Why Does Anyone Need Precinct-Level Election Results?

Kate Murray Samuel Baltz Charles Stewart III

May 2023*

Abstract

Precinct-level election results are an essential tool in the study of American elections, but they require substantial work on the part of both election administrators and academics to collect, standardize, and release on a statewide or national scale. To motivate these efforts, we describe the many important uses of these data, and also evaluate their limitations. Reviewing the major uses for precinct-level election results, we identify several areas where the added granularity of precinct data makes it possible to answer questions that cannot be fully addressed using just county- or district-level election data. Some areas are: creating and adjusting electoral districts; analyzing hyper-local phenomena like municipal regulations, public communication, or economic activity; the uses of voting technology; and statistical tests that require smaller units of geography or ones with reasonably uniform population sizes. We also identify some areas where more granular data may not be more useful: precinct data do not have demographics attached in the way that counties do, it can be difficult to use precinct data for problems that require geographic information (as precinct boundaries are not static), and some problems with aggregate data like ecological inference are not resolved when precinct-level results are used (though they may be less severe). Understanding the uses of precinct-level election results is important both for researchers and practitioners, since both groups put substantial time and resources into releasing these data, as well as choosing how they are shaped and which variables are included.

*This draft prepared for the Election Science, Reform, and Administration Conference, 2023.

Introduction

In addition to simply reporting an accurate tally of votes to the public, election results in the United States are useful for investigating a wide variety of political questions. Over decades, academics have used election returns to investigate salient voter issues and other political phenomena. Non-academics, including local election officials and politicians, have also utilized election data for important purposes. However, most of the election data used for analysis are those at high levels of aggregation: news organizations and governments primarily report election results at the state level, and the lowest level of aggregation widely available is at the district level, where returns are often reported by individual counties or townships. Precincts are the smallest geographic unit for which election results are reported, though data at this level of granularity is quite rare. This is primarily because collecting and publishing this data presents significant logistical challenges. The MIT Election Data and Science Lab (MEDSL) is one of several organizations that have taken up this task, and they have published standardized precinct-level election returns for all 50 states and the District of Columbia for every major election since 2016.

The primary motivation of this paper is to provide insight into how election data at the precinct level can be used by both researchers and practitioners for an array of political purposes. The first section contains a brief history of precinct data collection and an overview of the challenges involved with precinct data publication, followed by a literature review of how this data have been utilized in recent years. The final section details the various examples for which precinct-level election returns are useful, using data from MEDSL's extensive repository. The following section will outline scenarios for which more granular data aren't useful and explore their limitations.

Background

It's likely that the first effort to collect nation-wide precinct-level election data was the Record of American Democracy, which has data dating back to 1984 (King, Gary and Palmquist, Bradley and Adams, Greg and Altman, Micah and Beniot, Kenneth and Gay, Claudine and Lewis, Jeffrey B. and Mayer, Russ and Reinhardt, Eric 1997). Since then, various research institutions and non-profit organizations have taken up the task of collecting, cleaning, and publishing precinct returns: in addition to MEDSL, OpenElections OpenElections (2023) has compiled the majority of federal races for all 50 states since 2000, and the Voting and Election Science Team has collected precinct returns as well as shapefiles for numerous races since 2012 (Harvard Dataverse 2023). Additionally, the Redistricting Data Hub advocates using precinct data for redistricting purposes (for example, ensuring that precinct boundaries are drawn to allow for maximum competitiveness). Despite increased efforts to collect this data, certified election returns at the precinct-level remain quite rare. This is primarily because collecting, standardizing, and publishing these returns presents significant challenges. States have the freedom to choose when and in which formats to publish their election data, which often results in precinct-level returns that are incomplete or missing. Additionally, there are no universal naming or formatting conventions across states, so efforts to standardize precinct data requires extensive time and resources. With nearly 180,000 across the country and over ten million unique combinations of candidates and precincts in a general election, collecting, cleaning, and publishing these data becomes a significant undertaking.

