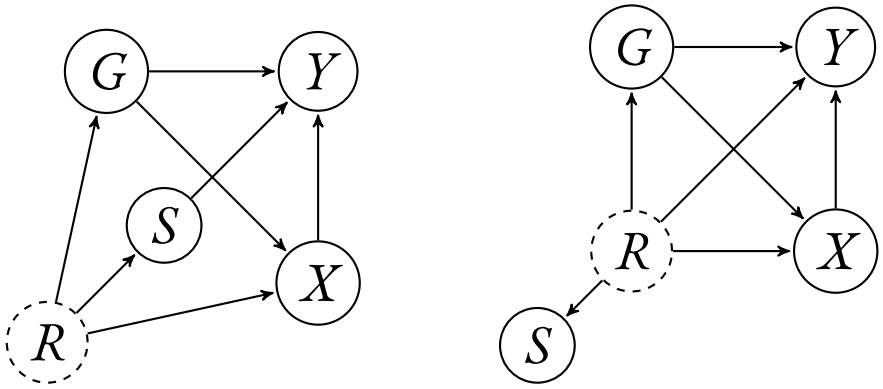
This result implies that when the BISG residuals $\mathbf{1}\{R = r\} - \mathbb{P}(R = r \mid G, X, S)$ are correlated with the outcome, estimates will be biased, even with infinite data. Conversely, conditional independence between individual's race and outcome given their surname, residence location, and other characteristics is sufficient to eliminate the asymptotic bias of the weighting estimator.

Assumption A3 (Conditional independence of outcome and race)**.** *For all i,*

$$Y_i \perp\!\!\!\perp R_i \mid G_i, X_i, S_i.$$

Figure 4.1(a) shows a causal directed acyclic graph (DAG) that satisfies Assumption A3 as well as Assumption A1. The dashed node border for R represents the fact that race is unobserved. The causal structure in Figure 4.1(a) implies the conditional independence relation $Y \perp\!\!\!\perp R \mid G, X, S$, because all paths from R to Y are blocked by G, X, or S. The key causal assumption of this DAG is that the effect of race R on the outcome Y must be entirely mediated by surname S, residence location G, and other observed characteristics X. This type of exclusion restriction may not be credible in many practical settings because race can affect the outcome through so many factors, biasing the weighting estimator.

But in other settings, Assumption A3 may be more plausible. For example, for a manager reviewing job applications on paper, race is typically unobserved, but the manager may be influenced by racial or gender cues in a candidate's name or address (Park et al., 2009; Åslund and Skans, 2012). So long as we observe all information used by the manager and use it in the BISG estimation, the weighting estimator would give an asymptotically unbiased estimate. Similarly, in evaluating the fairness of algorithmic decision-making,



(a) Assumption A3: Conditional independence of Y and R given (G, S, X)

(b) Assumption A4: Conditional independence of Y and S given (G, R, X)

Figure 4.1: Possible causal structures for which each of the labeled assumptions is satisfied, represented as a directed acyclic graph (DAG) where G is residence location, R is race, S is surname, X is observed covariates, and Y is the outcome. Race R is unobserved, which is signified by a dashed node boundary. Both DAGs also satisfy Assumption A1: Conditional independence of S and (G, X) given R.

as long as all the information used by the algorithm is incorporated into BISG, Assumption A3 would be appropriate. Outside of these cases, however, the weighting estimator is likely to be biased.

Finally, while the discussion in this section has been focused on the BISG methodology, the qualitative results and necessary assumptions carry over to other approaches which produce probabilistic predictions of individual race, such as those recently developed by Argyle and Barber (2022) and Decter-Frain (2022). Just as with BISG, well-calibrated probabilities are not generally sufficient to produce unbiased estimates of racial disparities using the outputs of these methods.

4.3 The Proposed Methodology

In this section, we propose an alternative identification strategy. Specifically, we show that racial disparity is identifiable if surname is conditionally independent of the outcome given race, residence geolocation, and other observed information, under the aforementioned assumptions that guarantee the nonparamet-

ric identification of race probabilities with BISG or its variants. We then develop a class of statistical models, called Bayesian Instrumental Regression for Disparity Estimation (BIRDiE), that estimate racial disparity under this identification condition by using surnames as a high-dimensional instrumental variable for race. These models take as inputs the BISG probabilities, and so can be easily applied on top of existing analysis pipielines, including with alternative probabilistic race prediction methodologies. We next discuss computation for BIRDiE models, as well as how the methodology can be extended to include an additional explanatory variable that was not used at the BISG stage. Finally, we show how to addresses the potential violations of the key identification assumptions, such as occurs with name-based discrimination.

4.3.1 Identification Strategy

To reduce the potential bias of the weighting estimator, we propose an alternative identification assumption that may be applicable when Assumption A3 is not credible. Specifically, we assume that surname, rather than race, satisfies the exclusion restriction conditional on (unobserved) race, residence location, and other observed characteristics.

Assumption A4 (Conditional independence of outcome and name)**.** *For all i,*

$$Y_i \perp\!\!\!\perp S_i \mid R_i, G_i, X_i.$$

Figure 4.1(b) shows one possible causal DAG that meets this assumption as well as Assumption A1. In this DAG, race can have a direct effect on the outcome Y as well as on residence location G and other observed characteristics X, while all paths from S to Y are blocked by G, X, or R.

This causal structure is plausible in many real-world settings because, unlike Assumption A3, Assumption A4 allows race to directly affect the outcome. For an outcome like party registration, Assumption A4

would mean that among White voters in a particular geographic region, voters named Smith would *a priori* be no more or less likely to identify with one party than voters named Thomas. In contrast, Assumption A3 would mean that among voters named Smith in a particular geographic region, White voters would be *a priori* be no more or less likely to identify with one party than Black voters. In this case, Assumption A3 is likely to be violated, while Assumption A4 is plausible. It is important to note a key trade-off between the two assumptions. While Assumption A4 rules out the possibility that surname directly affects the outcome (e.g., name-based discrimination), such a direct effect is allowed under Assumption A3. Section 4.3.6 revisits this important issue.

The appropriateness of each assumption depends on a specific application. In the above hiring example, if the manager reviews applicants anonymously, there will be no name-based discrimination and the assumption is likely to be satisfied. The assumption may be violated in other contexts, however. For example, in studying turnout, if campaigns use the surnames of individual voters to decide whether to mobilize them, Assumption A4 will be violated. Yet another possible violation of the assumption is the existence of unobserved confounder that affects both outcome and surname. The country of origin for an immigrant may represent such a confounder: where surnames are informative of country of origin, even within racial groups (as is often the case for Asian individuals), variations in outcomes by country of origin will likely violate A4. This reflects the limitations of the relatively coarse racial classifications used in BISG, as discussed above. Even in these cases, however, conditioning on (unobserved) race is likely to substantially reduce the magnitude of association between outcome and surnames. In Section 4.3.6, we show how to address this potential violation of Assumption A4.

We briefly note that Assumptions A3 and A4 are not necessarily mutually exclusive. For example, nei-

ther race or surname could have a direct causal effect on the outcome. In these cases, both the weighting estimator and our approach proposed below could give reasonable answers, though the data requirements of each may differ.

The following theorem shows that it is possible to nonparametrically identify racial disparities under Assumption A4. The proof of the theorem and all other results in this paper are deferred to the appendix.

Theorem 4.3 (Nonparametric Identification)**.** *For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix* $\mathbf{P}$ *with entries $p_{rs} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector* $\mathbf{b}$ *with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then under Assumption A4, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, G = g, X = x)$ are identified if and only if both* $\mathbf{P}$ *and the augmented matrix* $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$ *have rank $|\mathcal{R}|$.*

The essence of this identification result is the following simple observation. Under Assumption A4, we have, for all $y \in \mathcal{Y}, g \in \mathcal{G}, x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\mathbb{P}(Y = y \mid G = g, X = x, S = s) = \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x)\, \mathbb{P}(R = r \mid G = g, X = x, S = s). \tag{4.2}$$

The leftmost term is estimable from the data and corresponds to the vector $\mathbf{b}$ in Theorem 4.3, while the rightmost term is the BISG estimand and corresponds to the matrix $\mathbf{P}$ in Theorem 4.3. Lastly, the remaining term in the middle can be solved for, since Equation (4.2) holds across all combinations of Y, G, X, and S, leading to a large system of linear equations. Specifically, we have $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{S}|$ equations with $(|\mathcal{Y}| - 1) \times |\mathcal{G}| \times |\mathcal{X}| \times |\mathcal{R}|$ unknowns. Since $|\mathcal{R}| \ll |\mathcal{S}|$, we can identify these unknowns as long as the linear system has sufficient rank. Our result is closely related to causal identification based on proxy

variables in the presence of unmeasured confounding (Kuroki and Pearl, 2014; Miao et al., 2018; Knox et al., 2022). Here, we use surname as a proxy variable for (unobserved) race to identify racial disparities.

Together with Proposition 4.1, Theorem 4.3 implies that racial disparities can be identified under Assumptions A1, A2, and A4. The identifying equation (4.2) shows that $\mathbb{P}(Y = y \mid G = g, X = x, S = s)$ is linear in the BISG estimands $\mathbb{P}(R = r \mid G = g, X = x, S = s)$. Thus, it is natural to consider the following least-squares estimator of $\mathbb{P}(Y = y \mid R, G = g, X = x)$ under this alternative identification strategy,

$$\hat{\boldsymbol{\mu}}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x) = (\hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)}\hat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1}\hat{\mathbf{P}}_{\mathcal{I}(xg)}\,\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\},$$

where as above $\hat{\mathbf{P}}$ is the matrix of BISG probabilities, and $\mathcal{I}(xg)$ is the set of individuals i with $X_i = x$ and $G_i = g$. Here and throughout the paper, a dot will indicate a vector constructed over that index, so $\hat{\boldsymbol{\mu}}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ is a vector of conditional probabilities for a particular outcome level y across all racial groups in $\mathcal{R}$. By post-stratifying this estimator across the (G, X) cells, we arrive at an estimator of $\mathbb{P}(Y = y \mid R)$,

$$\hat{\boldsymbol{\mu}}^{(\text{p-ols})}_{Y|R}(y \mid r) = \sum_{x \in \mathcal{X}, g \in \mathcal{G}} (\hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)}\hat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1}\hat{\mathbf{P}}_{\mathcal{I}(xg)}\,\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\})_r q_{gx|r},$$

since $q_{gx|r} = \mathbb{P}(G = g, X = x \mid R = r)$ under Assumption A2. This estimator is unbiased, as the following theorem shows.

Theorem 4.4 (Unbiasedness of OLS Estimator)**.** *If Assumptions A1, A2, and A4 hold, and the identification conditions in Theorem 4.3 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,*

$$\mathbb{E}[\hat{\mu}^{(p\text{-}ols)}_{Y|R}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

It is worth comparing this OLS estimator with the weighting estimator $\hat{\mu}^{(\text{wtd})}_{y|r}$. The next theorem shows

that within the (G, X) cells these two estimators are guaranteed to disagree, unless either the BISG probabilities perfectly discriminate or the weighting estimator is constant across races. Unfortunately, these two conditions are almost never met in practice. This underscores the importance of selecting the appropriate assumption (Assumption A3 or A4) for a particular analysis, since they imply different estimators with different results.

Theorem 4.5 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators)**.** *For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\hat{\boldsymbol{\mu}}^{(wtd)}_{Y|R}(y \mid \cdot) = \hat{\boldsymbol{\mu}}^{(ols)}_{Y|R}(y \mid \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j \mid G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k \mid G_i, X_i, S_i) = 0$ and vice versa) or $\hat{\mu}^{(wtd)}_{Y|R}(y \mid j) = \hat{\mu}^{(wtd)}_{Y|R}(y \mid k)$.*

Despite potential advantages over the weighting estimator, the OLS estimator is not well-suited to estimation in practice, since it ignores the fact that the unknown parameters are probabilities and thus constrained to be nonnegative and sum to 1. As a result, in any particular sample, the estimator can produce impossible or contradictory estimates. This is particularly problematic because a large number of unique surnames make both $\mathbf{P}$ and $\mathbf{b}$ high-dimensional. To address this challenge, and to open the door to more flexible modeling, we next propose a Bayesian modeling approach that is based on our identification strategy and yet satisfies necessary constraints.

4.3.2 Bayesian Instrumental Regression for Disparity Estimation

The BIRDiE approach combines a user-specified complete-data outcome model $\pi(Y \mid R, G, X, \Theta)$, parametrized by Θ, with the BISG model in order to estimate the distribution $Y \mid R$ that is of interest. We first present

the general BIRDiE model and describe potential choices of complete-data outcome models before discussing our computational approach.

The BIRDiE posterior is obtained by applying Assumptions A1, A2, and A4 to the joint distribution $\pi(Y, R, G, X, S, \Theta)$:

$$\begin{aligned}
\pi(\Theta, \mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) &\propto \pi(\Theta, \mathbf{R}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \\
&\propto \pi(\Theta) \prod_{i=1}^{N} \pi(Y_i \mid R_i, G_i, X_i, \Theta) \pi(R_i \mid G_i, X_i, S_i) \\
&= \pi(\Theta) \prod_{i=1}^{N} \pi(Y_i \mid R_i, G_i, X_i, \Theta) \hat{P}_{iR_i}, \qquad (4.3)
\end{aligned}$$

where as above $\hat{\mathbf{P}}_i$ are the BISG probability estimates for individual i, which depend on Census data represented in $\mathbf{q}_{GX|R}$, $\mathbf{q}_{S|R}$, and $\mathbf{q}_R$, but not on the outcome-model parameters Θ.

To apply this general BIRDiE model to a particular analysis requires choosing a complete-data outcome model, given by the likelihood $\pi(Y_i \mid R_i, G_i, X_i, \Theta)$ and prior $\pi(\Theta)$. Since Y is discrete, a categorical regression model is appropriate for $\pi(Y_i \mid R_i, G_i, X_i, \Theta)$. The parametrization of the categorical regression will depend on the analyst's goals, computational resources, and prior beliefs about the structure of the problem. We present here several reasonable alternatives that trade off modeling flexibility and computational efficiency.

Complete-pooling model. The simplest possible model is one in which the relationship between Y and R does not vary with G or X. This model is parametrized by $\Theta = \{\boldsymbol{\theta}_r\}_{r \in \mathcal{R}}$, which describe the distribution of Y within every level of R:

$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Cat}_{\mathcal{Y}}(\boldsymbol{\theta}_{R_i})$$

$$\boldsymbol{\theta}_r \overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}),$$

where $\text{Cat}_{\mathcal{Y}}$ denotes a discrete (categorical) distribution on the set $\mathcal{Y}$. With known $\mathbf{R}$, the posterior of $\boldsymbol{\theta}_r$ is conjugate, a fact which will make computation under the EM scheme described in Section 4.3.3 below extremely efficient. Of course, this efficiency comes at the cost of a restrictive model that allows for no role of G and X. If in reality $\mathbb{P}(Y \mid R)$ does vary along these dimensions, it is possible that the posterior of $\boldsymbol{\theta}_r$ will not accurately estimate $\mathbb{P}(Y \mid R)$. In any case, if the analyst is interested in subgroup or small-area estimates of $\mathbb{P}(Y \mid R)$, the complete-pooling model will be of little use.

Saturated (no-pooling) model. At the other end of the spectrum from the complete-pooling model is a *saturated* or no-pooling model, which estimates a different distribution of $Y \mid R$ within every level of G and X:

$$Y_i \mid R_i, G_i, X_i, \Theta \sim \text{Cat}_{\mathcal{Y}}(\boldsymbol{\theta}_{R_i G_i X_i})$$

$$\boldsymbol{\theta}_{rgx} \overset{iid}{\sim} \text{Dirichlet}(\boldsymbol{\alpha}).$$

This model is closest to the OLS estimator, though it ensures that all probability estimates lie in $[0, 1]$. As with the complete-pooling model, the posterior of $\boldsymbol{\theta}_{rgx}$ is conjugate to its prior, and so computation can be made efficient. Additionally, this model allows for any arbitrary relationship between Y, R, G, and X. Since it is fully nonparametric, the posterior will converge to the true $\mathbb{P}(Y \mid R, G, X)$ with enough data in each (G, X) cell. However, in practice, the model can suffer from the curse of dimensionality: the number of (G, X) cells may be relatively large compared to the amount of available data, or even exceed it, especially since G can be quite large, covering many blocks or ZIP codes. In these cases, the prior will dominate the data in each cell, which could have a large biasing effect even on overall inferences about $\mathbb{P}(Y \mid R)$.

General mixed-effects model. As a compromise between the complete-pooling and no-pooling model, a partial pooling approach based on a multinomial mixed-effects model can be used. Properly specified, the mixed-effects model maintains the flexibility of the saturated model while avoiding its high bias and variance in finite samples.

$$
\begin{aligned}
Y_i \mid R_i, G_i, X_i, \Theta &\sim \text{Cat}_Y(g^{-1}(\boldsymbol{\mu}_{rgx})) \\
\mu_{rgxy} &= \mathbf{W}\boldsymbol{\beta}_{ry} + \mathbf{Z}\mathbf{u}_{ry} \\
\mathbf{u}_{ry} \mid \phi_{ry} &\sim \mathcal{N}\Big(0, \Sigma(\boldsymbol{\phi}_{ry})\Big) \\
\boldsymbol{\beta}_{ry} \overset{iid}{\sim} f_\beta, &\quad \boldsymbol{\phi}_{ry} \overset{iid}{\sim} f_\phi,
\end{aligned}
$$

where g^{-1} is a softmax or other link function, $\mathbf{W}$ and $\mathbf{Z}$ are matrices of fixed and random effects, respectively, $\boldsymbol{\phi}$ is a vector of random-effect parameters, and f_β and f_ϕ are some priors for the subscripted parameters. Some fixed or random effects could be shared across combinations of R and Y, though this could complicate computation. We recommend including X and especially G in the model as random effects, with hierarchical structure as appropriate. Such a structure partially pools estimates of $\mathbb{P}(Y \mid R, X, G)$ towards an overall estimate of $\mathbb{P}(Y \mid R)$, allowing the model to share information between geographic areas. This should prove especially useful in cases where some areas have few or no observations for certain racial groups.

We also recommend including group-level covariates as fixed effects, which will help share information across random effects and significantly improve generalization performance to unseen random effect levels (Buttice and Highton, 2013). For example, if G records counties, analysts could include racial and socioeconomic variables measured at the county level as predictors. This would help produce more accurate

estimates of $\mathbb{P}(Y \mid R, G)$ to the extent that variation in these probabilities is associated with these racial and socioeconomic variables. Ultimately the structure of this general model will have to be chosen based on the data and the relevant research question.

4.3.3 Computation

The posterior in Equation (4.3) contains the high-dimensional discrete nuisance parameter $\mathbf{R}$, which poses a challenge for computation. We suggest two approaches for handling $\mathbf{R}$, one suited to small sample sizes, and one suited to large sample sizes.

Small samples: Inference directly on the marginal likelihood. Since $\mathbf{R}$ is discrete, we can marginalize it out as follows:

$$\pi(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) = \sum_{\mathbf{r} \in \mathcal{R}^N} \pi(\Theta, \mathbf{r} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \propto \pi(\Theta) \prod_{i=1}^{N} \sum_{r \in \mathcal{R}} \pi(Y_i \mid r, G_i, X_i, \Theta) \hat{P}_{ir}. \tag{4.4}$$

This decouples the total number of parameters from the sample size. Additionally, Equation (4.4) has only continuous parameters, and so can be used with any general Bayesian inference procedure such as Markov chain Monte Carlo (MCMC). However, the moderate-to-high dimensionality in practical settings, even after integrating out $\mathbf{R}$, makes MCMC algorithms computationally too expensive outside of relatively small datasets.

Large samples: Expectation-Maximization. When the number of individuals exceeds a thousand or so, we propose an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) to calculate the maximum *a posteriori* (MAP) estimate of Θ for BIRDiE models. The EM algorithm alternates between an E-step which calculates the expected log posterior density Q, averaging over the missing $\mathbf{R}$, and an M-step which maximizes Q over values of Θ.

Specifically, given a current parameter estimate $\Theta^{(t)}$, the expected log posterior density can be written as

$$\begin{aligned}
Q(\Theta^{(t+1)} \mid \Theta^{(t)}) &= \mathbb{E}_t[\log \pi(\Theta^{(t+1)}, \mathbf{R} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})] \\
&= \sum_{\mathbf{r} \in \mathcal{R}^N} \pi(\mathbf{r} \mid \Theta^{(t)}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \log \pi(\Theta^{(t+1)}, \mathbf{r} \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S}) \\
&= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \Big\{ \Big(\log \pi(Y_i \mid r, G_i, , X_i, \Theta^{(t+1)}) + \log \hat{P}_{ir} \Big) \\
&\qquad \times \pi(R_i = r \mid \Theta^{(t)}, Y_i, G_i, X_i, S_i) \Big\} \\
&= C + \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \tilde{P}^{(t)}_{ir|Y} \log \pi(Y_i \mid r, G_i, X_i, \Theta^{(t+1)}),
\end{aligned}$$

where C is a constant that can be ignored as it does not depend on $\Theta^{(t+1)}$, and $\tilde{\mathbf{P}}^{(t)}_{|Y} = \pi(\mathbf{R} \mid \Theta^{(t)}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ are the BISG probabilities updated using Bayes' rule:

$$\tilde{P}^{(t)}_{ir|Y} = \frac{\pi(Y_i \mid r, G_i, , X_i, \Theta^{(t)}) \hat{P}_{ir}}{\sum_{r' \in \mathcal{R}} \pi(Y_i \mid r', G_i, , X_i, \Theta^{(t)}) \hat{P}_{ir'}}. \tag{4.5}$$

At the M-step, $Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ is straightforward to maximize, since it is just the log complete-data posterior, with likelihood weights given by the $\tilde{P}^{(t)}_{ir|Y}$. Additionally, if Θ can be partitioned into parameters which only affect individuals in each racial group (as is the case with all the models described in Section 4.3.2 above), then the maximization can be performed separately on each group of individuals.

A critical advantage of this EM scheme over working directly with the marginal likelihood is that the maximization in the M-step can be performed using sufficient statistics calculated as part of the E-step, rather than on all of the individual entries in the data. Since the M-step is usually the practical (if not also asymptotic) bottleneck in the computation, this is enormously helpful—the problem size scales with $|\mathcal{Y}| \times |\mathcal{X}| \times |\mathcal{G}|$ rather than with N. Specifically, notice that we can rewrite $Q(\Theta^{(t+1)} \mid \Theta^{(t)})$ (dropping

the unnecessary constant) as

$$\begin{aligned}\mathcal{Q}(\Theta^{(t+1)} \mid \Theta^{(t)}) &= \log \pi(\Theta^{(t+1)}) + \sum_{i=1}^{N} \sum_{r \in \mathcal{R}} \tilde{P}^{(t)}_{ir|Y} \log \pi(Y_i \mid r, G_i, X_i, \Theta^{(t+1)}) \\ &= \log \pi(\Theta^{(t+1)}) + \sum_{r \in \mathcal{R}} \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \sum_{g \in \mathcal{G}} \log \pi(y \mid r, g, x, \Theta^{(t+1)}) \left(\sum_{i \in \mathcal{I}(yxg)} \tilde{P}^{(t)}_{ir|Y} \right),\end{aligned}$$

where analogously to above $\mathcal{I}(yxg)$ is the set of individuals with $Y_i = y$, $X_i = x$, and $G_i = g$.

For BIRDiE models where the complete-data likelihood is conjugate to the prior, such as the complete- and no-pooling models, these sufficient statistics are used in the M-step anyway, and can be efficiently calculated during the E-step as well. In combination with the acceleration scheme described next, this allows the entire EM algorithm to be run to convergence on data with hundreds of thousands or millions of individuals in a matter of seconds.

While EM algorithms are highly stable, due to their monotonic increasing of the marginal likelihood, they are also often slow to converge (Laird, 1993). To address this, we propose, and include in our open-source software implementation, the use of fixed-point iteration accelerators such as Anderson acceleration or SQUAREM (Varadhan and Roland, 2008). These techniques can substantially reduce the overall computational time without meaningfully affecting the stability of inference.

Finally, it is important to note that the EM algorithm does not provide any uncertainty quantification. However, in large samples, sampling and model-based uncertainty is dominated by biases caused by even small violations of the underlying assumptions, a problem we discuss below in Section 4.3.6 and the accompanying appendix. For BIRDiE models with few parameters, such as the complete-pooling model, bootstrapping is computationally feasible and can be used to approximate the asymptotic covariance matrix of the MAP estimate.

4.3.4 Updated Individual Race Probabilities

The EM algorithm produces the updated individual race probabilities given in Equation (4.5). One feature of these updated probabilities is that it is appropriate to apply the weighting estimator to them to estimate disparities. This is because the updated probabilities condition on $\mathbf{Y}$, and so the asymptotic bias term in Theorem 4.2 becomes zero. In fact, the weighting estimate from to the updated probabilities is identical to the BIRDiE estimate. While more study is required, for downstream settings where weights are needed, generating these weights with BISG followed by BIRDiE will likely produce more accurate results than simply using BISG weights alone.

4.3.5 Additional Explanatory Variables

Often, researchers are interested in not just $\mathbb{P}(Y \mid R)$ but also $\mathbb{P}(Y \mid W, R)$, for some variable $W \in \mathcal{W}$ which is not part of the BISG predictors (X, G). For example, a lending firm auditing potential racial disparities in lending decisions would likely be interested both in how the rate of loan approval (Y) varies by race, but also how loan approval varies by race, conditional on a measure of creditworthiness (W). The unconditional disparities reflect realities of systemic racism and inequity, while the conditional disparities measure the fairness of the firm's lending decisions after controlling for these systemic factors. Such estimates could be used to compute various measures of algorithmic fairness, including calibration parity and false positive error rate balance. Another scenario is a policy evaluation study, where researchers are interested in how the impact of policy varies across racial groups. Such an anlysis requires incorporating an interaction between race and the treatment variable.

There are two main ways to perform such an analysis with our proposed methodology. The first, and perhaps simplest, is to apply the methodology to the combined variable $\underline{YW} \in \mathcal{Y} \times \mathcal{W}$. This will produce

estimates of $\mathbb{P}(Y, W \mid R)$, from which $\mathbb{P}(Y \mid W, R)$ can be straightforwardly calculated by appropriate normalization. This approach will work well if $|\mathcal{Y}|$ and $|\mathcal{W}|$ are both small, so that $|\mathcal{Y} \times \mathcal{W}|$ is of manageable size. If one of these variables has many levels, however, directly estimating the distribution of $Y, W \mid R$ could be less efficient, as it does not account for information about the marginal distributions $Y \mid R$ and $W \mid R$.

An alternative approach is to first apply the proposed methodology to estimate $\mathbb{P}(W \mid R)$. This allows for calculation of model-updated BISG probabilities $\tilde{\mathbf{P}}_{|W} = \pi(\mathbf{R} \mid \hat{\Theta}, \mathbf{W}, \mathbf{G}, \mathbf{X}, \mathbf{S})$, which are also computed as a byproduct of the EM algorithm described above. Then, the methodology can be applied again, using $\tilde{\mathbf{P}}_{|W}$ as the input probabilities rather than the original BISG probabilities, to estimate $\mathbb{P}(Y \mid W, R)$. This approach will likely perform better when W consists of multiple predictors or if either $|\mathcal{Y}|$ or $|\mathcal{W}|$ are large.

Both of these approaches require the following identifying assumption, which generalizes Assumption A4.

Assumption A5 (Conditional independence of outcome, predictor and name)**.** *For all i,* $(Y_i, W_i) \perp\!\!\!\perp S_i \mid R_i, G_i, X_i$*, or, equivalently,*

$$W_i \perp\!\!\!\perp S_i \mid R_i, G_i, X_i \quad \textit{and} \quad Y_i \perp\!\!\!\perp S_i \mid W_i, R_i, G_i, X_i.$$

In the lending example, this would translate to the assumption that a measure of creditworthiness is independent of last name after controlling for race, location, and covariates, and that lending decisions are independent of last names after controlling for creditworthiness, race, location, and covariates.

4.3.6 Addressing the Potential Violations of the Assumptions

BIRDiE crucially relies on Assumption A4 for identification. In addition, like the weighting and thresholding estimators, it also relies upon Assumptions A1 and A2, which are required for the BISG race probabilities to be accurate. Unfortunately, these assumptions may not exactly hold in practice, and are also not testable in observed data. In this section and Appendix D.3, we develop sensitivity analyses that assess how violations of these assumptions affect the estimates of racial disparities.

First, BIRDiE assumes that conditional on unobserved race and observed covariates, outcomes and surnames are independent. As discussed in Section 4.3.1, however, association between the outcome and country of origin or racial subgroups may lead to correlation between surnames and outcome even after controlling for race and geography. To address this problem, suppose that a low-dimensional summary statistic of surname, $f : \mathcal{S} \to \mathbb{R}^d$, $d \ll |\mathcal{S}|$, is available, where f may map each surname to a finer ethnic group within each racial category. For example, Imai is a Japanese name whereas McCartan is a name of Irish origin. If f can classify surnames into finer racial subgroups or countries of origin—even approximately—then it can be used to control for this channel of possible violations of Assumption A4. Formally, we relax Assumption A4 as follows.

Assumption A6 (Partial conditional independence of outcome and name)**.** *For all i,*

$$Y_i \perp\!\!\!\perp S_i \mid f(S_i), R_i, G_i, X_i.$$

The next theorem shows that it is still possible to nonparametrically identify racial disparities under Assumption A6 under the identification condition, which is only slightly stronger than for Theorem 4.3.

Theorem 4.6 (Nonparametric Identification Under Assumption A6)**.** *Let* $f : \mathcal{S} \to \mathbb{R}^d$, $d < |\mathcal{S}|$, *with*

range $f(\mathcal{S})$. *For any given* $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, *and* $y \in \mathcal{Y}$, *define a matrix* $\mathbf{P}$ *with entries* $p_{rs} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ *and a vector* $\mathbf{b}$ *with entries* $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. *Then under Assumption A6, and assuming knowledge of the joint distribution* $\mathbb{P}(R, G, X, S)$, *the conditional probabilities* $\mathbb{P}(Y = y \mid R, f(S) = z, G = g, X = x)$ *are identified if and only if both* $\mathbf{P}$ *and the augmented matrix* $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$ *have rank* $|\mathcal{R}|$.

As long as the dimension d of the surname summary statistic $f(S)$ is much smaller than the (usually large) number of surnames $|\mathcal{S}|$, racial disparities are likely to be identified under Theorem 4.6 when they are already identified under Theorem 4.3. Thus, Assumption A6 and Theorem 4.6 can be used in conjunction with carefully chosen f in order to probe likely failure modes of the more restrictive Assumption A4. If estimates are not much affected by the inclusion of $f(S)$, then researchers can be more confident in the plausibility of Assumption A4.