Literature Review

Over the last several decades, precinct-level election data has been utilized for voting and non-voting purposes alike. This literature review summarizes several of the research areas for which precinct data has been used, as well as examples of how this data has been applied by politicians and other practitioners.

Election and Voting-Related Applications of Precinct-Level Election Data

Election data has been used to explore a wide variety of academic questions, but precinct-level data has been utilized in specific ways. In his 2006 paper "Election Forensics: The Second-digit Benford's Law Test and Recent American Presidential Elections," Walter Mebane Mebane Jr. (2006) used precinct-level election returns to test whether second digits of vote totals appear with frequency as described by Benford's Law (a metric used to determine whether voting technology is flawed. This analysis has had profound implications for the narrative surrounding election fraud in recent years. Additionally, Karp and Banducci (2000) used Oregon's precinct data merged with census data to determine how all-mail voting would increase participation and alter the composition of the electorate. They found that all-mail voting would have the biggest impact for local and primary elections, and additionally, voting by mail is likely to have the biggest effect on turnout for those who are already likely to vote. Precinct data has also been used in research exploring the relationship between geography and political polarization. In "Spatial Scale and the Geographical Polarization of the American Electorate," Rhola et al. (2018) utilized precinct data for the 2008, 2012, and 2016 Presidential elections to determine whether political polarization is correlated with geography. Through their research, they found that polarization was greater at the precinct level than the county and state levels. In a thematically similar piece, Gregory Martin and Steven Webster's paper Martin and Webster (2020) exploring the relationship between geographic mobility and political preferences reported that location has a marginal effect on political preference. Political science research focusing on political representation, voter engagement, and identity have also relied on precinct data for their analyses. Pelissero et al. (2000) used precinct data from Chicago to assess the impact of Asian voters and political organizations on political participation. Their paper reports that larger Asian populations are correlated with greater voter registration but lower turnout. Miller and Crosnick (1999) utilized Ohio precinct data to determine whether the order in which candidates are listed on a ballot affects the share of votes they receive, discovering that candidates listed first on the ballot were advantaged by 2.5As for research on voting and identity, Hersh and Nall (2015) used geocoded precinct returns to explore the relationship between income and partisanship across the U.S. They found that income-based voting is heavily correlated with race, and correlation between income and partisanship differs based on geographic region and racial homogeneity. In their unique article, Gebru et al. (2017) collected data on automobiles from 50 million Google Street View images spanning 3,068 zip codes, 39,286 voting precincts, and 200 US cities to estimate social demographic information and voting preferences for each area. Using a logistic regression model to estimate income and voter preferences based on vehicle type, they found strong associations of Democratic precincts with sedans and Republican precincts with extended-cab pickup trucks. Another significant area of political science literature that has relied on precinct data is research on gerrymandering and redistricting reform. In the piece “Locating the Representational Baseline: Republicans in Massachusetts”, Duchin et al. (2019) investigate why Republican candidates do not have more favorable electoral outcomes in the state, discovering that these outcomes are not attributable to partisan gerrymandering but a true distributional disadvantage. Furthermore, Saxon (2020)) simulated voting districts using precinct level election returns to develop new software that identifies optimal arrangements for “compactness” according to 18 different definitions. In a related piece, Gurnee and Shmoys (2021) demonstrate that redistricting efforts that mainly focus on compactness (rather than political and demographic information) “explicitly optimize for arbitrary piecewise-linear definitions of fairness." Additionally, Henderson et al. (2018) use simulated House maps (created by precinct data) to analyze whether nonpartisan redistricting efforts make for more politically competitive districts, ultimately finding that, overall, this is not the case. In order to evaluate the efficacy of Pennsylvania's 2022 redistricting plan, Warshaw (2022) used precinct-level data from recent statewide elections to calculate the partisan bias in state house elections under the state's new plan. He found that the new redistricting arrangement facilitates a slight GOP bias, allowing the party to almost always win a majority of seats. Precinct data has even been used to develop complex mathematical models- Chikina et al. (2017) investigated how Markov chains can be applied to avoid bias in Congressional districting. In some cases, precinct data has also been useful for purposes outside of the academic sphere. In a 2014 article for the Journal of Economic Perspectives, Nickerson and Rogers (2014) detail the ways in which political campaigns have used precinct data to guide their electoral strategy and gain deeper insight into their constituencies. For instance, campaigns can merge Census data and precinct information onto relevant information in voter databases. Precinct data is also useful in determining which localities are most heavily contested, and this information is helpful for campaigns when considering where to invest the most resources.

Other Applications of Precinct-Level Election Data

While precinct-level returns have certainly been impactful in the election space, they have also had utility in other areas of research. Since 2020, precinct data has been utilized by public health researchers to investigate the widespread effects of the COVID-19 pandemic and evaluate public perceptions and responses. Using MEDSL precinct data to proxy for political partisanship, Barrios and Hochberg (2021), find that a higher percentage of Trump voters in a county is correlated with a lower perception of COVID risk (these findings replicated former work done by Allcott et al. (2020)). In a related project, researchers at the University of Michigan utilize the same data source to analyze beliefs, constraints, and risk preferences about COVID across identity groups, including gender and partisanship (though this source uses county-level totals) (Fan et al. 2020). Researchers have also used precinct-level data to analyze the impacts of policies and social movements. Laniyonu (2019) used precinct data as well as US Census data to explore the effects of local policing on voter turnout and candidate preference in New York City. This study found that areas with concentrated efforts of policing experienced different patterns in turnout over time, and stopping the intensity of policing was strongly associated with support Democratic mayoral candidates. As another example, researchers Mullin and Rubado (2017) aggregate precinct-level vote returns to water service areas in order to investigate the relationship of political partisanship to local policy responses to drought in Texas between 2010-2013. Trounstine (2020) maps precinct-level election data from several California cities onto demographic data to determine whether partisanship and racial identity affect voter support for redistricting development. Economists and data scientists have also utilized precinct-level election data in their work. Asquith et al. (2021) matched precinct vote totals to Census Voting Districts to investigate the connection between social capital determinants and local labor market networks. They found that neighborhoods with a larger Republican vote share (among other factors) are correlated with more active labor networks. As another example, Hayatpur et al. (2021) used MEDSL data as a “guided scenario” to present to study participants in their research about how virtual reality can be used to enhance spatial reasoning skills. As this section has demonstrated, precinct-level election data have had far-reaching implications for academic research as well as policy outcomes. The next portion of this paper will detail the specific reasons why this information is useful, utilizing data from MEDSL’s repository as well as other sources.

Benefits of Precinct Data

Exploring Local Variations

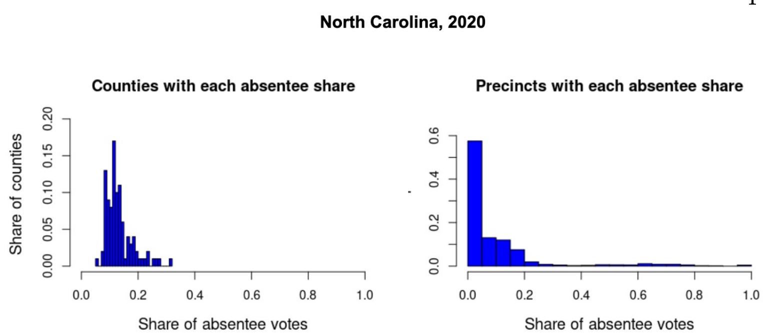
Put simply, one of the biggest benefits of using precinct-level data is that it allows for a more nuanced analysis of variations at the local level. Just as county and district-level results give one much more precise information about local phenomena than, say, state-level data, precinct data can be used to explain patterns at a local level much more holistically than data at a higher level of aggregation. A surprising amount of information about sub-county races is lost if one just reports at the county level.

To put this into perspective, Figure 1 displays the total number of office-county-state name combinations in MEDSL's precinct data for the given years compared to the total number of name combinations only contested at the precinct level. It should be noted that the share of sub-county races decreases over time simply because the lab did not collect as much local data as time went on. Even so, one would miss about 1 in 5 office categories contested in each county in 2018 if they used county data as opposed to precinct data. Put another way, over 20 percent of local data is lost reporting at a higher level of aggregation- data that could provide essential information to explaining characteristics and patterns among and within localities.