Second, bias can also arise from violations of the assumptions underlying the BISG methodology (Assumptions A1 and A2). Of course, this is not unique to the proposed methodology: violations of these assumptions will also affect the validity of other disparity estimators such as weighting or thresholding. However, since as discussed above the BISG assumptions may rarely hold exactly in practice, we provide in Appendix D.3 several results characterizing how the model's estimates are affected by bias in the BISG probabilities.

4.4 Empirical Validation

To better understand how BIRDiE performs in real-world contexts, we apply it to North Carolina voter registration data. This voter file contains individual-level self-reported race for almost all voters and hence the "ground truth" relationship between outcome and race is known. We compare the performance of

BIRDiE models against those of the weighting and thresholding estimators. We also evaluate how the estimation error depends on the level of geographic precision used in the BISG probabilities. Finally, we briefly demonstrate various extensions of the BIRDiE methodology: small-area estimates, improved individual race predictions, estimation conditional on an additional explanatory variable (as discussed in Section 4.3.5), and sensitivity analysis for potential assumption violation (Section 4.3.6).

4.4.1 North Carolina Voter File

Like most other Southern states, which have a history of disenfranchising minority voters, the state of North Carolina asks (and previously required) every voter to self-report their race upon registration. This data, along with voters' names, addresses, gender, party registration (if any), and voting history, is part of the voter file that the secretary of state makes publicly available. This feature makes the voter file an ideal validation setting for the proposed methodology. The outcome we examine here, party registration, is the product of many unobservable factors, and is known to differ across racial groups. Since self-reported race is available, inferences about these racial disparities using the estimators discussed here can be compared to the corresponding ground truth.

Estimation of party registration by race is of substantive interest as well, especially in the context of the Voting Rights Act of 1965 (VRA). The relationship between these variables is critical for understanding the impact of policy changes such as redistricting or election rules on compliance with the VRA, and for establishing legal standing to challenge these policies under the VRA. As many states do not ask for self-reported race during voter registration, and other states no longer require reporting race, methods like BIRDiE are important tools for evaluating VRA compliance.

We use a subset of the October 2022 voter file which could be linked to a proprietary voter file provided

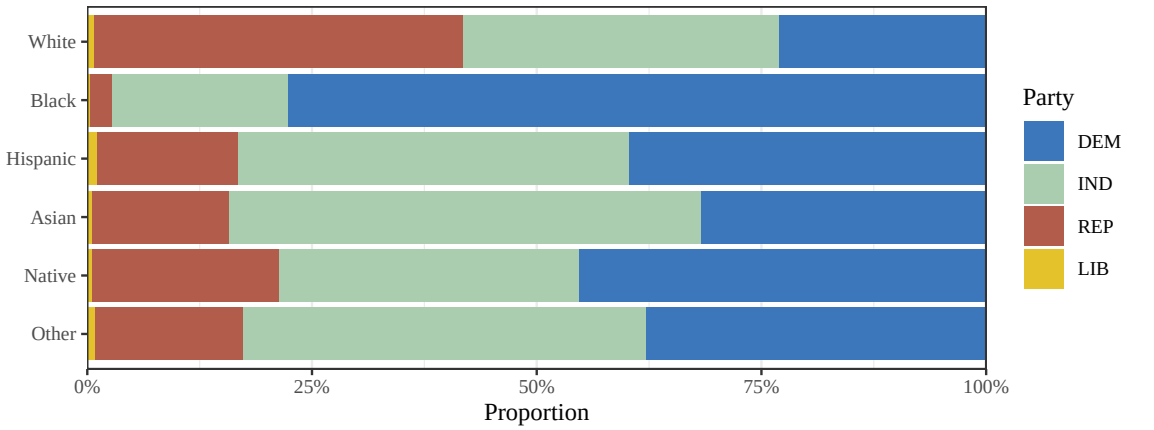


Figure 4.2: Distribution of party registration by race for a sample of 1,000,000 North Carolina voters. Parties are Libertarian (LIB), Republican (REP), Independent (IND), and Democratic (DEM).

by L2, Inc., a leading national non-partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters, and consultants for use in campaigns. The L2 file geocoded each address to a Census block, which allows for the finest block-level BISG predictions. We also removed any records without individual race information, since our goal is validation compared to some ground truth, rather than inference about the entire population of registered North Carolina voters. Altogether, 22.1% of records either had missing race information or could not be linked to the L2 file.

The overall merged voter file contains 5,754,912 voters, 71.1% of which are White, 21.1% of which are Black, and 7.8% of which belong to another race. To reduce computational burden, we further subsampled this file by selecting 1,000,000 records at random without replacement. This sample size is large enough to ensure that sampling error in the estimates is negligibly small.

Figure 4.2 shows the distribution of party registration by self-reported race in this subsample. White vot-

ers disproportionately register as Republicans, while Black voters disproportionately register as Democrats. This serves as the ground-truth in our validation analysis presented below.

4.4.2 The Model Setup

We first calculate BISG probabilities using 2010 Census data at the census block, tract, ZIP code tabulation area (ZCTA), and county level. Every record in the voter file contains county information, while roughly 13% of records are missing ZIP codes and 27% of records are missing blocks/tracts; when these finer geographic identifiers were missing, we used county-level Census tables in the BISG calculations.

The BISG probabilities are broadly accurate. Using the maximum a posteriori racial category as a prediction, we obtain the accuracy of 76.2% for the county probabilities, 78.4% for the ZIP code probabilities, 78.5% for the tract probabilities, and 79.6% for the block probabilities. An alternative measure of the quality of the BISG probabilities is the logarithmic score, a proper scoring rule which rewards precise and calibrated probabilistic estimates (higher values are better). The logarithmic scores for the BISG probabilities are −0.618 for counties, −0.587 for ZIP codes, −0.585 for tracts, and −0.607 for blocks. For comparison, the prior-only logarithmic score (i.e., using no name or geographic information) is −0.867. The worsened performance for block-level versus tract-level probabilities likely stems from the larger impact of census measurement error at smaller geographies, a problem that could be addressed using newer BISG methods such as those of Imai et al. (2022).

For a given set of BISG probabilities, we estimate the conditional distribution of each outcome variable given race using BIRDiE with both saturated pooling and multinomial mixed-effects models introduced above. We then compare the resulting estimates based on these BIRDiE models against those of the two existing estimators — the weighting estimator as well as a thresholding estimator that deterministically

assigns each individual the maximum *a posteriori* racial category. For the saturated BIRDiE model, we use geographic effects matching the geographic level used in the BISG probabilities (e.g., county effects for the county-level BISG probabilities), except for the block-level probabilities. Due to the large number of individual census blocks, we use tract-level effects instead for this particular model.

We use our software's default priors for both BIRDiE models. This is a uniform prior for the saturated model and a $\mathcal{N}(0, 1)$ prior on $\boldsymbol{\beta}$ and a Gamma$(2, 10)$ prior on the random intercept standard deviation for the mixed model. With standardized covariates, the prior on the fixed effects $\boldsymbol{\beta}$ is weakly informative, corresponding to a prior belief that a one-standard-deviation increase in the covariate level will generally correspond to a change in the linear predictor between -2 and 2. The weakly informative prior on the random intercept scale is centered at 0.2 and places most of its mass between 0.024 and 0.56. To give an idea of the computational efficiency of the proposed method, the maximum runtime of the saturated of the BIRDiE models fit to estimate party registration was 6.9 seconds (on a modern laptop with 8GB RAM), and the maximum runtime for the mixed model estimates was 22.6 seconds.

4.4.3 Estimates of Racial Disparity in Party Registration

We first examine the relative accuracy of the proposed methods in estimating the disparity between White and Black, and White and Hispanic voters, in party registration. For example, the true difference in Democratic registration between Black and White voters in the sample is 54.6 percentage points, meaning Black voters register Democratic at a much higher rate. However, the standard weighting approach produces an estimate of only 20.8 percentage points for this disparity—less than half the true value. The thresholding estimator, while slightly better, also misses the mark, with an estimate of 30.5 percentage points. In con-

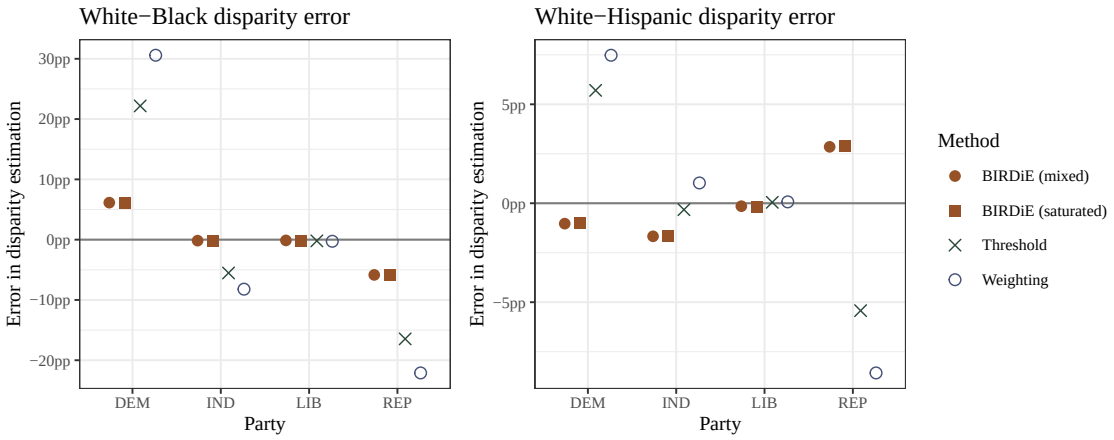


Figure 4.3: Error in the White-Black and White-Hispanic disparity estimates for party registration, by estimation method. All methods used block-level data for this figure; the results for other levels of geographic detail are generally similar. Estimation uncertainty is minimal and hence suppressed from the figure for clarity.

trast, the saturated BIRDiE model produces an estimate of 49.2 percentage points, and the mixed BIRDiE model also estimates 49.2. These estimates are only slightly lower than the ground truth.

Figure 4.3 compares the empirical performance of the BIRDiE models against that of the weighting and thresholding estimators across all of these possible disparity measurements, using the block-level BISG predictions. The true disparity is subtracted off from each estimate for ease of comparison, and so the figure plots the estimation error of racial disparity between two racial groups. For White–Black (left plot) and White-Hispanic (right plot) disparities in party registration, the BIRDiE models (solid circles and squares) substantially outperform the two commonly used estimators (open circles and crosses). For two major parties, both the weighting and thresholding estimators exhibit a substantial amount of estimation error, for example, exceeding 20 percentage points for the White-Black disparity for the Democratic party. In contrast, the two BIRDiE models yield a much smaller estimation error that ranges within several percentage

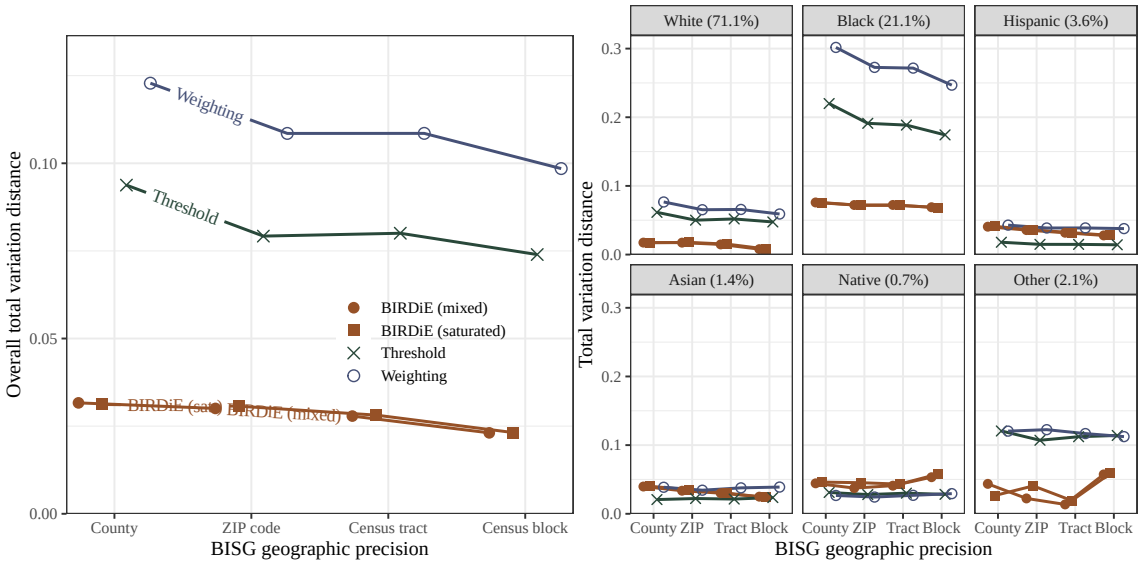


Figure 4.4: Total variation distance between the estimated and actual distribution of party registration, by estimation method and level of geographic detail used in the BISG predictions. The left plot shows the overall total variation distance, whlie the right plot decomposes it by racial group.

points for all racial disparity estimates. The saturated and mixed-effects BIRDiE models perform similarly with no discernible difference.

For a more comprehensive look at the error in the estimated partisanship-by-race distributions, we turn to the total variation (TV) distance, which is calculated as

$$d_{\text{TV}}(\hat{\boldsymbol{\mu}}_{Y|R}, \boldsymbol{\mu}_{Y|R}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} \sum_{r \in \mathcal{R}} |\hat{\mu}_{Y,R}(y, r) - \mu_{Y,R}(y, r)|,$$

where $\mu_{Y,R}$ is the joint distribution of Y and R. The TV distance is an upper bound on the error in *any* probability calculated from the estimated joint distribution, and as such is useful general-purpose measure of estimation error. The left plot of Figure 4.4 shows the TV distance for each estimator, not just for the block-level BISG estimates used in Figure 4.3 but also across the range of geographic levels used in the BISG calculation (x-axis). We find that both BIRDiE models substantially outperform every alternative at every

geographic level. In general, the estimates based on the BIRDiE models exhibit a total variation distance whose magnitude is about one third and one fourth of that for the thresholding and weighting estimators, respectively. As before, the saturated and mixed-effects BIRDiE models perform similarly. According to this measure, we find that finer geographic data provide only minor improvements in accuracy for the BIRDiE or conventional estimates. While possibly counterintuitive, this finding underscores the fact that calibrated BISG probabilities, rather than highly precise probabilities, are all that is needed for accurate disparity estimation.

It is also possible to measure the TV distance for each conditional distribution by race:

$$d_{\text{TV}}^{(r)}(\hat{\boldsymbol{\mu}}_{Y|r}, \boldsymbol{\mu}_{Y|r}) = \frac{1}{2} \sum_{y \in \mathcal{Y}} |\hat{\mu}_{Y,r}(y, r) - \mu_{Y,r}(y, r)|.$$

The right plot of Figure 4.4 shows these within-race TV distances, to illuminate how the estimators perform on each subgroup. In general, the BIRDiE models are more accurate than the weighting and thresholding estimators for the White and Black racial groups which together make up 96% of the sample. All estimators perform roughly equally well for Hispanic, Asian, and Native voters (though the thresholding estimator performs well for Hispanic voters), exhibiting relatively small estimation error. The weighting and thresholding estimators perform particularly poorly for Black voters and for the "Other" racial group.

4.4.4 Small-area Estimation

An advantage of the BIRDiE methodology is its explicit modeling of $\mathbb{P}(Y \mid R, G, X)$, which produces not only estimates of the marginal $\mathbb{P}(Y \mid R)$ but also subgroup estimates of how these conditional distributions vary across covariates and geographic areas. This section examines the accuracy of the saturated and mixed-effects BIRDiE models in recovering small-area relationships between party registration and race,

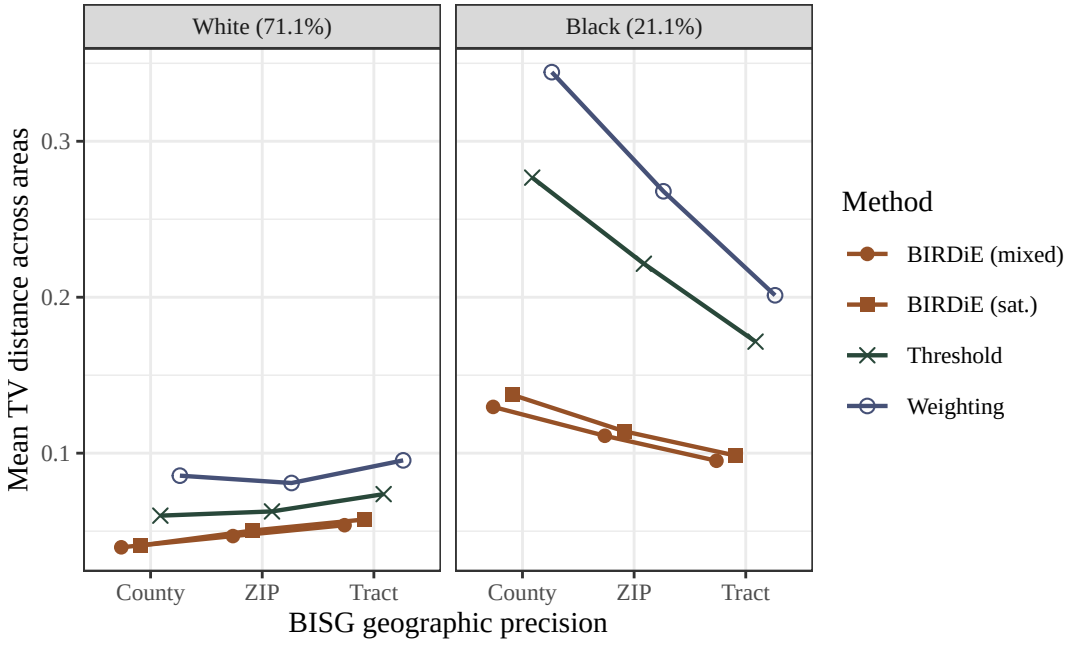


Figure 4.5: Accuracy of small-area estimates by race, as measured by the average total variation distance.

compared to standard methodology that simply applies the weighting and thresholding estimators within each geographic area. We study accuracy at the county, ZIP code, and tract level, using the BISG probabilities and BIRDiE models that were applied to each. Since in fitting the BIRDiE model to block-level BISG probabilities we used tract-level random intercepts, we do not present block-level estimates.

We evaluate the small-area estimates by calculating the mean total variation distance between the estimated and true conditional distributions of party registration by race (averaging across geographic areas). Figure 4.5 summarizes our results, which qualitatively track the patterns found overall in Figure 4.4. The two BIRDiE models exhibit substantially lower error than the weighting and thresholding estimators. Across all methods, the error is lower for White voters, who make up the bulk of the sample. Somewhat surprisingly, the amount of error does not appear to vary much for the BIRDiE models across different levels of geography—tract-level estimates are roughly as accurate as county-level estimates, on average.

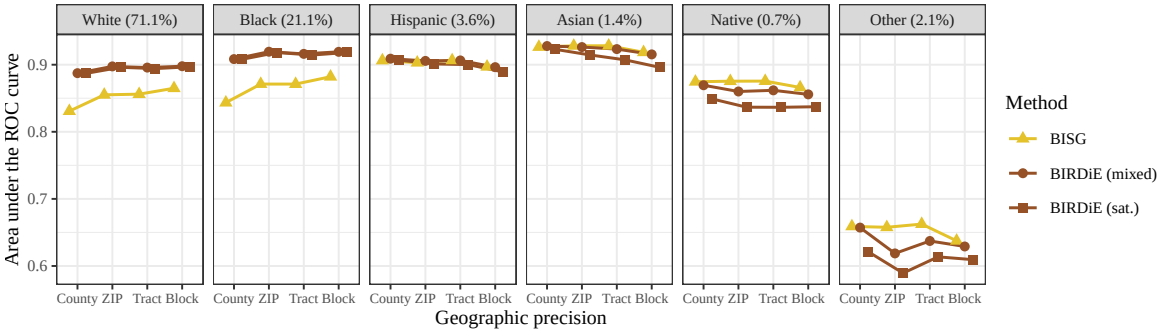


Figure 4.6: Race probability predictive accuracy, as measured by the area under the receiver operating characteristic (ROC) curve, for the input BISG probabilities as well as the BIRDiE-updated probabilities, by race and level of geographic precision. Larger values indicate more precise predicions.

Between the BIRDiE models, the mixed-effects model slightly outperforms the saturated model. This reflects the value in partially pooling estimates through the random effect structure.

4.4.5 Improved Individual Race Probabilities

As discussed in Section 4.3.4, we can use the conditional distribution $\mathbb{P}(Y \mid R, X, G)$ estimated with a BIRDiE model to create model-updated BISG probabilities $\tilde{\mathbf{P}}_{|Y} = \pi(\mathbf{R} \mid \hat{\Theta}, \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})$ by applying Bayes' rule. These updated probabilities may be more accurate than the original BISG probabilities.

For example, using the estimates produced by the mixed BIRDiE model applied to party registration with block-level BISG estimates, the MAP prediction accuracy increases from 79.6% with the input probabilities to 82.9% with the updated probabilities. These increases are significantly larger than the differences in accuracy between BISG probabilities using different levels of geographic precision.

The improvements are reflected in other accuracy measures as well. Figure 4.6 shows the accuracy of the predictions by race, as measured by the area under the receiver operating characteristic (ROC) curve. The updated probabilities are noticeably more accurate than the input probabilities for White and Black voters, about as accurate for Hispanic, Asian, and Native voters, and slightly less accurate for "Other" voters.

Since White and Black voters together constitute the vast majority of the sample, these patterns translate to improvement in the logarithmic scores as well: the score for the input probabilities is −0.607, and this increases to −0.570 with the updated probabilities.

4.4.6 Estimates Conditional on an Additional Variable

The North Carolina voter file also provides an opportunity to demonstrate the methodology described in Section 4.3.5 to produce estimates conditional on another predictor variable that is not used in the BISG probabilities. We will estimate party registration rates by race among voters and nonvoters in the 2020 election. Following the discussion in Section 4.3.5, we will compute these estimates two ways: (1) by estimating party registration and 2020 turnout jointly by race, and (2) by first estimating 2020 turnout by race, then estimating party registration by race and 2020 turnout.

We will use a multinomial mixed-effects BIRDiE model applied to the block-level BISG probabilities, with random intercepts by tracts, for all the estimation. Fitting this model to a combined 2018/2020 turnout variable (i.e., one with four levels: no/no, no/yes, and so on) produces estimates of the joint distribution of 2018 and 2020 turnout by race. Normalizing these probabilities within 2018 turnout and race groups produces estimates of 2020 turnout by race and 2018 turnout. The total variation distance between these estimates and the true distribution is 0.051, which indicates close agreement.

Figure 4.7 presents the errors in the estimates for this method (solid square), the two-step method (solid circle), and for the weighting (open circle) and thresholding (cross) estimators. The TV distance between these two-stage estimates and the true distribution is 0.052, very similar to the error in the joint-estimation approach. The two-step approach produces similar estimates to the joint-estimation approach, though the latter performs better in the "Other" category. As discussed in Section 4.3.5, while both approaches

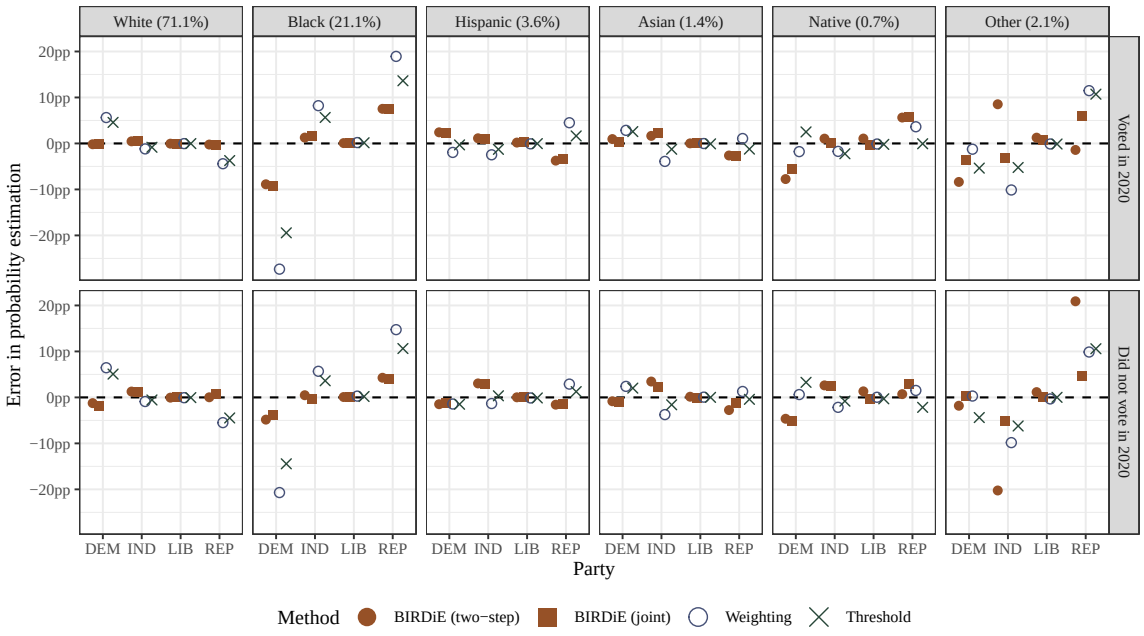


Figure 4.7: Error in the conditional expectation estimates for party registration by turnout and race, by estimation method. All methods used block-level data for this figure. Estimation uncertainty is minimal and hence suppressed from the figure for clarity.

produce highly accurate estimates in this example, we would expect the two-stage approach to be superior when one or both of the variables has more levels.

In contrast to the BIRDiE models, the weighting and thresholding estimates of party registration by 2020 turnout and race include large errors, especially for Black voters with all the errors exceeding 10 percentage points. The TV distance for the weighting estimator is 0.196, and the distance for the thresholding estimator is 0.147—around 3–4 times higher than for the estimates based on the BIRDiE models.

4.4.7 Sensitivity Analysis

Finally, we examine the sensitivity of our party registration estimates to potential violations of the key identifying Assumption A4, following the method outlined in Section 4.3.6 that is based on a low-dimensional summary statistic of surnames. We use a publicly available sample of 5% of the individual records for the 1930 Census (Ruggles et al., 2021), which contains individual names, individual and parental birthplace, and detailed race, ethnicity, and tribal codes. Since many regions of Asia, particularly Vietnam, experienced little emigration to the United States before 1930, we further supplement this data with around 3,000 Asian surnames classified into six regional subgroups: Chinese, Filipino, Indian, Japanese, Korean, Vietnamese, NHPI, and Other (Guage et al., 2023).

Using these subgroups and the 1930 birthplace and racial data, we can classify most surnames in the voter file into nine groups (see Appendix D.4 for a brief description of the groupings and the most common 50 surnames for each group). While somewhat arbitrary, these groups are chosen to combine countries of origin which had significant immigration to the United States during similar periods.

We first evaluate the plausibility of Assumption A4 by examining the correlation between the residuals of the BIRDiE model fit and indicator variables for each of the nine surname groups. Under Assumption

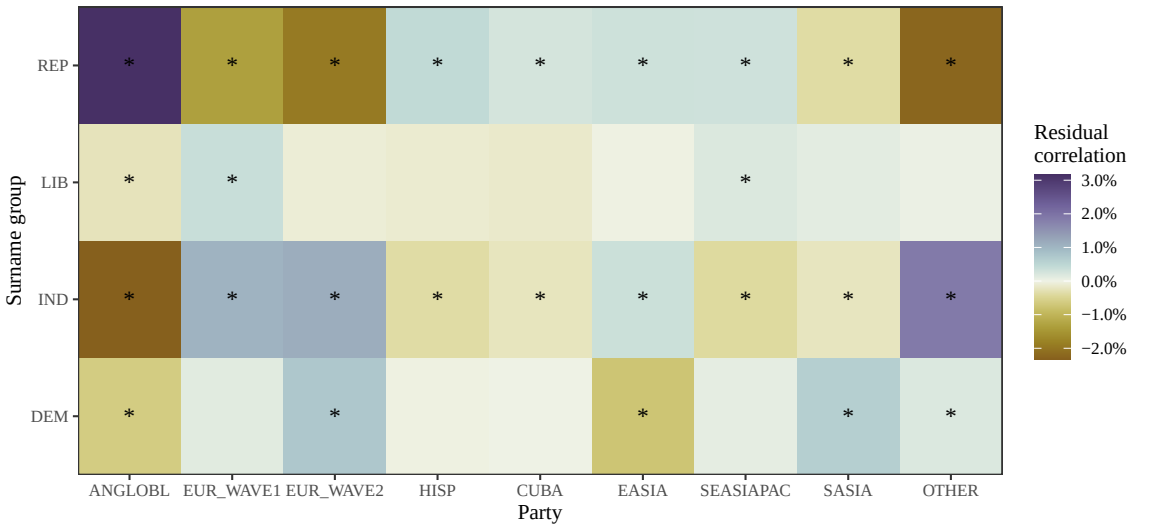


Figure 4.8: Residual correlation between party registration and nine surname groups, after controlling for race and location. Correlations whose 90% confidence intervals exclude zero are marked with an asterisk. See Appendix D.4 for details on the surname groups.

A4, this residual correlation should be zero everywhere. As Figure 4.8 shows, however, for many groups and party labels, the correlation is small but deviates from zero more than would be expected given only sampling variation. Here, we use the residuals from the county-level saturated model specification, but the results are not sensitive to this choice.

Notably, voters with names in the Anlglosphere and Black surname group, which includes surnames that are relatively more common among many-generation residents of the U.S., such as Smith, Williams, and Brown, are significantly less likely to register as Democrats and independents, and more likely to identify as Republicans, even after controlling for race and geography. Meanwhile, voters with names in the First and Second wave European immigration surname groups, which include surnames more common among 19th and 20th century immigrants from Europe, display the opposite pattern. Differences among surname groups designed to correlate with membership in various Asian subgroups are also visible. How-

ever, all of these correlations are quite small in magnitude, with most on the order of 0.01 or so. Thus, we expect our party-by-race estimates to be little affected by the inclusion of the surname group indicators.

Indeed, re-fitting the county-level saturated model with an additional surname group covariate produces nearly identical results, albeit at a moderately higher computational cost given the increased number of $(G, X, f(S))$ cells. This re-fitting requires Assumption A6, which relaxes Assumption A4. We find that the average party registration rate estimate changes only by 0.006, with the largest change being 0.021, for the rate of Democratic registration among Other voters. All in all, this analysis provides confidence that violations of Assumption A4 for the North Carolina voter file are likely minor and would have minimal effects on quantitaive and qualitative conclusions.