Year	**Total Office-County-State Combos**	**Combos Only Contested at the Sub-County Level**	**Sub-County Share**
2016[27]	43,856	18,209	42%
2018[28]	46,227	9,746	21%
2020[29]	30,747	4,269	14%

Exploring Vote Modes

In addition to simply gaining more accurate and nuanced information, precinct data can be useful for exploring geographic variation in the distribution of vote modes. Gaining insight into the way Americans vote is crucial to many political decisions, particularly when considering where and how resources should be allocated. To demonstrate this in practice, Figure 1 compares the share of absentee votes across all counties in North Carolina in 2020 with the share of absentee votes across all North Carolina precincts in the same year.

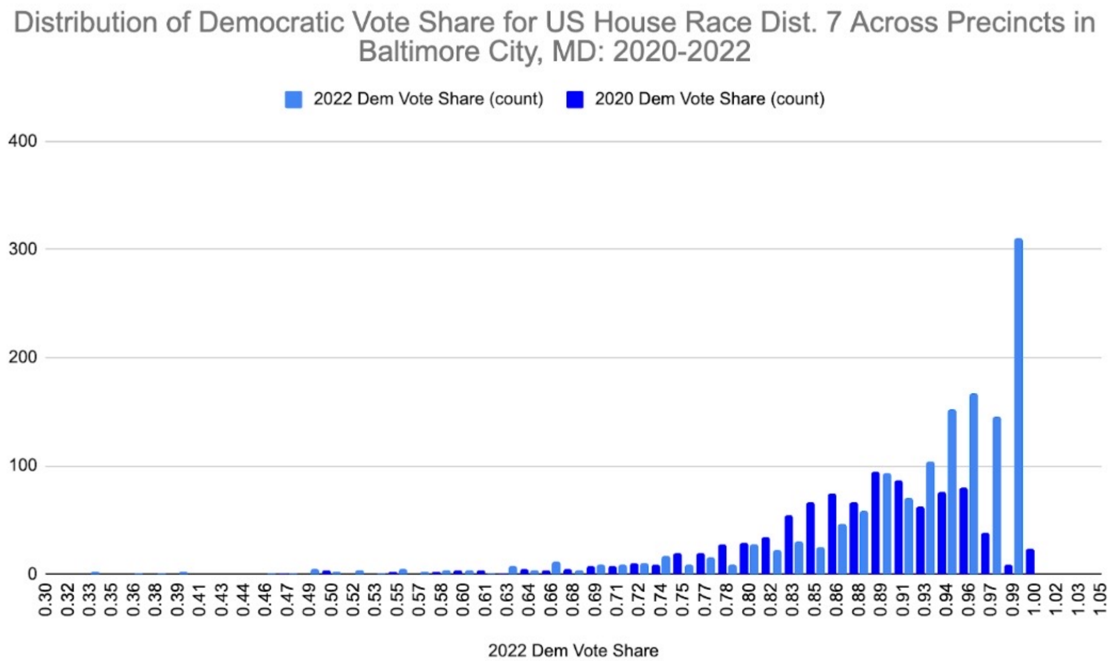


The story told by these graphs is an interesting one: the county-level spread on the left depicts the absentee vote share across all North Carolina counties as hovering between 10 and 20 percent (which is on par with what the national average has been since 2000) (Stewart III 2023). However, the precinct distribution on the right depicts a different spread, and for a particular reason: the small bar on the far right of the precinct graph demonstrates the share of North Carolina precincts containing an absentee vote share of 100 percent (not impossible, but also unlikely to occur in the real world). In fact, those data points are not real precincts at all; they are non-geographic precincts that the state has designated to "hold" all of the absentee votes in their counties. This, in turn, dramatically drives up the percentage of precincts with an absentee vote share of 0 percent, as evidenced by the tallest bar in the graph. Misleadingly, this bar actually represents plenty of precincts that have a much greater share of absentee voting. However, the precinct spread still displays important and accurate information about the distribution of data at this low level of aggregation, as the points in the middle are likely an accurate representation of the ground truth. This case serves not only as a prime example of how using precinct data provides more nuance than more aggregated data, it demonstrates the importance of being aware of external variables that may be affecting a distribution (in this case, the creation of non-geographic precincts). Though mode analysis at the precinct level is susceptible to skew (just as are data at other levels of aggregation), the information such analyses provides is valuable and essential for gaining insight into voting behavior.