4.5 Discussion

We have introduced a new identifying assumption and accompanying model, BIRDiE, and clarified other assumptions implicit in approaches to disparity estimation from proxy data. In many real-world applications, we believe the new model and identification condition are appropriate and will produce significantly improved estimates. However, there is no one-size-fits-all approach for the estimation of disparities. For example, the existence of name-based discrimination may violate our identification assumption especially when racial categories, for which data are available, are coarse. Careful consideration of the underlying causal and information structure is required to avoid making the incorrect conclusions. We also provide a sensitivity analysis that partially addresses this concern.

As our empirical demonstration showed, in realistic settings BIRDiE can substantially outperform existing estimators of racial disparities, both in aggregate and for small areas. The BIRDiE methodology

also produces improved BISG probabilities, and can be used to estimate disparities conditional on other variables. These additional features should prove helpful in practical settings.

Much work remains to be done in accurately and reliably measuring racial disparities. First, the BISG probabilities themselves can be improved. One approach to doing so is given by Imai et al. (2022), which accounts for some Census measurement error while remaining computationally tractable. Beyond the BISG probabilities, further empirical analyses could determine useful additional variables to condition on, which could allow analysts to weaken the required assumption. Finally, the BIRDiE model could be extended to directly model more complex types of outcome variables.

Appendices

This page intentionally left blank.

A

Appendices for Chapter 1

A.1 Proofs of Propositions

Lemma 1.1. *The probability of splitting a* valid *new district G_i from an existing area $\widetilde{G}_{i-1}$ using Algorithm 1 with parameter $k_i \geq K_i$ is*

$$q(G_i \mid \widetilde{G}_{i-1}, \text{pop}(V_i) \in [P_i^-, P_i^+]) \;=\; \frac{\tau(G_i)\tau(\widetilde{G}_i)}{\tau(\widetilde{G}_{i-1})k_i}|\mathcal{C}(G_i, \widetilde{G}_i)|. \tag{1.4}$$

Proof. Any spanning tree can be decomposed into two other trees and an edge joining them. Let $T \cup e \cup T'$ denote the spanning tree obtained by joining two other spanning trees, T and T', with an edge e. Then Equation (1.3) can be written as

$$q(G_i \mid \widetilde{G}_{i-1}) = \sum_{\substack{T^{(1)} \in \mathcal{T}(G_i) \\ T^{(2)} \in \mathcal{T}(\widetilde{G}_i)}} \sum_{e \in \mathcal{C}(T^{(1)}, T^{(2)})} q(G_i \mid T^{(1)} \cup e \cup T^{(2)})\, \tau(\widetilde{G}_{i-1})^{-1}.$$

Now, $q(G_i \mid T^{(1)} \cup e \cup T^{(2)})$ is determined by whether whether $e^* = e$, i.e., if e is the edge selected to be

cut. If e has d_e in the top k_i (if it induces one of the best k_i balanced splits), then it has a $1/k_i$ probability of being selected in step (c) and cut. If d_e is not in the top k_i, then this probability is zero.

Everything written to this point holds regardless of whether G_i is a valid district (i.e., satisfies $\text{pop}(V_i) \in [P_i^-, P_i^+]$). From here onwards we will restrict our attention to valid districts only. Notice that the forward-looking bounds P_i^- and P_i^+ are stricter than merely ensuring $\text{dev}(G_i) \leq D$. That is, conditional on $\text{pop}(V_i) \in [P_i^-, P_i^+]$, we must also have $\text{dev}(G_i) \leq D$.

Therefore, if a sorted edge e_j *in any spanning tree* induces such a balanced partition, we must have $j \leq K_i$, where as in the main text K_i counts the maximum number of such edges across all possible spanning trees. Thus, so long as we set $k_i \geq K_i$, we will have $d_e \leq D$.

Furthermore, across all spanning trees $T^{(1)} \in \mathcal{T}(G_i)$ and $T^{(2)} \in \mathcal{T}(\widetilde{G}_i)$, and connecting edges $e \in E(T^{(1)}, T^{(2)})$, the value of d_e is constant, since removing e induces the same districting. Combining these two facts, we have, conditional on satisfying the bounds P_i^- and P_i^+,

$$q(e^* = e \mid T^{(1)} \cup e \cup T^{(2)}, \text{pop}(V_i) \in [P_i^-, P_i^+]) \;=\; k_i^{-1},$$

which does not depend on $T^{(1)}$, $T^{(2)}$, or e. We may therefore write the conditional sampling probability as

$$\begin{aligned} q(G_i \mid \widetilde{G}_{i-1}, \text{pop}(V_i) \in [P_i^-, P_i^+]) \;&=\; \sum_{\substack{T^{(1)} \in \mathcal{T}(G_i) \\ T^{(2)} \in \mathcal{T}(\widetilde{G}_i)}} \sum_{e \in \mathcal{C}(T^{(1)}, T^{(2)})} \frac{1}{k_i \tau(\widetilde{G}_{i-1})} \\ &=\; \frac{\tau(G_i)\tau(\widetilde{G}_i)}{\tau(\widetilde{G}_{i-1}) k_i} |\mathcal{C}(G_i, \widetilde{G}_i)|, \end{aligned} \tag{A.1}$$

where as in the main text we let $\mathcal{C}(G, H)$ represent the set of edges joining nodes in a subgraph G to nodes in a subgraph H. □

Theorem 1.2. *Let $\pi_S = \sum_{j=1}^{S} w^{(j)} \delta_{[\xi^{(j)}]}$ be the weighted particle approximation generated by Algorithm 2. Then for all measurable h on unlabeled plans, as $S \to \infty$,*

$$\sqrt{S}(\mathbb{E}_{\pi_S}[h([\xi])] - \mathbb{E}_{\pi}[h([\xi])]) \xrightarrow{d} \mathcal{N}(0, V_{SMC}(h)),$$

for some asymptotic variance $V_{SMC}(h)$.

The proof proceeds by showing that the weights in Algorithm 2 are of a form derived from an existing SMC algorithm with an established central limit theorem.

Proof. We can associate our target measure $\pi([\xi])$ on unlabeled redistricting plans with a corresponding measure on labeled plans

$$\tilde{\pi}(\xi) \coloneqq \psi([\xi])^{-1} \pi([\xi]),$$

so that the pushfoward measure obtained by mapping $\xi \mapsto [\xi]$ recovers π.

Our SMC algorithm will operate on labeled plans, targeting $\tilde{\pi}$, so that the resulting plans, when considered as representatives of their corresponding unlabeled plans, will be representative of π in the sense given by the theorem statement. Recall that a labeled redistricting plan ξ is just a tuple of graph partitions $(G_1, G_2, \ldots, G_n)$. We begin by extending $\tilde{\pi}(\xi)$ to a series of measures on partial plans,

$$\begin{aligned}
\tilde{\pi}_i(G_1, G_2, \ldots, G_i) &:\propto \prod_{j=1}^{i} \frac{\tau(G_j)^{\rho} \tau(\widetilde{G}_j)}{\tau(\widetilde{G}_{j-1})} \mathbf{1}_{\mathrm{pop}(V_j) \in [p_j^-, p_j^+]} \\
&\propto \tau(\widetilde{G}_j) \prod_{j=1}^{i} \tau(G_j)^{\rho} \mathbf{1}_{\mathrm{pop}(V_j) \in [p_j^-, p_j^+]},
\end{aligned}$$

for $1 \leq i \leq n-2$, and where we have simplified the telescoping product in the second equality. Recall that the $\widetilde{G}_i$ are determined completely by $G_1, G_2, \ldots, G_i$.

For $i = n-1$, the above definition would yield

$$\tilde{\pi}_{n-1}(G_1, G_2, \ldots, G_i) \propto \tau(\widetilde{G}_{n-1}) \prod_{j=1}^{n-1} \tau(G_j)^{\rho} \mathbf{1}_{\mathrm{pop}(V_j) \in [P_j^-, P_j^+]}, = \tau(\xi)\tau(G_n)^{1-\rho} \mathbf{1}_{\mathrm{dev}(\xi) \leq D},$$

which is close to but not quite the target measure. So we instead define $\pi_{n-1} := \tilde{\pi}$; i.e., we add the additional terms $\exp(-J(\xi))$ and $\psi([\xi])$ and adjust for $\tau(G_n)^{1-\rho}$.

With these partial-plan measures defined, notice that the incremental weight $w_i^{(j)}$ for partial plans with $1 \leq i \leq n-2$ and $\mathrm{pop}(V_j) \in [P_j^-, P_j^+]$ may be written as

$$\begin{aligned} w_i^{(j)} &= \tau(G_i^{(j)})^{\rho-1} \frac{k_i}{|\mathcal{C}(G_i^{(j)}, \widetilde{G}_i^{(j)})|} \\ &= \frac{\tau(G_i^{(j)})^{\rho} \tau(\widetilde{G}_i^{(j)})}{\tau(\widetilde{G}_{i-1}^{(j)})} \left(\frac{\tau(G_i^{(j)})\tau(\widetilde{G}_i^{(j)})}{\tau(\widetilde{G}_{i-1}^{(j)})} \frac{|\mathcal{C}(G_i^{(j)}, \widetilde{G}_i^{(j)})|}{k_i} \right)^{-1} \\ &= \frac{\tilde{\pi}_i(G_1, \ldots, G_i)}{\tilde{\pi}_{i-1}(G_1, \ldots, G_{i-1}) q(G_i \mid \widetilde{G}_{i-1}, \mathrm{pop}(V_i) \in [P_i^-, P_i^+])}. \end{aligned} \tag{A.2}$$

For the final weighting at split $i = n-1$, the incremental weight (i.e., not including the residual previous weights $\left(\prod_{i=1}^{n-2} w_i^{(j)}\right)^{1-\alpha}$ given by step (c) of Algorithm 2 is

$$\begin{aligned} &\exp\left(-J(\xi^{(j)})\right) w_{n-1}^{(j)} \psi([\xi])^{-1} \left(\tau(G_{n-1}^{(j)})\right)^{\rho-1} \\ &\qquad = \frac{\tilde{\pi}_{n-1}(G_1, \ldots, G_{n-1})}{\tilde{\pi}_{n-2}(G_1, \ldots, G_{n-2}) q(G_{n-1} \mid \widetilde{G}_{n-2}, \mathrm{pop}(V_{n-1}) \in [P_{n-1}^-, P_{n-1}^+])}, \end{aligned}$$

since this weight includes exactly the same additional terms as $\tilde{\pi}_{n-1}$ mentioned above. So in fact Equation (A.2) holds for all $1 \leq i \leq n-1$.

These incremental weights are precisely those of the SMC partial rejection control algorithm of Peters et al. (2012) (see also LeGland and Oudjane (2005)), with the weights set to zero for invalid samples and the partial rejection threshold set to the minimum possible nonzero weight. So after pushing forward to

unlabeled plans, we gain immediately the theorem proved in that work (its Equation 6), viz., that for all measurable h on unlabeled plans and as $S \to \infty$, we have

$$\sqrt{S}(\mathbb{E}_{\pi_S}[h([\xi])] - \mathbb{E}_{\pi}[h([\xi])]) \xrightarrow{d} \mathcal{N}(0, V_{\text{SMC}}(h)),$$

for asymptotic variance $V_{\text{SMC}}(h)$ given by Equation 7 of the same work. □

A.2 Additional Validation Example

This section reports the results of another validation study applied to a 50-precinct map taken from the state of Florida. As in Section 1.5, we use the efficient enumeration algorithm of Fifield et al. (2020b) to obtain all possible redistricting maps with three and four contiguous districts, and use these plans as a baseline to validate the proposed algorithm. The left plot of Figure A.1 below shows the validation map.

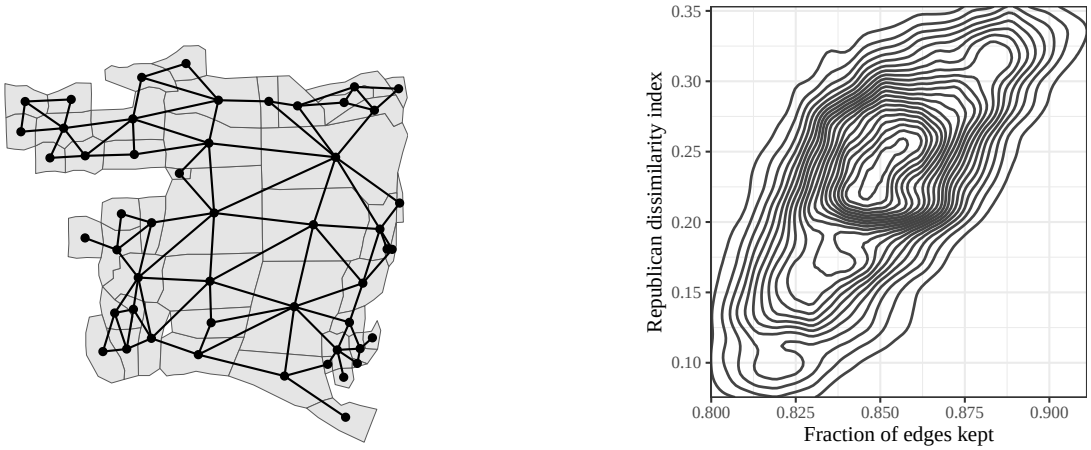


Figure A.1: The 50-precinct Florida map used for validation (left) and the joint distribution of Republican dissimilarity and compactness on the map over all partitions into four districts with $\text{dev}(\xi) \leq 0.10$ (right).

There are 112,515,494 partitions of the map into four districts, 33,635 of which have $\text{dev}(\xi) \leq 0.10$. We evaluate the accuracy of the proposed algorithm, and compare it to the same MCMC algorithm used

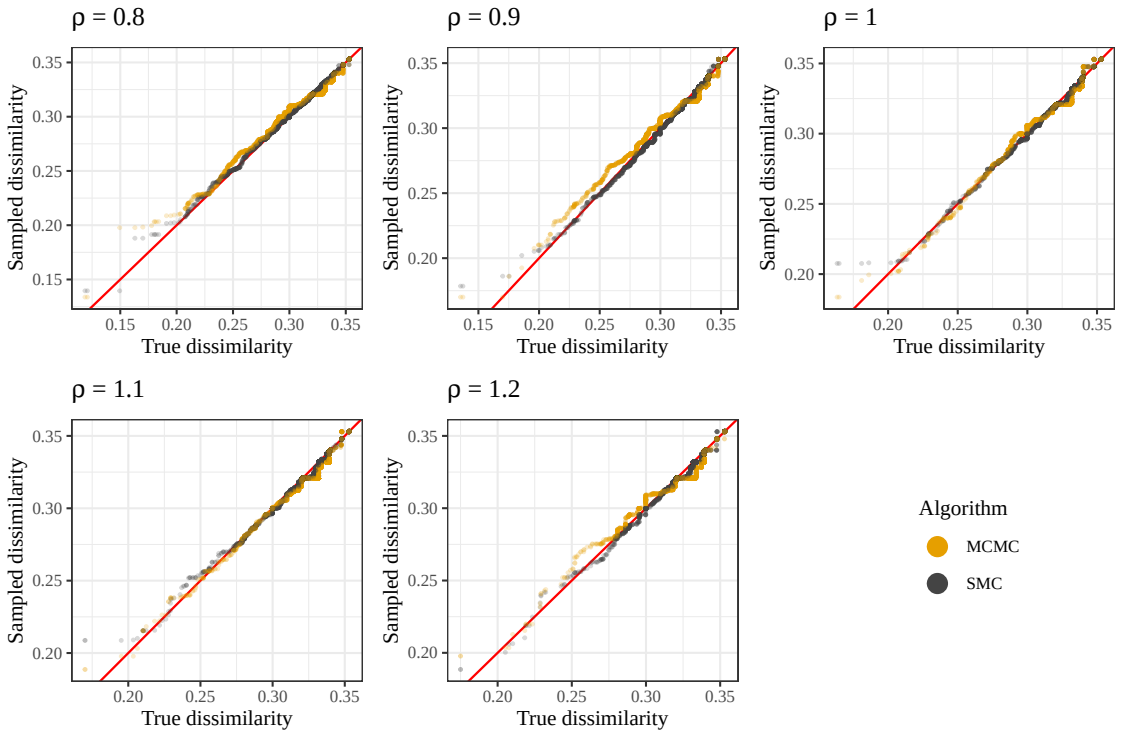


Figure A.2: Quantile-quantile plots for 1,500 MCMC and SMC samples of Republican dissimilarity across a range of target distributions with different compactness parameters ρ.

in Section 1.6, across a range of target distributions, with ρ running from 0.8 to 1.2 The right plot of Figure A.1 shows the joint distribution of compactness and the summary statistic we use for the validation, the Republican dissimilarity index (Massey and Denton, 1988). With this validation map, Republican dissimilarity is reasonably sensitive to the compactness of districts. This makes dissimilarity a good test statistic for comparing distributions that differ primarily in their average compactness.

For each target distribution, we sample 1,500 redistricting plans from both the SMC and MCMC algorithms, and repeat the sampling four times in order to produce $\hat{R}$ estimates. The MCMC algorithm is initialized with a random SMC-drawn map and first run for 500 warm-up iterations. To validate and

compare the samples, we reweight the enumerated redistricting plans by $\tau(\xi)^\rho$, and then produce quantile-quantile plots of the Republican dissimilarity index, which are shown in Figure A.2.

Across the range of ρ values, the agreement between the SMC sample and target distributions is excellent, even in the lower tail. The $\hat{R}$ values for the SMC algorithm were also all less than 1.003. The MCMC algorithm fares well in general but has noticeable bias for $\rho = 0.9$ and $\rho = 1.2$, and has some additional misses for the other target distributions as well, which manifest as small protuberances in the quantile-quantile plot. The $\hat{R}$ values for the MCMC algorithm were all less than 1.004, indicating that the overall location and scale of the Republican dissimilarity were estimated well, but $\hat{R}$ is not designed to capture these smaller-scale deviations from the target distribution.

Here and in Section 1.5, we validated and compared the SMC and MCMC algorithms with several summary statistics. This reflects the applied use case for redistricting analysis. But it is also informative to study how well the SMC algorithm can target the actual distribution of plans themselves. Of course, we cannot expect the algorithm to perform well in this regard if there are fewer samples then there are plans, which is the case in almost all real-world problems. So we subset the enumerated plans to those with $\text{dev}(\xi) \leq 0.01$, of which there are just 38. We first generate 10,000 samples from the SMC algorithm, targeting a distribution with $\rho = 1$. We measure the discrepancy between the sampled distribution of the 38 plans and the enumerated set weighted to $\tau(\xi)$ with the total variation distance, which is 0.0178 for this sample. Increasing the sample size to 50,000 decreases the total variation distance to 0.0149.

A.3 Algorithm Implementation Details

A.3.1 Estimating K_i

A natural approach to estimating K_i is to draw a moderate number of spanning trees $\mathcal{T}_i \subseteq \mathcal{T}(\widetilde{G}_i)$ and compute $ok(T)$ for each $T \in \mathcal{T}_i$. The sample maximum, or the sample maximum plus some small buffer amount, would then be an estimate of the true maximum $\widehat{K}_i$ and an appropriate choice of k_i. In practice, we find little noticeable loss in algorithmic accuracy even if $k_i < K_i$. The following proposition theoretically justifies this finding. As above, d_e represents the population deviation of the district induced by removing edge e from a spanning tree.

Proposition A.1. *The probability $q(e = e^* \mid \mathcal{F})$, i.e., the probability that an edge e is selected to be cut at iteration i, given that the tree T containing e has been drawn, and that e would induce a valid district, satisfies*

$$\max\left\{0, q(d_e \leq d_{e_{k_i}} \mid \mathcal{F})\left(1 + \frac{1}{k_i}\right) - 1\right\} \leq q(e = e^* \mid \mathcal{F}) \leq \frac{1}{k_i},$$

where $\mathcal{F}$ is the σ-field generated by $\{T, \text{pop}(V_i) \in [P_i^-, P_i^+]\}$.

Proof. We can write

$$q(e = e^* \mid \mathcal{F}) \;=\; q(e = e^*, d_e \leq d_{e_{k_i}} \mid \mathcal{F}) \;=\; \frac{1}{k_i} q(d_e \leq d_{e_{k_i}} \mid \mathcal{F}),$$

This holds because the edge e will not be cut unless $d_e \leq d_{e_{k_i}}$, i.e., if e is among the top k_i edges. We then have immediately that $q(e = e^* \mid \mathcal{F}) \leq k_i^{-1}$. Additionally, using the lower Fréchet inequality, we find the lower bound

$$q(e = e^* \mid \mathcal{F}) \;=\; q(e = e^*, d_e \leq d_{e_{k_i}} \mid \mathcal{F})$$

$$
\begin{aligned}
&\geq \max\left\{0, q(e = e^* \mid \mathcal{F}) + q(d_e \leq d_{e_{k_i}} \mid \mathcal{F}) - 1\right\} \\
&= \max\left\{0, \frac{1}{k_i} q(d_e \leq d_{e_{k_i}} \mid \mathcal{F}) + q(d_e \leq d_{e_{k_i}} \mid \mathcal{F}) - 1\right\} \\
&= \max\left\{0, q(d_e \leq d_{e_{k_i}} \mid \mathcal{F})\left(1 + \frac{1}{k_i}\right) - 1\right\}. \qquad \square
\end{aligned}
$$

If $k_i \geq K_i$, then $q(e = e^* \mid \mathcal{F})$ is exactly k_i^{-1}, a fact which is used in the proof of Proposition 1.1. This result, which is proved using a simple Fréchet bound, shows that as long as $q(d_e \leq d_{e_{k_i}} \mid \mathcal{F})$ is close to 1, using k_i^{-1} in Proposition 1.1 is a good approximation to the true sampling probability.

Having sampled $\mathcal{T}_i$, we can compute for each value of k the sample proportion of trees where a randomly selected edge e among the top k of edges of the tree is also among the top k for the other trees—in effect estimating $q(d_e \leq d_{e_{k_i}} \mid \mathcal{F})$. We may then choose k_i to be the smallest k for which this proportion exceeds a pre-set threshold (e.g., 0.99). We have found that this procedure, repeated at the beginning of each sampling stage, efficiently selects k_i without compromising the ability to sample from the target distribution.

A.3.2 Calculating $\psi([\xi])$

To calculate $\psi([\xi])$, we first observe that sequentially valid labelings of a particular unlabeled plan are in bijective correspondence with certain increasing sequences of connected subgraphs of the quotient graph G/ξ. Specifically, as noted in the text, for every $1 \leq i \leq n-1$, the subgraph $\mathcal{A}_i \coloneqq \{i+1, i+2, \ldots, n\}$ of G/ξ is connected if and only if ξ is a sequentially valid labeling. Let $j = n - i$; then the sequence $\mathcal{A}_j$ is increasing. Going from $\mathcal{A}_j$ to $\mathcal{A}_{j+1}$ for any j involves adding a vertex in G/ξ which is adjacent to $\mathcal{A}_j$. This fact provides an easy scheme to generate the number of sequentially valid labelings: for each vertex v in G/ξ, let $\mathcal{A}_1 = \{v\}$. Then pick a neighbor of $\mathcal{A}_1$ and add it to the set to form $\mathcal{A}_2$. Continue in this fashion

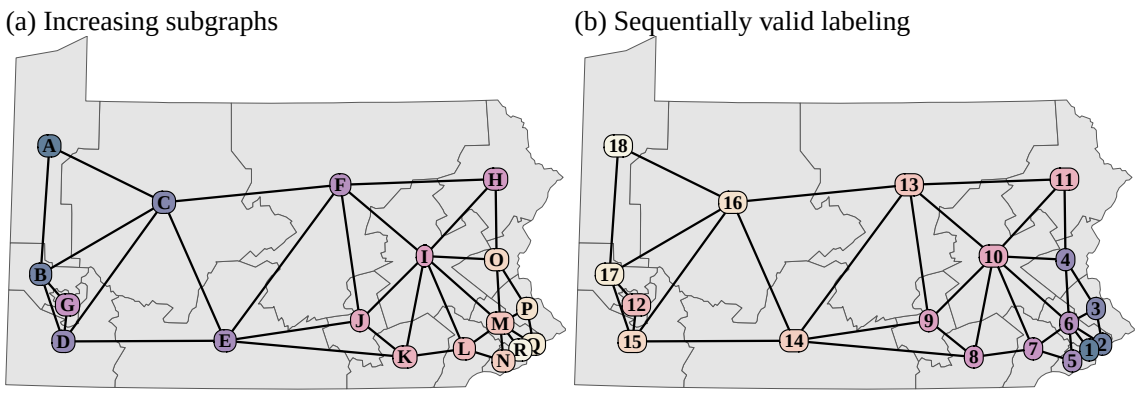


Figure A.3: Schematic of the process to generate a sequentially valid labeling, using the district-level graph for the Pennsylvania court-imposed plan. Subgraphs are built in alphabetical order in panel (a): $\{A\}, \{A, B\}, \ldots, \{A, B, \ldots, Q\}$. This corresponds to a sequentially valid labeling shown in panel (b).

until $\mathcal{A}_{n-1}$ contains all but one vertex of G/ξ. Undoing our bijection, this remaining vertex is labeled 1; the vertex added between $\mathcal{A}_{n-2}$ and $\mathcal{A}_{n-1}$ is labeled 2, and so on.

Figure A.3 illustrates this scheme on district-level graph for the Pennsylvania court-imposed plan. We pick an arbitrary vertex, denoted by "A" in Figure A.3(a). Then we pick a neighbor, denoted "B," and add it to the subgraph. We continue this way, adding the vertices indicated by the alphabetical ordering, until we have covered the whole district-level graph. Then the last vertex added, "R," is labeled 1 in the district labeling, as shown in Figure A.3(b). Vertex "Q" is labeled 2, and so on, until vertex "A" is labeled 18. One can easily check that this labeling is sequentially valid—at every point, the region of the map corresponding to unlabeled districts is contiguous.

As Section 1.4.4 mentioned, we adopt different strategies for calculating $\psi([\xi])$ when $n \leq 13$ and $n > 13$. When there are no more than 13 districts, we simply recurse down the tree of possibilities for generating all sequentially valid plans. In the example of Figure A.3, we have 18 options for the first vertex

in the subgraph. Once we have picked vertex "A", then we could add "B" or "C". If we add "B", then we could add "C", "D", or "G".

Clearly this tree grows quite large very quickly. Fortunately, there are many duplicate nodes—in our example, adding vertex B and then C in sequence produces the same subgraph as if we had added C and then B. Thus by memoizing our counting function we can significantly reduce the number of tree branches we must explore in full.

When $n > 13$, it is no longer computationally feasible to perform the recursion for each sampled plan in what will often be a large number of sampled plans. In these cases (it should be noted that for the 2020 redistricting cycle, only nine states had more than 13 districts), we can only estimate the number of sequentially valid labelings. To do so, we generate a large number of random sequentially valid labelings by following the scheme outlined above, picking vertices of G/ξ to add to each $\mathcal{A}_j$ uniformly at random. Since we observe the number of candidate vertices to add at each stage, and because each set of choices produces a unique sequentially valid labeling, we can compute the probability of drawing each sequentially valid labeling directly.

Denote the probability distribution created by this sampling scheme by p_{sv}, and the (not-probability) measure which assigns mass 1 to all $n!$ relabelings by μ_{all}. Our goal here is to estimate

$$\int \mathbf{1}\{\sigma \text{ is sequentially valid}\} d\mu_{\text{all}}(\sigma)$$

the total number of sequentially valid labelings. This can be done with our sample from p_{sv}, since it is supported on precisely the same set for which the indicator function takes a value of 1. We need only take the mean of the inverse of the probability of each sampled labeling.

This importance sampling estimate is quite accurate, since the proposal support matches the target sup-

port exactly, and there is not too much variation in the proposal probabilities. The estimate can be made arbitrarily good by increasing the number of importance samples. This is demonstrated in Figure A.4, which shows how the bias in importance sampling estimates varies with the number of importance samples and the overall size of the district-level graph. In our software, we use on the order of 1,500 samples for $n = 14$ by default, and more when n is larger (e.g., around 3,000 for the $n = 28$ Florida congressional districts). These numbers are chosen to ensure that the relative standard error (also known as the *coefficient of variation*) of the $\psi([\xi])$ estimate is controlled to a reasonable amount, something we demonstrate next for the case of Pennsylvania.

For the district-level graph shown above in A.3, we have $\log \psi([\xi]) = 29.999$, calculated exactly using the recursive procedure described above. Using the importance sampling procedure with 2,000 samples, we estimate a value of 29.949. This corresponds to a relative error of 0.051. We can also calculate the relative standard error of this example estimate with the delta method, which yields 0.052. In other words, the importance sampling error in the estimate of $\psi([\xi])$ is on the order of 5%. This is also small on the scale of the variation in $\psi([\xi])$ across plans: for the six comparison plans used in the main text, $\log \psi([\xi])$ ranges from 28.456 to 30.317. So the error in the importance sampling estimate ($|29.949 - 29.999| = 0.05$) is just 2.7% of the range of $\log \psi([\xi])$ across these six plans.

Finally, we note that the procedures used here to calculate $\psi([\xi])$ can also be applied to partial plans. Notating this formally is more difficult, but a partial plan can be viewed as a (unbalanced) plan with fewer districts. Calculating ψ for these partial plans and incorporating them into the SMC weights in Algorithm 2 improves the sampling efficiency without changing the target distribution, as long as each partial

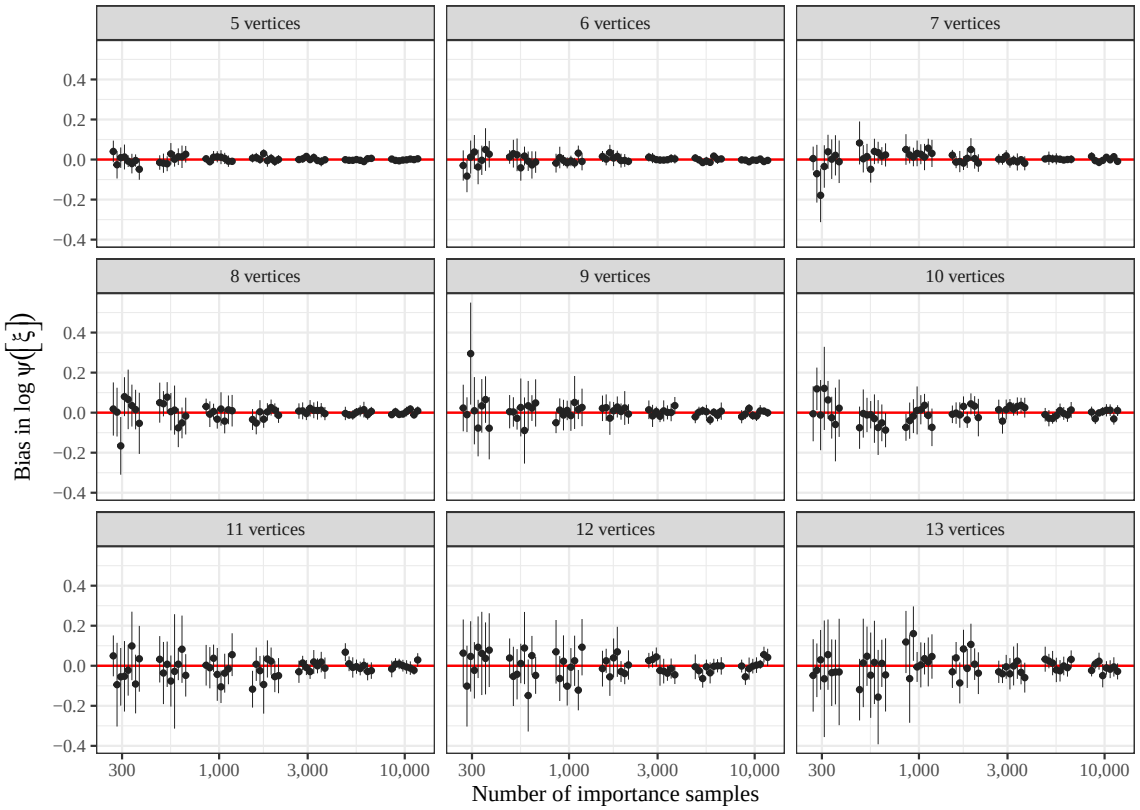


Figure A.4: Bias in importance sampling estimates of $\log \psi([\xi])$ by the number of importance samples and the number of vertices in the district-level graph. District-level graphs were selected as random subsets of the graph shown in A.3. Eight replicate experiments were performed for each importance sample size and graph size. The vertical bars for each point span two standard errors in either direction. Standard errors were calculated with the delta method.