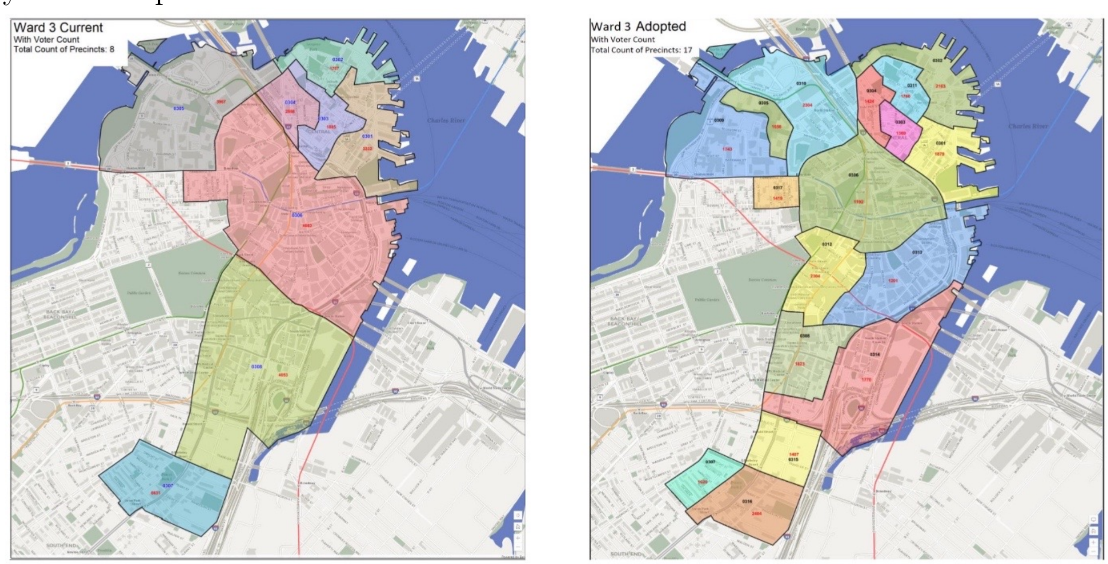
1 Limitations of Precinct Data

Longitudinal Analysis

One primary way in which precinct-level election returns are not useful for research is when one is interested in observing patterns over time. This is simply because precinct boundaries are not static; when precinct boundaries change, so too do the characteristics within precincts. Everything from the number of voters within each precinct to the partisan makeup of these areas changes with each redistricting cycle. Figure 2 depicts a recent example of the effects of redistricting, showing the difference in the Democratic vote share in U.S. House Race District 7 across all of Baltimore City, Maryland's precincts for 2020 and 2022. Baltimore underwent re-precincting between 2020 and 2022, and this likely accounts for a lot of the changes observed in the vote share. In particular there is a considerable difference in the number of precincts that have a Democratic vote share between 94 and 99 percent. While it's possible that other factors were contributing to this significant increase over time, the majority of the change was most likely simply due to boundaries changing.

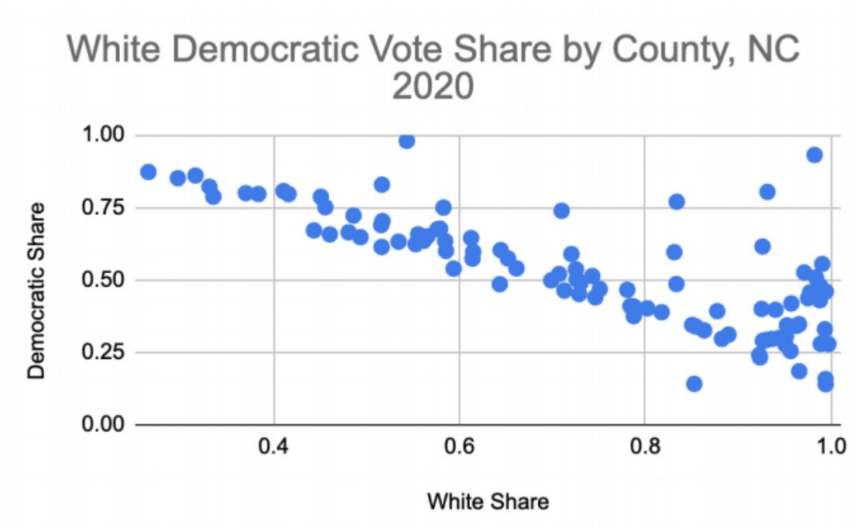


Another example of how redrawing boundary lines can affect precinct analyses can be seen in Figure 3. These images are aerial views of Boston's third voting ward, with the map on the left showing the division of precincts within the ward in 2020 and the right map showing how it looks currently after having undergone reprecincting. Having gained nine precincts in the redistricting process, it would be unhelpful to do any sort of longitudinal analysis at the precinct level in this case.



Ecological Inference

A crucial thing to be aware of when working with precinct data is the all-too-common ecological inference problem (also known as the cross-level problem). This is where one is interested in behavior of a certain population but only data at a higher level of aggregation is available and then that data must be used to make inferences about the population of interest (Tam Cho and Manski 2008). In the case of election data, there are often discrepancies at the county and precinct levels. Figures 4 and 5 demonstrate ecological inference by comparing the partisan vote share by race in North Carolina in 2020.



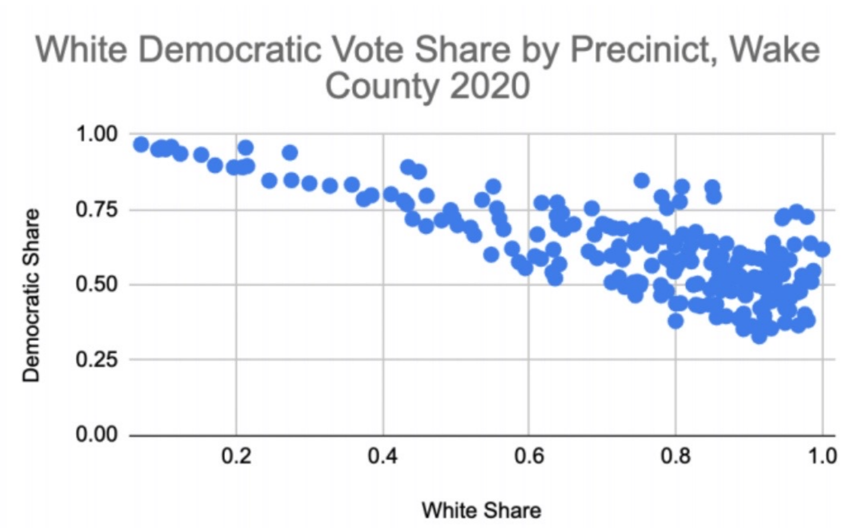


Figure 4 shows the distribution of the share of white Democratic voters across all counties in North Carolina in 2020, and Figure 5 shows the distribution of the share of white Democratic voters across all precincts within Wake County (one of the state's largest.) Both spreads look similar, but if one were to use county level data to try and make inferences about race and partisanship at the precinct level, they'd be missing important nuances. For example, there is a large cluster of localities in the precinct distribution that are majority white but only about 50 percent Democrat that are not present in the county data. The county spread depicts that it is more common for heavily white areas to be more politically polarized.