ψ's contribution to the weights is canceled in the following iteration. This is the approach taken by our software implementation.

A.3.3 Stabilizing Importance Weights

When $\rho \neq 1$ or when the constraints imposed by J are severe, there can be substantial variance in the importance sampling weights. For large maps with $\rho = 0$, for instance, the weights will generally span hundreds if not thousands of orders of magnitude. This reflects the general computational difficulty in sampling uniformly from constrained graph partitions. As Najt et al. (2019) show, sampling of node-balanced graph

partitions is computationally intractable in the worst case. In such cases, the importance sampling estimates will be highly variable, and resampling based on these weights may lead to degenerate samples with only one unique map.

When the importance weights are variable but not quite so extreme, we find it useful to truncate the normalized final weights (such that their mean is 1) from above at a value $w_{\max}$ at the end of sampling. The theoretical basis for this maneuver is provided by Ionides (2008), who proved that as long as $w_{\max} \to \infty$ and $w_{\max}/S \to 0$ as $S \to \infty$, the resulting estimates are consistent and have bounded variance (since the truncation occurs only after the final SMC step, these conclusions, which were made in the context of importance sampling, carry over.) One such choice we have found to work well for the weights generated by this sampling process is $w_{\max} = S^{0.4}/100$, though for particular maps other choices of exponent and constant multiplier may be superior.

Truncation is no panacea, however. As with any method that relies on importance sampling, it is critical to examine the distribution of importance weights to ensure that they will yield acceptable resamples.

A.3.4 Computational Complexity

The two asymptotically slowest steps of the SMC algorithm are computing $\tau(G_i)$ for every district G_i and drawing a spanning tree using Wilson's algorithm for each iteration. All other steps, such as computing d_e and $|\mathcal{C}(G_i^{(j)}, \widetilde{G}_i^{(j)})|$, are linear in the number of vertices, and are repeated at most once per iteration.[1] Computing $\tau(G_i)$ requires computing a determinant, which currently has computational complexity $O(|V_i(\xi)|^{2.373})$ though most implementations are $O(|V_i(\xi)|^3)$. Since this must be done for each district of size roughly m/n, the total complexity for sampling one plan is $O(n \cdot (m/n)^{2.373})$. For the

[1] To compute d_e, we walk depth-first over the tree and store, for each node, the total population of that node and the nodes below it. This allows for $O(1)$ computation of d_e for all edges.

spanning trees, the expected runtime of Wilson's algorithm is the mean hitting time of the graph, which is $O(m^2)$ in the worst case. So the total complexity for each sample is roughly $O(nm^2 + m^{2.373}n^{-1.373})$ (ignoring the random rejection procedure). Note that when $\rho = 1$, we need not compute $\tau(G_i)$, and the total complexity is roughly $O(nm^2)$.

This page intentionally left blank.

B

Appendices for Chapter 2

B.1 Correlation Between Partisan Gerrymandering Metrics Across Multiple States

Figure B.1 expands on Figure 2.3 in the main text with pairwise scatter plots and correlation summaries.

Table B.1 below reports the correlation between differential harm and the partisan gerrymandering metrics used in Section 2.6 for eight states which cover a wide range of partisan leans and number of seats:

- Utah, with 4 districts and a Democratic vote share baseline of 33%
- Mississippi, with 4 districts and a Democratic vote share baseline of 43%
- Iowa, with 4 districts and a Democratic vote share baseline of 46%
- North Carolina, with 14 districts and a Democratic vote share baseline of 50%
- Wisconsin, with 8 districts and a Democratic vote share baseline of 51%
- New Jersey, with 12 districts and a Democratic vote share baseline of 57%
- Massachusetts, with 9 districts and a Democratic vote share baseline of 61%

- California, with 52 districts and a Democratic vote share baseline of 65%

The correlations are calculated across 5,000 simulated plans for each state. The simulations are provided by the 50-State Simulation Project (McCartan et al., 2022). In particular, the New Jersey simulations are not the same as those used in Section 2.6.

	Correlation with differential harm and:				
State	Exp. Dem. seats	Efficiency gap	Partisan bias	Mean-median	Declination
Utah	-0.983	0.957	-0.893	-0.945	-0.876
Mississippi	-0.772	0.207	0.031	-0.096	-0.656
Iowa	-0.597	0.449	-0.185	0.184	0.365
North Carolina	-0.943	0.740	0.520	0.704	0.519
Wisconsin	-0.816	0.621	0.191	0.652	0.374
New Jersey	-0.789	0.459	0.082	-0.014	0.381
Massachusetts	-0.826	-0.069	-0.158	0.159	NA
California	-0.590	0.616	-0.199	-0.302	0.359

Table B.1: Correlations between five gerrymandering metrics and partisan differential harm, for simulated districting plans from eight states. Declination is NA for Massachusetts since there are no Republican-leaning districts at baseline under any simulated plan.

B.2 Similar Two-party Election Model Validation

This analysis, performed by the author, is excerpted and adapted from Kenny et al. (2023). The notation may differ from the main text.

We model the two-party Democratic vote share in each district y_{it} for an election t as

$$\text{logit}(y_{it}) = \alpha_i + \beta_t + \varepsilon_{it}, \tag{B.1}$$

$$\beta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\beta^2),$$

$$\varepsilon_{it} \overset{iid}{\sim} \text{t}_\nu(0, \sigma_\varepsilon^2),$$

where α_i is district's baseline partisanship, β_t is election-to-election national swing, and ε_{it} is district-election specific error.

It is important to validate the electoral model because fitting a random effect to past House elections is not the same as predicting a partisan baseline using past presidential results. Figure B.2 shows the results of this validation. We compare the fitted random effects for the 2010 decade-districts to the baseline that would be obtained by aggregating our precinct-level partisan baseline to 2010 district boundaries. The agreement is generally excellent, despite the baseline only using information on two of the three presidential elections conducted during this period. The few large deviations indicate particularly strong candidate or regional effects (like Democrats holding the Montana at-large district, for instance).

It is difficult to further validate the predictive accuracy model, given the paucity of precinct-level election data with national coverage. However, others have found the combination of heavy-tailed error and the Stochastic Uniform Partisan Swing model to produce well-calibrated predictions Ebanks et al. (2022). We also evaluate the predictions for 2018, both *a priori* (i.e., averaging over possible national environments) and conditional on the actual shift from our 2016–2020 baseline. Figure B.3 shows these predictions, displayed as the number of seats Democrats are expected to win out of the 255 in our 2018 dataset. For these predictions, we do not use the fitted district random effects (y-axis of Figure B.2), but rather the precinct-level partisan baseline (x-axis of Figure B.2). The predictions are not fully out-of-sample, because 2018 data informs the overall estimates of σ_β, σ_ε, and ν. The actual number of seats won by Democrats falls well within the range expected under the model. Notice also the wider variation in the leftmost panel of Figure B.3 due to the uncertainty in the national environment.

We fit the model with the `brms` R package Bürkner (2017), with a weakly informative $t_3(0, 2.5)$ prior on the Intercept and all scale parameters, and a $\text{Gamma}(2, 0.1)$ prior on ν. Estimates for the β_t are shown in

Figure B.4. There is no evidence for heteroskedasticity corresponding to larger or smaller national swings over time. This supports our use of historical data back to 1976.

We use the posterior medians of these hyperparameters from the fitted model in our electoral predictions. Combined with the baseline vote estimate $\hat{\alpha}_{ip}$ for a specific plan p, the model yields the posterior predictive distribution of the vote share y_{it} in district i. Averaging over this distribution gives us the probability that the district for any plan is won by a Republican or a Democrat.

B.3 Ecological Inference Model Estimates for Alabama

Posterior medians of turn_{jr} and supp_{jr} from the model described in Section 2.7 are plotted in Figure B.5. These estimates broadly track with the findings of the voting rights expert retained in *Merrill*.

B.4 Additional Sampling Analysis for Alabama

Here, we perform the same analysis as in Section 2.7, but using a randomly sampled set of plans as a counterfactual, rather than the four plans used in litigation. Specifically, we use 2,000 plans provided by the 50-State Simulation Project (McCartan et al., 2022). These plans are designed to discourage packing of Black voters and maintain at least the same number of majority-opportunity districts as the enacted 2020 plan.

Figure B.6 summarizes the results, which are mostly similar to the results in in Section 2.7. Across the state, the estimated average harm is 6.9%, which translates to around 144,000 voters. Of these harmed voters, around 56,000 are Black, while 88,000 are White. Similarly, 73,000 are Democratic, while only 71,000 are Republican. Normalizing by group sizes, the expected harm for a Black Democrat is 9.6%, for a Black Republican is 16.8%, for a White Democrat is 9.7%, and for a White Republican is 5.2%. So the

challenged plan differentially harms Democratic voters, but does not differentially harm Black Democrats versus White Democrats. However, it it especially harms Black Republicans, who have an estimated differential harm of 7.2% versus Black Democrats and of 11.6% versus White Republicans.

B.5 Validation of Latent Factor Ecological Inference Model for Boston Primary Elections

To test the suitability of the model for representing the outcomes of hypothetical future elections, we use the model fitted to the primary elections to predict the outcome of the mayoral general election in November 2021. Michelle Wu and Annissa Essaibi-George advanced to the general from the preliminary election. We fix the model's loading matrix $\mathbf{L}$ to the posterior mean of their estimated loadings. We draw s_{jk} from a Beta$(20\bar{s}, 20(1-\bar{s}))$ distribution, where $\bar{s}$ is each candidate's share of the two-candidate vote from the preliminary election (59.7% Wu – 40.3% Essaibi George).

The model's predictions (posterior medians) for the turnout and vote share in each precinct are shown in Figure B.7. Considering the differences between a primary and general election, and the additional campaign time elapsed between the elections, the predictions are remarkably accurate. The predicted turnout has a Spearman correlation of 0.84 with the actual turnout, and the predicted and actual vote shares have a correlation 0.87. The turnout and support levels are also in relatively close agreement.

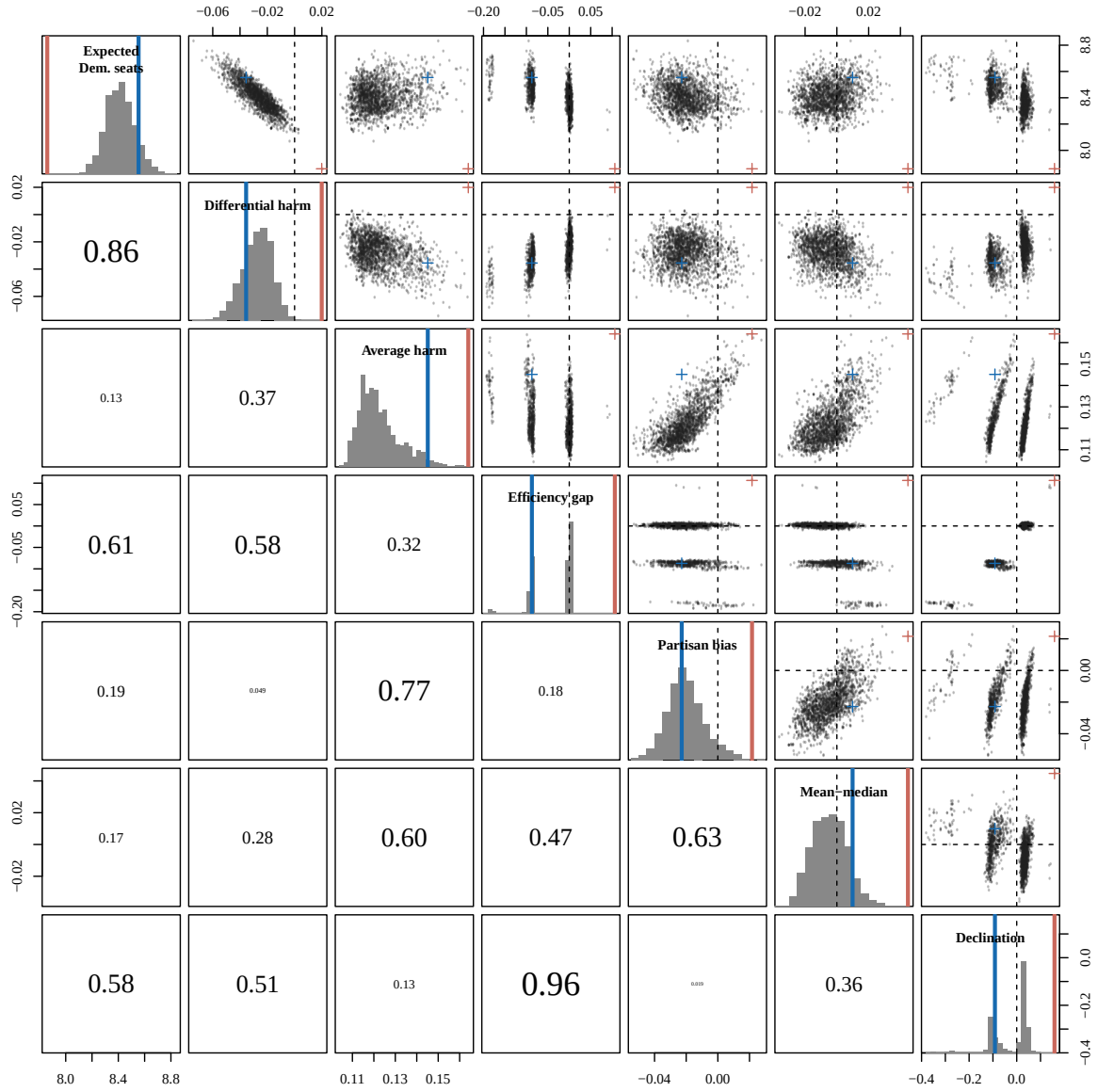


Figure B.1: Scatterplots for all pairs of measures. The Democratic and Republican gerrymanders are shown with the blue and red lines and crosses. The dashed line indicates a value of zero. Below the diagonal, numbers indicate the absolute value of the correlation between variables.

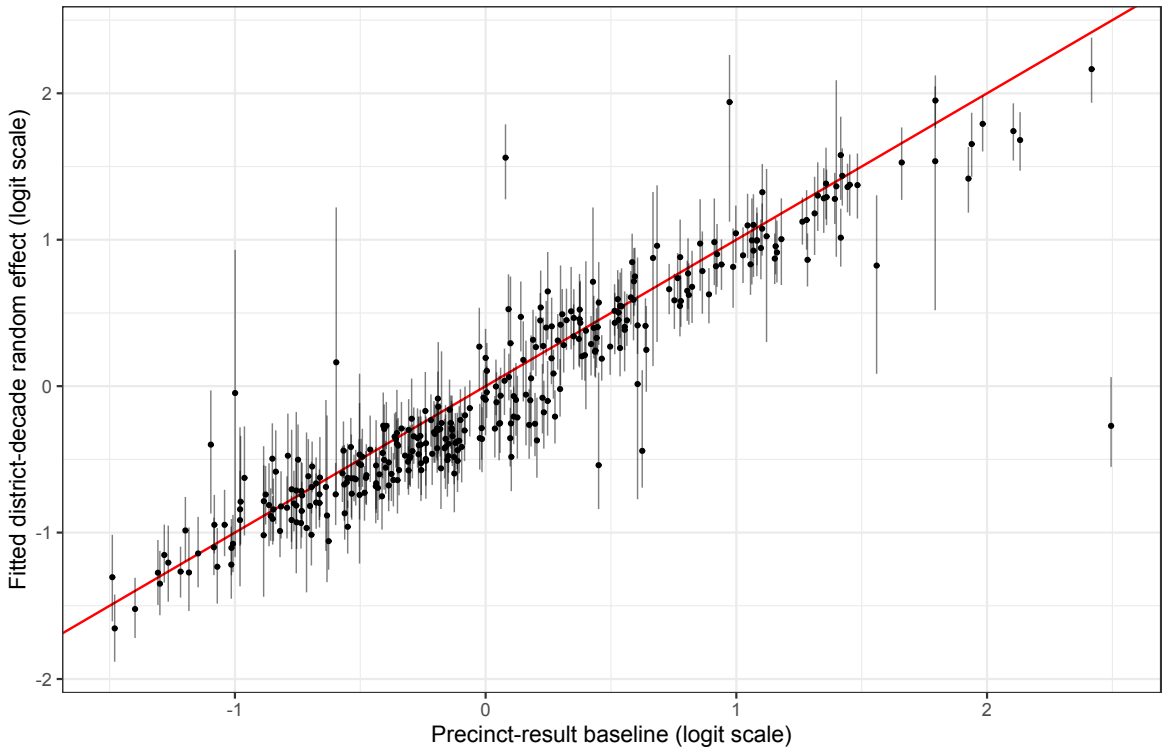


Figure B.2: Posterior point predictions and intervals for decade-district random effects, versus the 2016–2020 partisan baseline calculated from precinct-level presidential returns. The red line indicates perfect agreement.

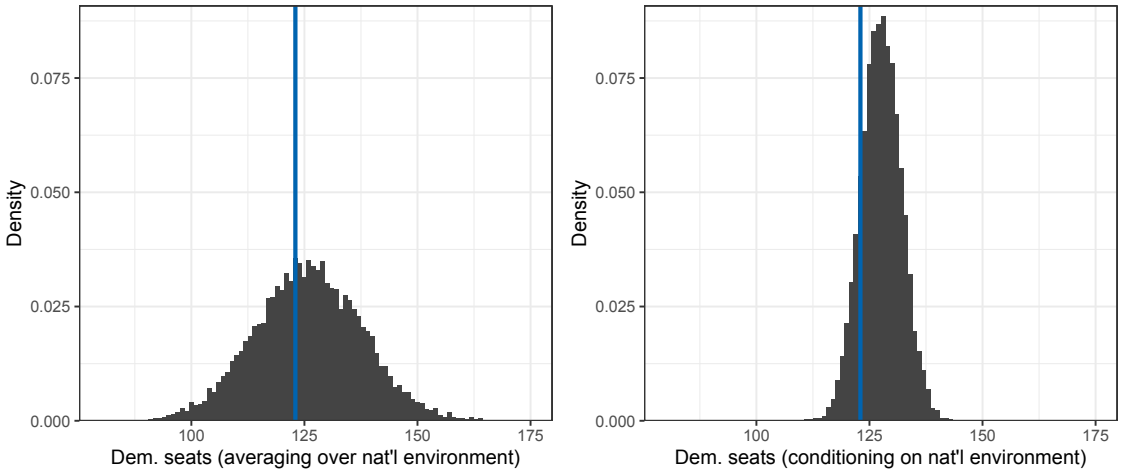


Figure B.3: Posterior predictive distribution and actual results for 255 2018 House elections in the historical dataset. On the left, predictions are unconditional; on the right, they are conditional on the national environment.

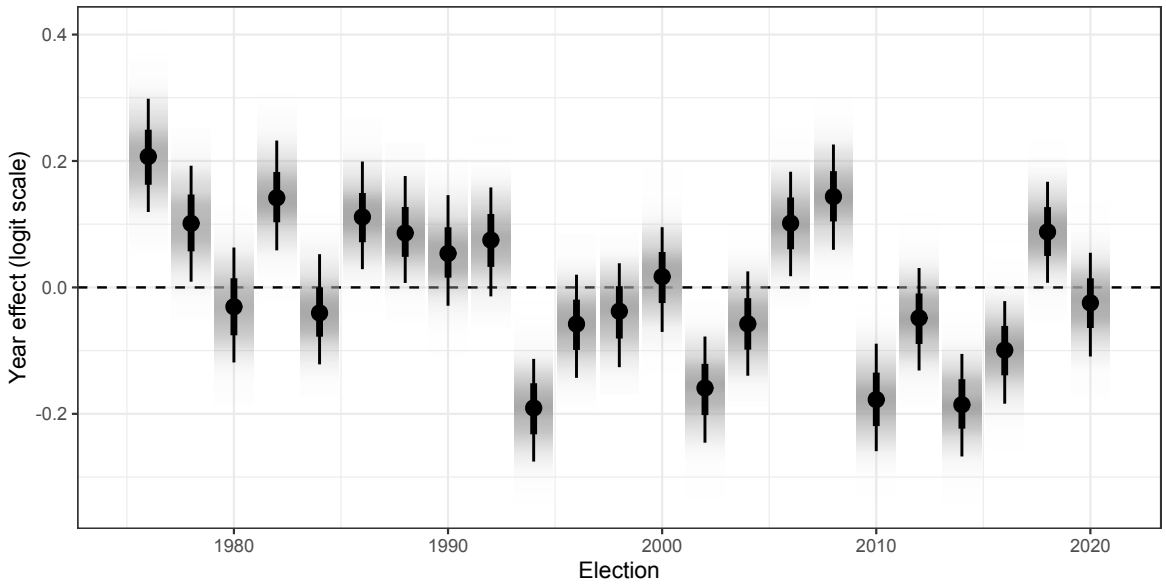


Figure B.4: Posterior distribution of β_t for each election year, from a hierarchical model fit to historical data. The β_t coefficients are on a logit scale, so a value of 0.1 corresponds roughly to a 2.5 percentage point vote share overperformance by Democrats.

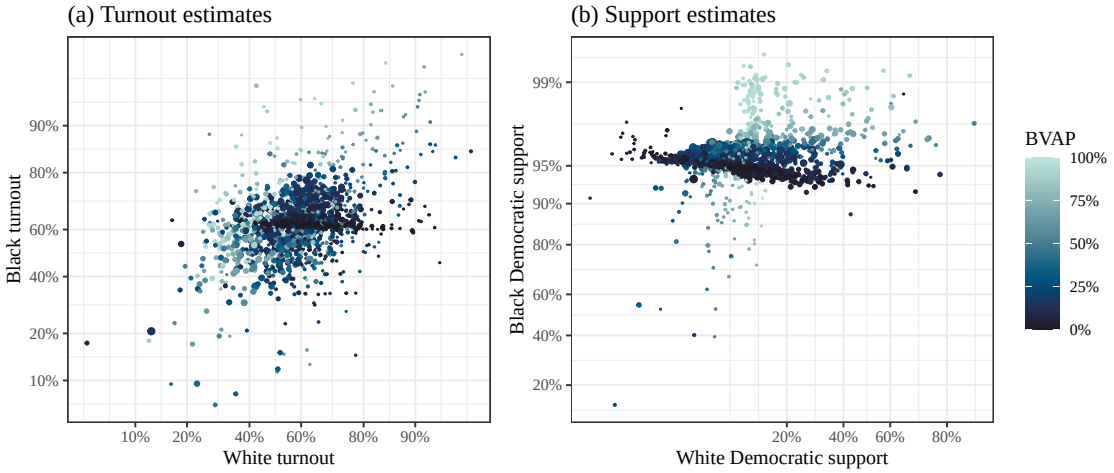


Figure B.5: Model estimates of precinct-level turnout by race (a) and Democratic support by race (b). Points are sized in proportion to their voting-age population, and colored according to their Black share of the voting-age population. All axes are transformed to a logit scale.

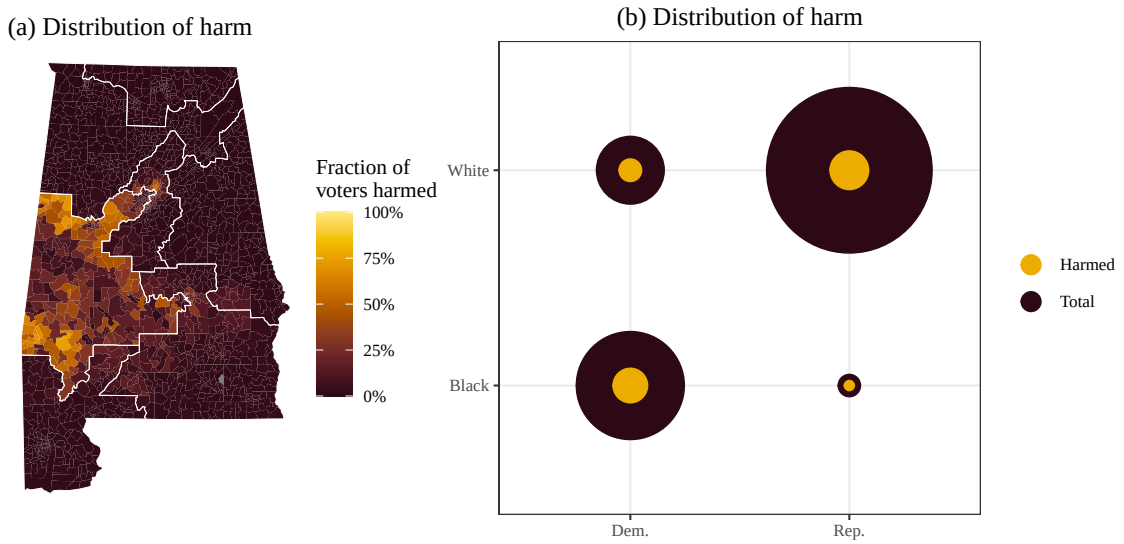


Figure B.6: On the left, the expected fraction of voters in each precinct which are harmed by the 2020 enacted congressional districts versus the sampled counterfactual set. On the right, the expected number of harmed voters in each race/partisan group, relative to the total number of such voters.

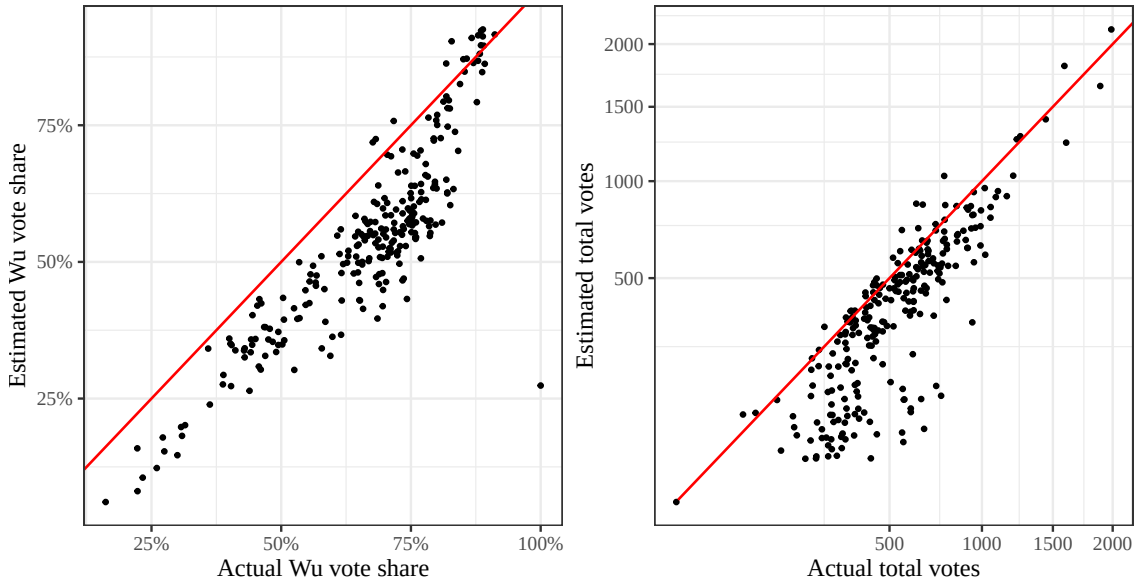


Figure B.7: Model predictions of general election vote shares and turnout, compared to actual general election results. Red reference line indicates perfect agreement.

This page intentionally left blank.

C

Appendices for Chapter 3

C.1 Survey administration

Consent was obtained from survey participants prior to participation in the study. Participants were aware that they were taking part in a research study. No deception was used in this study. Respondents were not compensated for participation and were informed at the time of consent that there would be no compensation and that participation was completely voluntary.

C.2 Summary Statistics

Tables C.1 and C.2 contain summary statistics for the full and control samples, respectively.

C.3 Variable Descriptions and Model Specification

Table C.1: Survey Sample Summary Statistics

	Miami (n = 474)		NYC (n = 453)		Phoenix (n = 1,600)		Pooled (n = 2,527)	
	Avg.	St. Dev.	Avg.	St. Dev.	Avg.	St. Dev.	Avg.	St. Dev.
Democrat	0.54	0.50	0.66	0.47	0.44	0.50	0.50	0.50
Republican	0.37	0.48	0.27	0.45	0.48	0.50	0.42	0.49
Biden	0.64	0.48	0.74	0.44	0.56	0.50	0.61	0.49
Age	60.08	13.25	59.30	12.23	62.76	12.57	61.64	12.72
Female	0.44	0.50	0.46	0.50	0.48	0.50	0.47	0.50
White	0.69	0.46	0.74	0.44	0.88	0.33	0.82	0.39
Income	106,321	42,254	119,237	42,013	109,828	42,073	111,271	42,318
College	0.80	0.40	0.86	0.35	0.74	0.44	0.77	0.42
Years Residence	15.65	9.99	22.17	13.77	19.85	70.35	19.48	56.48
Homeowner	0.86	0.34	0.73	0.44	0.91	0.28	0.87	0.33
Married	0.65	0.48	0.59	0.49	0.68	0.47	0.66	0.47
Children in Home	0.35	0.48	0.38	0.49	0.29	0.45	0.32	0.47

Table C.3: Variable description and model specification.

Variable	Description	Interaction(s)	Baseline only?
Church	Block contains church		
Distance to church	Logarithm of distance to nearest church from block, in meters		
Park	Block contains park		
School	Block contains school	Children	
Distance to school	Logarithm of distance to nearest school from block, in meters	Children	
Children	Respondent has children at home	School, distance to school	*
Same block group	Block in same block group as respondent		
Same tract	Block in same tract as respondent		
Same road region	Block in same region bounded by major roads and railroads as respondent		
Population	Square root of the block population divided by 10,000		
Area	Square root of the area in square miles		

Table C.3 (Continued)

Variable	Description	Interaction(s)	Baseline only?
Fraction same race	Fraction of block of the same race as respondent (White, Black, Hispanic, Other)	Minority	*
Minority	Whether respondent is non-white, or Hispanic of any race	Fraction same race	*
Fraction same party	Fraction of block of the same party as respondent	Party	*
Party	Self-reported political party: Democratic, Republican, or independent	Fraction same party	*
Fraction same ownership	Fraction of block with same home ownership or rental status as respondent	Homeowner	*
Homeowner	Whether respondent owns their home	Fraction same ownership	*
Fraction same education	Fraction of block with same education as respondent	Education	*
Education	Respondent education; either "college" or "no college"	Fraction same education, Income	*
Income	Logarithm of median income of block group	Education	*
Age	Respondent age group: 0–40, 41–55, 56–65, 66-75, or 76+		*
Retired	Whether respondent is retired		*
Tenure	Square root of how long respondent has lived in current residence		*

C.4 Full and baseline model estimates

Tables C.4 and C.5 contain posterior summaries for all model coefficients on the original model scale. These models were fit using data from 471 survey respondents, consisting of 474,655 individual block-level observations.

Table C.2: Survey Sample Summary Statistics – Control Sample

	Miami (n = 474)		NYC (n = 453)		Phoenix (n = 1,600)		Pooled (n = 2,527)	
	Avg.	St. Dev.	Avg.	St. Dev.	Avg.	St. Dev.	Avg.	St. Dev.
Democrat	0.54	0.50	0.66	0.47	0.44	0.50	0.50	0.50
Republican	0.37	0.37	0.27	0.27	0.48	0.48	0.42	0.42
Vote Biden 2020	0.64	0.48	0.74	0.44	0.56	0.50	0.61	0.49
Age	60.08	13.25	59.30	12.23	62.76	12.57	61.64	12.72
Female	0.44	0.50	0.46	0.50	0.48	0.50	0.47	0.50
White	0.75	0.44	0.75	0.43	0.88	0.32	0.84	0.37
Income (1,000s)	106.56	41.78	119.44	41.55	109.99	41.73	111.46	41.92
College	0.80	0.40	0.86	0.35	0.74	0.44	0.77	0.42
Years Residence	15.65	9.99	22.17	13.77	17.42	10.90	17.94	11.49
Homeowner	0.86	0.34	0.73	0.44	0.91	0.28	0.87	0.33
Married	0.65	0.48	0.59	0.49	0.68	0.47	0.66	0.47
Children in Home	0.35	0.48	0.38	0.49	0.29	0.45	0.32	0.47

Table C.4: Full model estimates.

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
(Intercept)	Miami	-7.08	9.12	-21.71	-7.29	8.12
(Intercept)	NYC	-4.90	7.71	-17.17	-5.06	7.82
(Intercept)	Phoenix	-10.58	6.62	-21.14	-10.51	0.40
Church	Miami	0.00	0.06	-0.11	0.00	0.10
Church	NYC	0.11	0.05	0.04	0.11	0.19
Church	Phoenix	0.15	0.05	0.07	0.15	0.23
Distance	Miami	0.03	0.02	0.00	0.03	0.07
Distance	Miami	0.12	0.03	0.07	0.12	0.17
Distance	NYC	0.12	0.02	0.08	0.12	0.15
Distance	NYC	0.01	0.03	-0.03	0.01	0.06
Distance	Phoenix	0.09	0.01	0.07	0.09	0.11
Distance	Phoenix	0.11	0.02	0.09	0.11	0.14
Park	Miami	-0.07	0.09	-0.22	-0.07	0.07
Park	NYC	0.15	0.05	0.07	0.15	0.23
Park	Phoenix	0.03	0.04	-0.04	0.03	0.10
School	Miami	0.34	0.14	0.13	0.34	0.57
School	NYC	-0.15	0.13	-0.36	-0.15	0.05
School	Phoenix	0.07	0.08	-0.06	0.07	0.20
Children	Miami	0.74	3.09	-4.26	0.75	5.68
Children	NYC	-0.77	2.73	-5.10	-0.81	3.67

Table C.4 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
Children	Phoenix	-0.36	2.21	-4.07	-0.29	3.26
Same block group	Miami	0.13	0.07	0.01	0.13	0.25
Same block group	NYC	0.21	0.09	0.07	0.21	0.36
Same block group	Phoenix	0.14	0.03	0.08	0.14	0.20
Same tract	Miami	-0.34	0.10	-0.50	-0.34	-0.18
Same tract	NYC	0.07	0.13	-0.13	0.07	0.28
Same tract	Phoenix	0.05	0.09	-0.10	0.05	0.20
Same road region	Miami	-0.52	0.05	-0.60	-0.52	-0.44
Same road region	NYC	0.13	0.04	0.07	0.13	0.19
Same road region	Phoenix	0.25	0.03	0.20	0.25	0.31
Population	Miami	-0.08	0.26	-0.51	-0.08	0.35
Population	NYC	-1.19	0.20	-1.52	-1.19	-0.87
Population	Phoenix	-1.06	0.19	-1.36	-1.08	-0.75
Area	Miami	-0.34	0.08	-0.47	-0.34	-0.20
Area	NYC	-0.14	0.10	-0.31	-0.14	0.03
Area	Phoenix	-0.16	0.04	-0.23	-0.16	-0.09
Fraction same race	Miami	-0.37	0.05	-0.46	-0.37	-0.29
Fraction same race	NYC	-0.12	0.05	-0.21	-0.12	-0.04
Fraction same race	Phoenix	-0.33	0.03	-0.38	-0.33	-0.28
Fraction same party	Miami	-0.18	0.08	-0.31	-0.18	-0.04
Fraction same party	NYC	-0.23	0.07	-0.34	-0.24	-0.11
Fraction same party	Phoenix	-0.35	0.08	-0.47	-0.34	-0.22
Fraction same ownership	Miami	-0.19	0.14	-0.42	-0.20	0.06
Fraction same ownership	NYC	-0.27	0.13	-0.48	-0.27	-0.07
Fraction same ownership	Phoenix	-0.18	0.11	-0.36	-0.18	0.00
Fraction same education	Miami	0.04	0.14	-0.19	0.04	0.27
Fraction same education	NYC	-0.53	0.10	-0.70	-0.53	-0.37
Fraction same education	Phoenix	-0.94	0.09	-1.08	-0.94	-0.80
Income	Miami	-0.10	0.05	-0.18	-0.10	-0.02
Income	NYC	-0.13	0.04	-0.20	-0.13	-0.05
Income	Phoenix	0.20	0.04	0.14	0.20	0.26
Age	Miami	0.00	0.16	-0.27	0.01	0.26
Age	NYC	0.00	0.13	-0.21	0.00	0.21
Age	Phoenix	0.00	0.11	-0.18	0.00	0.19
Education = no college	Miami	-2.09	3.08	-7.31	-2.19	2.98
Education = no college	NYC	0.19	2.93	-4.69	0.20	4.82
Education = no college	Phoenix	-0.26	2.39	-4.11	-0.29	3.61
Retired	Miami	0.60	3.25	-4.57	0.55	5.87
Retired	NYC	0.54	3.08	-4.47	0.48	5.65
Retired	Phoenix	0.13	2.81	-4.30	0.10	5.07
Tenure	Miami	-0.09	0.93	-1.69	-0.09	1.43

Table C.4 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
Tenure	NYC	-0.04	0.90	-1.56	-0.05	1.47
Tenure	Phoenix	0.11	0.80	-1.22	0.11	1.43
Party = ind	Miami	-0.20	4.44	-7.44	-0.38	7.04
Party = ind	NYC	0.16	4.44	-7.07	0.12	7.62
Party = ind	Phoenix	-0.36	3.69	-6.45	-0.45	5.71
Party = rep	Miami	0.07	2.72	-4.63	0.15	4.39
Party = rep	NYC	-0.50	2.48	-4.55	-0.43	3.76
Party = rep	Phoenix	-0.01	1.95	-3.08	-0.07	3.23
Minority	Miami	-0.50	3.38	-6.10	-0.44	5.09
Minority	NYC	0.00	2.53	-4.17	0.09	3.97
Minority	Phoenix	0.09	3.03	-5.04	0.14	4.95
Homeowner	Miami	0.20	4.94	-7.43	0.22	8.26
Homeowner	NYC	-0.64	2.90	-5.34	-0.62	3.95
Homeowner	Phoenix	0.10	3.69	-5.74	0.22	6.40
School * children	Miami	-0.31	0.21	-0.65	-0.30	0.04
School * children	NYC	0.39	0.17	0.10	0.39	0.68
School * children	Phoenix	0.31	0.13	0.09	0.30	0.53
Children * distance	Miami	-0.03	0.04	-0.11	-0.03	0.04
Children * distance	NYC	0.07	0.04	0.01	0.07	0.14
Children * distance	Phoenix	0.08	0.03	0.03	0.08	0.13
Same tract * same road region	Miami	0.18	0.11	0.00	0.17	0.35
Same tract * same road region	NYC	-0.25	0.14	-0.46	-0.25	-0.02
Same tract * same road region	Phoenix	-0.63	0.09	-0.77	-0.63	-0.48
Fraction same race * minority	Miami	0.30	0.09	0.15	0.30	0.45
Fraction same race * minority	NYC	0.27	0.12	0.07	0.27	0.46
Fraction same race * minority	Phoenix	0.00	0.23	-0.38	0.01	0.37
Fraction same party * party = ind	Miami	-0.09	0.20	-0.41	-0.09	0.25
Fraction same party * party = ind	NYC	-0.03	0.31	-0.53	-0.05	0.49
Fraction same party * party = ind	Phoenix	0.72	0.15	0.47	0.72	0.97
Fraction same party * party = rep	Miami	-0.26	0.13	-0.48	-0.25	-0.04
Fraction same party * party = rep	NYC	-0.04	0.13	-0.27	-0.03	0.17
Fraction same party * party = rep	Phoenix	0.08	0.09	-0.07	0.07	0.23
Fraction same ownership * homeowner	Miami	0.35	0.17	0.05	0.35	0.63
Fraction same ownership * homeowner	NYC	0.42	0.16	0.14	0.42	0.69
Fraction same ownership * homeowner	Phoenix	0.04	0.12	-0.16	0.04	0.24
Fraction same education * educ = no college	Miami	0.48	0.25	0.05	0.48	0.91
Fraction same education * educ = no college	NYC	1.41	0.20	1.09	1.41	1.75
Fraction same education * educ = no college	Phoenix	0.90	0.20	0.59	0.89	1.24
Income * education = no college	Miami	0.17	0.09	0.03	0.17	0.32
Income * education = no college	NYC	-0.10	0.08	-0.24	-0.11	0.03
Income * education = no college	Phoenix	0.00	0.08	-0.13	0.00	0.14

Table C.4 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
Alpha	Miami	1.46	0.05	1.39	1.46	1.54
Alpha	NYC	1.46	0.05	1.39	1.47	1.55
Alpha	Phoenix	1.27	0.03	1.23	1.27	1.31

Table C.5: Baseline model estimates.

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
(Intercept)	Miami	-7.76	1.07	-9.55	-7.72	-6.04
(Intercept)	NYC	-8.08	1.15	-9.98	-8.07	-6.19
(Intercept)	Phoenix	-8.55	0.97	-10.11	-8.58	-6.93
Church	Miami	0.01	0.06	-0.09	0.01	0.11
Church	NYC	0.13	0.05	0.05	0.13	0.22
Church	Phoenix	0.20	0.05	0.13	0.20	0.28
Distance	Miami	0.03	0.02	-0.01	0.03	0.06
Distance	Miami	0.09	0.02	0.06	0.09	0.13
Distance	NYC	0.12	0.02	0.09	0.12	0.16
Distance	NYC	0.05	0.02	0.01	0.05	0.08
Distance	Phoenix	0.11	0.01	0.08	0.11	0.13
Distance	Phoenix	0.13	0.01	0.11	0.13	0.15
Park	Miami	-0.05	0.09	-0.19	-0.05	0.10
Park	NYC	0.18	0.05	0.10	0.18	0.26
Park	Phoenix	0.04	0.04	-0.03	0.04	0.11
School	Miami	0.23	0.11	0.05	0.24	0.41
School	NYC	0.05	0.10	-0.11	0.05	0.21
School	Phoenix	0.22	0.06	0.12	0.21	0.33
Same block group	Miami	0.10	0.07	-0.02	0.10	0.22
Same block group	NYC	0.18	0.09	0.02	0.18	0.33
Same block group	Phoenix	0.14	0.03	0.08	0.14	0.19
Same tract	Miami	-0.33	0.10	-0.49	-0.33	-0.17
Same tract	NYC	0.14	0.12	-0.06	0.14	0.35
Same tract	Phoenix	0.09	0.09	-0.06	0.08	0.23
Same road region	Miami	-0.55	0.05	-0.63	-0.55	-0.48
Same road region	NYC	0.16	0.04	0.10	0.16	0.23
Same road region	Phoenix	0.27	0.03	0.22	0.27	0.32
Population	Miami	-1.18	0.22	-1.53	-1.19	-0.82
Population	NYC	-1.76	0.18	-2.07	-1.76	-1.49
Population	Phoenix	-2.56	0.17	-2.85	-2.55	-2.29
Area	Miami	-0.25	0.08	-0.37	-0.24	-0.12
Area	NYC	-0.08	0.11	-0.27	-0.07	0.10

Table C.5 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
Area	Phoenix	-0.12	0.04	-0.19	-0.12	-0.05
Same tract * same road region	Miami	0.16	0.10	-0.01	0.15	0.32
Same tract * same road region	NYC	-0.32	0.14	-0.55	-0.32	-0.09
Same tract * same road region	Phoenix	-0.69	0.09	-0.83	-0.69	-0.55
Alpha	Miami	1.45	0.05	1.37	1.45	1.53
Alpha	NYC	1.38	0.05	1.31	1.38	1.47
Alpha	Phoenix	1.28	0.03	1.24	1.28	1.32

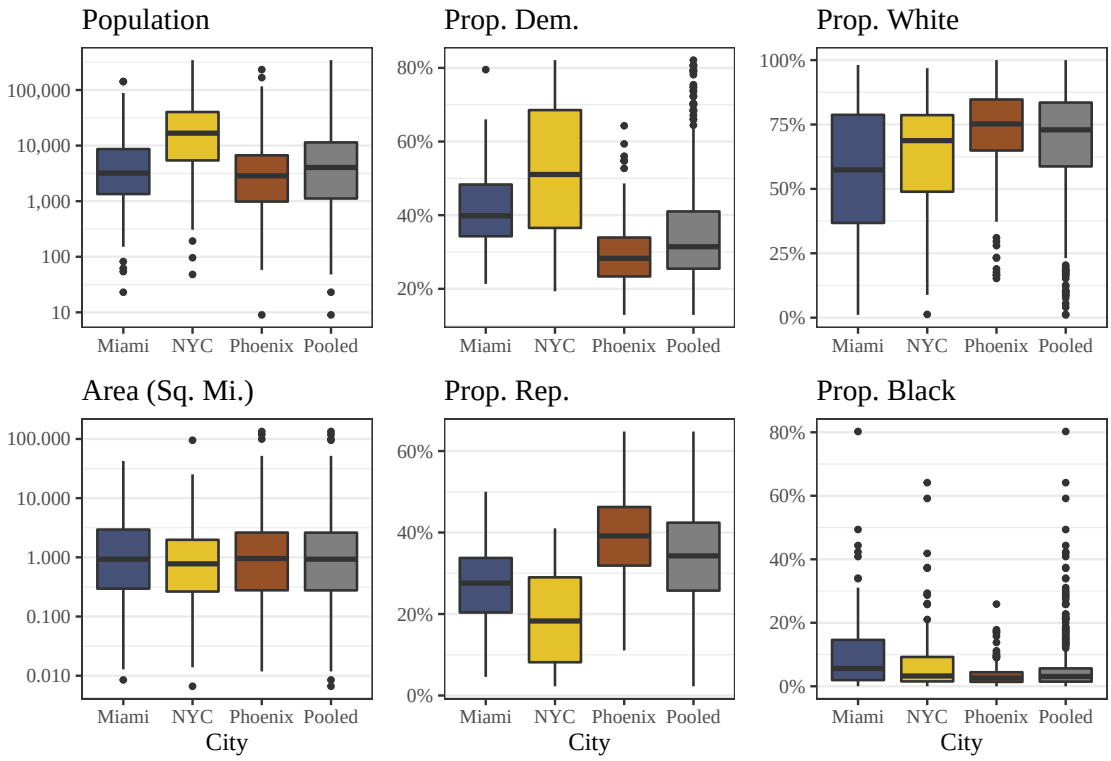


Figure C.1: Summary statistics for respondent neighborhoods in the control sample.

C.5 Local Diversity and Aggregate-level Outcomes

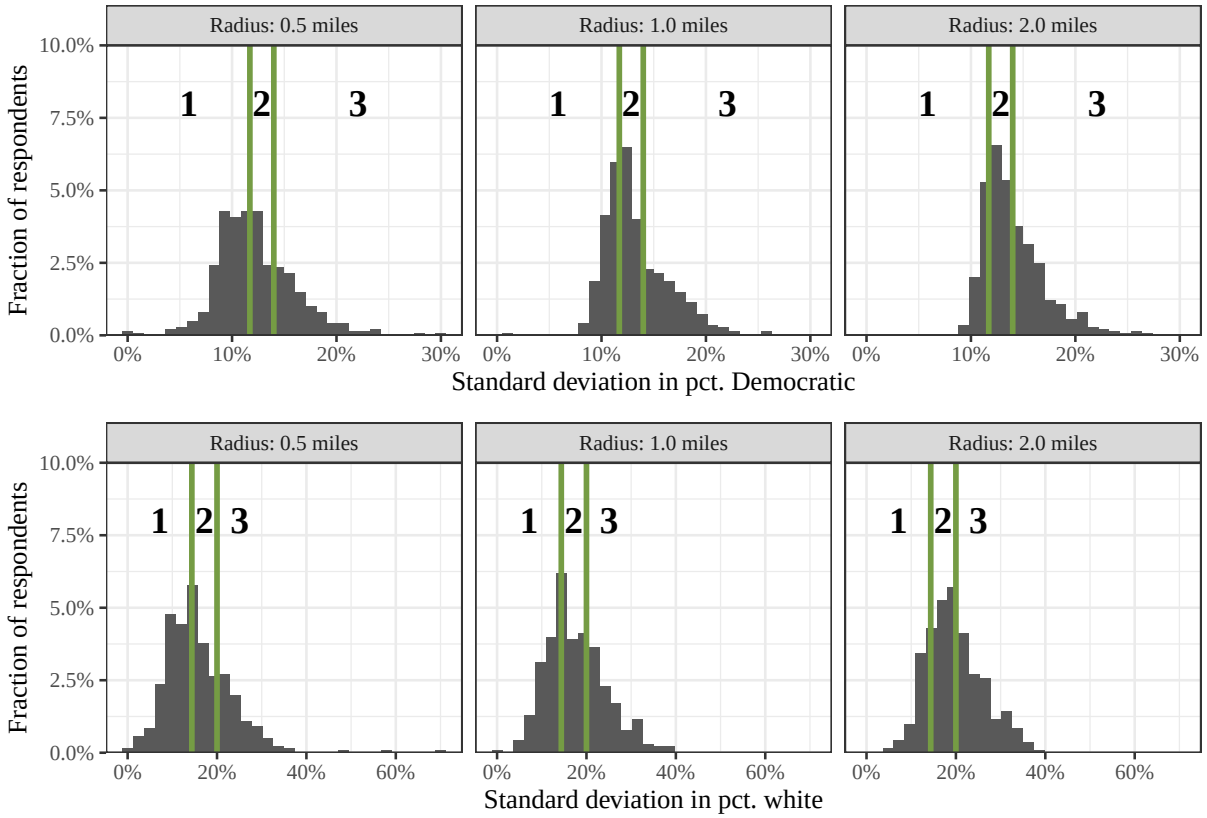


Figure C.2: Local partisan and racial diversity for each respondent, as measured by the block-level standard deviation of the respective variables for blocks within a 0.5, 1.0, and 2.0-mile radius. Terciles are indicated by vertical lines and bold labels.

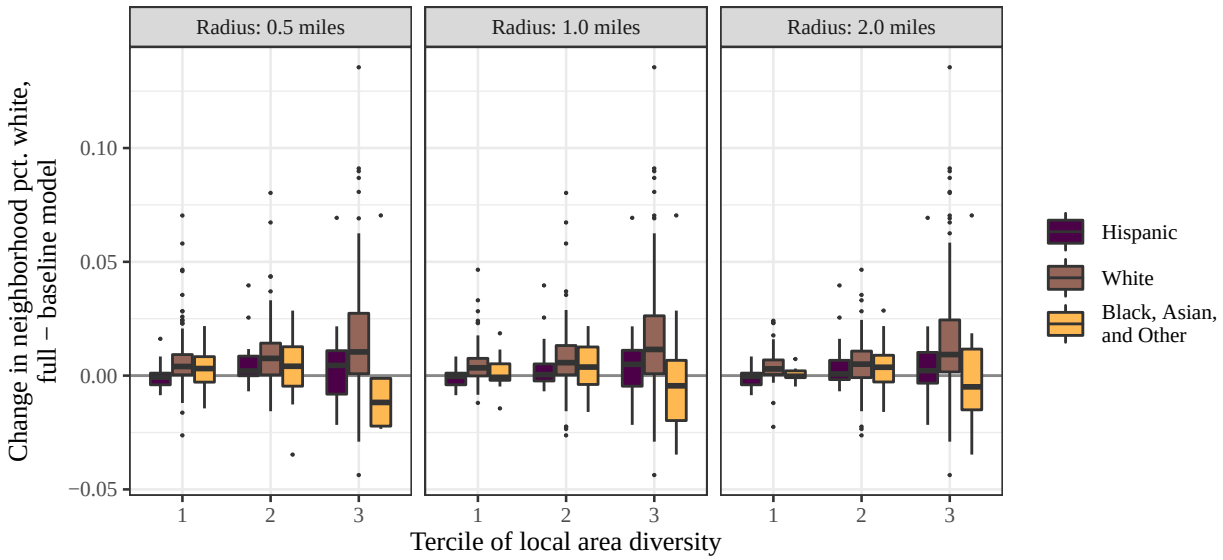


Figure C.3: Change in neighborhood fraction white from baseline to full model, broken out by tercile of local area diversity, across a range of measurement radii.

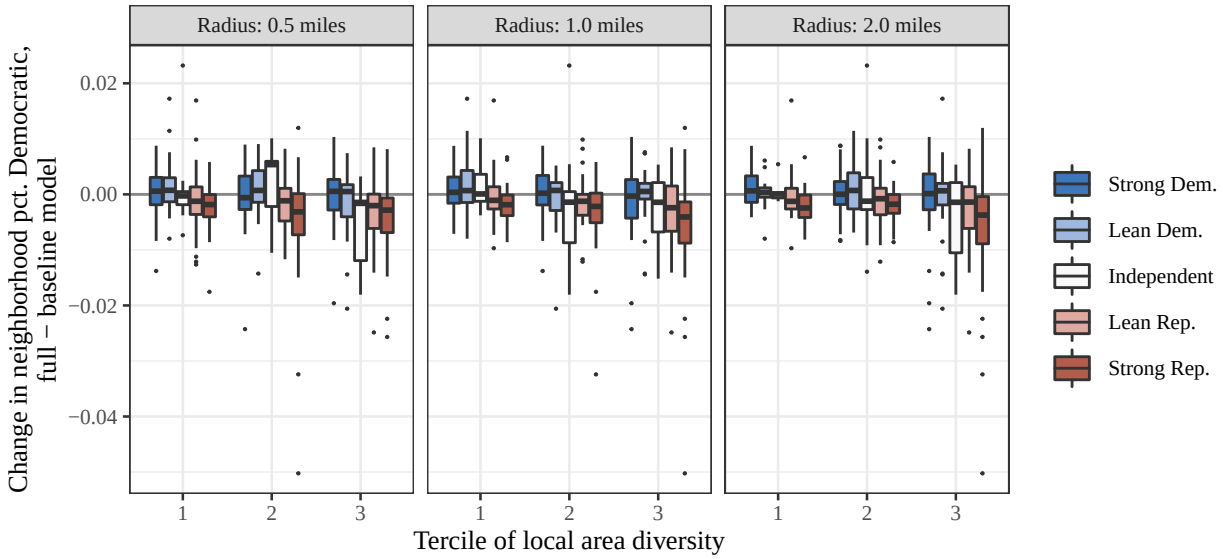


Figure C.4: Change in neighborhood fraction Democratic from baseline to full model, broken out by tercile of local area diversity, across a range of measurement radii.

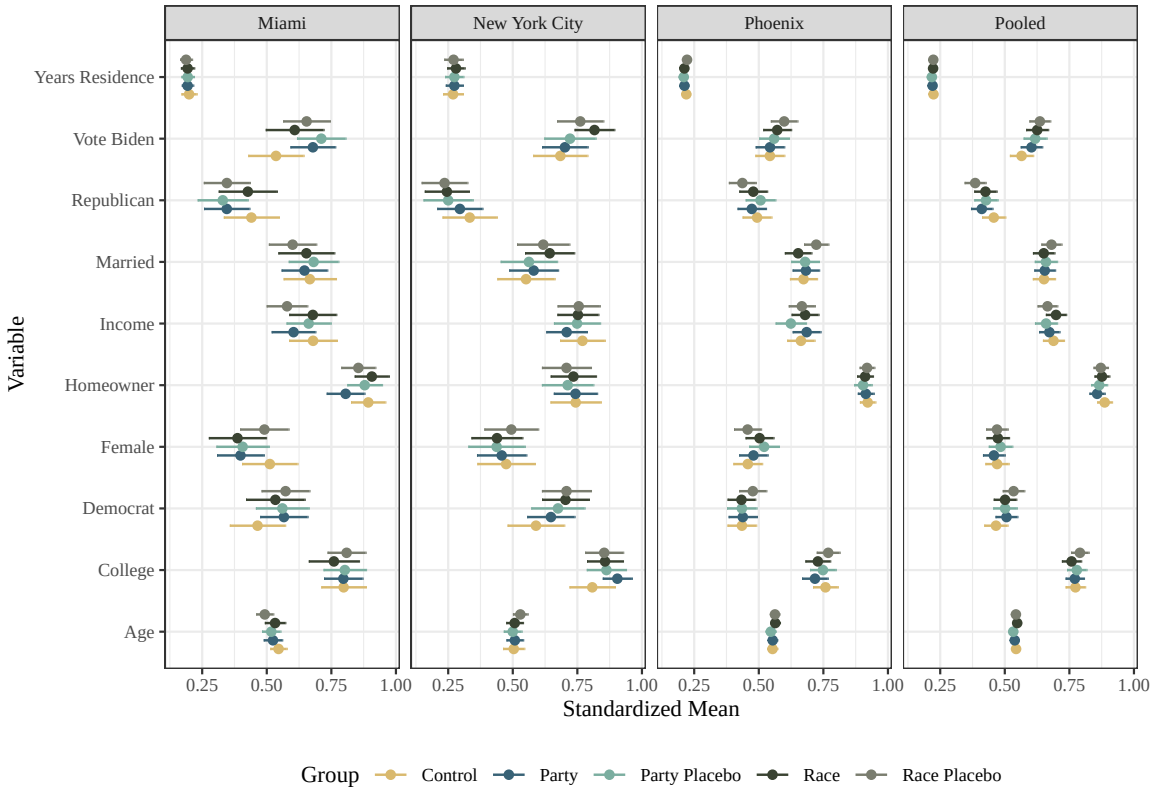


Figure C.5: Average levels of pre-treatment variables by treatment group. Variables are standardized to be between 0 and 1.

C.6 Experiment

Table C.7 contains posterior summaries for all model coefficients on the original model scale. These models were fit separately to each city:

- Miami: 474 survey respondents (91,569 individual block-level observations)
- New York: 453 survey respondents (85,113 individual block-level observations)
- Phoenix: 1,600 survey respondents (285,365 individual block-level observations)

Table C.6: Treatment Effect on Usable Neighborhoods

	Dependent variable: Usable Neighborhood		
	Miami	New York City	Phoenix
Party Condition	0.035	0.082	0.032
	(0.052)	(0.057)	(0.036)
Party Placebo Condition	0.065	−0.031	−0.007
	(0.055)	(0.058)	(0.037)
Race Condition	−0.004	0.094	0.045
	(0.056)	(0.059)	(0.037)
Race Placebo Condition	0.081	0.026	0.022
	(0.054)	(0.058)	(0.036)
Age	−0.004**	−0.0004	−0.004**
	(0.002)	(0.002)	(0.001)
College	0.020	0.045	0.047
	(0.043)	(0.050)	(0.027)
Democrat	0.118	−0.009	0.050
	(0.064)	(0.078)	(0.047)
Female	0.011	0.010	0.006
	(0.035)	(0.037)	(0.024)
Homeowner	0.015	−0.079	0.030
	(0.051)	(0.048)	(0.038)
Income	0.001	0.002**	0.0004
	(0.0004)	(0.0005)	(0.0003)
Married	0.045	−0.059	0.005
	(0.039)	(0.043)	(0.027)
Republican	0.141*	−0.033	0.016
	(0.067)	(0.079)	(0.044)
Vote Biden	0.109	0.071	0.093*
	(0.062)	(0.063)	(0.039)
Years Residence	0.002	0.003	0.002
	(0.002)	(0.002)	(0.001)
Constant	0.254*	0.168	0.418**
	(0.115)	(0.135)	(0.083)
Observations	1,468	1,193	4,028
R^2	0.007	0.005	0.0004
Adjusted R^2	0.005	0.001	−0.001
Residual Std. Error	0.467 (df = 1463)	0.485 (df = 1188)	0.490 (df = 4023)
F Statistic	2.735** (df = 4; 1463)	1.366 (df = 4; 1188)	0.430 (df = 4; 4023)
Note:			*p<0.05; **p<0.01

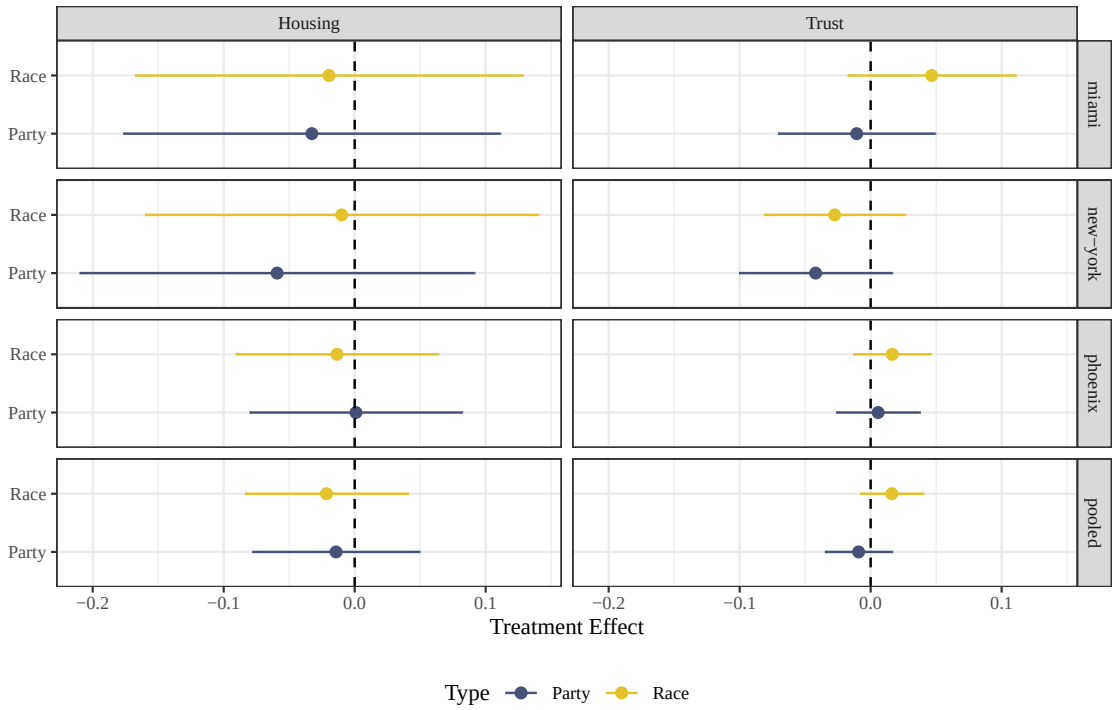


Figure C.6: Treatment effects on housing and trust survey outcomes.