Conclusion

Precinct-level election returns have been an asset to a wide range of political projects, for those directly relating to elections as well as other areas. This paper has outlined the history of precinct-level data collection in the United States and has contributed a substantial literature review on the ways in which these data have been applied in recent decades. Furthermore, we have identified further areas of study and types of analyses for which granular election data can be useful, including observing patterns in voting technology and geographic patterns at the local level. We have also explored scenarios where the utility of these data may be limited, namely when tracking and analyzing changes over time. Precinct-level election returns require substantial effort to collect and standardize, which largely explains the relative rarity of accurate datasets. We hope that this paper has demonstrated the unique contributions these data can make in the election space and beyond, and by extension, that efforts to carry out this work are worth supporting and preserving.

Works Cited

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. "Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic". *Journal of Public Economics*, vol. 191, 2020, p. 104254. https://doi.org/https://doi.org/10.1016/j.jpubeco.2020.104254.

Asquith, Brian, Judith K. Hellerstein, Mark J. Kutzbach, and David Neumark. "Social capital determinants and labor market networks". *Journal of Regional Science*, vol. 61, no. 1, 2021, pp. 212–60. https://doi.org/https://doi.org/10.1111/jors.12508.

Barrios, John M., and Yael V. Hochberg. "Risk perceptions and politics: Evidence from the COVID-19 pandemic". *Journal of Financial Economics*, vol. 142, no. 2, 2021, pp. 862–79. https://doi.org/https://doi.org/10.1016/j.jfineco.2021.05.039.

Chikina, Maria, Alan Frieze, and Wesley Pegden. "Assessing significance in a Markov chain without mixing". *Proceedings of the National Academy of Sciences*, vol. 114, no. 11, 2017, pp. 2860–64. https://doi.org/https://doi.org/10.1073/pnas.1617540114.

Duchin, Moon, Taissa Gladkova, Eugene Henninger-Voss, Ben Klingensmith, Heather Newman, and Hannah Wheelen. "Locating the Representational Baseline: Republicans in Massachusetts". *Election Law Journal: Rules, Politics, and Policy*, vol. 18, no. 4, 2019, pp. 388–401. https://doi.org/https://doi.org/10.1089/elj.2018.0537.

Fan, Ying, A. Yesim Orhun, and Dana Turjeman. "Heterogeneous Actions, Beliefs, Constraints and Risk Tolerance During the Covid-19 Pandemic". *NBER Working paper*, w27211, 2020. papers.ssrn.com/sol3/papers.cfm?abstract_id=3621796.

Gebru, Timnet, Jonathan Krause, Yilun Wang, Duyen Chen, Erez Lieberman Aiden, and Li Fei-Fei. "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States". *Proceedings of the National Academy of Sciences*, vol. 114, no. 50, 2017, pp. 13108–13. https://doi.org/https://doi.org/10.1073/pnas.1700035114.

Gurnee, Wes, and David B. Shmoys. "Fairmandering: A Column Generation Heuristic for Fairness-optimized Political Districting". *Proceedings of the 2021 SIAM Conference on Applied and Computational Discrete Algorithms*, 2021, pp. 88–99. https://doi.org/https://doi.org/10.1137/1.9781611976830.9.

Harvard Dataverse. Voting and Election Science Team. 2023. dataverse.harvard.edu/dataverse/electionscience?q=&types=dataverses%3Adatasets&sort=dateSort&order=desc&page=1.

Hayatpur, Devamardeep, Haijun Xia, and Daniel Wigdor. "DataHop: Spatial Data Exploration in Virtual Reality". Association for Computing Machinery, 2021, pp. 818–28, https://doi.org/https://doi.org/10.1145/3379337.3415878.

Henderson, John A., Brian T. Hamel, and Aaron M. Goldzimer. "Gerrymandering Incumbency: Does Nonpartisan Redistricting Increase Electoral Competition?" *The Journal of Politics*, vol. 80, no. 3, 2018. https://doi.org/https://doi.org/10.1017/pan.2019.45.

Hersh, Eitan, and Clayton Nall. "The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records". *American Journal of Political Science*, vol. 60, no. 2, 2015, pp. 289–303. https://doi.org/https://doi.org/10.1111/ajps.12179open_in_new.