Table C.7: Full experimental sample model estimates.

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
(Intercept)	Miami	-8.29	6.60	-19.23	-8.05	2.34
(Intercept)	NYC	-6.38	14.64	-30.90	-6.48	17.71
(Intercept)	Phoenix	-11.35	0.38	-11.98	-11.35	-10.72
church	Miami	0.01	0.03	-0.04	0.01	0.07
church	NYC	0.09	0.02	0.05	0.09	0.12
church	Phoenix	0.20	0.02	0.16	0.20	0.24
distance	Miami	0.06	0.01	0.04	0.06	0.07
distance	Miami	0.20	0.02	0.17	0.20	0.22
distance	NYC	0.05	0.01	0.04	0.05	0.07
distance	NYC	0.09	0.01	0.07	0.09	0.11
distance	Phoenix	0.12	0.01	0.11	0.12	0.13
distance	Phoenix	0.06	0.01	0.04	0.06	0.07
park	Miami	-0.12	0.04	-0.19	-0.12	-0.05
park	NYC	0.13	0.02	0.10	0.13	0.16
park	Phoenix	0.13	0.02	0.09	0.13	0.17
school	Miami	0.33	0.07	0.22	0.33	0.44
school	NYC	0.06	0.05	-0.02	0.06	0.15
school	Phoenix	0.06	0.04	0.00	0.06	0.12
children	Miami	1.40	2.39	-2.62	1.37	5.34
children	NYC	0.14	5.09	-8.19	0.33	8.49
children	Phoenix	-1.04	0.15	-1.28	-1.04	-0.79
same block group	Miami	0.01	0.03	-0.04	0.01	0.06
same block group	NYC	0.04	0.04	-0.02	0.04	0.11
same block group	Phoenix	-0.03	0.02	-0.06	-0.03	0.00
same tract	Miami	-0.22	0.05	-0.31	-0.22	-0.15
same tract	NYC	-0.04	0.06	-0.14	-0.03	0.06
same tract	Phoenix	-0.25	0.06	-0.36	-0.25	-0.15
same road region	Miami	-0.29	0.02	-0.33	-0.29	-0.26
same road region	NYC	0.00	0.02	-0.04	0.00	0.03
same road region	Phoenix	-0.18	0.02	-0.20	-0.18	-0.15
population	Miami	-1.63	0.12	-1.84	-1.62	-1.42
population	NYC	-0.92	0.09	-1.06	-0.92	-0.78
population	Phoenix	-1.70	0.10	-1.86	-1.70	-1.55
area	Miami	-0.10	0.03	-0.16	-0.10	-0.05
area	NYC	-0.05	0.05	-0.13	-0.05	0.04
area	Phoenix	-0.17	0.02	-0.21	-0.17	-0.13
age	Miami	0.01	0.11	-0.18	0.01	0.20
age	NYC	0.01	0.12	-0.18	0.00	0.20
age	Phoenix	0.01	0.01	0.01	0.01	0.02
education = No College	Miami	-3.14	5.77	-12.69	-3.16	6.34
education = No College	NYC	-1.11	8.77	-15.37	-1.20	13.02
education = No College	Phoenix	-1.10	1.09	-2.89	-1.10	0.69
retired	Miami	0.20	2.80	-4.25	0.16	4.62
retired	NYC	0.01	2.56	-4.12	0.02	4.27

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
retired	Phoenix	-0.04	0.12	-0.24	-0.04	0.15
tenure	Miami	-0.04	0.89	-1.49	-0.02	1.43
tenure	NYC	-0.02	1.19	-1.99	-0.01	1.85
tenure	Phoenix	0.02	0.02	-0.02	0.02	0.06
party = IND	Miami	-0.09	8.24	-13.59	-0.22	13.76
party = IND	NYC	-0.10	8.04	-12.76	-0.13	13.05
party = IND	Phoenix	-0.39	0.40	-1.05	-0.39	0.27
party = REP	Miami	-0.29	5.11	-8.70	-0.26	8.24
party = REP	NYC	-0.94	7.34	-12.86	-0.73	11.15
party = REP	Phoenix	-0.06	0.21	-0.40	-0.06	0.28
minority	Miami	0.20	5.48	-8.71	0.18	8.97
minority	NYC	0.14	4.60	-7.61	0.18	7.49
minority	Phoenix	-0.21	0.29	-0.70	-0.21	0.27
homeowner	Miami	1.28	4.66	-6.50	1.56	8.70
homeowner	NYC	-0.28	5.32	-9.05	-0.22	8.73
homeowner	Phoenix	-0.19	0.31	-0.70	-0.19	0.32
school * children	Miami	-0.50	0.11	-0.67	-0.50	-0.32
school * children	NYC	-0.10	0.08	-0.22	-0.10	0.03
school * children	Phoenix	0.29	0.07	0.17	0.29	0.40
children * distance	Miami	-0.16	0.02	-0.19	-0.16	-0.11
children * distance	NYC	-0.07	0.02	-0.10	-0.07	-0.04
children * distance	Phoenix	0.17	0.02	0.14	0.17	0.20
same tract * same road region	Miami	0.11	0.05	0.03	0.11	0.19
same tract * same road region	NYC	-0.17	0.06	-0.28	-0.17	-0.07
same tract * same road region	Phoenix	-0.17	0.06	-0.27	-0.17	-0.07
Fraction same race * C group	Miami	-0.38	0.06	-0.48	-0.38	-0.28
Fraction same race * C group	NYC	-0.29	0.05	-0.37	-0.29	-0.21
Fraction same race * C group	Phoenix	-0.44	0.03	-0.49	-0.44	-0.39
Fraction same race * P group	Miami	-0.42	0.06	-0.51	-0.42	-0.33
Fraction same race * P group	NYC	-0.54	0.05	-0.63	-0.54	-0.45
Fraction same race * P group	Phoenix	-0.32	0.03	-0.37	-0.32	-0.28
Fraction same race * PH group	Miami	-0.17	0.06	-0.26	-0.17	-0.07
Fraction same race * PH group	NYC	-0.18	0.05	-0.26	-0.19	-0.10
Fraction same race * PH group	Phoenix	-0.46	0.04	-0.52	-0.46	-0.40
Fraction same race * R group	Miami	-0.28	0.06	-0.38	-0.28	-0.18
Fraction same race * R group	NYC	-0.43	0.06	-0.53	-0.43	-0.35
Fraction same race * R group	Phoenix	-0.40	0.03	-0.45	-0.40	-0.35
Fraction same race * RH group	Miami	-0.37	0.06	-0.46	-0.37	-0.28
Fraction same race * RH group	NYC	-0.45	0.05	-0.52	-0.45	-0.36
Fraction same race * RH group	Phoenix	-0.32	0.03	-0.37	-0.32	-0.28
C group * Fraction same party	Miami	-0.22	0.10	-0.38	-0.22	-0.07
C group * Fraction same party	NYC	-0.48	0.07	-0.59	-0.48	-0.36
C group * Fraction same party	Phoenix	-0.42	0.09	-0.56	-0.42	-0.28
P group * Fraction same party	Miami	-0.07	0.08	-0.20	-0.07	0.07
P group * Fraction same party	NYC	-0.32	0.07	-0.42	-0.32	-0.21

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
P group * Fraction same party	Phoenix	-0.27	0.08	-0.40	-0.27	-0.15
PH group * Fraction same party	Miami	0.09	0.08	-0.04	0.09	0.22
PH group * Fraction same party	NYC	-0.35	0.06	-0.46	-0.35	-0.25
PH group * Fraction same party	Phoenix	-0.15	0.09	-0.29	-0.15	-0.01
R group * Fraction same party	Miami	-0.18	0.07	-0.31	-0.18	-0.06
R group * Fraction same party	NYC	-0.39	0.06	-0.49	-0.39	-0.29
R group * Fraction same party	Phoenix	-0.58	0.08	-0.71	-0.58	-0.45
RH group * Fraction same party	Miami	-0.31	0.07	-0.43	-0.31	-0.19
RH group * Fraction same party	NYC	-0.22	0.06	-0.32	-0.22	-0.12
RH group * Fraction same party	Phoenix	-0.21	0.07	-0.32	-0.21	-0.10
C group * Fraction same ownership	Miami	-0.43	0.17	-0.71	-0.43	-0.14
C group * Fraction same ownership	NYC	-0.43	0.14	-0.66	-0.43	-0.19
C group * Fraction same ownership	Phoenix	-0.40	0.13	-0.62	-0.40	-0.18
P group * Fraction same ownership	Miami	0.20	0.13	-0.01	0.19	0.42
P group * Fraction same ownership	NYC	-0.43	0.14	-0.65	-0.42	-0.20
P group * Fraction same ownership	Phoenix	0.34	0.14	0.11	0.34	0.58
PH group * Fraction same ownership	Miami	-0.64	0.24	-1.03	-0.64	-0.23
PH group * Fraction same ownership	NYC	-0.42	0.11	-0.62	-0.42	-0.25
PH group * Fraction same ownership	Phoenix	0.16	0.13	-0.04	0.16	0.37
R group * Fraction same ownership	Miami	1.36	0.76	0.14	1.37	2.62
R group * Fraction same ownership	NYC	-0.08	0.13	-0.30	-0.08	0.14
R group * Fraction same ownership	Phoenix	0.41	0.15	0.17	0.41	0.66
RH group * Fraction same ownership	Miami	-0.29	0.23	-0.64	-0.28	0.07
RH group * Fraction same ownership	NYC	-0.07	0.12	-0.27	-0.07	0.11
RH group * Fraction same ownership	Phoenix	-0.55	0.17	-0.82	-0.55	-0.27
C group * Fraction same education	Miami	0.01	0.17	-0.27	0.00	0.29
C group * Fraction same education	NYC	-0.55	0.10	-0.73	-0.55	-0.37
C group * Fraction same education	Phoenix	-1.37	0.09	-1.52	-1.37	-1.22
P group * Fraction same education	Miami	-0.22	0.13	-0.44	-0.22	0.00
P group * Fraction same education	NYC	0.13	0.09	-0.02	0.13	0.27
P group * Fraction same education	Phoenix	-1.07	0.09	-1.21	-1.07	-0.92
PH group * Fraction same education	Miami	0.19	0.16	-0.09	0.19	0.46
PH group * Fraction same education	NYC	-0.42	0.09	-0.56	-0.42	-0.28
PH group * Fraction same education	Phoenix	-0.18	0.10	-0.36	-0.18	-0.01
R group * Fraction same education	Miami	-0.23	0.16	-0.50	-0.24	0.03
R group * Fraction same education	NYC	0.15	0.08	0.02	0.15	0.28
R group * Fraction same education	Phoenix	-1.13	0.10	-1.29	-1.13	-0.96
RH group * Fraction same education	Miami	-0.80	0.13	-1.02	-0.80	-0.57
RH group * Fraction same education	NYC	-0.07	0.09	-0.21	-0.07	0.07
RH group * Fraction same education	Phoenix	-0.80	0.08	-0.94	-0.80	-0.67
C group * income	Miami	-0.19	0.06	-0.29	-0.19	-0.08
C group * income	NYC	-0.13	0.04	-0.20	-0.13	-0.06
C group * income	Phoenix	0.19	0.03	0.14	0.19	0.24
P group * income	Miami	0.06	0.06	-0.03	0.07	0.15
P group * income	NYC	-0.16	0.05	-0.24	-0.16	-0.08

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
P group * income	Phoenix	0.18	0.03	0.13	0.18	0.23
PH group * income	Miami	-0.07	0.06	-0.16	-0.07	0.03
PH group * income	NYC	-0.10	0.04	-0.17	-0.10	-0.04
PH group * income	Phoenix	0.14	0.03	0.09	0.14	0.19
R group * income	Miami	-0.19	0.06	-0.28	-0.18	-0.09
R group * income	NYC	-0.14	0.04	-0.21	-0.14	-0.07
R group * income	Phoenix	0.15	0.03	0.09	0.15	0.20
RH group * income	Miami	0.05	0.05	-0.03	0.05	0.13
RH group * income	NYC	-0.20	0.05	-0.28	-0.21	-0.13
RH group * income	Phoenix	0.17	0.03	0.12	0.17	0.21
education = No College * P group	Miami	2.38	7.50	-10.15	2.49	14.74
education = No College * P group	NYC	-7.46	10.26	-24.47	-7.60	9.96
education = No College * P group	Phoenix	-0.74	1.46	-3.15	-0.74	1.66
education = No College * PH group	Miami	5.25	8.33	-8.54	5.42	18.77
education = No College * PH group	NYC	2.05	24.83	-37.91	1.97	43.38
education = No College * PH group	Phoenix	9.30	1.65	6.59	9.30	12.02
education = No College * R group	Miami	-4.96	8.90	-19.49	-4.99	9.13
education = No College * R group	NYC	-5.89	15.22	-30.60	-5.88	18.73
education = No College * R group	Phoenix	3.87	1.44	1.50	3.87	6.25
education = No College * RH group	Miami	12.14	8.40	-1.81	12.40	26.00
education = No College * RH group	NYC	1.58	7.59	-10.40	1.50	14.16
education = No College * RH group	Phoenix	-3.58	1.36	-5.81	-3.58	-1.34
party = IND * P group	Miami	-0.18	11.03	-18.72	-0.25	18.02
party = IND * P group	NYC	0.67	14.33	-21.86	1.03	23.47
party = IND * P group	Phoenix	0.73	0.54	-0.15	0.73	1.61
party = REP * P group	Miami	0.18	6.73	-10.87	0.45	10.79
party = REP * P group	NYC	0.94	9.10	-13.17	0.45	16.11
party = REP * P group	Phoenix	0.11	0.29	-0.36	0.11	0.58
party = IND * PH group	Miami	1.68	10.99	-16.29	2.13	19.37
party = IND * PH group	NYC	1.36	17.49	-27.25	1.52	30.11
party = IND * PH group	Phoenix	1.03	0.60	0.04	1.03	2.02
party = REP * PH group	Miami	0.93	7.29	-11.68	0.96	12.80
party = REP * PH group	NYC	0.01	10.42	-17.69	-0.07	16.80
party = REP * PH group	Phoenix	-0.14	0.29	-0.63	-0.14	0.34
party = IND * R group	Miami	1.01	15.51	-23.95	0.85	27.06
party = IND * R group	NYC	-0.13	12.50	-20.35	-0.05	20.44
party = IND * R group	Phoenix	1.11	0.54	0.23	1.11	1.99
party = REP * R group	Miami	0.80	7.54	-11.92	0.85	12.46
party = REP * R group	NYC	0.73	8.40	-13.12	0.74	14.45
party = REP * R group	Phoenix	0.15	0.29	-0.32	0.15	0.62
party = IND * RH group	Miami	1.22	11.27	-17.23	1.05	20.33
party = IND * RH group	NYC	-0.39	12.43	-20.99	-0.17	19.71
party = IND * RH group	Phoenix	-0.02	0.53	-0.89	-0.02	0.84
party = REP * RH group	Miami	0.64	6.74	-10.60	0.47	11.83
party = REP * RH group	NYC	1.18	9.91	-14.83	1.35	17.98

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
party = REP * RH group	Phoenix	0.29	0.28	-0.18	0.29	0.75
minority * P group	Miami	-0.84	6.73	-12.01	-0.85	10.09
minority * P group	NYC	0.10	6.04	-9.86	-0.01	9.76
minority * P group	Phoenix	0.26	0.42	-0.42	0.26	0.95
minority * PH group	Miami	-0.38	7.02	-11.59	-0.08	11.17
minority * PH group	NYC	-0.29	7.98	-13.06	-0.05	12.28
minority * PH group	Phoenix	-0.08	0.42	-0.76	-0.08	0.61
minority * R group	Miami	0.14	7.97	-13.60	0.18	13.33
minority * R group	NYC	0.00	6.52	-11.03	0.13	10.28
minority * R group	Phoenix	0.06	0.41	-0.61	0.06	0.73
minority * RH group	Miami	-0.62	6.84	-12.33	-0.62	11.06
minority * RH group	NYC	-0.24	6.73	-11.44	-0.17	11.15
minority * RH group	Phoenix	0.24	0.39	-0.40	0.24	0.87
homeowner * P group	Miami	-1.78	4.59	-9.64	-1.75	5.68
homeowner * P group	NYC	0.39	4.37	-6.57	0.45	7.23
homeowner * P group	Phoenix	0.44	0.40	-0.21	0.44	1.10
homeowner * PH group	Miami	-1.33	4.74	-9.25	-1.51	6.75
homeowner * PH group	NYC	-0.57	7.94	-13.87	-0.48	12.29
homeowner * PH group	Phoenix	0.43	0.42	-0.25	0.43	1.11
homeowner * R group	Miami	-0.56	5.21	-9.18	-0.56	8.22
homeowner * R group	NYC	0.17	4.27	-6.73	0.07	7.30
homeowner * R group	Phoenix	0.51	0.40	-0.15	0.51	1.17
homeowner * RH group	Miami	-2.20	4.66	-9.89	-2.14	5.30
homeowner * RH group	NYC	0.68	4.86	-7.72	0.63	8.83
homeowner * RH group	Phoenix	-0.03	0.40	-0.68	-0.03	0.63
minority * Fraction same race * C group	Miami	0.43	0.11	0.26	0.43	0.60
minority * Fraction same race * C group	NYC	0.44	0.14	0.23	0.44	0.67
minority * Fraction same race * C group	Phoenix	-0.19	0.27	-0.63	-0.19	0.25
minority * Fraction same race * P group	Miami	0.35	0.11	0.17	0.34	0.53
minority * Fraction same race * P group	NYC	0.34	0.15	0.10	0.34	0.58
minority * Fraction same race * P group	Phoenix	0.34	0.23	-0.05	0.34	0.72
minority * Fraction same race * PH group	Miami	-0.10	0.09	-0.25	-0.10	0.05
minority * Fraction same race * PH group	NYC	0.12	0.12	-0.06	0.12	0.31
minority * Fraction same race * PH group	Phoenix	0.34	0.13	0.13	0.34	0.55
minority * Fraction same race * R group	Miami	0.16	0.12	-0.04	0.16	0.36
minority * Fraction same race * R group	NYC	0.45	0.19	0.14	0.45	0.77
minority * Fraction same race * R group	Phoenix	0.66	0.13	0.44	0.66	0.88
minority * Fraction same race * RH group	Miami	0.17	0.09	0.01	0.17	0.33
minority * Fraction same race * RH group	NYC	1.40	0.17	1.13	1.40	1.68
minority * Fraction same race * RH group	Phoenix	0.11	0.14	-0.11	0.11	0.34
party = IND * C group * Fraction same party	Miami	-0.19	0.24	-0.58	-0.20	0.20
party = IND * C group * Fraction same party	NYC	0.05	0.27	-0.40	0.06	0.51
party = IND * C group * Fraction same party	Phoenix	0.43	0.18	0.13	0.43	0.73
party = REP * C group * Fraction same party	Miami	-0.37	0.16	-0.63	-0.37	-0.11
party = REP * C group * Fraction same party	NYC	0.25	0.14	0.02	0.24	0.48

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
party = REP * C group * Fraction same party	Phoenix	0.07	0.10	-0.09	0.07	0.24
party = IND * P group * Fraction same party	Miami	-0.41	0.22	-0.78	-0.42	-0.04
party = IND * P group * Fraction same party	NYC	-0.40	0.58	-1.40	-0.39	0.50
party = IND * P group * Fraction same party	Phoenix	-0.07	0.17	-0.35	-0.07	0.21
party = REP * P group * Fraction same party	Miami	-0.09	0.15	-0.32	-0.09	0.14
party = REP * P group * Fraction same party	NYC	-0.33	0.19	-0.64	-0.34	-0.02
party = REP * P group * Fraction same party	Phoenix	-0.01	0.09	-0.15	-0.01	0.14
party = IND * PH group * Fraction same party	Miami	-0.17	0.34	-0.73	-0.17	0.39
party = IND * PH group * Fraction same party	NYC	0.56	0.30	0.08	0.55	1.04
party = IND * PH group * Fraction same party	Phoenix	-0.17	0.29	-0.64	-0.17	0.30
party = REP * PH group * Fraction same party	Miami	-0.88	0.17	-1.16	-0.88	-0.61
party = REP * PH group * Fraction same party	NYC	0.60	0.15	0.35	0.60	0.85
party = REP * PH group * Fraction same party	Phoenix	-0.32	0.10	-0.49	-0.32	-0.15
party = IND * R group * Fraction same party	Miami	0.27	0.81	-1.04	0.25	1.63
party = IND * R group * Fraction same party	NYC	-0.28	0.42	-0.99	-0.27	0.38
party = IND * R group * Fraction same party	Phoenix	0.45	0.25	0.04	0.45	0.86
party = REP * R group * Fraction same party	Miami	-0.13	0.15	-0.39	-0.13	0.11
party = REP * R group * Fraction same party	NYC	-0.04	0.17	-0.31	-0.03	0.23
party = REP * R group * Fraction same party	Phoenix	0.25	0.09	0.10	0.25	0.41
party = IND * RH group * Fraction same party	Miami	-0.02	0.31	-0.53	-0.01	0.49
party = IND * RH group * Fraction same party	NYC	-0.21	0.24	-0.60	-0.21	0.18
party = IND * RH group * Fraction same party	Phoenix	-0.70	0.16	-0.96	-0.70	-0.44
party = REP * RH group * Fraction same party	Miami	-0.49	0.15	-0.75	-0.49	-0.24
party = REP * RH group * Fraction same party	NYC	-0.07	0.22	-0.44	-0.07	0.28
party = REP * RH group * Fraction same party	Phoenix	-0.31	0.09	-0.45	-0.31	-0.17
homeowner * C group * Fraction same ownership	Miami	0.80	0.21	0.46	0.80	1.13
homeowner * C group * Fraction same ownership	NYC	0.72	0.18	0.44	0.71	1.02
homeowner * C group * Fraction same ownership	Phoenix	0.32	0.15	0.08	0.32	0.56
homeowner * P group * Fraction same ownership	Miami	-0.32	0.17	-0.61	-0.32	-0.05
homeowner * P group * Fraction same ownership	NYC	0.47	0.17	0.20	0.47	0.75
homeowner * P group * Fraction same ownership	Phoenix	-0.60	0.15	-0.85	-0.60	-0.35
homeowner * PH group * Fraction same ownership	Miami	0.39	0.26	-0.03	0.39	0.82
homeowner * PH group * Fraction same ownership	NYC	0.50	0.14	0.28	0.50	0.72
homeowner * PH group * Fraction same ownership	Phoenix	-0.14	0.14	-0.38	-0.14	0.09
homeowner * R group * Fraction same ownership	Miami	-0.97	0.77	-2.23	-0.98	0.27
homeowner * R group * Fraction same ownership	NYC	0.11	0.17	-0.15	0.11	0.39
homeowner * R group * Fraction same ownership	Phoenix	-0.46	0.16	-0.73	-0.46	-0.19
homeowner * RH group * Fraction same ownership	Miami	0.29	0.25	-0.10	0.29	0.69
homeowner * RH group * Fraction same ownership	NYC	0.11	0.15	-0.14	0.11	0.36
homeowner * RH group * Fraction same ownership	Phoenix	0.47	0.18	0.18	0.47	0.76
education = No College * C group * Fraction same educ	Miami	0.82	0.31	0.30	0.83	1.33
education = No College * C group * Fraction same educ	NYC	1.54	0.23	1.17	1.54	1.90
education = No College * C group * Fraction same educ	Phoenix	1.45	0.23	1.07	1.45	1.84
education = No College * P group * Fraction same educ	Miami	0.19	0.34	-0.37	0.20	0.73
education = No College * P group * Fraction same educ	NYC	1.25	0.37	0.64	1.25	1.84

Table C.7 (Continued)

Coefficient	City	Mean	Std. Dev.	Q5	Median	Q95
education = No College * P group * Fraction same educ	Phoenix	1.76	0.23	1.38	1.76	2.13
education = No College * PH group * Fraction same educ	Miami	0.02	0.35	-0.55	0.01	0.57
education = No College * PH group * Fraction same educ	NYC	1.98	0.34	1.42	1.98	2.53
education = No College * PH group * Fraction same educ	Phoenix	1.13	0.29	0.66	1.13	1.60
education = No College * R group * Fraction same educ	Miami	2.02	0.35	1.46	2.03	2.60
education = No College * R group * Fraction same educ	NYC	-0.58	0.32	-1.12	-0.58	-0.07
education = No College * R group * Fraction same educ	Phoenix	0.42	0.22	0.05	0.42	0.79
education = No College * RH group * Fraction same educ	Miami	-0.50	0.36	-1.08	-0.51	0.12
education = No College * RH group * Fraction same educ	NYC	0.50	0.26	0.09	0.50	0.93
education = No College * RH group * Fraction same educ	Phoenix	1.46	0.18	1.16	1.46	1.76
education = No College * C group * income	Miami	0.26	0.11	0.08	0.26	0.44
education = No College * C group * income	NYC	-0.01	0.09	-0.16	-0.01	0.14
education = No College * C group * income	Phoenix	0.07	0.09	-0.07	0.07	0.22
education = No College * P group * income	Miami	0.02	0.13	-0.20	0.02	0.23
education = No College * P group * income	NYC	0.66	0.15	0.40	0.66	0.90
education = No College * P group * income	Phoenix	0.07	0.08	-0.06	0.07	0.21
education = No College * PH group * income	Miami	-0.22	0.12	-0.42	-0.21	-0.02
education = No College * PH group * income	NYC	-0.09	0.17	-0.36	-0.09	0.19
education = No College * PH group * income	Phoenix	-0.76	0.10	-0.93	-0.76	-0.59
education = No College * R group * income	Miami	0.66	0.13	0.45	0.65	0.86
education = No College * R group * income	NYC	0.69	0.14	0.47	0.69	0.92
education = No College * R group * income	Phoenix	-0.25	0.08	-0.38	-0.25	-0.12
education = No College * RH group * income	Miami	-0.74	0.13	-0.96	-0.75	-0.53
education = No College * RH group * income	NYC	-0.10	0.11	-0.28	-0.10	0.08
education = No College * RH group * income	Phoenix	0.35	0.07	0.24	0.35	0.46
alpha	Miami	1.12	0.02	1.09	1.12	1.14
alpha	NYC	1.27	0.02	1.24	1.27	1.30
alpha	Phoenix	1.24	0.01	1.22	1.24	1.25

This page intentionally left blank.

D

Appendices for Chapter 4

D.1 Proofs of Propositions

Theorem 4.3 (Nonparametric Identification)**.** *For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, define a matrix* $\mathbf{P}$ *with entries* $p_{rs} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ *and a vector* $\mathbf{b}$ *with entries* $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$*. Then under Assumption A4, and assuming knowledge of the joint distribution* $\mathbb{P}(R, G, X, S)$*, the conditional probabilities* $\mathbb{P}(Y = y \mid R, G = g, X = x)$ *are identified if and only if both* $\mathbf{P}$ *and the augmented matrix* $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$ *have rank* $|\mathcal{R}|$*.*

Proof. Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y \mid R, G, X$, we have, for all $y \in \mathcal{Y}, g \in \mathcal{G}, x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned}\mathbb{P}(Y = y \mid G = g, X = x, S = s) &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x, S = s)\, \mathbb{P}(R = r \mid G = g, X = x, S = s) \\ &= \sum_{r \in \mathcal{R}} \mathbb{P}(Y = y \mid R = r, G = g, X = x)\, \mathbb{P}(R = r \mid G = g, X = x, S = s).\end{aligned}$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, G = g, X = x)$. These parameters are identified if and only if this system has a unique solution, i.e. if the coefficient matrix $\mathbf{P}$ has rank $|\mathcal{R}|$ and so does the augmented matrix $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$. □

Theorem 4.4 (Unbiasedness of OLS Estimator)**.** *If Assumptions A1, A2, and A4 hold, and the identification conditions in Theorem 4.3 are satisfied, then for all $y \in \mathcal{Y}$ and $r \in \mathcal{R}$,*

$$\mathbb{E}[\hat{\mu}^{(p\text{-}ols)}_{Y|R}(y \mid r)] = \mathbb{P}(Y = y \mid R = r).$$

Proof. Fix $y \in \mathcal{Y}$ and define $m_{gxr} = \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x)]$. Then under Assumptions A1, A2, and A4,

$$\begin{aligned}\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, S = s] &= \sum_{r \in \mathcal{R}} \mathbb{E}[\mathbf{1}\{Y = y\} \mid R = r, G = g, X = x)]\, \mathbb{P}(R = r \mid G = g, X = x, S = s) \\ &= \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r,\end{aligned}$$

where as in the main text $\hat{\mathbf{p}}$ is the (random) vector of BISG probabilities. In fact, since the right-hand side depends on S only through $\hat{\mathbf{p}}$, we have

$$\mathbb{E}[\mathbf{1}\{Y = y\} \mid G = g, X = x, \hat{\mathbf{p}}] = \sum_{r \in \mathcal{R}} m_{gxr} \hat{p}_r.$$

So the conditional expectation of $\mathbf{1}\{Y = y\}$ given X, G, and the BISG probabilities $\hat{\mathbf{p}}$ is linear in those probabilities, with coefficients m_{gxr}. Consequently, the OLS estimate $\hat{\boldsymbol{\mu}}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ will be unbiased for m_{gxr}, by the standard results.

Now, we can expand $\mathbb{P}(Y = y \mid R = r)$ as

$$\mathbb{P}(Y = y \mid R = r) = \sum_{x \in \mathcal{X}, g \in \mathcal{G}} \mathbb{P}(Y = y \mid R = r, G = g, X = x)\, \mathbb{P}(G = g, X = x \mid R = r) = \sum_{x \in \mathcal{X}, g \in \mathcal{G}} m_{gxr} q_{gx|r}.$$

Since $\hat{\boldsymbol{\mu}}^{(\text{ols})}_{Y|RGX}(y \mid \cdot, g, x)$ is unbiased for m_{gxr}, by the linearity of expectation the poststratified estimator $\hat{\mu}^{(\text{p-ols})}_{Y|R}(y \mid r)$ is unbiased for $\mathbb{P}(Y = y \mid R = r)$. □

Theorem 4.5 (Necessary and Sufficient Condition for Equality of the Weighting and OLS Estimators)**.** *For any $y \in \mathcal{Y}$, $g \in \mathcal{G}$ and $x \in \mathcal{X}$, within the set of individuals with $G_i = g$ and $X_i = x$, we have that $\hat{\boldsymbol{\mu}}^{(wtd)}_{Y|R}(y \mid \cdot) = \hat{\boldsymbol{\mu}}^{(ols)}_{Y|R}(y \mid \cdot)$ if and only if for every pair $j, k \in \mathcal{R}$, either the BISG probabilities perfectly discriminate (i.e., $\mathbb{P}(R_i = j \mid G_i, X_i, S_i) > 0$ implies $\mathbb{P}(R_i = k \mid G_i, X_i, S_i) = 0$ and vice versa) or $\hat{\mu}^{(wtd)}_{Y|R}(y \mid j) = \hat{\mu}^{(wtd)}_{Y|R}(y \mid k)$.*

Proof. Fix a $y \in \mathcal{Y}, g \in \mathcal{G}$ and $x \in \mathcal{X}$. The weighting estimator of $\mathbb{P}(Y = y \mid R = r)$ within the set of individuals with $G_i = g$ and $X_i = x$ may be written

$$\hat{\mu}^{(\text{wtd})}_{Y|RGX}(y \mid r, g, x) = \frac{\hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)r} \mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\}}{\hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)r} \mathbf{1}} = \frac{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\}) \right\|}{\left\| \text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}}(\mathbf{1}) \right\|},$$

the ratio of the projected length of the outcome vector $\mathbf{1}\{\mathbf{Y} = y\}$ and the constant vector $\mathbf{1}$ onto $\hat{\mathbf{P}}_{\cdot r}$. We can write the OLS estimator as

$$\hat{\boldsymbol{\mu}}^{(\text{ols})}_{Y|R} = (\hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)} \hat{\mathbf{P}}_{\mathcal{I}(xg)})^{-1} \hat{\mathbf{P}}^{\top}_{\mathcal{I}(xg)} \mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\} = \text{coord}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)}}(\text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\}),$$

where $\text{coord}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)}}$ is the function that returns the coordinates of its input vector in the $\hat{\mathbf{P}}_{\mathcal{I}(xg)}$ basis (by assumption $\hat{\mathbf{P}}_{\mathcal{I}(xg)}$ has rank $|\mathcal{R}|$ and so its columns are linearly independent). To make the comparison even easier, notice that we can break the projection $\text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}}$ into two steps, writing it instead as $\text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}} = \text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}} \circ \text{proj}_{\hat{\mathbf{P}}}$. Letting $\mathbf{Y}_{\text{proj}} = \text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)}}(\mathbf{1}\{\mathbf{Y}_{\mathcal{I}(xg)} = y\})$, then, we can rewrite our

estimators as

$$\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid r) = \frac{\left\|\text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}}(\mathbf{Y}_{\text{proj}})\right\|}{\left\|\text{proj}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)r}}(\mathbf{1})\right\|} \quad \text{and} \quad \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid r) = \text{coord}_{\hat{\mathbf{P}}_{\mathcal{I}(xg)}}(\mathbf{Y}_{\text{proj}})_r.$$

Now, since the individual BISG probabilities are nonnegative and sum to 1, a pair $j, k \in \mathcal{R}$ of races has perfectly discriminating BISG probabilities if and only if the corresponding columns of $\hat{\mathbf{P}}$ are orthogonal, i.e., $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = I$. Begin by writing $\mathbf{Y}_{\text{proj}}$ in terms of the $\hat{\mathbf{P}}_{\mathcal{I}(xg)}$ basis, so

$$\mathbf{Y}_{\text{proj}} = \sum_{j\in\mathcal{R}} c_j \hat{\mathbf{P}}_{\mathcal{I}(xg)j},$$

and thus $\hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j) = c_j$. Without loss of generality, suppose the c_j are numbered as $c_1 \geq c_2 \geq \cdots \geq c_{|\mathcal{R}|}$. We can also expand $\mathbf{1}$ in the same basis. Since the individual probabilities must sum to one, in fact we have $\mathbf{1} = \sum_{j\in\mathcal{R}} \hat{\mathbf{P}}_{\mathcal{I}(xg)j}$.

For the forward direction, we assume $\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid j) = \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j) = c_j$; multiplying out the denominator of the weighting estimator, we have $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\mathbf{Y} = c_j\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\mathbf{1}$ for all j; substituting the basis expansions of $\mathbf{Y}_{\text{proj}}$ and $\mathbf{1}$, this yields

$$\sum_{k\in\mathcal{R}} c_k\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = \sum_{k\in\mathcal{R}} c_j\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k}, \quad \text{so} \quad \sum_{k\in\mathcal{R}}(c_j - c_k)\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = 0.$$

Now fix $j \in J_1 = \arg\max_j c_j$; this relation still holds, but now every term in the sum is nonnegative and in particular $c_j > c_k$ for all $k \notin J_1$. Therefore we must have $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = 0$ for all $k \notin J_1$. Then fix $j \in J_2 = \arg\max_{j\notin J_1} c_j$; since $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)l} = 0$ for all $l \in J_1$, every term in the sum is still nonnegative and in particular $c_j > c_k$ for all $k \notin J_1 \cup J_2$. Therefore we must have $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = 0$ for all $k \notin J_1 \cup J_2$. Proceeding this way through all sets of common values in the c_j we find that for all $j, k \in \mathcal{R}$, either $c_j = c_k$ or $\hat{\mathbf{P}}^\top_{\mathcal{I}(xg)j}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = 0$.

For the reverse direction, fix $j \in \mathcal{R}$ and let $J = \{k \in \mathcal{R} : c_k = c_j\}$, so that by assumption $\hat{\mathbf{P}}_{\mathcal{I}(xg)j}^{\top}\hat{\mathbf{P}}_{\mathcal{I}(xg)k} = 0$ for all $k \notin J$. Then by the above basis expansion, $\hat{\mu}^{(\text{ols})}_{Y|RGX}(y \mid j) = c_j$, and

$$\hat{\mu}^{(\text{wtd})}_{Y|R}(y \mid j) = \frac{\sum_{k\in\mathcal{R}} c_k \hat{\mathbf{P}}_{\mathcal{I}(xg)j}^{\top}\hat{\mathbf{P}}_{\mathcal{I}(xg)k}}{\sum_{k\in\mathcal{R}} \hat{\mathbf{P}}_{\mathcal{I}(xg)j}^{\top}\hat{\mathbf{P}}_{\mathcal{I}(xg)k}} = \frac{c_j \sum_{k\in J} \hat{\mathbf{P}}_{\mathcal{I}(xg)j}^{\top}\hat{\mathbf{P}}_{\mathcal{I}(xg)k}}{\sum_{k\in J} \hat{\mathbf{P}}_{\mathcal{I}(xg)j}^{\top}\hat{\mathbf{P}}_{\mathcal{I}(xg)k}} = c_j = \hat{\mu}^{(\text{ols})}_{Y|R}(y \mid j). \qquad \square$$

Theorem 4.6 (Nonparametric Identification Under Assumption A6)**.** *Let $f : \mathcal{S} \to \mathbb{R}^d$, $d < |\mathcal{S}|$, with range $f(\mathcal{S})$. For any given $g \in \mathcal{G}$, $x \in \mathcal{X}$, $z \in f(\mathcal{S})$, and $y \in \mathcal{Y}$, define a matrix $\mathbf{P}$ with entries $p_{rs} = \mathbb{P}(R = r \mid G = g, X = x, S = s)$ and a vector $\mathbf{b}$ with entries $b_s = \mathbb{P}(Y = y \mid G = g, X = x, S = s)$. Then under Assumption A6, and assuming knowledge of the joint distribution $\mathbb{P}(R, G, X, S)$, the conditional probabilities $\mathbb{P}(Y = y \mid R, f(S) = z, G = g, X = x)$ are identified if and only if both $\mathbf{P}$ and the augmented matrix $\begin{pmatrix}\mathbf{P} & \mathbf{b}\end{pmatrix}$ have rank $|\mathcal{R}|$.*

Proof. The argument is identical to the proof of Theorem 4.3.

Applying the law of total probability and our conditional independence relation $S \perp\!\!\!\perp Y \mid f(S), R, G, X$, we have, for all $y \in \mathcal{Y}, g \in \mathcal{G}, x \in \mathcal{X}$, and $s \in \mathcal{S}$,

$$\begin{aligned}\mathbb{P}(Y = y \mid G = g, X = x, S = s) &= \sum_{r\in\mathcal{R}} \mathbb{P}(Y = y \mid R = r, f(S) = s, G = g, X = x, S = s) \\ &\qquad \times \mathbb{P}(R = r \mid G = g, X = x, S = s) \\ &= \sum_{r\in\mathcal{R}} \mathbb{P}(Y = y \mid R = r, f(S) = s, G = g, X = x) \\ &\qquad \times \mathbb{P}(R = r \mid G = g, X = x, S = s).\end{aligned}$$

The left-hand side is estimable from the data and the rightmost term $\mathbb{P}(R = r \mid G = g, X = x, S = s)$ is assumed known. So for each $y \in \mathcal{Y}$, $z \in f(\mathcal{S})$, $g \in \mathcal{G}$, and $x \in \mathcal{X}$, this relation is a linear system in unknown parameters $\mathbb{P}(Y = y \mid R = r, f(S) = s, G = g, X = x)$. These parameters are identified if

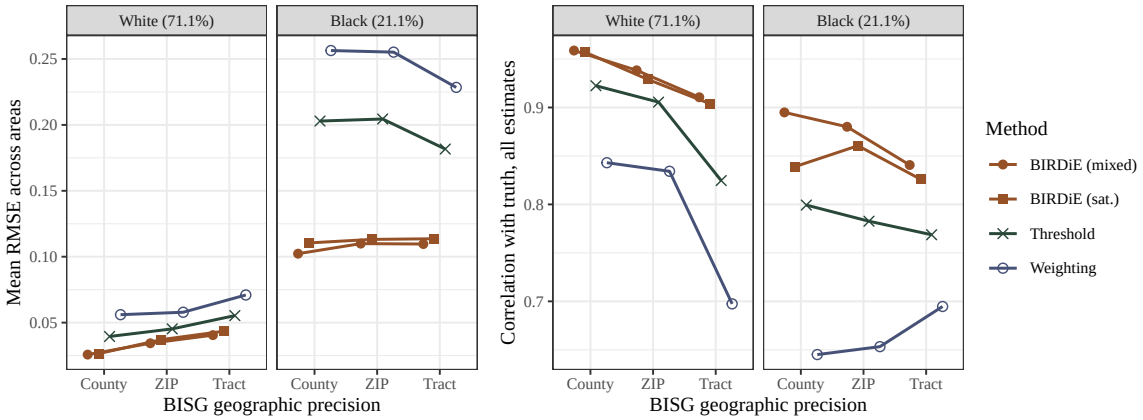


Figure D.1: Accuracy of small-area estimates by race, as measureed by root-mean-square error (RMSE; lower is better) and the correlation between the estimates and ground truth (higher is better).

and only if this system has a unique solution, i.e. if the coefficient matrix $\mathbf{P}$ has rank $|\mathcal{R}|$ and so does the augmented matrix $\begin{pmatrix} \mathbf{P} & \mathbf{b} \end{pmatrix}$. □

D.2 Additional Small-area Accuracy Evaluation

We evaluate the small-area estimates with two additional measures. First, we calculate the root-mean-square error (RMSE) of the estimated conditional probabilities by race within each geographic area, and then average this across all geographic areas. This captures the overall accuracy of the estimates. Second, to measure how well each method captures relative differences between geographic areas, we calculate the correlation between the estimated and true conditional probabilities across all geographic areas by race. As in the main text, we remove area-race cells with fewer than 5 voters. A set of estimates which uniformly underestimates the proportion of Black voters which are registered Democrats, but which otherwise correctly orders geographic areas according to their proportion of Black Democrats, will score high on the correlation measure but also higher in RMSE.

Figure D.1 summarizes our results, which closely track the findings of Section 4.4.4. The BIRDiE mod-

els are more accurate at all geographic levels and for both Black and White voters. The weighting estimator performs the worst of all the methods.

D.3 Sensitivity Analysis

D.3.1 Local sensitivity analysis

In this section, we develop a sensitivity analysis that assesses how the bias in BISG race probabilities affect the estimates of racial disparities. In particular, we consider a setting where Assumptions A1 and A2 may be violated but Assumption A4 still holds. For example, consider the existence of unobserved confounder that affects some or all of the variables except the outcome, i.e., (R, S, G, X). This leads to the violation of Assumption A1, but Assumption A4 continues to be satisfied so long as such unobserved confounder does not affect the outcome. Unfortunately, even inaccurate BISG predictions can still lead to biased estimates of racial disparities.

Specifically, if either the Census data are inaccurate, or the conditional independence relation does not hold $S \not\perp\!\!\!\perp G, X \mid R$, then the BISG predictions $\hat{\mathbf{P}}$ will differ from the "true" individual race probabilities $\mathbf{P}^* = \mathbb{P}(R \mid G, X, S)$. Our goal is to quantify how an error in these probabilities $\mathbf{P}^* - \hat{\mathbf{P}}$ shifts the posterior and hence the estimates of racial disparities.

Denote by $\pi_{\boldsymbol{\delta}}$ the posterior constructed using the error-corrected BISG race probabilities $\hat{\mathbf{P}}_i + \boldsymbol{\delta}_i$ as the input probabilities for the model (see Equation (4.3)), where $\pi_{\boldsymbol{\delta}^*}$ is the true posterior with $\boldsymbol{\delta}_i^* := \mathbf{P}_i^* - \hat{\mathbf{P}}_i$. Estimating how π and $\pi_{\boldsymbol{\delta}}$ differ in general is difficult, but we focus on the settings where $\boldsymbol{\delta}$ is small enough to make a linear approximation appropriate. In sum, we aim to quantify how the small error in BISG probabilities can alter the estimates of racial disparities.

For clarity, in this section we will use $\boldsymbol{\theta}_{rG_iX_iY_i}$ to denote the model parameter or function thereof that

represents $\pi(Y_i \mid R_i = r, G_i, X_i)$. This mirrors the notation of most of the specific models discussed in Section 4.3.2 above. Then define the following perturbation weight, which represents the ratio of posterior based on the biased and error-corrected BISG race probabilities:

$$w(\Theta, \boldsymbol{\delta}^*) := \prod_{i=1}^{N} \left(1 + \frac{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \boldsymbol{\delta}_i^*}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}\right) \propto \frac{\pi_{\boldsymbol{\delta}^*}(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})}{\pi(\Theta \mid \mathbf{Y}, \mathbf{G}, \mathbf{X}, \mathbf{S})},$$

Then, using a local linear approximation, we write the bias for a particular quantity of interest $g(\Theta)$ as

$$\begin{aligned} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}[g(\Theta)] - \mathbb{E}_{\pi}[g(\Theta)] &= \left.\frac{d\,\mathbb{E}_{\pi_{\boldsymbol{\delta}}}[g(\Theta)]}{d\boldsymbol{\delta}}\right|_{\boldsymbol{\delta}=0}^{\top} \boldsymbol{\delta}^* + o(\|\boldsymbol{\delta}^*\|) \\ &= \mathrm{Cov}_{\pi}\left(g(\Theta), \left.\frac{d \log w(\Theta, \boldsymbol{\delta})}{d\boldsymbol{\delta}}\right|_{\boldsymbol{\delta}=0}\right)^{\top} \boldsymbol{\delta}^* + o(\|\boldsymbol{\delta}^*\|), \end{aligned} \tag{D.1}$$

where the second equality is due to Theorem 2.1 of (Giordano et al., 2018; see also the idea of *local sensitivity* from Gustafson, 1996).

With this representation, we can bound the total error in $\mathbb{E}_{\pi}[g(\Theta)]$ for sufficiently small $\boldsymbol{\delta}$ as the following theorem shows.

Theorem D.1 (Bias Bound). *Define* $\tilde{\vartheta}_{ir} := \frac{\theta_{rG_iX_iY_i}}{\theta_{\cdot G_iX_iY_i}^\top \hat{\mathbf{P}}_i}$. *Then for any input probabilities with total error* $\|\boldsymbol{\delta}^*\|^2 = \sum_{i=1}^{n} \|\boldsymbol{\delta}_i^*\|^2 \leq \Delta^2$,

$$|\,\mathbb{E}_{\pi^*}[g(\Theta)] - \mathbb{E}_{\pi}[g(\Theta)]| \lesssim \Delta \left\| \mathrm{Cov}_{\pi}(g(\Theta), \tilde{\vartheta}) \right\|, \tag{D.2}$$

as $\Delta \to 0$.

Proof. This is immediate from (D.1) once we compute

$$\begin{aligned} \frac{d \log w(\Theta, \boldsymbol{\delta})}{d\delta_{ir}} &= \frac{d}{d\delta_{ir}} \sum_{i=1}^{N} \log\left(1 + \frac{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \boldsymbol{\delta}_i}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}\right) \\ &= \frac{d}{d\delta_{ir}} \log\left(1 + \frac{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \boldsymbol{\delta}_i}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^\top \hat{\mathbf{P}}_i}\right) \end{aligned}$$

$$= \frac{1}{1 + \frac{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^{\top} \boldsymbol{\delta}_i}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^{\top} \hat{\mathbf{P}}_i}} \times \frac{\boldsymbol{\theta}_{r G_i X_i Y_i}}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^{\top} \hat{\mathbf{P}}_i}$$

$$= \frac{\boldsymbol{\theta}_{r G_i X_i Y_i}}{\boldsymbol{\theta}_{\cdot G_i X_i Y_i}^{\top} (\hat{\mathbf{P}}_i + \boldsymbol{\delta}_i)}$$

and evaluate at $\boldsymbol{\delta} = 0$, since the worst-case bias for a fixed total error can be obtained by having the maximum allowable $\boldsymbol{\delta}$ point in the direction of the gradient of $\log w(\Theta, \boldsymbol{\delta})$. □

The theorem shows that once researchers choose the amount of total error Δ, then the bound on the shift in a quantity of interest can be computed readily from posterior draws. It is important to note that both $\boldsymbol{\delta}^*$ and $\text{Cov}_\pi(g(\Theta), \tilde{\vartheta})$ are vectors whose dimension depends on the sample size N. Thus, all else being equal, their norms will each grow as $\sqrt{N}$. However, each entry $\text{Cov}_\pi(g(\Theta), \tilde{\vartheta}_{ir})$ will tend to shrink as N increases, since each observation exerts less leverage on the overall posterior. Thus the overall impact of the sample size on the bound in Equation (D.2) may depend on specific features of the data. In particular, and as should be expected, the error is not guaranteed to vanish as $N \to \infty$. Practitioners should evaluate Equation (D.2) under a range of plausible Δ to understand how robust their findings are to worst-case linear violations of Assumptions $\mathcal{A}1$ and $\mathcal{A}2$.

D.3.2 OLS sensitivity analysis

For a different understanding of the effect of a particular δ, we can derive a result on the error in conditional probability estimates under the OLS estimator and particular configurations of $\boldsymbol{\delta}$. Unlike Theorem D.1, this result holds across all sizes of $\boldsymbol{\delta}$, and not just asymptotically as $\|\boldsymbol{\delta}\| \to 0$. However, it applies to the OLS estimator, which, while unbiased, we do not recommend in practice. Despite this difference, we expect many of the qualitative conclusions to hold for BIRDiE models. The effect of any particular $\boldsymbol{\delta}$ can of course be calculated directly by re-fitting the model to new race probabilities.

Here, we will work with a fixed $y \in \mathcal{Y}$ and among the subset of individuals with a particular $g \in \mathcal{G}$ and $x \in \mathcal{X}$. Then for notational simplicity we let $\hat{\boldsymbol{\mu}}^{(\text{ols})}$ be the vector of estimates of $\mathbb{P}(Y = y \mid R, G = g, X = x)$, and $\boldsymbol{\mu}$ the corresponding true probabilities. Similarly, we write $\hat{\mathbf{P}}$ for the matrix of individual race probability estimates for the subset of individuals with $g \in \mathcal{G}$ and $x \in \mathcal{X}$; elsewhere in the text this would be notated $\hat{\mathbf{P}}_{\mathcal{I}(xg)}$

Proposition D.2 (OLS Bias from incorrect $\hat{\mathbf{P}}$)**.** *Let Assumption A4 hold. If the OLS estimator $\hat{\boldsymbol{\mu}}^{(ols)}$ is calculated using race probabilities $\hat{\mathbf{P}}$ which differ from the true probabilities $\mathbf{P}^* = \hat{\mathbf{P}} + \boldsymbol{\delta}$, then its bias satisfies*

$$\mathbb{E}[\hat{\boldsymbol{\mu}}^{(ols)}] - \boldsymbol{\mu} = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \boldsymbol{\delta} \boldsymbol{\mu}$$

Proof. We can write the OLS estimate as

$$\hat{\boldsymbol{\mu}}^{(\text{ols})} = (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \mathbf{1}\mathbf{Y} = y.$$

As shown in Theorem 4.4, under Assumption A4, $\mathbb{P}(Y = y \mid R = r, G = g, X = x)$ is linear in the true $\mathbf{P}^*$. Thus letting $\varepsilon = \mathbf{1}\mathbf{Y} = y - \mathbb{P}(Y = y \mid R = r, G = g, X = x)$, we can substitute and find

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{(\text{ols})} &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\mathbf{P}^* \boldsymbol{\mu} + \varepsilon) \\
&= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \Big((\hat{\mathbf{P}} + \boldsymbol{\delta}) \boldsymbol{\mu} + \varepsilon \Big).
\end{aligned}$$

Taking an expectation, since $\mathbb{E}[\varepsilon] = 0$ we find

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\mu}}^{(\text{ols})}] &= (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top \Big((\hat{\mathbf{P}} + \boldsymbol{\delta}) \boldsymbol{\mu} \Big) \\
&= \boldsymbol{\mu} + (\hat{\mathbf{P}}^\top \hat{\mathbf{P}})^{-1} \hat{\mathbf{P}}^\top (\boldsymbol{\delta} \boldsymbol{\mu});
\end{aligned}$$

rearrangement yields the result. □

Informally, for BISG error $\boldsymbol{\delta}$ to cause problems with the OLS estimate, two things must happen. First, within individuals it must be "correlated" (i.e., have nonzero inner product) with the true conditional probabilities $\boldsymbol{\mu}$. Since $\boldsymbol{\delta}_i$ must always sum to zero, practically, this means that positive BISG errors must tend to occur in racial groups which have a relatively high occurrence of outcome y: $\mathbb{P}(Y = y \mid R = r, G = g, X - x) > \mathbb{P}(Y = y \mid G = g, X - x)$. Second, the vector $\boldsymbol{\delta}\boldsymbol{\mu}$ (where each entry measures this "correlation" between errors and relative frequencies) must be correlated the BISG probabilities themselves $\hat{\mathbf{P}}$. For example, if $\boldsymbol{\delta}_i\boldsymbol{\mu}$ is positive and tends to be larger for individuals with a high BISG probability of being Hispanic, then the overall OLS estimator the conditional probability of $Y = y$ among Hispanics will be biased upwards.

While Proposition D.2 applies within a (G, X) cell, if the same conditions hold across all (G, X) combinations, then the overall poststratified estimator will be similarly biased.

D.4 Surname Groupings for North Carolina Robustness Analysis

We classify every surname in the voter file into one of nine groups, each containing surnames from one or more of 22 surname groups that we provide in the replication data and software. These groups are organized mainly around different regions of the world and different waves of immigration to the United States. To create the surname groups, each individual in the 1930 Census data was classified into one of the 22 groups. Then among the set of individuals with each surname, the group with the highest number of individuals relative to the whole population was assigned to that surname. For example, while most people named "Smith" fall into the Anglosphere group (containing 3rd or more generation White U.S. residents as of 1930, as well as immigrants from the U.K., Canada, Australia, etc.), there are relatively more Smiths

among Black people than any other of the 22 groups. Thus "Smith" is assigned to the Black surname group. The full code for creating the 22 surname groupings from the 1930 Census data is available in the replication materials.

Because of the demographics of the United States, as well as limitations of the source data there is more geographic specificity in the surname groupings for some regions (e.g., Europe) than for others (e.g., South America and Africa). We collapse the 22 surname groups to nine for the robustness analysis in Section 4.4 based on the demographics of North Carolina specifically and to minimize the computational burden of performing the robust analysis.

The 50 roughly most frequent surnames in each group, along with a brief description of the group, are listed below. We stress that for the purposes of sensitivity analysis, the surname groups need only be correlated with countries of origin and racial subgroups. Perfect alignment is neither possible nor necessary.

Anlglosphere and Black surname group. Surnames which are relatively more prevalent among 3rd-or-more generation White U.S. residents and Black U.S. residents in 1930.

1. SMITH	11. LEWIS	21. HALL	31. COLLINS	41. COX
2. WILLIAMS	12. ROBINSON	22. CAMPBELL	32. STEWART	42. WARD
3. BROWN	13. WALKER	23. MITCHELL	33. MORRIS	43. RICHARDSON
4. JONES	14. ALLEN	24. CARTER	34. COOK	44. WATSON
5. DAVIS	15. WRIGHT	25. ROBERTS	35. ROGERS	45. BROOKS
6. TAYLOR	16. SCOTT	26. PHILLIPS	36. MORGAN	46. WOOD
7. MOORE	17. HILL	27. EVANS	37. COOPER	47. JAMES
8. JACKSON	18. GREEN	28. TURNER	38. BAILEY	48. BENNETT
9. WHITE	19. ADAMS	29. PARKER	39. REED	49. GRAY
10. CLARK	20. BAKER	30. EDWARDS	40. HOWARD	50. HUGHES

First wave European immigration surname group. Surnames associated with German, Nordic, and Irish immigrants.

1. JOHNSON	11. BURNS	21. CARROLL	31. SCHULTZ	41. HIGGINS
2. ANDERSON	12. OLSON	22. RILEY	32. PEARSON	42. OCONNOR
3. NELSON	13. WAGNER	23. BURKE	33. BARRETT	43. QUINN
4. MURPHY	14. MEYER	24. LARSON	34. BECK	44. SWANSON
5. PETERSON	15. SCHMIDT	25. CARLSON	35. POWERS	45. FITZGERALD
6. KELLY	16. RYAN	26. OBRIEN	36. LEONARD	46. CHRISTENSEN
7. SULLIVAN	17. DUNN	27. LYNCH	37. BENSON	47. MANNING
8. MURRAY	18. KELLEY	28. HANSON	38. LYONS	48. MCLAUGHLIN
9. MCDONALD	19. HANSEN	29. WEBER	39. MCCARTHY	49. DOYLE
10. KENNEDY	20. CUNNINGHAM	30. WALSH	40. ERICKSON	50. BRADY

Second wave European immigration surname group. Surnames associated with Eastern European, Italian, Jewish, Russian, Greek, and other Southern European immigrants.

1. FOX	11. ZIMMERMAN	21. KLINE	31. KATZ	41. NICHOLAS
2. NICHOLS	12. KLEIN	22. BERGER	32. MARINO	42. ROSENBERG
3. HOFFMAN	13. GROSS	23. STEIN	33. BRUNO	43. ROSSI
4. NEWMAN	14. GOODMAN	24. RAYMOND	34. MOSER	44. SINGER
5. SCHNEIDER	15. SHERMAN	25. FRIEDMAN	35. GOLDSTEIN	45. ABRAMS
6. KELLER	16. WOLF	26. LEVY	36. GOLDBERG	46. ACKERMAN
7. GREGORY	17. KRAMER	27. NOVAK	37. KAPLAN	47. HELLER
8. SCHWARTZ	18. NICHOLSON	28. KAUFMAN	38. KESSLER	48. STERN
9. COHEN	19. WEISS	29. LEVINE	39. ROMANO	49. SCHAFER
10. BECKER	20. RUSSO	30. LEHMAN	40. FINK	50. SHAPIRO

East Asian surname group. Surnames associated with Chinese, Japanese, and Korean immigrants.

1. LEE	11. BOWEN	21. HORNE	31. HAN	41. LIANG
2. YOUNG	12. LIU	22. XIONG	32. LAU	42. SUN
3. WONG	13. PAUL	23. LIM	33. MA	43. JUNG
4. WANG	14. CHAN	24. TANG	34. PUCKETT	44. ZHOU
5. PARK	15. TODD	25. CHO	35. CHIN	45. GEE
6. MAY	16. ZHANG	26. CHENG	36. GIL	46. ZHAO
7. JOSEPH	17. LANG	27. KANG	37. XU	47. SHIN
8. LOWE	18. YU	28. LAW	38. SONG	48. OHARA
9. CHANG	19. CHOI	29. CRAFT	39. KAY	49. ZHU
10. LIN	20. MOON	30. NG	40. STROUD	50. YEE

South Asian surname group. Surnames associated with Indian and Southwest Asian immigrants.