Karp, Jeffrey A., and Susan A. Banducci. "Going Postal: How All-Mail Elections Influence Turnout". *Political Behavior*, vol. 22, 2000, pp. 223–329. https://doi.org/https://doi.org/10.1023/A:1026662130163.

King, Gary and Palmquist, Bradley and Adams, Greg and Altman, Micah and Beniot, Kenneth and Gay, Claudine and Lewis, Jeffrey B. and Mayer, Russ and Reinhardt, Eric. The Record of American Democracy, 1984-1990. 1997. road.hmdc.harvard.edu/pages/road-documentation.

Laniyonu, Ayobami. "The Political Consequences of Policing: Evidence from New York City". *Political Behavior*, vol. 41, 2019, pp. 527–58. https://doi.org/https://doi.org/10.1007/s11109-018-9461-9.

Martin, Gregory J., and Steven W. Webster. “Does Residential Sorting Explain Geographic Polarization?” *Political Science Research and Methods*, vol. 8, no. 2, 2020, pp. 215–31. https://doi.org/https://doi.org/10.1017/psrm.2018.44.

Mebane Jr., Walter R. “Election Forensics: Vote Counts and Benford’s Law”. *Summer Meeting of the Political Methodology Society*, 2006. www-personal.umich.edu/~wmebane/pm06.pdf.

Miller, Joanne, and Jon Crosnick. “The Impact of Candidate Name Order on Election Outcomes”. *The Public Opinion Quarterly*, vol. 62, no. 3, 1999, pp. 291–330. *JSTOR*, www.jstor.org/stable/2749662.

Mullin, Megan, and Meghan E. Rubado. “Local Response to Water Crisis: Explaining Variation in Usage Restrictions During a Texas Drought”. *Urban Affairs Review*, vol. 53, no. 4, 2017, pp. 752–74. https://doi.org/https://doi.org/10.1177/1078087416657199.

Nickerson, David W., and Todd Rogers. “Political Campaigns and Big Data”. *Journal of Economic Perspectives*, vol. 28, no. 2, 2014, pp. 51–74. https://doi.org/https://doi.org/10.1073/pnas.1617540114.

OpenElections. FAQs. 2023. openelections.net/faq/.

Pelissero, John P., Timothy B. Krebs, and Shannon Jenkins. “Asian-Americans, Political Organizations, and Participation in Chicago Electoral Precincts”. *Urban Affairs Review*, vol. 35, 2000, pp. 750–69. journals.sagepub.com/doi/pdf/10.1177/10780870022184660.

Rhola, Ryne, Ron Johnson, Kelvyn Jones, and David Maley. “Spatial scale and the geographical polarization of the American electorate”. *Political Geography*, vol. 65, 2018, pp. 117–22. https://doi.org/https://doi.org/10.1016/j.polgeo.2018.05.010.

Saxon, James. “Reviving Legislative Avenues for Gerrymandering Reform with a Flexible, Automated Tool”. *Political Analysis*, vol. 28, no. 3, 2020, pp. 372–94. https://doi.org/https://doi.org/10.1017/pan.2019.45.

Stewart III, Charles. How We Voted in 2022: A Topical Look at the Survey of the Performance of American Elections. 2023. electionlab.mit.edu/sites/default/files/2023-05/How-We-Voted-In-2022.pdf.

Tam Cho, Wendy K., and Charles F. Manski. *The Oxford Handbook of Political Methodology*. Oxford Univeristy P, 2008, p. 547.

Trounstine, Jessica. "The Geography of Inequality: How Land Use Regulation Produces Segregation". *American Political Science Review*, vol. 114, no. 2, 2020, pp. 443–55. https://doi.org/https://doi.org/10.1017/S0003055419000844.

Warshaw, Christopher. "An Evaluation of the Partisan Fairness of the Pennsylvania Legislative Reapportionment Commission's Proposed State House Districting Plan". *Working paper*, 2022, pp. 1–26. www.redistricting.state.pa.us/resources/Press/2022-01-14%20Warshaw%20Testimony.pdf.