1. WILSON	11. CARR	21. DAVID	31. MOHAMED	41. SAMUEL
2. THOMAS	12. SINGH	22. HOWE	32. BOGGS	42. SEWELL
3. PATEL	13. BISHOP	23. HAHN	33. KUMAR	43. HASSAN
4. STEVENS	14. MANN	24. GOOD	34. WESTON	44. SADLER
5. WOODS	15. FRANCIS	25. JOHN	35. BEATTY	45. PINTO
6. SHAW	16. GILL	26. OSBORN	36. SWAIN	46. MAJOR
7. FERGUSON	17. YATES	27. ABRAHAM	37. GOMES	47. BARNHART
8. RAY	18. MARSH	28. RODRIGUES	38. JACOB	48. CARMICHAEL
9. WILLIS	19. ROY	29. PEREIRA	39. TOLBERT	49. MUHAMMAD
10. GEORGE	20. KAUR	30. SHARMA	40. PAYTON	50. GUPTA

Southeast Asian and Pacific surname group. Surnames associated with Southeast Asian and Pacific Islander immigrants, including Vietnamese and Filipino immigrants.

1. MILLER	11. SILVA	21. HUANG	31. PHAN	41. HOANG
2. MARTIN	12. SANTOS	22. WU	32. VO	42. CASH
3. KING	13. GREENE	23. ROWE	33. VU	43. BUI
4. NGUYEN	14. LI	24. BAUTISTA	34. LU	44. CHU
5. KIM	15. LE	25. HOUSTON	35. NGO	45. SINCLAIR
6. LONG	16. YANG	26. LAM	36. TAN	46. SORIANO
7. TRAN	17. LITTLE	27. HUYNH	37. HONG	47. ZHENG
8. CHEN	18. MORAN	28. HO	38. DANG	48. LESLIE
9. WEBB	19. PHAM	29. CHUNG	39. DO	49. ANGEL
10. GORDON	20. RAMSEY	30. TRUONG	40. LY	50. DUONG

Non-Cuban Hispanic surname group. Surnames associated with Mexian and Latin American immigrants, not including Cuban immigrants, and Puerto Rican residents.

1. GARCIA	11. RIVERA	21. MENDOZA	31. CASTRO	41. ALVARADO
2. RODRIGUEZ	12. GOMEZ	22. RUIZ	32. FERNANDEZ	42. DELGADO
3. MARTINEZ	13. DIAZ	23. CASTILLO	33. VARGAS	43. PENA
4. HERNANDEZ	14. CRUZ	24. GONZALES	34. GUZMAN	44. CONTRERAS
5. LOPEZ	15. REYES	25. VASQUEZ	35. MENDEZ	45. SANDOVAL
6. PEREZ	16. MORALES	26. ROMERO	36. MUNOZ	46. GUERRERO
7. SANCHEZ	17. GUTIERREZ	27. MORENO	37. SALAZAR	47. RIOS
8. RAMIREZ	18. ORTIZ	28. HERRERA	38. GARZA	48. ESTRADA
9. TORRES	19. RAMOS	29. MEDINA	39. SOTO	49. ORTEGA
10. FLORES	20. CHAVEZ	30. AGUILAR	40. VAZQUEZ	50. NUNEZ

Cuban surname group. Surnames associated with Cuban immigrants.

1. GONZALEZ	11. SUAREZ	21. CRANE	31. SARGENT	41. MARRERO
2. ALVAREZ	12. CONNER	22. FRYE	32. GORE	42. VALDES
3. JIMENEZ	13. SANTANA	23. PARRA	33. ZIEGLER	43. OLIVA
4. BOWMAN	14. DECKER	24. MAYO	34. TOMLINSON	44. MCCLENDON
5. DAVIDSON	15. SKINNER	25. DAVIES	35. LOWRY	45. QUEEN
6. ACOSTA	16. ABBOTT	26. BLANCO	36. PAGAN	46. MCCORD
7. MOLINA	17. GARRISON	27. WITT	37. LORD	47. CRESPO
8. MIRANDA	18. PONCE	28. CARRASCO	38. CARBAJAL	48. CORNEJO
9. CASTANEDA	19. PALACIOS	29. ALONSO	39. BETANCOURT	49. DUMAS
10. BALL	20. SLOAN	30. HAINES	40. PATINO	50. BUENO

"Other" surname group. Surnames not associated with one of the other categories, including those associated with later Western European immigration, Middle Eastern & North African-associated surnames, Native-associated surnames and Afro-Carribean-associated surnames.

1. PERRY	11. WELCH	21. SIMON	31. FRANCO	41. MCKENZIE
2. HENRY	12. DAY	22. CUMMINGS	32. HAMMOND	42. BEIL
3. HUNT	13. STANLEY	23. CHANDLER	33. CLARKE	43. COCHRAN
4. ROSE	14. HOPKINS	24. SHARP	34. WATERS	44. NASH
5. PIERCE	15. LAMBERT	25. BARBER	35. FRANK	45. BRYAN
6. PETERS	16. NORRIS	26. GRIFFITH	36. ANDRADE	46. MEYERS
7. KNIGHT	17. WALTERS	27. PACHECO	37. LLOYD	47. CARSON
8. RICHARDS	18. STEELE	28. CROSS	38. FRENCH	48. WILKINSON
9. MORRISON	19. BUSH	29. GOODWIN	39. OWEN	49. ATKINSON
10. JACOBS	20. WOLFE	30. MULLINS	40. CHARLES	50. VINCENT

This page intentionally left blank.

References

Abott, C. and Magazinnik, A. (2020). At-large elections and minority representation in local government. *American Journal of Political Science*, 64(3):717–733.

Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M., and Zhuravlev, P. (2022). The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harvard Data Science Review*, (Special Issue 2). https://hdsr.mitpress.mit.edu/pub/7evz361i.

Akitaya, H. A., Korman, M., Korten, O., Souvaine, D. L., and Tóth, C. D. (2022). Reconfiguration of connected graph partitions via recombination. *Theoretical Computer Science*.

Alao, R., Bogen, M., Miao, J., Mironov, I., and Tannen, J. (2021). How Meta is working to assess fairness in relation to race in the US across its products and systems. Technical report, Technical Report. Meta AI.

Anderson, M. and Fienberg, S. E. (1999). *Who Counts?: The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation.

Argyle, L. and Barber, M. (2022). Misclassification and bias in predictions of individual ethnicity from administrative records.

Åslund, O. and Skans, O. N. (2012). Do anonymous job application procedures level the playing field? *ILR Review*, 65(1):82–107.

Autry, E., Carter, D., Herschlag, G., Hunter, Z., and Mattingly, J. (2020). Multi-scale merge-split Markov chain Monte Carlo for redistricting. *Working paper*.

Baker *v.* Carr (1962). 369 U.S. 186, 82 S. Ct. 691, 7 L. Ed. 2d 663.

Bangia, S., Graves, C. V., Herschlag, G., Kang, H. S., Luo, J., Mattingly, J. C., and Ravier, R. (2017). Re-

districting: Drawing the line. *arXiv preprint arXiv:1704.03360*.

Barreto, M. A. (2007). ¡Sí se puede! Latino candidates and the mobilization of Latino voters. *American Political Science Review*, 101(3):425–441.

Becker, A., Duchin, M., Gold, D., and Hirsch, S. (2021). Computational redistricting and the Voting Rights Act. *Election Law Journal: Rules, Politics, and Policy*, 20(4):407–441.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Bernstein, M. and Duchin, M. (2017). A formula goes to court: Partisan gerrymandering and the efficiency gap. *Notices of the AMS*, 64(9):1020–1024.

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., and Ho, D. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1479–1503.

Bozkaya, B., Erkut, E., and Laporte, G. (2003). A tabu search heuristic and adaptive memory procedure for political districting. *European journal of operational research*, 144(1):12–26.

Brown, D. A. (2022). *The whiteness of wealth: How the tax system impoverishes Black Americans–and how we can fix it*. Crown.

Brown, J. R. (2021). Partisan conversion through neighborhood influence: How voters adopt the partisanship of their neighbors and reinforce geographic polarization. Working Paper.

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

Brown, J. R., Enos, R. D., Feigenbaum, J., and Mazumder, S. (2021). Childhood cross-ethnic exposure predicts political behavior seven decades later: Evidence from linked administrative data. *Science Advances*, 7(24):eabe8432.

Brunell, T. (2010). *Redistricting and representation: Why competitive elections are bad for America*. Routledge.

Budiman, A., Cilluffo, A., and Ruiz, N. G. (2019). Key facts about Asian origin groups in the US. *Washington, DC: Pew Research Center*.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical*

Software, 80:1–28.

Burns *v.* Richardson (1966). 384 U.S. 73, 86 S. Ct. 1286, 16 L. Ed. 2d 376.

Buttice, M. K. and Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Political analysis*, 21(4).

Cain, B. E. and Hopkins, D. A. (2002). Mapmaking at the grassroots: The legal and political issues of local redistricting. *Election Law Journal*, 1(4):515–530.

Calvin *v.* Jefferson County (2015). 172 F. Supp. 3d 1292.

Cannon, S., Duchin, M., Randall, D., and Rule, P. (2022). Spanning tree methods for sampling graph partitions. *arXiv preprint arXiv:2210.01401*.

Carter, D., Herschlag, G., Hunter, Z., and Mattingly, J. (2019). A merge-split proposal for reversible Monte Carlo Markov chain sampling of redistricting plans. *arXiv preprint arXiv:1911.01503*.

Chaskin, R. J. (1997). Perspectives on neighborhood and community: A review of the literature. *Social Service Review*, 71(4):521–547.

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135.

Chen, J. (2017). Expert report of Jowei Chen, Ph.D. Expert witness report in League of Women Voters *v.* Commonwealth.

Chen, J. (2018). Fair lending needs explainable models for responsible recommendation. *arXiv preprint arXiv:1809.04684*.

Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348.

Chen, J. and Rodden, J. (2013). Unintentional gerrymandering: Political geography and electoral bias in legislatures. *Quarterly Journal of Political Science*, 8(3):239–269.

Chetty, R. and Hendren, N. (2018a). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162.

Chetty, R. and Hendren, N. (2018b). The impacts of neighborhoods on intergenerational mobility II:

County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228.

Chikina, M., Frieze, A., and Pegden, W. (2017). Assessing significance in a Markov chain without mixing. *Proceedings of the National Academy of Sciences*, 114(11):2860–2864.

Chikina, M., Frieze, A., and Pegden, W. (2019). Understanding our Markov chain significance test: A reply to Cho and Rubinstein-Salzedo. *Statistics and Public Policy*, 6(1):50–53.

Cho, W. K. T. and Liu, Y. Y. (2018). Sampling from complicated and unknown distributions: Monte Carlo and Markov chain Monte Carlo methods for redistricting. *Physica A: Statistical Mechanics and its Applications*, 506:170–178.

Cho, W. K. T. and Rubinstein-Salzedo, S. (2019). Understanding significance tests from a non-mixing Markov chain for partisan gerrymandering claims. *Statistics and Public Policy*, 6(1):44–49.

Cho, W. T. and Manski, C. F. (2008). *Oxford Handbook of Political Methodology*, chapter Cross-Level/Ecological Inference, pages 547–569. Oxford University Press.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Christ, O., Schmid, K., Lolliot, S., Swart, H., Stolle, D., Tausch, N., Al Ramiah, A., Wagner, U., Vertovec, S., and Hewstone, M. (2014). Contextual effect of positive intergroup contact on outgroup prejudice. *Proceedings of the National Academy of Sciences*, 111(11):3996–4000.

Cirincione, C., Darling, T. A., and O'Rourke, T. G. (2000). Assessing South Carolina's 1990s congressional districting. *Political Geography*, 19(2):189–211.

Common Cause *v.* Lewis (2019). 834 S.E.2d 425 (N.C.: Supreme Court).

Conway, D. A. and Roberts, H. V. (1983). Reverse regression, fairness, and employment discrimination. *Journal of Business & Economic Statistics*, 1(1):75–85.

Coulton, C. J., Jennings, M. Z., and Chan, T. (2013). How big is my neighborhood? individual and contextual effects on perceptions of neighborhood scale. *American Journal of Community Psychology*, 51(1):140–150.

Coulton, C. J., Korbin, J., Chan, T., and Su, M. (2001). Mapping residents' perceptions of neighborhood boundaries: A methodological note. *American Journal of Community Psychology*, 29(2):371–383.

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons, 2nd edition.

Davidson *v.* City of Cranston (2016). 188 F. Supp. 3d 146.

de Kadt, D. and Sands, M. L. (2021). Racial isolation drives racial voting: Evidence from the new south africa. *Political Behavior*, 43(1):87–117.

Decter-Frain, A. (2022). How should we proxy for race/ethnicity? Comparing Bayesian Improved Surname Geocoding to machine learning methods. *arXiv preprint arXiv:2206.14583*.

DeFord, D., Dhamankar, N., Duchin, M., Gupta, V., McPike, M., Schoenbach, G., and Sim, K. W. (2020). Implementing partisan symmetry: Problems and paradoxes. *arXiv preprint arXiv:2008.06930*.

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.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

DeLuca, K. and Curiel, J. A. (2022). Validating the applicability of Bayesian inference with surname and geocoding to congressional redistricting. *Political Analysis*, pages 1–7.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Dinesen, P. T. and Sønderskov, K. M. (2015). Ethnic diversity and social trust evidence from the microcontext. *American Sociological Review*, 80(3):550–573.

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer, New York.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Dube, M. P. and Clark, J. T. (2016). Beyond the circle: Measuring district compactness using graph theory. In *Annual Meeting of the Northeastern Political Science Association*.

Duchin, M. (2018). Outlier analysis for Pennsylvania congressional redistricting.

Duchin, M., Gladkova, T., Henninger-Voss, E., Klingensmith, B., Newman, H., and Wheelen, H. (2018). Locating the representational baseline: Republicans in massachusetts.

Ebanks, D., Katz, J. N., and King, G. (2022). If a statistical model predicts that common events should

occur only once in 10,000 elections, maybe it's the wrong model.

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008). A new method for estimating race/ethnicity and associated disparities where administrative records lack self-reported race/ethnicity. *Health services research*, 43(5p1):1722–1736.

Elmendorf, C. S., Quinn, K. M., and Abrajano, M. A. (2016). Racially polarized voting. *The University of Chicago Law Review*, pages 587–692.

Elzayn, H., Smith, E., Hertz, T., Ramesh, A., Fisher, R., Ho, D. E., and Goldin, J. (2023). Measuring and mitigating racial disparities in tax audits.

Enos, R. D. (2015). What the demolition of public housing teaches us about the impact of racial threat on political behavior. *American Journal of Political Science*.

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

Evenwel *v.* Abbott (2016). 136 S. Ct. 1120, 578 U.S. 54, 194 L. Ed. 2d 291.

Fifield, B., Higgins, M., Imai, K., and Tarr, A. (2020a). Automated redistricting simulation using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 29(4):715–728.

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

Fiscella, K. and Fremont, A. M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41(4p1):1482–1500.

Fischer, C. S. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. University of Chicago Press. Google-Books-ID: H6ni9D3_ScAC.

Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.

Fowler, J. H. and Christakis, N. A. (2010). Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences*, 107(12):5334–5338.

Gay, C. (2001). The effect of black congressional representation on political participation. *American Political Science Review*, 95(3):589–602.

Gay, C. (2006). Seeing difference: The effect of economic disparity on black attitudes toward Latinos. *American Journal of Political Science*, 50(4):982–997.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and King, G. (1994). A unified method of evaluating electoral systems and redistricting plans. *American Journal of Political Science*, pages 514–554.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Gill *v.* Whitford (2018). 138 S. Ct. 1916, 201 L. Ed. 2d 313, 585 U.S. __.

Gilliland, J. F. (1969). Racial gerrymandering in the deep south. *Ala. L. Rev.*, 22:319.

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness and variational Bayes. *Journal of Machine Learning Research*, 19(51).

Goodman, L. A. (1953). Ecological regressions and behavior of individuals. *American Sociological Review*.

Greene, W. H. (1984). Reverse regression: The algebra of discrimination. *Journal of Business & Economic Statistics*, 2(2):117–120.

Greiner, D. J. (2006). Ecological inference in Voting Rights Act disputes: Where are we now, and where do we want to be. *Jurimetrics*, 47:115.

Greiner, J. D. and Quinn, K. M. (2009). $R \times C$ ecological inference: Bounds, correlations, flexibility and transparency of assumptions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):67–81.

Griffith, E. C. (1907). *The rise and development of the gerrymander*. Scott, Foresman.

Grofman, B. (1983). Measures of bias and proportionality in seats-votes relationships. *Political Methodology*, pages 295–327.

Grofman, B. and Handley, L. (1991). Identifying and remedying racial gerrymandering. *JL & Pol.*, 8:345.

Guage, C., Goldin, J. S., Ouyang, D., and Ho, D. E. (2023). Enabling data disaggregation of Asian subgroups: A dataset of Wikipedia names. *Working Paper*.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Guth, L., Nieh, A., and Weighill, T. (2020). Three applications of entropy to gerrymandering. *arXiv preprint arXiv:2010.14972*.

Habyarimana, J., Humphreys, M., Posner, D. N., and Weinstein, J. M. (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review*, 101(4):709–725.

Hajnal, Z. and Trounstine, J. (2005). Where turnout matters: The consequences of uneven turnout in city politics. *The Journal of Politics*, 67(2):515–535.

Halberstam, M. (2012). Process failure and transparency reform in local redistricting. *Election Law Journal*, 11(4):446–471.

Herschlag, G., Ravier, R., and Mattingly, J. C. (2017). Evaluating partisan gerrymandering in Wisconsin. *arXiv preprint arXiv:1709.01596*.

Hipp, J. R., Faris, R. W., and Boessen, A. (2012). Measuring 'neighborhood': Constructing network neighborhoods. *Social networks*, 34(1):128–140.

Hjorth, F., Dinesen, P. T., and Sonderkov, K. M. (2021). The content and correlates of subjective local contexts. Working Paper.

Hood III, M., Morrison, P. A., and Bryan, T. M. (2018). From legal theory to practical application: A how-to for performing vote dilution analyses. *Social Science Quarterly*, 99(2):536–552.

Hooper, W. (1881). The method of statistical analysis. *Journal of the Statistical Society of London*, 44(1):31–49.

Hopkins, D. J. (2010). Politicized places: Explaining where and when immigrants provoke local opposition. *American Political Science Review*, 104(1):40–60.

Hopkins, D. J. (2011). Flooded communities: Explaining local reactions to post-katrina migrants. *Political Research Quarterly*, 65(2):443–459.

Huckfeldt, R. and Sprague, J. (1987). Networks in context: The social flow of political information. *American Political Science Review*, 81(4):1197–1216.

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

Imai, K., Lu, Y., and Strauss, A. (2008). Bayesian and likelihood inference for 2× 2 ecological tables: an incomplete-data approach. *Political Analysis*, 16(1):41–69.

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully Bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):1–10.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Issacharoff, S. and Karlan, P. S. (1998). Standing and Misunderstanding in Voting Rights Law. *Harvard Law Review*, 111(8):2276.

Kamakura, W. A., Mazzon, J. A., and De Bruyn, A. (2006). Modeling voter choice to predict the final outcome of two-stage elections. *International Journal of Forecasting*, 22(4):689–706.

Katz, J., King, G., and Rosenblatt, E. (2020). Theoretical foundations and empirical evaluations of partisan fairness in district-based democracies. *American Political Science Review*, 114(1):164–178.

Katz, J. N., King, G., and Rosenblatt, E. (2021). The essential role of statistical inference in evaluating electoral systems: A response to DeFord et al. *Political Analysis*, pages 1–11.

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

Kenny, C. T., McCartan, C., Fifield, B., and Imai, K. (2020). redist: Computational algorithms for redistricting simulation. https://CRAN.R-project.org/package=redist.

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*, In Press.

Kermani, A. and Wong, F. (2021). Racial disparities in housing returns. Technical report, National Bureau of Economic Research.

King, G. (1989). Representation through legislative redistricting: A stochastic model. *American Journal of Political Science*, pages 787–824.

King, G. (1997). A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.

King, G. and Browning, R. X. (1987). Democratic representation and partisan bias in congressional elections. *American Political Science Review*, 81(4):1251–1273.

King, G., Tanner, M. A., and Rosen, O. (2004). *Ecological inference: New methodological strategies*. Cambridge University Press.

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.

Kostochka, A. V. (1995). The number of spanning trees in graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):269–274.

Kuriwaki, S., Ansolabehere, S., Dagonel, A., and Yamauchi, S. (2021). The geography of racially polarized voting: Calibrating surveys at the district level.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Laird, N. (1993). The EM algorithm. In *Computational Statistics*, volume 9 of *Handbook of Statistics*, pages 509–520. Elsevier.

League of Women Voters *v.* Commonwealth (2018). 178 A. 3d 737 (Pa: Supreme Court).

Lee, A. and Whiteley, N. (2018). Variance estimation in the particle filter. *Biometrika*, 105(3):609–625.

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

LeGland, F. and Oudjane, N. (2005). A sequential particle algorithm that keeps the particle system alive. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE.

Liu, J. S., Chen, R., and Logvinenko, T. (2001). A theoretical framework for sequential importance sampling with resampling. In *Sequential Monte Carlo methods in practice*, pages 225–246. Springer.

Liu, J. S., Chen, R., and Wong, W. H. (1998). Rejection control and sequential importance sampling. *Journal of the American Statistical Association*, 93(443):1022–1031.

Liu, Y. Y., Cho, W. K. T., and Wang, S. (2016). PEAR: a massively parallel evolutionary computation approach for political redistricting optimization and analysis. *Swarm and Evolutionary Computation*, 30:78–92.

Macmillan, W. (2001). Redistricting in a GIS environment: An optimisation algorithm using switching-points. *Journal of Geographical Systems*, 3(2):167–180.

Magleby, D. B. and Mosesson, D. B. (2018). A new approach for developing neutral redistricting plans. *Political Analysis*, 26(2):147–167.

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

Mattingly, J. C. and Vaughn, C. (2014). Redistricting and the will of the people. *arXiv preprint arXiv:1410.8796*.

McCartan, C. (2023). birdie: Bayesian instrumental regression for disparity estimation. https://github.com/CoryMcCartan/birdie.

McCartan, C., Brown, J. R., and Imai, K. (2021). Measuring and modeling neighborhoods. *arXiv preprint arXiv:2110.14014*.

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

McCartan, C. and Imai, K. (2023). Sequential Monte Carlo for sampling balanced and compact redistricting plans. *Annals of Applied Statistics*, Accepted.

McCartan, C. and Kenny, C. T. (2022). Individual and differential harm in redistricting. *SocArXiv preprint nc2x7*.

McCartan, C., Kenny, C. T., Simko, T., Garcia III, G., Wang, K., Wu, M., Kuriwaki, S., and Imai, K. (2022). Simulated redistricting plans for the analysis and evaluation of redistricting in the united states. *Scientific Data*, 9(1):689.

McCartney, W. B., Orellana, J., and Zhang, C. (2021). Sort selling: Political polarization and residential choice. Working Paper.

McDonald, M. D. and Best, R. E. (2015). Unfair partisan gerrymanders in politics and law: A diagnostic applied to six cases. *Election Law Journal*, 14(4):312–330.

McKay, B. D. (1981). Spanning trees in random regular graphs. In *Proceedings of the Third Caribbean Conference on Combinatorics and Computing*. Citeseer.

Mehrotra, A., Johnson, E. L., and Nemhauser, G. L. (1998). An optimization based heuristic for political districting. *Management Science*, 44(8):1100–1114.

Merrill *v.* Milligan (2022). 142 S. Ct. 879, 595 U.S __.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993.

NAACP *v.* East Ramapo Central School District (2020). No. 17 Civ. 8943 (CS)(JCM).

Najt, L., Deford, D., and Solomon, J. (2019). Complexity and geometry of sampling connected graph partitions. *arXiv preprint arXiv:1908.08881*.

National Conference of State Legislatures (2021). Redistricting criteria. Available at https://www.ncsl.org/research/redistricting/redistricting-criteria.aspx.

Nuamah, S. A. and Ogorzalek, T. (2021). Close to home: Place-based mobilization in racialized contexts. *American Political Science Review*, 115(3):757–774.

O'Loughlin, J. (1982). The identification and evaluation of racial gerrymandering. *Annals of the Association of American Geographers*, 72(2):165–184.

Olsson, J. and Douc, R. (2019). Numerically stable online estimation of variance in particle filters. *Bernoulli*, 25(2):1504–1535.

Onnela, J.-P., Arbesman, S., González, M. C., Barabási, A.-L., and Christakis, N. A. (2011). Geographic constraints on social network groups. *PLOS ONE*, 6(4):1–7.

Paddison, R. (1983). *The fragmented state: The political geography of power*. Blackwell Oxford.

Park, J., Malachi, E., Sternin, O., and Tevet, R. (2009). Subtle bias against Muslim job applicants in personnel decisions. *Journal of Applied Social Psychology*, 39(9):2174–2190.

Park, R., Burgess, E., and Sampson, R. (1925). *The City*. Heritage of Sociology Series. University of Chicago Press.

Perez-Truglia, R. (2017). Political conformity: Event-study evidence from the United States. *The Review of Economics and Statistics*, 100.

Peters, G. W., Fan, Y., and Sisson, S. A. (2012). On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. *Statistics and Computing*, 22(6):1209–1222.

Polsby, D. D. and Popper, R. D. (1991). The third criterion: Compactness as a procedural safeguard against partisan gerrymandering. *Yale Law & Policy Review*, 9(2):301–353.

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.

Rawls, J. (2004). A theory of justice. In *Ethics*, pages 229–234. Routledge.

Reeskens, T. and Hooghe, M. (2008). Cross-cultural measurement equivalence of generalized trust. evidence from the european social survey (2002 and 2004). *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, 85(3):515–532.

Reynolds *v.* Sims (1964). 377 U.S. 533, 84 S. Ct. 1362, 12 L. Ed. 2d 506.

Rucho *v.* Common Cause (2019). 139 S. Ct. 2484, 204 L. Ed. 2d 931, 588 U.S. __.

Ruggles, S., Fitch, C. A., Goeken, R., Hacker, J. D., Nelson, M. A., Roberts, E., Schouweiler, M., and Sobek, M. (2021). IPUMS ancestry full count data: 1930 5% sample, version 3.0.

Sampson, R. J. (1988). Local friendship ties and community attachment in mass society: A multilevel systemic model. *American Sociological Review*, 53(5):766–779.

Sampson, R. J., Morenoff, J. D., and Gannon-Rowley, T. (2002). Assessing neighborhood effects: Social processes and new directions in research. *Annual Review of Sociology*, 28(1):443–478.

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*, 40:559–579.

Shaw *v.* Reno (1993). 509 U.S. 630, 113 S. Ct. 2816, 125 L. Ed. 2d 511.

Siegel-Hawley, G. (2013). Educational gerrymandering? race and attendance boundaries in a demographically changing suburb. *Harvard Educational Review*, 83(4):580–612.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.0.

Stephanopoulos, N. O. (2014). The South after *Shelby County*. *The Supreme Court Review*, 2013(1):55–134.

Stephanopoulos, N. O. and McGhee, E. M. (2015). Partisan gerrymandering and the efficiency gap. *U. Chi. L. Rev.*, 82:831.

Storper, M. (2013). *Keys to the City: How Economics, Institutions, Social Interaction, and Politics Shape Development*. Princeton University Press.

Strmic-Pawl, H. V., Jackson, B. A., and Garner, S. (2018). Race counts: Racial and ethnic data on the U.S. census and the implications for tracking inequality. *Sociology of Race and Ethnicity*, 4(1):1–13.

Suttles, G. D. (1972). *The social construction of communities / Gerald D. Suttles*. University of Chicago Press Chicago.

Tam Cho, W. K. (2017). Measuring partisan fairness: How well does the efficiency gap guard against sophisticated as well as simple-minded modes of partisan discrimination? *University of Pennsylvania Law Review Online*, 166(1):2.

Thomas, A., Gelman, A., King, G., and Katz, J. N. (2013). Estimating partisan bias of the electoral college under proposed changes in elector apportionment. *Statistics, Politics, and Policy*, 4(1):1–13.

Thornburg *v.* Gingles (1986). 478 U.S. 30, 106 S. Ct. 2752, 92 L. Ed. 2d 25.

Trounstine, J. (2015). Segregation and inequality in public goods. *American Journal of Political Science*.

Tufte, E. R. (1973). The relationship between seats and votes in two-party systems. *American Political Science Review*, 67(2):540–554.

Tutte, W. T. (1984). *Graph Theory*. Addison-Wesley.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124ï¿œ–1131.

United Jewish Organizations *v.* Carey (1977). 430 U.S. 144, 97 S. Ct. 996, 51 L. Ed. 2d 229.

U.S. Census Bureau (2014). Frequently occurring surnames from the 2010 census. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html. Accessed: 2022-03-25.

U.S. Census Bureau (2022). Census Bureau releases estimates of undercount and overcount in the 2020 census. *March 10, 2022, Release Number CB22-CN.02*.

Van Ryn, M. and Fu, S. S. (2003). Paved with good intentions: do public health and human service providers contribute to racial/ethnic disparities in health? *American journal of public health*, 93(2):248–255.

Varadhan, R. and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scandinavian Journal of Statistics*, 35(2):335–353.

Vargas, R., Cano, C., Del Toro, P., and Fenaughty, B. (2021). The racial and economic foundations of municipal redistricting. *Social Problems*.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2019). Rank-normalization,

folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC. *arXiv preprint arXiv:1903.08008*.

Velez, Y. R. and Wong, G. (2017). Assessing contextual measurement strategies. *The Journal of Politics*, 79(3):1084–1089.

Veomett, E. (2018). Efficiency gap, voter turnout, and the efficiency principle. *Election Law Journal: Rules, Politics, and Policy*, 17(4):249–263.

Vieth *v.* Jubelirer (2004). 178 A. 3d 737 (Pa: Supreme Court).

Wakefield, J. (2004). Ecological inference for 2×2 tables. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(3):385–425.

Warrington, G. S. (2018). Introduction to the declination function for gerrymanders.

Williams, D. R. and Jackson, P. B. (2005). Social sources of racial disparities in health. *Health affairs*, 24(2):325–334.

Wilson, D. B. (1996). Generating random spanning trees more quickly than the cover time. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 296–303.

Wong, C. (2010). *Boundaries of Obligation in American Politics: Geographic, National, and Racial Communities*. Cambridge University Press, New York.

Wong, C., Bowers, J., Rubenson, D., Fredrickson, M., and Rundlett, A. (2020). Maps in people's heads: Assessing a new measure of context. *Political Science Research and Methods*, 8(1):160–168.

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.

Wu, L. C., Dou, J. X., Sleator, D., Frieze, A., and Miller, D. (2015). Impartial redistricting: A Markov Chain approach. *arXiv preprint arXiv:1510.03247*.

Yule, G. U. (1905). The introduction of the words "statistics," "statistical" into the English language. *Journal of the Royal Statistical Society*, pages 391–396.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180.